London School of Economics and Political Science

Adjusting for Unobserved and Observed Heterogeneity in Survey-Based Performance Indicators

An Application to Adult Social Care in England

Juliette Nicola Malley

A thesis submitted to the Department of Social Policy at the London School of Economics for the degree of Doctor of Philosophy, London, January, 2017

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party.

I declare that my thesis consists of 97,615 words (including tables).

Abstract

This thesis is concerned with the statistical adjustment of survey-based indicators to account for unobserved and observed sources of heterogeneity. Recent years have seen a growth in the use of survey-based indicators to measure performance, but questions have been raised over their legitimacy due to high levels of nonresponse, particularly among certain groups, and the influence of factors unrelated to organisational performance, which complicate their interpretation. In light of this, this thesis uses a range of methods that go beyond those ordinarily applied to performance assessment, to explore the role that nonresponse and factors unrelated to performance, i.e. case-mix, have on indicators. The empirical analysis focuses on the Adult Social Care Outcomes Framework (ASCOF) indicators drawn from the English Adult Social Care Survey. The core concerns of this thesis are whether (i) nonresponse and (ii) adjusting for factors beyond the control of organisations affects the interpretation of indicator scores. Nonresponse has a limited effect on inferences about performance, but conclusions depend on the method used to explore the effects of nonresponse, the level of nonresponse, the importance of unobserved factors and the value placed on accuracy over intelligibility of indicators. Adjustment for case-mix has an important effect on the interpretation of indicators, but the adjustment method used was less critical for inference, at least where the aim is to compare organisations. This thesis suggests that the accuracy of some of the ASCOF indicators would be improved by adjusting for case-mix and, possibly, for nonresponse. It is important for future studies to explore the effect of nonresponse on indicators. Policymakers may also wish to consider amending the survey design to improve its representativeness of the adult social care population. Future studies of survey-based performance indicators would benefit from using a wider range of methods similar to those applied here.

Preface

The idea for this thesis emerged from a stream of work that I was involved in within the Costs, Quality and Outcomes Policy Research Programme at PSSRU at the University of Kent. As part of this research programme that was funded by the Department of Health, I led a series of studies under the guidance of Professor Ann Netten that looked broadly at the measurement of quality within user experience surveys. This work culminated in the development of the Adult Social Care Survey (ASCS)[1], which forms the main dataset for this thesis. During this time I also worked on the development of the ASCOT measure[2], an equivalent of the EQ-5D for social care, which features heavily in this thesis. I have benefited enormously from Professor Ann Netten's enthusiasm for this area of research and her support over the years, which have shaped my understanding of quality and outcomes in social care.

I conceived the idea for the analyses in this thesis, but I am very grateful to Professor Ann Netten and Dr Jose-Luis Fernandez for helping me to secure the funding for this research as part of the Quality and Outcomes Research Unit (QORU) Policy Research Programme grant from the Department of Health. A number of the chapters in this thesis are based on outputs from the QORU programme. Chapters 4, 6, 7 and 8 are based on working papers jointly authored with Dr Jose-Luis Fernandez[3], which were circulated to officials at the Department of Health. Parts of Chapter 2 draw on a peer-reviewed journal

---

[1] Malley, J. & Netten, A. (2008), *Measuring User Experience of Social Care Services: A Discussion of Three Approaches. A Report to the Department of Health. Discusssion Paper 2529*. Canterbury, Personal Social Services Research Unit; Malley, J. & Netten, A. (2009), *Putting People First: Development of the Putting People First User Experience Survey. Discussion Paper 2637.* Canterbury, Personal Social Services Research Unit; Malley, J., Caiels, J., Fox, D., McCarthy, M., Smith, N., Beadle-Brown, J., Netten, A. & Towers, A.-M. (2010), *A Report on the Development Studies for the National Adult Social Care User Experience Survey, PSSRU Discussion Paper 2721.* Canterbury, Personal Social Services Research Unit, University of Kent; Malley, J. (2010), *A Comparison of Options for Performance Indicators from the Adult Social Care Survey (ASCS), PSSRU Discussion Paper 2736.* Canterbury, Personal Social Services Research Unit, University of Kent.

[2] Netten, A., P. Burge, J. Malley, D. Potoglou, A.-M. Towers, J. Brazier, T. Flynn and J. Forder (2012). "Outcomes of Social Care for Adults: Developing a preference-weighted measure." *Health Technology Assessment 16(16).*; Malley, J., A.-M. Towers, A. Netten, J. Brazier, J. Forder and T. Flynn (2012) "An assessment of the construct validity of the ASCOT measure of social care-related quality of life with older people." *Health and Quality of Life Outcomes, 10.* DOI: 10.1186/1477-7525-10-21.

[3] Malley, J. & Fernandez, J. (2012), Patterns and effects of nonresponse in the English Adult Social Care Survey. *PSSRU Discussion Paper 2841*, London, Personal Social Services Research Unit, London School of Economics and Political Sciences; Malley, J. & Fernandez, J. (2014), Generating adjusted indicators from social care survey data. *PSSRU Discussion Paper 2873*, London, Personal Social Services Research Unit, London School of Economics and Political Sciences.

article jointly authored with Dr Jose-Luis Fernandez[4] . In more recent years I have also benefited greatly from working alongside Professor Julien Forder on the Identifying the Impact of Adult Social Care (IIASC)[5] study. This study evolved from the early findings from this thesis and was designed to address some of the limitations in the ASCS data. The findings from the IIASC study will be used to case-mix adjust indicators within the Adult Social Care Outcomes Framework (ASCOF).

I am extremely grateful to my supervisors Dr Jose-Luis Fernandez and Professor Martin Knapp for their guidance and helpful discussions on the content and shape of the thesis, and particularly to Dr Jose-Luis Fernandez for his guidance on the statistical aspects. I am also indebted to them and to Professor Julien Forder for making it possible for me to focus on my thesis during the few months prior to submission. I am extremely grateful to the social care user survey team at NHS Digital, and in particular Pete Broughton and Katherine Robbins, for their perseverance in helping me to renegotiate the data sharing agreement with NHS Digital for the ASCS data. We got there in the end!

Thanks are also due to my colleagues, particularly the residents of COW 4.06 past and present, who have patiently listened to my musings, provided insight and comments on earlier drafts of chapters, and kept me going even when it seemed the end would never come into sight! There are too many people to list, but a special mention is due to Mike Clarke, Francesco D'Amico, Cate Henderson, Bo Hu, Derek King, Madeline Stevens, Ann-Marie Towers, Lisa Trigg and Valentina Zigante who have all in various ways gone beyond what is required of a room-mate and colleague.

I could not have completed this thesis without the support of my family and friends. Although they are probably not aware of it, my children – Leila (aged five) and Frederick (aged two) – have provided a constant source of emotional support. Coming home to their happy, care-free faces every evening always makes me smile however tired and grumpy I feel. My deepest thanks go to my husband Paddy who, I think it is fair to say, has at times in this last year felt the noose of this thesis round his neck more keenly than have I. He has provided emotional and practical support to me and has managed to keep on top of his own work, albeit that has probably meant more nights than usual spent burning the midnight oil. My heartfelt thanks also goes to Laura Pizarro Cochancela who in the last few months to

---

[4] Malley, J. & Fernández, J.-L. (2010), Measuring quality in social care services: theory and practice. *Annals of Public and Co-operative Economics,* 81, 4: 559-582.

[5] Forder, J., J. Malley, S. Rand, F. Vadean, K. Jones and A. Netten,. (2016). *IIASC report: Interpreting outcomes data for use in the Adult Social Care Outcomes Framework (ASCOF). Discussion Paper 2892*. University of Kent, Canterbury, PSSRU.

submission entirely reorganised her schedule to provide care for my children. Many other friends and family have stepped in over the years, often at the last minute, to help with childcare, and my thanks goes in particular to Carolina Guiloff, my mother Glenda Malley, and my Mother-in-Law, Anne Halliday.

Lastly, to my parents – Glenda and Michael Malley – this thesis is dedicated to you. For as long as I can remember you have always believed in me, encouraged and supported me with my studies. Thank you.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ADL	Activity of daily living
AIC	Akaike Information Criterion
ASC	Adult Social Care
ASCOF	Adult Social Care Outcomes Framework
ASCOT	Adult Social Care Outcomes Toolkit
ASCS	Adult Social Care Survey
BIC	Bayesian Information Criterion
CAHPS	Consumer Assessment of Healthcare Providers and Systems surveys
CASSR	Council with Adult Social Services Responsibilities
CCA	Complete case analysis
CQC	Care Quality Commission
DH	Department of Health (for England and Wales)
EP	Explanatory power
EQ-5D	Euroqol EQ-5D health outcomes instrument
FE	Fixed effects regression
FR	Fractional response regression
IADL	Instrumental activity of daily living
IIASC	Identifying the Impact of Adult Social Care
IPW	Inverse propensity weighting
LD	Learning disability
LA	Local Authority
MAR	Missing at random
MCAR	Missing completely at random
MH	Mental health
MI	Multiple imputation
NMAR	Not missing at random
MNL	Multinomial logistic (regression)
NHS	National Health Service
OLS	Ordinary least squares
PI	Performance indicator
PF	Production function

POW	Production of welfare
PSD	Physically and sensorily disabled (social care client group)
QoL	Quality of life
RE	Random effects regression
RMSE	Root mean squared error
SCRQoL	Social care-related quality of life
SD	Standard deviation
SE	Standard error
SM	Substance misuse
SSUSG	Social Services User Survey Group
SSV	Simplified significant variables
SV	Significant variables
VP	Vulnerable person

Chapter 1
Introduction

Introduction

The analysis and interpretation of performance indicators (PIs) often involves comparisons between organisations (Smith 1990). Yet, such comparisons do not provide a straightforward basis for inferences about good and poor performance. Two reasons lie behind this. First, one must contend with sampling error. Secondly, judging performance from observational data is complicated by the possibility that observed and unobserved factors beyond the direct control of the organisation may also influence performance. Since the value of PIs is premised on a correspondence between differences in scores and differences in organisational performance, the possibility that scores are confounded by other sources of variation raises serious questions about their validity. This thesis is concerned with understanding the sources of observed and unobserved heterogeneity in survey-based PIs and finding strategies to address them in order to improve the validity of inferences about performance.

Validity (that indicators capture what they are intended to measure) is an important quality for any PI (Campbell et al. 2002, Streiner et al. 2014). Where differences in scores do not reflect real differences in performance, end-users of the PIs may draw the wrong conclusions with potentially negative consequences for performance, or may disregard the PIs altogether. Evidence is emerging that *perceptions* of the validity of indicators are also important. If managers are to use performance information for performance improvement, they need to be convinced of the usefulness of the data (Kroll 2015). Problems with the quality of the data and the possibility that differences in indicator scores are due to factors other than performance will militate against the use of data for performance improvement.

The difficulty of drawing causal interpretations about the performance of organisations from observational data has long been recognised as a problem for performance assessment studies. Over the years various statistical techniques have been proposed to analyse such data, such as risk-adjustment and standardisation (Bird et al. 2005, Iezzoni 2013, Nicholl et al. 2013). These techniques aim to account for observed differences in the population characteristics of each organisation (often referred as 'case-mix'), such that comparisons between organisations can be made on a fairer basis. Risk-adjustment is particularly important where outcomes data are used to measure performance

because of the need to account for the effect that pre-existing conditions have on 'outcomes', i.e. on results from treatment (Shahian and Normand 2008). Performance assessment using outcomes data is therefore more complicated, because it is frequently difficult to observe the baseline state (Black and Jenkinson 2009, Devlin and Appleby 2010, Smith and Street 2013).

Where survey data are used for performance assessment, the number of reasons to doubt the validity of the PIs grows. High levels of nonresponse to surveys used for performance assessment purposes (referred to hereafter as 'performance surveys') have led many to voice concerns over the representativeness of the data and to question the validity of inferences regarding performance (Black and Jenkinson 2009, Hutchings et al. 2012, Kroll et al. 2014, Peters et al. 2014b). Statistical techniques that make assumptions about how unobserved cases may have responded based on the observed data can be used to draw more robust conclusions in the presence of missing data (Little and Rubin 2002). However, only a couple of studies have looked at the value of such techniques in the context of performance assessment with survey-based PIs and they have reached opposing conclusions (Elliott et al. 2005, Gomes et al. 2016). The potential for proxy respondents and assistance from others to complete the survey to affect the validity of responses is another source of concern (Black and Jenkinson 2009, Kroll et al. 2014). As well as raising doubts about the validity of survey-based PIs, these additional sources of error also make survey data more difficult to analyse than other types of performance data.

Despite these concerns and analytical challenges, nowadays survey-based indicators play an important role in performance measurement systems. The initial growth of surveys within the public sector was connected to the consumer movement (Mold 2010, Mold 2011). As consumerist ideas gained political traction over the latter part of the 20th century, it became accepted wisdom that public services needed to become more consumer-oriented (Mold 2011). This perspective provided a strong critique of early sets of PIs, which were seen to focus too heavily on aspects that mattered to politicians, like economy and crude efficiency, and too little on aspects, like quality, that mattered to users of public services, presenting a disconnect between the rhetoric of quality improvement and the approach to performance measurement (National Consumer Council 1983, Pollitt 1988, Pfeffer and Coote 1991, Boyne 2002). Consumer surveys were seen as important tools for driving the system towards being more responsive to the needs and wishes of users (Griffiths 1988, Department of Health 1998) and were co-opted by managers as tools for performance management.

Particularly in the early days of survey-based indicators, many managers questioned the wisdom of using survey data at all to inform performance improvement, on the basis that measures from surveys of citizens and service users seem to contradict indicators from administrative sources (Stipak 1979, Brown and Coulter 1983, Parks 1984, Rao et al. 2006). Questions were also raised over the suitability of the questions for assessing performance, which led to improvements in the methods used to gather users' views (Fitzpatrick 1991, Carr-Hill 1992, Sitzia and Wood 1997, Cleary 1998, Cleary 1999, Coulter 2006). Momentum gathered behind survey-based PIs as the UK government embraced first the 'quality'[6] movement (Pfeffer and Coote 1991, Kirkpatrick and Martinez Lucio 1995b, Bovaird and Löffler 2003, Talbot 2010) and then the 'outcomes-based management'[7] movement (Talbot 2010, Wimbush 2011), both of which highlighted the experiential nature of service provision and the value of users' views for understanding these central, but intangible, benefits of services (Pollitt 1988, Chang et al. 2002, Coulter 2006). As a consequence, very few questions are raised today over the relevance of service users' views and their importance for understanding organisational performance. Indeed, the sheer number of survey-based PIs and their centrality to the current performance measurement frameworks for health and social care are testament to how far attitudes regarding survey-based indicators have changed since their inception.

As survey-based PIs have moved from the fringes of performance measurement systems to the centre-ground, the aforementioned questions about the validity of survey-based PIs have gained greater prominence among policy-makers and in the literature. The current focus of the performance frameworks on outcomes means that much of this discussion is centred on the type of survey-based indicators, known (in the health care context) as the patient-reported outcome measure (PROM). PROMs present a particular challenge for performance assessment as they are subject to all the problems of survey-based PIs and those of outcomes measures. Since PROMs are a very recent introduction, there has been little research into how to address the effects of nonresponse and differences in case-mix between organisations, and this is particularly true for the social care sector. There have been numerous calls for more research to look carefully at how PROMs are collected, analysed and interpreted to maximise their value for performance assessment and improvement (Black and Jenkinson 2009, Dawson et al. 2010, Devlin and Appleby

---

[6] This movement is sometimes referred to as 'excellence in service delivery' and 'culture' (Talbot 2010).
[7] This movement is sometimes referred to as 'results-based management' and 'outcomes-based accountability' (Heinrich 2003, Talbot 2010, Wimbush 2011).

2010, Parkin and Devlin 2012, Valderas et al. 2012, Black 2013, Smith and Street 2013, Coulter et al. 2014, Kroll et al. 2014).

In light of the importance of survey-based PIs to performance assessment in England and the questions surrounding their validity, this thesis addresses the following research question: *if survey-based indicators are to provide valid evidence concerning performance for routine use, what strategies should be used to address the confounding influence of observed and unobserved heterogeneity?* Three hypotheses flow from the foregoing discussion with respect to this question. First, differences in the characteristics of people served by organisations will explain variations in PIs. Secondly, survey nonresponse will be influenced by the characteristics of people served by organisations and will therefore explain variations in PIs. Thirdly, these two sources of variations in PIs – nonresponse and case-mix -- will have an effect on performance assessment. I test these hypotheses empirically using data from the English performance measurement system for adult social care (ASC) and explore strategies to address these sources of heterogeneity.

The aim of this thesis is to provide evidence to inform policy. Specifically it is my intention that this analysis will allow policymakers to answer the question of whether it is necessary to address heterogeneity arising from nonresponse and case-mix in ASC. Where it seems necessary to address these sources of heterogeneity, the analysis from this thesis should help policymakers to identify appropriate strategies to minimise the threats to the validity of performance assessment from nonresponse and differences in case-mix. As the empirical analysis uses data from ASC, many of the policy recommendations will be specific to the ASC context. Given, however, that the questions raised here about the validity of survey-based PIs are not specific to the ASC indicators, conclusions regarding the value of the methods used in this thesis should have relevance to the wider context, including the health care sector.

In the rest of this chapter I provide a background to the concepts, methods and data that I use in this thesis. I begin by considering the purpose of performance measurement, since this determines how PIs are to be used and forms the backdrop against which policymakers will assess the value of introducing strategies to improve the validity of PIs. I then provide a brief overview of how I approach the research question and highlight the methods that I apply in this thesis. Since these ideas are developed more fully in the subsequent conceptual and methods chapters, I keep this discussion brief. A key aim of this section, however, is to consider how I can assess whether alternative strategies improve the validity of survey-based PIs. I also reflect on the other attributes of PIs, aside

from validity, that policymakers may need to consider when deciding whether to address the confounding influence of observed and unobserved heterogeneity. Since I use data from ASC for the empirical analysis, in the third section I provide some background to the performance measurement system within ASC in England. I discuss the way performance is conceptualised in the current performance measurement framework and the survey, known as the Adult Social Care Survey (ASCS), which is used to populate the survey-based PIs. Finally, I provide an outline of the thesis.

The purpose of performance measurement in public services: theory, evidence and practice in England

The primary objective of performance measurement is performance improvement. This is a laudable and uncontroversial objective, but not a simple one to pursue in practice. To improve performance, one needs first to define it and operationalise it for measurement; next data needs to be collected and analysed; and finally, if performance improvement is to follow, the results need to be interpreted and managers need to act on the information (Deming 1994). At each stage of this process there are hurdles to overcome. In this section I briefly review some of the debates around performance measurement, focusing on the themes of how to define performance and the use of PIs in theory and in practice. While I draw on the international literature in this discussion, I illustrate the debates with examples from UK practice and survey-based PIs, where possible. My aim is to locate performance measurement in its wider policy context and to draw out the contingent relationship between the use of indicators and their technical attributes.

Public sector performance is often referred to as "elusive" (Stewart and Walsh 1994). The point of such a comment is usually not to deny the 'reality' of performance[8] but to recognise that how we define performance is legitimately a matter for debate. Where organisations are within the public sector, politics and the public interest demand that organisations also have concern for public values, such as probity, social equity and transparency (Talbot 2010). The existence of multiple goals, which may conflict with each other, and multiple stakeholders who may disagree over the objectives to pursue presents problems for performance assessment. Which and whose objectives should be assessed? Performance assessment in the public sector raises questions about values and political

[8] Social constructionists, however, would endorse this view (see e.g. Nutley et al. 2003).

choices, but as long as we are clear about the "variety of motives and interests" then this should not hinder an objectives-based approach (Knapp 1984, p. 16).

In the UK, the New Labour government (1997-2010) attempted to address the problem of competing objectives by designing performance measurement systems using a 'balanced scorecard' approach (Kaplan and Norton 1992, Chang et al. 2002, McAdam and Walker 2003). Thus in the performance measurement frameworks of that era (known as the performance assessment framework, PAF) different dimensions of performance were identified and a number of indicators were chosen for each dimension. For example, in the social care PAF the dimensions were national priorities and strategic objectives, cost and efficiency, effectiveness of service delivery and outcomes, quality of services for users and carers, and fair access (Department of Health 1998). (The current performance measurement framework has four dimensions, with slightly different definitions, as I discuss below.) Since the dimensions of performance are defined in terms of objectives, they do not map neatly to the interests of different stakeholders, such as service managers, the public and central government. Each objective is not necessarily relevant to each stakeholder; conversely some objectives may be relevant to multiple stakeholders. Nevertheless, the interests of all stakeholders are represented by the inclusion of multiple indicators within each dimension of performance (Chang et al. 2002, McAdam et al. 2005).

Reflecting the variety of objectives and stakeholders, each PI can serve a number of different purposes. Drawing on the concept of 'administrative arguments' developed by Hood and Jackson (1991), Talbot (2007) describes a number of arguments for performance measurement including, "accountability", "user choice", "customer service", "efficiency", "results, effectiveness and "what works"", "resource allocation" and "public value". The reason I raise the 'administrative argument' concept here is because it emphasises the doctrinal nature of performance measurement. As Hood and Jackson (1991) explain, administrative arguments are doctrines of public administration, underpinned by justifications. They set out the rationale for certain forms of activity and the mechanism by which that activity will achieve the desired goals. Many of the arguments support performance improvement, albeit emphasising different aspects of performance, e.g. efficiency versus consumer-related aspects. Others, however, are unconnected to it, e.g. accountability is argued to be necessary to support the efficient functioning of a democracy (Heckman and Smith 1995). Where performance information is collected for reasons of accountability it serves a "political" rather than a "purposeful" use, where by purposeful I mean it is used for performance improvement (Moynihan 2009).

Most performance measurement systems are designed to support multiple goals, which can have consequences for the extent to which PIs are used purposefully. For example, two studies (Clarkson 2010, Micheli and Neely 2010) that analyse the use of performance information collected under the New Labour government concluded that despite government rhetoric encouraging the use of PIs for improvements in service delivery, it was this use that was least enabled by the choice of measures and the performance regime. Rather, the performance measurement system for health, social care and police services seemed designed to support accountability, and to help central government to control the behaviour of locally-run organisations. This undermined the potential for local organisations to use the data for service improvement, in part because where performance measures are used to control behaviour a range of perverse and unintended consequences follow.

Perverse uses are often a consequence of high-stakes systems, such as the one that operated under the New Labour government. The New Labour regime combined high-level scrutiny of key targets with a punitive inspection regime and rewards for success (James 2004, Hood 2006, Hood 2007). Although the regime did produce improvements in key areas, gaming, 'creative compliance' and a variety of other perverse and unintended effects were widely documented (Bevan and Hood 2006b, Bevan and Hood 2006a, Hood 2006, Hood 2007). These perverse uses were a result of the way the high stakes associated with the achievement of key targets shaped incentives for certain behaviours. The evidence suggests that if performance information is to be used for performance improvement, a supportive environment is necessary. This should facilitate the use of information through leadership and stakeholder involvement in the definition of the measures, and provide the necessary resources to support the use of information (McAdam and Walker 2003, Moynihan and Pandey 2010, Kroll 2015).

There is little research looking at the factors that drive the use of performance surveys for improvement, but the evidence from the few studies that have considered this question identifies similar factors to those already mentioned (Coulter et al. 2014). As Coulter et al. (2014) lament, the most striking finding is how underexploited the datasets of users' views are and how little evidence there is that performance surveys have led to improvements in care. A recent study (Heath et al. 2015) that investigated the use of social care surveys for performance improvement found that managers experienced difficulties understanding how to appropriately analyse, interpret and use the data for performance improvement. Participants cited concerns over the validity of survey-based indicators due

to nonresponse, and, particularly for the PROMs data, difficulties interpreting the meaning of scores and attributing differences in scores to differences in service effectiveness, as barriers to use of the survey data. Managers felt they lacked support, leadership, statistical skills and the appropriate resources to address these problems with the data.

This study by Heath et al. (2015) demonstrates the clear need for strategies to address the validity of the ASC survey-based PIs. It also highlights the importance of considering the skills and resources of end-users when presenting solutions to the technical problems inherent in survey-based PIs: solutions need to be appropriate for end-users with little statistical expertise and few resources to undertake statistical analysis otherwise it is unlikely that they will be adopted. These findings hint at the need to balance improvements in the validity of indicators against other considerations, such as the usability of the data.

Improving the validity of survey-based PIs: key concepts and an overview of methods

My intention here is to set out the broad approach I take to the empirical analysis and the questions that guide the separate analyses. More detailed methods are given in Chapter 3. I also reflect on how the analysis informs the policy questions raised by this thesis. The concept of validity – the extent to which it is affected by confounding factors and the extent to which it is improved by strategies to address nonresponse and case-mix – is critical to addressing these policy questions. As the preceding discussion suggests, however, validity is not the only criterion that matters to policymakers when they make choices about PIs. Therefore, I consider here what other attributes of PIs are important and how policymakers can use the evidence from this analysis to inform decisions about whether and how to address nonresponse and case-mix.

As I have previously stated, statistical analysis is an essential tool to address the effects of case-mix over performance (Bird et al. 2005). In this thesis I use risk-adjustment (Shahian and Normand 2008, Iezzoni 2013), and an adaptation of this approach based on the theory of production relations, to address the potentially confounding effects of case-mix on social care PROMs. Risk-adjusted comparisons between organisations are generally considered to provide more valid inferences about performance than those obtained from comparisons of organisations based on the raw PIs (Iezzoni 2013). Nevertheless, risk-adjustment is not a panacea: the statistical models and the method itself are based on a number of assumptions, some of which are testable, but many of which are

not. There is also disagreement among experts about the most appropriate statistical models to use to address features of the data that affect modelling assumptions. In some instances, the choice of method can affect results (Goldstein and Spiegelhalter 1996, DeLong et al. 1997, Shahian et al. 2001, Hannan et al. 2005, Glance et al. 2006, Li et al. 2009, Eijkenaar and van Vliet 2014), which can introduce further uncertainty to inferences about performance (Li et al. 2009). It is therefore important to compare and contrast different approaches to risk-adjustment to provide greater certainty over the validity of the adjusted indicators (Li et al. 2009, Iezzoni 2013).

The dependence of case-mix adjustment on statistical modelling and the potential sensitivity of case-mix adjustment to choices over methods lead to two sub-questions that guide the empirical analysis:

i. What is the most appropriate method for modelling ASC outcomes?
ii. What is the effect of adjusting for case-mix on inferences about performance and does the method for implementing case-mix adjustment matter?

I address question (i) in Chapters 6 and 7 and question (ii) in Chapter 8.

The nonresponse literature identifies two strategies to address nonresponse to surveys. First, strategies can be implemented at the design stage to increase response rates (see e.g. Groves et al. 2000, Dillman et al. 2009). Secondly, post-hoc statistical adjustments can be introduced at the analysis stage (see e.g. Rubin 1976, Rubin 1987, Little and Vartivarian 2003, Horton and Kleinman 2007). The empirical analysis in this thesis should provide evidence to inform both of these strategies for addressing nonresponse, thus leading to two further empirical questions:

iii. Who is missing from performance surveys and how can response rates be improved?
iv. What is the effect of nonresponse on inferences about performance and does the method used to address differences in samples due to nonresponse matter?

The empirical analysis for question (iii) is carried out in Chapter 4 and that for question (iv) in Chapter 5.

The empirical analysis is framed with the aim of improving the validity of the survey-based PIs, but as I have already suggested this is not the only criterion that matters to policymakers. Campbell et al. (2002) identify five attributes that are important for any PI, which are set out in Box 1. Validity is critically important for the reasons I have already outlined, but it must be balanced against the other four attributes. Where data are collected for performance assessment purposes feasibility is important. Strategies that

increase the validity of the PIs, but have unreasonable consequences for the costs of data collection or the timeliness of the data, are likely to be considered unacceptable. Equally, statistical analysis that is opaque and complicated may deter people from using the measures. Policymakers need to carefully balance all of these attributes when deciding whether to introduce any strategies that are designed to improve the validity of PIs. To provide policy recommendations, it is therefore important to have some sense of the extent to which the proposed strategies to address heterogeneity arising from nonresponse and case-mix will improve the validity of PIs.

Box 1: Desirable attributes of indicators†

Acceptability: whether the indicator is considered suitable, relevant and usable to the assessor and those being assessed.
Feasibility: whether it is possible to collect valid, reliable and consistent data in a timely and cost-effective manner.
Reliability: whether the indicator is accurate enough for its purpose.
Sensitivity: whether the indicator has the capacity to detect changes or meaningful differences in performance.
Validity: whether the indicator accurately measures what it purports to measure.

† Adapted from Campbell et al. (2002)

This raises the question of how to determine whether the strategies proposed to address nonresponse and case-mix provide more valid inferences about performance? Since there is no yardstick against which to judge 'true' performance I assess the validity of the adjusted indicators from a theoretical perspective. The strength of the study's conclusions therefore rests on the evidence concerning the appropriate specification of the statistical model, the plausibility of the modelling assumptions and the accurate implementation of the methods (Cartwright 2007, Li et al. 2009, Clarke et al. 2015). Such an approach has been suggested and implemented by various authors, using the concepts of 'face validity', 'content validity', 'construct validity', 'convergent validity' and 'predictive validity' from psychometric theory[9] to assess the statistical models and the PIs (Li et al. 2009, Iezzoni 2013). Although assessments of this kind cannot provide a definitive answer

[9] See, for example, the following texts on scale and measure development (Nunnally 1967, De Vellis 2003).

to the question of whether the adjusted PIs are more valid than the raw PIs, they do provide evidence regarding the extent to which the strategies improve the validity of inferences regarding performance.

Performance measurement in adult social care: key concepts and practice

The empirical analysis in this thesis uses data from the Adult Social Care Survey (ASCS), which is used to populate the PIs in the ASC outcomes framework (ASCOF), the performance measurement framework for ASC. Since these data are used throughout this thesis, in this section I first describe how performance is conceptualised and measured in the ASCOF and then I describe the design of the ASCS. I begin by providing a brief background to ASC provision in England and the characteristics of users of these services, as the nature of ASC services and the characteristics of its users has consequences for the objectives of social care and how performance is defined in the ASC context.

A brief description of the adult social care system in England

ASC in England is often characterised as a 'safety-net' system since it supports only those with very severe needs who are unable to meet the costs of their care (Fernandez et al. 2009). Access to publicly-funded ASC is through an assessment of care needs which is coordinated by the local authority (LA) adult social services department. The philosophy underpinning the system is that the primary responsibility for social care rests with individuals and their families. ASC services are strictly means-tested and, even where people pass the means-test for publicly-funded care, LAs direct services towards those who live alone and do not receive informal care (Pickard 2001) (Netten and Davies 1990, Department of Health 2010a, Pickard et al. 2012). Due to continued pressures on local government finances, and therefore ASC budgets, between 2005/6 and 2012/13 the numbers receiving publicly-funded ASC declined by around 26 per cent (Fernandez et al. 2013b) – a trend that is continuing (Humphries et al. 2016). This has resulted in the increased concentration of care on those with very high levels of need with little means to pay privately for care (Fernandez et al. 2013b, Burchardt et al. 2015).

Although ASC users are all relatively income-poor and have a high level of social care need, the nature of service provision means they have diverse social care needs. ASC is usually provided to frail older people or people with long-term conditions, whose (often deteriorating) health or condition results in impairments in activities of daily living (ADL). Users of ASC services therefore include frail older people, people with dementia,

intellectual disabilities, mental health or substance misuse problems, as well as people with a range of physical or sensory impairments. Reflecting this diversity in the health conditions and social care needs of people served, ASC includes a range of interventions that provide assistance with personal care tasks (such as dressing and bathing) and domestic tasks (such as shopping and preparing meals) (Comas-Herrara et al. 2010). Examples of ASC interventions include short-term services, such as 're-ablement'[10] services, equipment and adaptations, and long-term services, such as domiciliary care, meals services and community-based day centres. Social care also includes placements in nursing homes and 'personal' care homes, where the latter do not provide round-the-clock nursing care.

Understanding social care performance

As I have previously described, the performance measurement frameworks in England take an objectives-based approach to defining performance. Relevant objectives for ASC include those related to user and carer well-being, such as providing good quality care and improving quality of life (QoL); social goals, including social cohesion, community development and social inclusion; political goals, such as controlling public expenditure; and economic goals, such as improved efficiency, improved equity in the distribution of benefits and burdens and the smoother operation of markets (Knapp et al. 2001). Even with a balanced scorecard approach it is not possible to measure all of these objectives, so performance measurement systems tend to emphasise some more than others. Historically, attention has been given to political goals, such as reducing the cost of provision, and economic goals, such as efficiency (Challis et al. 2006, Clarkson 2010). The current framework places an emphasis on user and carer well-being goals (Department of Health 2014).

Focusing on the user well-being objectives, social care services attempt to do three things: (i) maximise the QoL of individuals who are not fully capable of long-term self-care, (ii) maximise the ability of individuals to live independently and (iii) prevent or slow the decline in health (Qureshi and Nicholas 2001, Netten 2011). Importantly, although the second of these goals is rehabilitative, in general social care services do not seek to improve or remedy the underlying impairment; rather they seek to compensate the person for the impact of their impairments on their physical and mental functioning. Services

[10] This is a form of rehabilitation service.

provide continuous support, often on a daily basis, over a period of many years. Most social care therefore does not produce "change" outcomes, such as those associated with lessening the person's underlying level of impairment. Instead social care outcomes are mostly "maintenance" outcomes, where the aim is to maintain the person's QoL in the context of an unavoidable deterioration in the physical and mental functioning of the service user (Qureshi and Nicholas 2001).

The ongoing nature of most social care provision combined with the often very personal nature of the care provided means that the way social care is delivered is usually considered to be important by users (Nocon and Qureshi 1996b, Raynes 1998, Bamford et al. 1999, Qureshi and Nicholas 2001, Raynes et al. 2001, Patmore 2004, Francis and Netten 2004). Qureshi and Nicholas (2001) refer to these valued aspects as "process outcomes", with analogy to the economist's concept of "procedural utility" (Ryan 1999, Frey et al. 2004, Frey and Stutzer 2005). Both terms capture the idea that people not only value the final outcome, but also the conditions and processes by which that outcome is achieved. Evidence from studies of social care services suggest users value the experience of care delivery because of the intimate nature of the care provided and the way the care worker becomes involved in the person's daily life (Nocon and Qureshi 1996b, Bamford et al. 1999, Qureshi and Henwood 2000, Qureshi and Nicholas 2001, Netten 2011, Netten et al. 2012a).

The Adult Social Care Outcomes Framework: operationalising performance and the purpose of measurement

The Department of Health (DH) describes the ASCOF as its "main tool for setting direction and strengthening transparency in adult social care" (Department of Health 2014, p. 8). Despite it being a central government-led policy, the ASCOF was "co-produced" with the social care sector in a bid to create a sense of ownership of ASCOF by the sector (Department of Health 2014). It has around 19 PIs[11] organised around four domains: enhancing QoL for people with care needs, delaying and reducing the need for care and support, ensuring that people have a positive experience of care and support, and safeguarding adults whose circumstances make them vulnerable and protecting them from avoidable harm. Of the 19 PIs in the 2015/16 Framework shown in Box 2, ten come from

[11] New PIs are being developed so the number of PIs changes on an annual basis.

social care surveys and in two domains the survey indicators are the overarching measure (see asterisked items).

In creating the ASCOF, the government of the day was keen to demarcate the ways in which this performance measurement framework was different from the one it replaced. As well as the outcomes focus and sense of sectoral ownership, the Coalition Government stressed the importance of developing a culture of "sector-led"[12] improvement in place of top-down, target-based, performance management (Department of Health 2010b, Department of Health 2010d, Department of Health 2011a). The ASCOF is therefore designed more as an "intelligence" system, where there is no fixed interpretation of the data (Hood 2007).

Box 2: Survey-based indicators in the ASCOF, 2015/16

Domain 1: enhancing QoL for people with care needs
1A. Social care-related quality of life *
1B. Proportion of people who use services who have control over their daily lives
1D. Carer-reported quality of life
1I. Proportion of people who use services and their carers, who reported that they had as much social contact as they would like.

Domain 3: ensuring that people have a positive experience of care and support
3A. Overall satisfaction of people who use services with their care and support *
3B. Overall satisfaction with social services of carers *
3C. The proportion of carers who report that they have been included or consulted in discussions about the person they care for
3D. The proportion of people who use services and carers who find it easy to find information about support

Domain 4: safeguarding adults whose circumstances make them vulnerable and protecting them from avoidable harm
4A. The proportion of people who use services who feel safe
4B. The proportion of people who use services who say that those services have made them feel safe and secure

*overarching measure

The ASCOF is largely justified in terms of the potential for the PIs to be used locally: as the DH states, the ASCOF should provide "robust, nationally comparable information on the outcomes and experiences of local people" that LAs can use in a range

[12] The 'Towards Excellence in Adult Social Care' (TEASC) programme, which is a partnership between various organisations, including the Department of Health and the Association of Directors of Adult Social Services, led by the Local Government Association, provides oversight and supports improvement (Local Government Association 2013).

of ways to support performance (Department of Health 2012, p. 5). The DH provides examples of the ways in which LAs can use ASCOF indicators locally. These include monitoring the success of local interventions, informing strategic policy and leadership, identifying and sharing best practice through regional benchmarking efforts, and strengthening accountability to the local population. These examples suggest that the DH sees ASCOF as a tool for performance improvement and accountability. While there is fairly good evidence that LAs are using the ASCOF indicators for local accountability, by publishing ASCOF indicators in annual Local Accounts[13], there is less evidence of purposeful uses, particularly of survey-based PIs as I have discussed (Heath et al. 2015). When considering how to address the effects of nonresponse and case-mix on the validity of survey-based PIs, it is important to attend to all potential uses for the measures.

The Adult Social Care Survey

Currently two surveys are used to populate the ASCOF: the Adult Social Care Survey (ASCS), which is a survey of all ASC service users, and the Survey of Adult Carers' Experiences (SACE), which is a survey of adult carers who are in contact with ASC services. In this thesis I look only at indicators from the ASCS and draw on the data from the first wave (2010-11) of this survey. I describe the ASCS methods here; details of the sample, including distributional statistics are given in the relevant empirical chapters.

The ASCS is an annual national survey of social care service users aged 18 years and over, which aims to capture information on the outcomes experienced by users from social care services. The design of the survey drew heavily on the experience of conducting the user experience surveys, which were the direct predecessors to this survey (Malley and Netten 2008, Malley and Netten 2009, Malley et al. 2010). The ASCS is primarily a postal survey and a separate sample is drawn each year, so there is no longitudinal component. The survey uses what can be described as a form of cluster-randomised sampling, whereby each LA with responsibility for social care (referred to hereafter as councils with adult social services responsibilities, CASSRs) in England selects a random sample of ASC users from their records. The sample is therefore drawn from the population of ASC users who are in receipt of full or partially publicly-funded

[13] Local accounts are accounts of the "quality and outcome priorities which the council has chosen, in consultation with its partners, and the progress it has made in achieving them during the past year" (Department of Health 2010b, para 4.5). They are tools for improving the transparency of local government to local populations.

services.[14] Having drawn the sample, each CASSR is charged with conducting the survey for its locality.

Survey design decisions (i.e. to conduct a postal survey, to cover all ASC users, to not have a longitudinal component and to have CASSRs manage the survey) present a number of challenges for performance assessment. Notably, these decisions have consequences for the consistency of data collection processes between CASSRs, the quality of the data, and the analysis of PIs for performance assessment. Choices about the ASCS design were made for a mixture of reasons, including affordability, practicality and political reasons (Department of Health 2009). Policymakers recognise that these design choices compromise the quality of the data and its suitability for performance assessment. For this reason, steps are taken within the design of the ASCS to limit inconsistencies in the management of the survey and improve data quality. Interestingly, the types of design limitations found in the ASCS are fairly common across performance surveys. For example, the NHS surveys are all postal surveys; the GP patient survey and the US Consumer Assessment of Health Patient Surveys (CAHPS) both cover diverse populations; and in the CAHPS provider organisations manage the survey process.

As part of the steps taken to limit inconsistencies and improve data quality, NHS Digital[15], which is charged with the overall management of the ASCS and the collation of results for publication[16], provides a detailed guidance document and issues various tools along with the questionnaires (e.g. NHS Information Centre 2010). The guidance covers all aspects of the survey process, i.e. ordering and formatting of the questionnaire, choosing the sample, when and how to administer the survey, including following up non-respondents and setting up a system for booking in returned forms, maximising response rates, inputting and checking the data for accuracy. It effectively sets rules around survey design and management that all CASSRs are supposed to follow.

---

[14] Eligibility is defined as "a person receiving services on 30 September who had the capacity to consent to take part in the survey" (NHS Information Centre 2010, para 14.2). Where service receipt is defined as a person who is receiving one or more services provided or commissioned by social services which are part of a care plan following a Community Care Assessment and the care being received is managed by the CASSR. Services provided or commissioned by an NHS health partner under section 75 arrangements are also included. In fact the definition of eligibility has slightly changed following the recent developments to the data collections in social care. Since the data I use all predates these changes, however, I describe the surveys as they relate to the data used in this thesis. These changes anyway do not substantially alter the key features of the sample, so this discussion is correct.

[15] NHS Digital is a DH sponsored organisation that collects and reports statistics for the NHS and social care.

[16] Initially this was the responsibility of the DH, but following a reorganisation of the DH, this and many of the statistical and analytical functions were transferred to a new arms-length body the Health and Social Care Information Centre, now known as NHS Digital.

The guidance does, however, allow CASSRs some discretion over how to implement various aspects of the survey process. The purpose of this is largely to enable them to tailor local surveys to the different needs of ASC users in their area in order to improve response rates. To capture variation between CASSRs in data collection processes, CASSRs collect and report 'paradata'[17], i.e. data about implementation of the survey. Therefore, along with the questionnaire data, CASSRs return information about, for example, the use of incentives, alternative modes of delivery for accessing hard-to-reach people, how often and with what methods nonrespondents are chased, sampling frame exclusions, survey timing, and addition and modification of questions. These paradata can be used to understand the extent of variations across CASSRs and, as I do in Chapter 4, explore the effects of variations in survey design and implementation on response rates.

Consistency between CASSRs is also achieved through the use of a standardised questionnaire[18] with a set of 24 mandatory questions. The mandatory questions include the eight-item ASCOT-SCT4 social care outcome measure (Netten et al. 2011, Netten et al. 2012a), which is used to populate the social care-related quality of life (SCRQoL) PI[19] and several other PIs in domains one and four (see Box 2). There is also a global satisfaction with services question and a question about access to information about services, which are used to populate indicators in domain three of ASCOF (see Box 2). Various other questions are included in the questionnaire to interpret the ASCOT items (Malley and Netten 2009, Malley et al. 2010), including measures of health outcomes (self-perceived health and a question on pain and anxiety/depression), functional disability (activities of daily living, ADL[20], and an instrumental activity of daily living, IADL[21]) and other needs-related characteristics of service users, such as self-perceived appropriateness of design of

---

[17] Paradata is a term coined by Couper (1998).
[18] The standard version of the ASCS and all the variant versions can be downloaded from the NHS Digital website at this address: http://content.digital.nhs.uk/socialcare/usersurveyguide1011 [21/12/2016].
[19] The eight SCRQoL items can be scored using a set of weights from a combined best worst scaling-time trade off approach to generate a social care equivalent to the QALY (Netten et al. 2011, Potoglou et al. 2011, Netten et al. 2012a). Since these weights were not available when the ASCS and ASCOF were being developed, in the ASCOF the response options are scored from zero (the worst response) to three (the best response) with the final score being the sum of all items (Department of Health 2011a).
[20] The exact ADL questions include: ability to get around indoors, get in/out of a bed/chair, manage to feed self, deal with finances and paperwork, manage to wash all over using either a bath or shower, manage to get dressed / undressed, manage to use toilet / WC, and manage to wash face and hands. These are the Katz (1963, 1970) ADLs with the omission of the continence ADL, which for ethical reasons cannot be asked in a self-completion format.
[21] The inclusion of the manage finances/paperwork IADL (Lawton and Brody 1969), is designed to capture cognitive and intellectual impairment.

the home, self-perceived ability to get around the local area and regular receipt of practical help. In addition, questions are asked to capture differences in reporting, including questions asking about whether the questionnaire was completed by a proxy respondent, or with assistance and the source and type of assistance offered. There is also a question asking whether the person purchased additional support using private funds. I use these data to explore the effects of case-mix on performance assessment in Chapters 6, 7 and 8.

The diversity of the ASC population means that if the aim is to be inclusive and to ensure that, as far as is possible, the service user responds to the questionnaire without any help, a single questionnaire is not appropriate (Qureshi and Rowlands 2004, Malley et al. 2010, Malley and Fernández 2010). While the focus of the questions on social care outcomes means that they are relevant to all ASC service users, the questionnaire requires some modification to make it accessible to people with learning disabilities and to accommodate a difference in the meaning of key terms like 'home' in alternative caring environments. Consequently, there are four different versions of the questionnaire: a standard version, a care home version, an Easy Read version for people with learning disabilities and an Easy Read version for people with learning disabilities living in an institutional care setting (Malley et al. 2010, The NHS Information Centre 2010).

While the care home version does not have any substantial differences from the standard version, the Easy Read version is quite different. Although it covers the same content, the language is changed throughout to be simpler, in some places there are fewer response options and there are illustrations to aid understanding. Despite the differences between the Easy Read and standard version official publications combine all responses (see e.g. Department of Health 2014). It is clearly necessary to make some fairly strong assumptions about the correspondence of questions and response options to combine responses across versions. It is beyond the scope of this thesis to critique this practice. I do, however, reflect on the challenges the different versions (and sample diversity) pose for case-mix adjustment and how I overcome this in Chapters 6, 7 and 8.

Along with responses to the questionnaire items (available only for respondents), CASSRs also provide 'auxiliary data' from their records for each sample member, irrespective of response status. The variables available from CASSR records are limited, but include socio-demographic variables (such as sex, age group, ethnicity and so on), service receipt (e.g. residential care, home care, and so on), client group and the 'personal

budget' provided to the user which captures the total cost[22] of the service package. These data provide some basic information about nonrespondents and can be used to explore who is missing from the ASCS, as I do in Chapter 4, and the effects of adjusting for nonresponse on performance assessment as I do in Chapter 5.

Outline of the thesis

This thesis is organised as follows. The two chapters subsequent to this one provide the theoretical framework for the thesis and set out the methods that I will use in the empirical chapters. In Chapter 2, I develop a conceptual framework for analysing nonresponse and the effect of case-mix on performance that is tailored to the characteristics of ASC and to the particular features of performance surveys. In Chapter 3, I describe the tools for the analysis, including the methods used for addressing case-mix, for addressing nonresponse and for examining the effects of adjustment for both case-mix and nonresponse on inferences about performance.

Chapters 4 to 8 are empirical and map onto the sub-questions for the research set out above. Thus, Chapter 4 addresses the question of who is missing from performance surveys and how response rates can be improved (sub-question iii). In Chapter 5, I explore what analytical approaches can be used to address nonresponse to performance surveys and the effects of applying these analytical approaches on inferences about performance (sub-question iv). The three subsequent chapters look at the effects of case-mix. Chapters 6 and 7 explore the suitability of different methods for modelling social care outcomes (sub-question i). Chapter 6 looks at a variety of different specifications for models of social care outcomes within the familiar risk-adjustment framework; Chapter 7 compares the risk-adjustment models for social care outcomes to models based on productions relations theory. In Chapter 8, I assess the effect of adjusting for case-mix on inferences about performance and consider whether the case-mix adjustment method makes a difference to conclusions about the relative performance of CASSRs (sub-question ii).

Finally, in Chapter 9, I return to the core research question and policy questions raised by this thesis. I summarise the main findings from Chapters 4 through 8 in light of the core research question and consider the limitations of the analyses for the conclusions

---

[22] This is not meant to be an estimate of the economic cost of services, rather the price that is charged by the provider for the services provided. It is intended that this would include both the amount paid by the user (where user charges are incurred on service provision) and that by the council. In instances where the individual has a personal budget in place, this would be the total assessed, and agreed, budget for that person.

from this study. In the final section of this chapter, I conclude by discussing the implications of this thesis for policy and for research, more generally. I also offer some directions for future research.

The appendices contain more detail concerning the methods and extended analyses. They are provided for completeness only.

Chapter 2
A Conceptual Framework for Addressing Case-Mix and Nonresponse When Using Survey-Based Indicators to Assess Performance

Introduction

There are many sources of observed and unobserved heterogeneity in survey-based indicators that may confound inferences about the performance of organisations. This thesis is concerned with understanding the consequences for performance assessment of two such sources of heterogeneity – nonresponse and differences in the case-mix of people served by organisations – and finding strategies to address them in order to improve the validity of the PIs. Statistical analysis is the main tool used in this thesis to explore the effects of nonresponse and of case-mix on performance assessment. The purpose of this chapter is to outline the conceptual framework I use for the empirical analysis.

I draw on different disciplines to outline a conceptual framework for each of these two sources of heterogeneity in survey-based PIs. To understand the importance of variations in the case-mix of the population served by CASSRs on PIs, I draw on the economic theory of production relations, as adapted for the social care context in the production of welfare framework (Davies and Knapp 1981). This theory provides insight into how factors both within and beyond the control of CASSRs are likely to influence variations in social care outcomes. To explore the role of nonresponse in performance assessment, I draw on theories of survey participation from the sociological and psychological literature. These theories set out the factors that influence survey participation so providing a conceptual framework for exploring participation in surveys used for performance assessment.

The chapter is organised in three sections. In the first section, I review the three ways in which performance has been measured in surveys. Although this topic deviates from the purpose of this chapter, it is important context for the conceptual framework that explores the role of variations in the case-mix of people served by CASSRs. This is because the different types of indicators capture different aspects of well-being and, as I argue, have different consequences for the analysis of the effects of case-mix on PIs. In the second section, I set out the conceptual framework for exploring variations in social care outcomes. In the third section, I outline a conceptual framework for exploring participation in surveys used for performance assessment.

A note on the measurement of performance using survey-based indicators

How performance is measured within survey measures has important consequences for the types of factors that need to be considered for case-mix adjustment. Despite the 'outcomes' moniker, the ASCOF and its health and public health counterparts actually include a range of different types of indicators. In the ASCOF, there are broadly three different types of survey-based indicators – satisfaction measures, so-called patient-reported experience measures (PREMs), and so-called patient-reported outcomes measures (PROMs)[23]. The measures have different conceptual roots. Consequently, each measure provides an alternative perspective on user well-being and, by extension, focuses on a separate aspect of social care outcomes. I therefore reflect briefly on what is being measured by each type of indicator and consider the consequences for performance assessment.

Satisfaction measures

The rationale for using customer satisfaction as a measure of performance has its roots in marketing theory (Neely et al. 1995, Sitzia and Wood 1997, Heinrich 2003, Kelly 2005). For services sold in the private sector, it is argued that the concept of customer satisfaction replicates "the virtues of the marketplace" (Gormley and Weimer 1999, p. 198 quoted in Heinrich 2003). Dissatisfaction leads to exit, a loss of custom, revenue and, therefore, profits among lower quality providers. Consequently, customer satisfaction reflects performance, as better organisations will be better at satisfying their customers with greater efficiency and effectiveness than their competitors (Neely et al. 1995).[24] It follows naturally from this that the 'bottom line' for managers should be to increase customer satisfaction since it is indicative of customer market behaviour (Kelly 2005).

In the public sector, however, this logic is found wanting, because there is usually no single 'customer'. The customer is instead a hybrid of the service user/consumer and citizen/payer (Dixit 2002, Kelly 2005). Market failure means that individuals who would be expected to use public services often lack the knowledge or economic means to access them. The government steps in in such circumstances to correct inefficiencies in the market, with the consequence that many public services are either entirely funded or subsidised through

---

[23] The concept of 'patient-reported' is not that appropriate in the social care context given that users of services are not referred to as patients. Given the familiarity of the term, and to save coining a new less familiar term to describe the same idea, I use this term for ASC.

[24] The terms effectiveness and efficiency are used precisely to refer, in the case of the former, "to the extent to which customer requirements are met", and, in the latter, to "how economically the firm's resources are utilized when providing a given level of customer satisfaction" (Neely et al. 1995, p. 1228).

taxation. This means that most users of public services do not pay directly for the services they consume. Depending on the nature of market failure and the nature of the service, public services are also often distributed on the basis of needs and/or means, so that at any one time many of the (tax-)payers are not consuming the services. While consumers and citizens are both legitimate stakeholders, only consumers have direct experience of the service. Questioning citizens about their satisfaction with services is likely to elicit "non-attitudes", where people express opinions about subjects they know little about (Converse 1970, Stipak 1979). Non-attitudes are of little use from a performance improvement perspective. For performance management it is therefore more appropriate to focus on *consumer* satisfaction. Indeed, in areas of the public sector where services are distributed according to need, such as the health, social care and education sectors, the focus of research has been consumer satisfaction (Griffiths 1988, Fitzpatrick 1991, Carr-Hill 1992, Williams 1994, Sitzia and Wood 1997, Cleary 1998, Qureshi and Rowlands 2004, Clemes et al. 2008).

The shift from customer to consumer, however, leads to the unravelling of the logic that equates satisfaction with overall performance. Instead of measuring performance in public sector services, satisfaction measures now tend to be treated as measures of quality. Empirical research supports this interpretation, showing that satisfaction is influenced by a number of factors including expectations and perceptions of quality (Linder-Pelz 1982, Parasuraman et al. 1985, Erevelles and Leavitt 1992, Oliver 1997, Sitzia and Wood 1997). This makes it hard to propose that satisfaction should be singularly pursued, as what increases consumer satisfaction – namely improved service quality – is not always in the best interests of citizen-taxpayers (Kelly 2005). Where managers use satisfaction to improve performance, they need to balance improvements against other factors, such as costs. Due in part to the lack of conceptual clarity, satisfaction surveys have declined in use, particularly in the health and social care sector (Cleary 1999, Coulter 2006, Elwyn et al. 2007). Many health and social care surveys, however, continue to include single items asking about global satisfaction with services. As discussed in Chapter 1, the ASCS has such a question which is used to populate the user experience domain of ASCOF (see Box 2).

Patient-reported experience measures (PREMs)

PREMs are measures of the consumer's experience of service quality and include measures such as users' perceptions of promptness of the service and the behaviour of staff towards them. They became popular in performance management circles as a consequence of the influence of management gurus like Deming (1986) who called upon managers to strive for

and to measure quality (Pfeffer and Coote 1991, Kirkpatrick and Martinez Lucio 1995a, Bovaird and Löffler 2003, Talbot 2010). The nature of service goods means that the consumer experience is seen as critical to understanding quality. The simultaneous consumption and production and the high labour content, mean they have few "search" properties, or tangible aspects of quality that can be easily observed, measured and therefore known *ex ante*. Instead service goods are replete with "experience" properties, which can only be determined subsequent to purchasing the good or during consumption and require first-hand contact to establish the most valued characteristics (Nelson 1970, Parasuraman et al. 1985). From a performance management perspective, it is argued that PREMS are superior to satisfaction measures because they capture the distinct aspects of the quality of service delivery and require little analysis to guide performance improvement (Cleary 1999, Coulter 2006, Elwyn et al. 2007).

PREMs were popular in the UK during the early 2000s when the national performance surveys were first being developed. They formed the basis for the content of the health and social care surveys (Jenkinson et al. 2002b, Qureshi and Rowlands 2004, Coulter 2006). Although various generic instruments are available for measuring service quality, such as the SERVQUAL[25] and SERVPERF (Parasuraman et al. 1991, Cronin Jr and Taylor 1992, Donnelly et al. 1995, Wisniewski 1996, Cleary 1999, Coulter 2006), these instruments have been criticised for failing to capture important aspects of the service, thus undermining their value (Buttle 1996, Shiu et al. 1997). For this reason, in health and social care, service-specific instruments have been developed with engagement from patients and service users to define the elements of the service that matter (Jenkinson et al. 2002a, Qureshi and Rowlands 2004, Darby et al. 2005, Sangl et al. 2007, Mead et al. 2008, Triemstra et al. 2010). Consequently, there are many instruments that are tailored to the specific features of distinct service forms. This is seen as a limitation of PREMs as it makes it virtually impossible to compare quality across different forms of service provision and therefore to survey diverse populations, who may receive many types of services. The current ASCS and ASCOF include only one question on the experience of service delivery (see Box 2, Chapter 1), which was included only because it was of relevance to all ASC users (Malley and Netten 2008, Malley and Netten 2009).

[25] Although this measure uses a satisfaction framework, it does claim to be a measure of service quality. The service quality 'gap' is understood as the difference between expectations and satisfaction.

Patient-reported outcome measures (PROMs)

PROMs are typically multi-attribute measures derived from responses to a battery of standardised questions about various aspects of QoL. They first emerged in the context of health services research, where they are more commonly referred to as multi-attribute health status or health-related QoL measures (Black and Jenkinson 2009). PROMs were developed for the purpose of evaluating health interventions, where what mattered was the effect the intervention had on health symptoms and QoL more broadly (Valderas et al. 2012, Black 2013, Jenkinson and Fitzpatrick 2013). As the UK government embraced outcomes-based management in the mid-2000s, it started to rethink performance measurement frameworks and consider how they could be more outcomes-focused. PROMs were adopted by government to populate the emerging outcomes frameworks. In social care, the ASCOT measure (Netten et al. 2011, Netten et al. 2012a) was chosen as the ASCOF PROM (Malley and Netten 2009). It also feeds all the other domain one ASCOF QoL indicators (see Box 2, Chapter 1). The majority of the ASCOF survey-based indicators are of the PROMs-type.

A key problem with using PROMs for performance assessment is that they generally do not capture the outcome of treatment; rather they measure the QoL-outcome state of the person (Black and Jenkinson 2009, Black 2013). Where such measures are used as part of a research study to evaluate an intervention, it is the experimental design and subsequent analysis that provides the basis for drawing inferences about outcomes. The choice of measure is important insofar as it is sensitive to the anticipated effects of the service. This is by no means a trivial matter; there are important debates to be had over the value of 'generic' measures like ASCOT, which have broad applicability as they are designed to capture the effect of a broad range of ASC services, over 'specific' measures such as the PDQ-39 (Jenkinson et al. 1997), which is designed to be sensitive to the effect of treatments for Parkinson's disease. This question is, however, qualitatively different to the challenge of determining the effectiveness of services from observational survey data, particularly where only cross-sectional data is available as is the case for ASC. The primary problem is to find an appropriate way to account for the effect of pre-existing conditions (or 'needs-related characteristics' using terminology that is more appropriate for the ASC context) and other characteristics that affect the service user's capacity to benefit from services that does not introduce bias into the estimates of relative organisational performance. This is critical since the benefit of PROMs over other measures is premised on the idea that they capture the effectiveness of services.

The need to estimate effectiveness in addition to addressing the challenge of drawing causal interpretations about the relative performance of organisations complicates the use of PROMs for performance assessment purposes (Black and Jenkinson 2009, Black 2013). Structural models, which set out the theoretical relationships between needs-related characteristics, social care inputs and social care outcomes *a priori*, provide a route to estimate effectiveness for PROMs. From such structural models, it is possible to develop statistical models that can provide estimates the effectiveness of social care that are sufficient to draw "policy-relevant conclusions" regarding effectiveness provided the statistical models are internally valid (Cartwright 2007, Clarke et al. 2015). In the next section, I describe how the Production of Welfare (POW) model (Davies and Knapp 1981, Knapp 1984) can be modified for this purpose.

A theoretical model for exploring the effectiveness of social care and its application to the ASCS data

My purpose here is (i) to set out the theoretical relationships between needs-related characteristics, social care inputs and social care outcomes to inform the statistical modelling of effectiveness and (ii) to identify the sets of factors that need to be considered to address variations in case-mix between organisations. I draw on the POW framework for this purpose. Although the POW framework is grounded in the economic theory of production relations, which seeks to understand the relationship between the factors of production (i.e. inputs such as labour and capital) and the product of the productive process, it can serve as a framework for understanding the broader set of relationships described above. This is because the POW framework builds on theoretical insights and empirical evidence from many distinct fields of social science to provide a theory of production relations for adult social care (Davies and Knapp 1981, Knapp 1984). In particular, Davies and Knapp (1981) draw on this literature to set out how factors exogenous to production, i.e. the needs-related characteristics that are relevant for addressing case-mix, influence the production of social care outcomes at a given point in time. The key causal and non-causal relationships, as conceptualised within the POW framework, are illustrated in a simplified form in Figure 1.

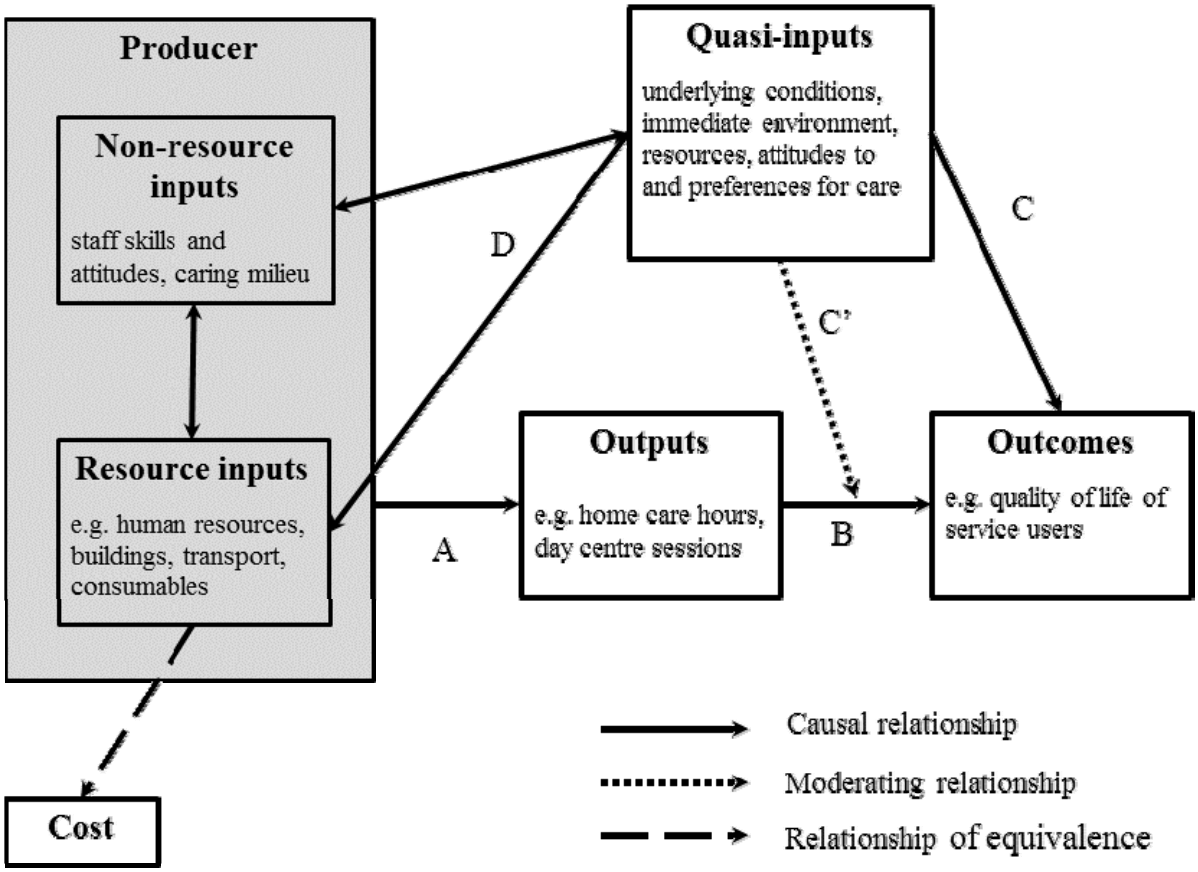

Figure 1: The production of welfare framework

Adapted from Knapp (1984) and Fernandez (2005)

The main message of Davies and Knapp (1981) is that in the ASC context the relationship between the factors of production and outcomes is influenced to a large extent, and in several different ways, by a variety of relatively intangible factors over which the producer (administrator or policy-maker) has limited, and in some cases no, control. To articulate this point, they distinguish three kinds of inputs, referred to as "resource inputs", "non-resource inputs" and "quasi-inputs". *Resource inputs* correspond to the conventional inputs or factors of production, over which the producer has a high degree of control. *Non-resource inputs* are the intangible factors, such as staff attitudes and the caring milieu in a care home, over which the producer has less control than is the case for resource inputs. They determine the quality of the service, e.g. aspects like its institutional or homelike feel, but cannot be easily measured. *Quasi-inputs* refer to other intangible factors such as the underlying health conditions of service users and their care preferences that are largely exogenously determined. As illustrated in Figure 1, the role quasi-inputs have in moderating and determining the relationship between resource inputs and care outcomes is central to the POW framework. Indeed, it is a basic premise of the approach that outcomes "are

determined by the levels and modes of combination of the resource and non-resource inputs (which are mainly under the control of the administrator or policy-maker, albeit sometimes only after the elapse of a considerable period of time), given the exogenously determined values of the quasi-inputs" (Davies and Knapp 1981, p. 8).

We can draw several important observations from Figure 1 that are relevant to the task of estimating effectiveness from cross-sectional data on the observed outcome state and addressing variations in case-mix. First, it suggests that we can think of the effectiveness of ASC services as having a quality (non-resource inputs) and a quantity (resource inputs) component for each 'technology', i.e. home care, day centres, etc. We can therefore write the effectiveness of ASC interventions, E, as given by the function,

$$E = f(\boldsymbol{x}_i, \boldsymbol{q}_i; \boldsymbol{\beta}_1) + \varepsilon_i, \qquad (1)$$

where $\boldsymbol{x}_i$, is a vector describing the quantity of each technology provided and, $\boldsymbol{q}_i$, is a vector describing the various components of care quality. This relationship is reflected in the arrows A and B in Figure 1, which capture the conventional production process, i.e. the relationship between the factors of production (including non-resource inputs) and outcomes.

Second, quasi-inputs exert both a causal (arrow C) and moderating (arrow C') effect over the relationship between social care inputs and social care outcomes, potentially confounding the relationship between inputs and outcomes and by extension inferences about the performance of organisations. The role of quasi-inputs can by understood by considering how the level of functional ability of a person receiving care affects their likely outcomes. Arrow C represents the fact that a person with greater functional ability is likely to achieve better outcomes from care than someone with low levels of functional ability despite potentially less service input simply because they are starting from a better position. This is illustrated in Figure 2 by the difference AB in the absence of any service input, which, were functional ability not to be accounted for, would be wrongly attributed to the effect of services. Arrow C' captures the fact that "exactly identical resources, service configurations and caring environments will affect different people in different ways" (Knapp 1984, p. 32). This has been demonstrated empirically, with studies showing that returns from increasing inputs start diminishing for more functionally able individuals at lower levels of input than for less functionally able people (Davies et al. 2000b, Fernández 2005). This effect is illustrated in Figure 2 which shows a less steep gradient for the high functional ability group

compared to the low functional ability group at the same level of service input. This represents differences in the marginal productivity of services for different user groups.

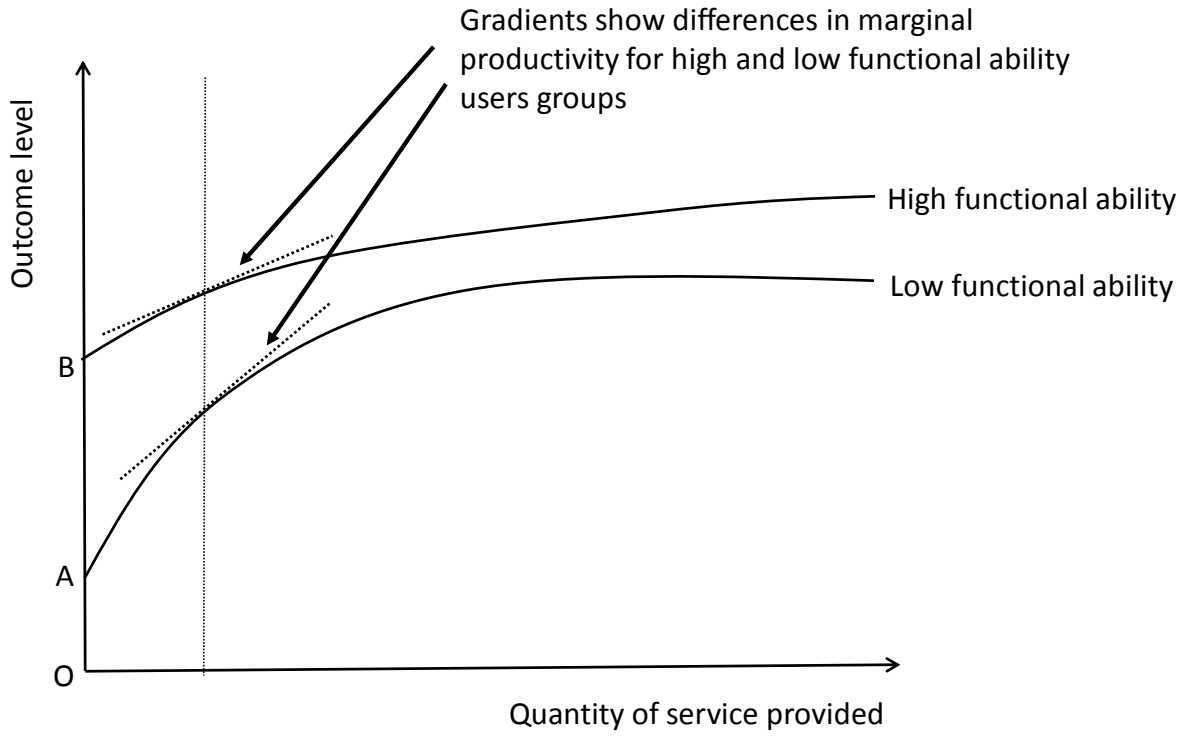

Figure 2: Example showing the effect of functional ability on the relationship between resource inputs and outcomes

Consequently, the QoL-outcome state that we observe for individuals in the cross-sectional ASCS data using the ASCOT PROM can be thought of as a function of the effectiveness of social care, the impact of quasi-inputs and the relationship between quasi-inputs and effectiveness. Building on equation (11), a model for the observed outcome state (as gathered from the ASCS data) can be written as follows,

$$y_i = \alpha + f(\boldsymbol{x}_i, \boldsymbol{q}_i; \boldsymbol{\beta}_1) + \boldsymbol{\beta}_2 \boldsymbol{z}_i + f\big((\boldsymbol{x}_i, \boldsymbol{q}_i).\boldsymbol{z}_i; \boldsymbol{\beta}_3\big) + \varepsilon_i, \qquad (2)$$

where y_i is the observed QoL-outcome state, the term $f(\boldsymbol{x}_i, \boldsymbol{q}_i; \boldsymbol{\beta}_1)$ represents the effectiveness of care services, and expresses it as a function of a vector of care technologies with quantity, $\boldsymbol{x}_i$, and quality, $\boldsymbol{q}_i$. Quasi-inputs are represented in vector form by the term, $\boldsymbol{z}_i$, with the vector of coefficients, $\boldsymbol{\beta}_2$, representing the direct effect of these factors on the observed QoL-outcome state. The term, $f\big((\boldsymbol{x}_i, \boldsymbol{q}_i).\boldsymbol{z}_i; \boldsymbol{\beta}_3\big)$, is an interaction between the

effectiveness of care services and quasi-inputs, reflecting the marginal productivity of services for different user groups. The independently-distributed error term is given by ε_i. This equation in (2) suggests that if we want to avoid bias from case-mix in the estimation of performance using the observed QoL-outcome we need to control for quasi-inputs, $\boldsymbol{z}_i$.

The third observation of importance is the effect that quasi-inputs have on the choice of resource inputs, as illustrated by arrow D. In England the specifics of the care package are determined through a needs assessment, which is usually carried out by a social worker employed by the CASSR. The aim of the assessment is to identify the social care needs of the individual and develop an agreed vision of the outcomes the individual would like to achieve in a care plan (Glendinning et al. 2008b). In general more services are given to people with greater needs, but interventions are also "customised", like other service products, to meet the consumer's requirements and preferences for types of services, timing and so on (Lovelock and Gummesson 2004, Ferguson 2012). Customisation is often referred to as 'personalisation' in the UK health and social care sector and the extent to which services are personalised to meet people's individual preferences is considered to be a defining feature of the quality of the intervention, reflecting the skills of the social worker (Lymbery 2010, Netten et al. 2012b). This therefore suggests that the adequacy and mix of technologies reflects a further aspect of the quality of provision, namely, to adapt a phrase from Davies and Challis (1986), the ability to match resources to needs and preferences.

The role of quasi-inputs in determining the choice of resource inputs means that *the very same factors that determine a person's capacity to benefit from services also determine the care package they receive*. Consequently, resource inputs and quasi-inputs are likely to be highly correlated, i.e. $Corr(\boldsymbol{x}_i, \boldsymbol{z}_i) \neq 0$. This has two consequences for the estimation of effectiveness and adjustment for case-mix. First, it means that a standard risk-adjustment model that estimated the QoL-outcome state simply as a function of quasi-inputs, would underestimate the effectiveness of social care and underestimate the negative effect of quasi-inputs on QoL-outcomes. Since CASSR policies regarding the allocation of services and market management also govern the way social workers match resources to the needs and preferences of people the degree to which the effectiveness of social care is underestimated may vary across CASSRs so biasing estimates of relative performance. For this reason, in Chapter 3 I suggest that it may be preferable to include an indicator capturing the quantity of care package received in the risk-adjustment model changing it into a 'production function' model. While this solves the underestimation of effectiveness, and adjusts for the

confounding effect of observable quasi-inputs on outcomes, it brings a new problem of endogeneity bias due to the confounding influence of unobservables. The researcher cannot observe all the factors that social workers consider in their needs assessment, meaning that some unobserved needs-related factors, which are also correlated with quantity of the care package delivered, will be in the error term (Forder and Caiels 2011b, Forder et al. 2014). These omitted variables result in endogeneity bias in the estimate of the service effect because the unobserved needs-related characteristics of cases in the counterfactual groups (i.e. receiving a different quantity of service) may be different to those cases in the treatment group (i.e. receiving the quantity of service, X_i). I discuss this issue further in Chapter 3, where I develop the empirical approach for the modelling of social care outcomes in Chapters 6 and 7 using the ASCS data.

Before moving on to consider the sets of variables that are exogenous to the production of social care outcomes and should therefore be considered as quasi-inputs in ASC, I reflect on the role of external conditions and why they are omitted from Figure 1. By external conditions, I am referring to the wider social, political and economic environment that has a bearing on the behaviour of producers. Often referred to as 'external constraints', these factors have been shown to exert an important influence over LA performance (Andrews 2004, Andrews et al. 2005, Haubrich and McLean 2006, McLean et al. 2007, Gutiérrez-Romero et al. 2008, Gutiérrez Romero et al. 2010). In the ASC context, economic factors such as local market conditions, and competition from other sectors pushing up the price of social care labour, have been shown to affect the ability to recruit and retain suitable staff (Knapp and Smith 1985). Environmental factors, including spillover effects from neighbouring organisations have also been shown to affect LA performance (Revelli 2006, Moscone et al. 2007, Fernandez and Forder 2015). External conditions are excluded from Figure 1 partly for the sake of simplicity, but also because they enter into the production process as constraints on the choice and availability of resources. Including these factors in an empirical model would complicate the estimation of the relationship between inputs and outcomes. From a policy perspective it is more informative to assess the effect of these area-level factors on individual outcomes in a second stage, having established the relationship between inputs and outcomes.

The sets of exogenous factors influencing social care effectiveness as applied to the ASCS data

As Davies and Knapp (1981) are careful to point out the critical question of which factors are endogenous or exogenous to the investigation depends on your point of view and the time scale of investigation. In ASC, the long-term nature of most care complicates the decision over which factors are exogenous and which are endogenous. Much research that seeks to model PROMs for the purpose of performance assessment uses the concept of 'pre-existing conditions' or 'baseline circumstances' to account for factors that put people 'at risk' of worse outcomes, but are outside the control of the service (Smith and Street 2013, Nuttall et al. 2015). The concept is of limited use in ASC because people are likely to have 'presented' to the service many years ago and to have experienced numerous life events during that time that will have impacted on their QoL, but have nothing to do with the efficacy or otherwise of ASC, e.g. the death of a spouse. Instead in ASC the concept of 'social care need' is often used to differentiate the QoL-outcome from antecedent factors (Davies and Knapp 1981, Netten et al. 2012a).

Distinguishing social care needs from social care outcomes, is not straightforward, as many 'needs-related characteristics' can also be considered outcomes from care in certain circumstances. I use the approach taken in the development of the ASCOT measure of social care outcomes. ASCOT adopts a conceptualisation of QoL based on a large body of research into 'lay theories' of what outcomes matter to ASC users to propose a measure of social care-related quality of life (SCRQoL) (Nocon and Qureshi 1996a, Nocon and Qureshi 1996b, Qureshi et al. 1998, Bamford et al. 1999, Qureshi and Nicholas 2001, Miller et al. 2008). Eight domains emerged from this work, which are: safety, personal cleanliness and comfort, food and nutrition, accommodation cleanliness and comfort, control over daily life, social participation and involvement, occupation and dignity. Three groups of factors – the underlying health and disability of users, the physical environment and financial well-being – that were identified in the literature as important for good QoL were excluded from the SCRQoL concept. Netten et al. (2012a) argued that these aspects are not core outcomes of social care but were more closely related to social care need, since they are taken into consideration by social workers in the needs assessment. These three 'needs-related' factors, along with two other sets of factors, are considered here to be quasi-inputs to the production process. I discuss each of these groups in turn and their relevance for case-mix adjustment. The first group of factors are the *underlying health and disabling conditions* of the person, which directly affect their QoL and constrain their capacity to benefit from services.

Functional disability, as measured by ADL scales, is considered to be the core indicator of social care need. Research shows that ADLs are important predictors of QoL and they are regarded as critical to understanding variations in the QoL of ASC service users (Davies and Knapp 1981, Hellström and Hallberg 2001, Kane 2001, Kane et al. 2003, Hellström et al. 2004, Degenholtz et al. 2006, Glendinning et al. 2008a, Vaarama 2009, van Leeuwen et al. 2014, Shippee et al. 2015). Health conditions and health status, including mental health problems such as depressed mood and symptoms such as pain, have also been shown to be important predictors of QoL in ASC populations (Hellström et al. 2004, van Leeuwen et al. 2014) and in older populations more generally (Bowling et al. 2002, Bowling and Gabriel 2007). Factors with these effects are often referred to as 'risk' factors in the healthcare literature (see e.g. Iezzoni 2013) as they describe the pre-existing characteristics of individuals that affect the likelihood of a good outcome.

Where care is provided in private households, the *internal (home) and external physical environment* have important effects on social care need and are the second set of quasi-inputs. Research has shown that homes with poor accessibility exacerbate the effects of health conditions and frailty on functional disability (Iwarsson et al. 2007), while more accessible homes can lessen functional dependence and in some instances reduce the quantity of ASC and informal care inputs (Connell et al. 1993, Trickey et al. 1994, Gitlin et al. 1999, Gitlin et al. 2001, Heywood and Turner 2007, Tanner et al. 2008, Hwang et al. 2011). The suitability of housing and the external physical environment has also been shown to be related to QoL-outcomes (Gitlin et al. 1999, Heywood 2004, Heywood and Turner 2007, Nygren et al. 2007, Hutchings et al. 2008, Vaarama 2009, van Leeuwen et al. 2014), including the person's sense of safety and control (Scheidt and Norris-Baker 2003) and ability to sustain social relationships (Sixsmith and Sixsmith 2008, Tanner et al. 2008). Decades of research have also shown the importance of the physical environment for social care outcomes in residential care (Davies and Knapp 1981, Lawton 1983). In this setting, however, the physical environment is considered part of the fabric of the service – an aspect of quality and therefore not a quasi-input.

The *personal characteristics and motivations* of the service user form the third group of quasi-inputs. In ASC, all services users will 'co-produce' to a lesser or greater extent their own outcomes (Baldock 1997, Byford and Sefton 2003). The service user may sometimes be constructive and sometimes obstructive and such behaviours could be influenced by the skills and character of care staff. The less recent literature tends to conceptualise this type of involvement from users in terms of personality types. For example, research has identified

the importance of an optimistic and pessimistic attitude in determining care outcomes (Lawton 1983, Bowling et al. 2002, Kane et al. 2003). But this theoretical approach downplays the potential for staff to engender more constructive involvement from service users. More recent literature considers the involvement of users in terms of concepts like 'self-care ability', which can be measured by 'patient activation status'. Patient activation status has been shown to be a strong predictor of patient experience in the USA (Heller et al. 2009) and related to health-related outcomes (Greene and Hibbard 2012). The more actively involved an individual is in their care the more likely it is that their behaviours, attitudes and values will affect the outcomes from a given intervention, confounding the relationship between inputs and outcomes (Byford and Sefton 2003, Trukeschitz 2011). Where service users hold Direct Payments or Personal Budgets and there is so-called "deep" personalisation (Leadbeater 2004), the role of service users in co-designing the components and shape of the intervention is likely to further confound this relationship, as the form of inputs is no longer determined by social care staff.

The fourth group of factors consist of *resources arising from "social capital" and "economic capital"* (Bourdieu 1986). Social capital can be understood as the intra- and extra-familial social networks of ASC users and economic capital can be thought of as the financial resources available to the individual. These sources of capital can be drawn upon when required, for example to 'buy-in' help or to provide additional sources of help, to improve the productivity of the household unit. Importantly, my interest here is not in the relationship between social or economic capital and QoL, as has been the focus of much of the broader QoL literature (Bowling 1995, Bowling and Windsor 2001, Bowling et al. 2002, Smith et al. 2004). It is not the availability of these sources of capital that confounds inferences regarding the effectiveness of services, but the additional resources derived from them. These additional resources are drawn into the production process and substitute for or complement formal services.

It is useful to consider the example of how persons from intra- and extra-familial networks may become involved in the production process in more detail to understand the issues in the ASC context. While ASC is ostensibly provided to the disabled or frail individuals to address their needs, decisions about provision of ASC in fact take place at the household level. As Netten and Davies explain:

> once an individual's ability to function and produce commodities is reduced, other people, both in and outside the household, may get drawn in to substitute or provide additional basic commodities for that individual and household. The form that this extension of the productive unit takes will depend on the individual's role in the

> production process previous to disability, and the relationship, proximity and knowledge of potential contributors or carers. Once people contribute substantially, be it by direct production of commodities for the household or management of production, they become part of the extended productive unit or informal care network. (Netten and Davies 1990, p. 338).

Where formal care is provided alongside informal care, there is likely to be fluidity between the tasks completed by informal and formal carers[26] as they respond to the fluctuating and competing needs of the individual and the wider household (Netten and Davies 1990). An important consequence of this is that the informal care network acts as an additional input to care outcomes, confounding the relationship between inputs and outcomes.

A fifth set of factors applies where survey data are used to measure outcomes and I refer to them as *reporting-related factors*. This group of factors reflects the fact that the responses people give to survey questions can be influenced by the way in which the survey is delivered. Survey research has demonstrated clear effects from mode of administration (Dillman et al. 2009). Research has also shown that proxy respondents evaluate QoL differently to how respondents evaluate their own QoL (Epstein et al. 1989, Pierre et al. 1998, Andresen et al. 2001, Perry and Felce 2002, Janssen et al. 2005, Schmidt et al. 2010) and where respondents are assisted to answer the questionnaire this can also affect responses (Elliott et al. 2008). The nature of the ASC population means that a large proportion of the population receive help to answer the questionnaire (over 55 per cent in the 2010-11 ASCS) and there is a substantial proportion of people whose responses are given by a proxy respondent (over five per cent in the 2010-11 ASCS) (The Information Centre for Health and Social Care 2012b). Although not quasi-inputs in a strict sense, such reporting-related factors form a further set of variables that can affect the interpretation of QoL-outcomes.

In summary, the five types of factors that should be considered when adjusting QoL-outcomes for case-mix are: personal characteristics and motivations, underlying health and disabling conditions, immediate physical environment, additional resources arising from social and economic capital and reporting-related factors. I now move on to discuss the conceptual framework for survey participation.

---

[26] There is a large body of evidence demonstrating that informal care substitutes for formal care (Lo Sasso and Johnson 2002, Van Houtven and Norton 2004, Van Houtven and Norton 2008, Bonsang 2009), although less evidence of formal care substituting for informal care (Hanley et al. 1991, Davies et al. 1998, Pickard 2012).

Theories of survey participation and their application to the Adult Social Care Survey

Survey nonresponse is a special case of missing data, with a large theoretical and empirical literature dedicated to understanding the reasons why people choose to participate in surveys. This literature tends to treat nonresponse probabilistically, as a function of characteristics of the potential respondent, the various features of the survey and the relationship between these different aspects (Groves and Cialdini 1991, Groves et al. 1992). A theoretical understanding of these relationships is necessary to guide the modelling of nonresponse to surveys. Such models are valuable not only for adjusting survey estimates at the analysis stage (e.g. Rubin 1976, Rubin 1987, Little and Vartivarian 2003, Horton and Kleinman 2007) but also for informing strategies to improve response rates (see e.g. Groves et al. 2000, Dillman et al. 2009). In the following discussion I review the main theories of nonresponse and explore the extent to which these can be used to explain participation in performance surveys. My aim is to develop a model that can be used to explore nonresponse to the ASCS.

Theories of survey participation

A number of theories seek to explain survey participation. Prominent examples are social exchange theory (Dillman et al. 2009), which proposes that a rational calculation of social costs and benefits occurs when people decide to participate in a survey; cognitive heuristics (Groves et al. 1992), which by contrast explains participation decisions in terms of short-cuts in thinking based on factors such as liking a field worker; and social disorganisation (House and Wolf 1978, Goyder et al. 1992, Couper and Groves 1996), which asserts that in areas characterised by social disorganisation (for example with high residential mobility, ethnic heterogeneity and low socioeconomic status) there are weakened social ties and a resulting weakening of cultural norms, such as being helpful and having trust in institutions, that leads to lower response rates. All of these theories are supported by evidence from empirical research into the factors explaining nonresponse.

As Goyder (2006) observes, the theoretical literature is not well integrated, so he proposes bringing the theories described above together within the framework of social exchange theory. The overarching framework for the decision-making process is the calculation of the costs and benefits of participation. The difference between the theory suggested by Goyder and the social exchange theory of Dillman is that Goyder allows for the decision-making process to vary in terms of the "amount of decision making" and to be influenced explicitly by "cultural factors". This adaptation of social exchange theory enables Goyder to integrate insights from theories of cognitive heuristics, which suggest that

sometimes the decision to participate is a fairly shallow decision, based on cues and other mental short-cuts. It also allows Goyder to accommodate findings from social disorganisation theory, which emphasises the influence of cultural factors in the decision-making process. By bringing the theories together within the social exchange framework, Goyder retains the insights into survey design from social exchange theory, and at the same time increases the explanatory power of the model.

It is useful to reflect on how social exchange theory can be used to improve survey design since it is largely for this reason that it has been so influential. Social exchange theory posits that decisions to participate can be manipulated by "tailoring" the survey design to make it more attractive to potential respondents. There is evidence that by implementing strategies to establish *trust* (e.g. through sponsorship, advance token of appreciation), increasing the *benefits* of participation (e.g. making the questionnaire interesting, giving tangible rewards), and decreasing the *costs* of participation (e.g. avoiding subordinating language, making it convenient to respond) response rates can be improved (Dillman et al. 2009). This evidence combined with insights from other theories, highlights the need to focus on the characteristics of both individuals and the survey to identify factors that are likely to influence participation decisions (Groves and Cialdini 1991, Groves et al. 2000, Groves and Peytcheva 2008, Dillman et al. 2009). Based on the literature five sets of factors can be identified as influencing participation decisions (Y). These are the characteristics of individuals ($\boldsymbol{X}$), features of the survey design ($\boldsymbol{S}$), household characteristics ($\boldsymbol{H}$), interviewer attributes ($\boldsymbol{T}$), and the social environment ($\boldsymbol{N}$) (Groves et al. 1992, Groves and Heeringa 2006). Including a stochastic error component, ε, survey participation can be written as a function of these factors as follows:

$$\mathrm{Y} = f(\boldsymbol{X}, \boldsymbol{S}, \boldsymbol{H}, \boldsymbol{T}, \boldsymbol{N}, \varepsilon) \tag{3}$$

A corollary of the concept of tailoring is that where surveys differ from one another in their target population and overall design, they will need a different mix of strategies to improve response rates. For that reason there are likely to be important differences between surveys in exactly the types of variables that are relevant for investigating survey participation. I, therefore, turn now to explore the types of factors that are likely to be important for explaining participation in the ASCS.

A framework for modelling missingness to the ASCS

Performance surveys are national surveys, but they are quite different, in terms of both survey design and the population of interest, from most national surveys that are studied in the nonresponse literature. Performance surveys are primarily postal surveys not face-to-face interviews, so factors such as interviewer attributes will not be relevant. Additionally, the focus of sampling is on individuals rather than households and the sampling frame is usually not the general population but the records of service recipients. This means that household characteristics, although not necessarily less relevant, will be unobserved in the sampling frame (except where they are captured in relation to the characteristics of the individual). Furthermore, as is the case in the ASCS, some performance surveys are conducted by the organisations in charge of the service being assessed. Effectively, there is not one but many surveys, and where these surveys are run differently this is likely to have consequences for response propensity. Consequently, a theory of nonresponse to the ASCS needs to take into account these features of the survey design and the special characteristics of the survey population, as well as the role of the social environment.

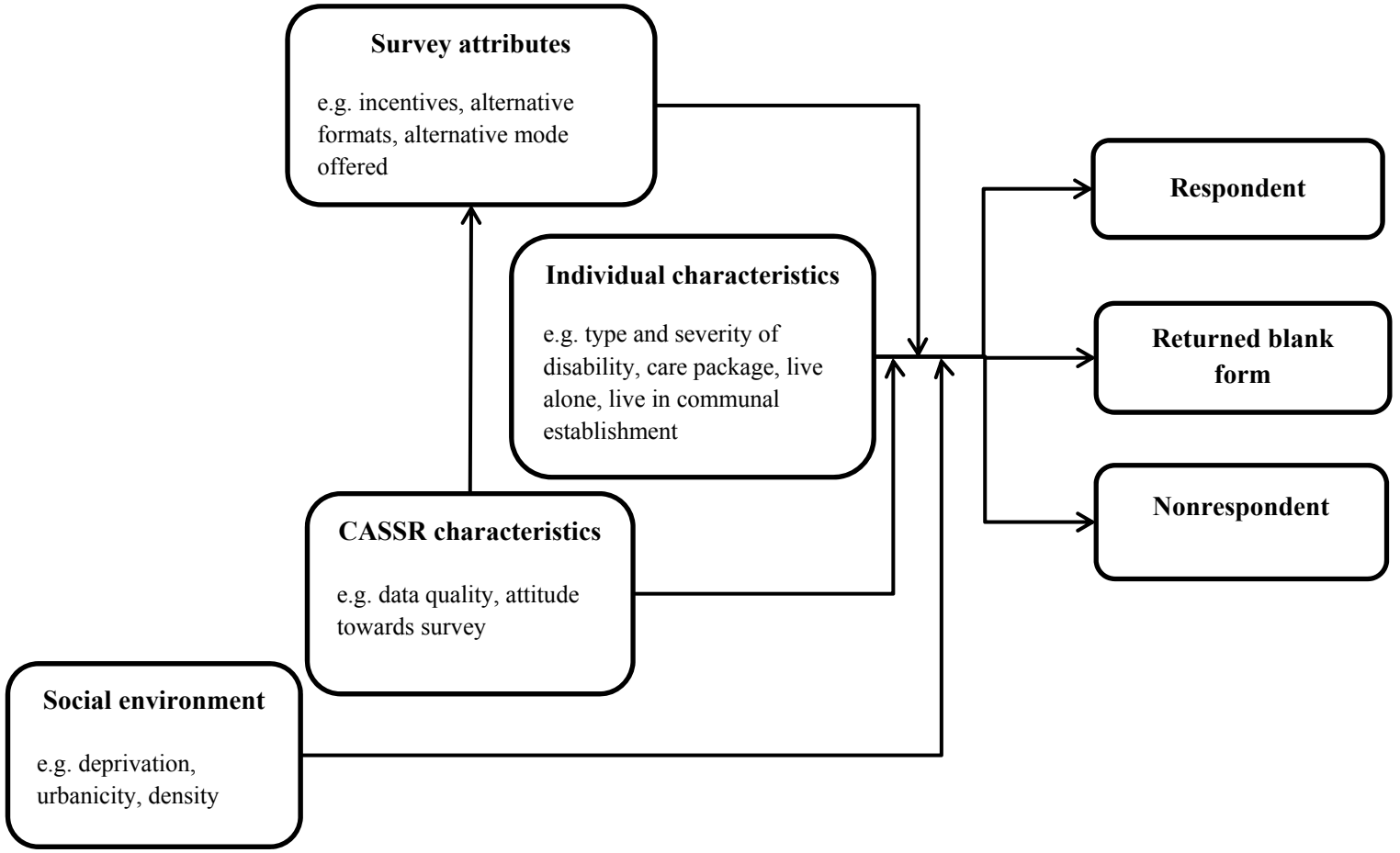

Figure 3: Illustration of the sets of factors influencing response propensity within the ASCS

The relationship between the three sets of factors – features of the survey design (survey attributes), individual characteristics and the social environment – and survey participation is illustrated in Figure 3. I discuss the relevance of each of these three factors to participation in the ASCS in turn. Importantly, I also discuss the role of CASSRs in shaping the relationships between the three characteristics and response propensity. The aim of this discussion is to develop a rationale for inclusion of variables in the empirical modelling in Chapter 4.

Features of the survey design

Evidence suggests that nonresponse bias is less severe in postal compared to face-to-face interview surveys, which has led to speculation that different mechanisms may explain nonresponse for the two modes (Groves and Peytcheva 2008). Postal surveys also differ from interviewer-led surveys in the types of nonresponse that can be observed. In interviewer-led surveys it is usually possible to distinguish refusals from non-contact and from people who are unable to respond, for example, due to cognitive impairment. For postal surveys, the different types of nonresponse usually cannot be distinguished from one another.

The ASCS is slightly unusual in this respect as nonrespondents fall into one of two categories: returned blank questionnaire or failed to return the questionnaire. The questionnaire instructions direct people to return a blank questionnaire if either of the following two conditions apply: they cannot complete this questionnaire either on their own, or by giving answers for someone else to record; or they do not want to participate and do not want a reminder questionnaire sent to them. Consequently, people who return a blank questionnaire will be a mixture of those who refuse to participate and those who are unable to participate (but are able to, or have someone to help them to, return the blank questionnaire). People who fail to return the questionnaire are likely to be a mixture of all types of nonrespondent, i.e. people who fail to receive the survey request (for example, because of non-delivery, or interception), refuse to participate, are unable to respond, as well as those who run out of time or forget to respond. These two nonrespondent groups are likely to be different in their motivations and therefore their characteristics. Thus, although there is a good deal of overlap between the groups in terms of the factors that are likely to explain the reasons for nonresponse, it seems important to explore whether there are differences in the factors explaining the two nonresponse outcomes.

Theories of nonresponse suggest that organisations charged with managing the survey have an important effect on response rates. On the one hand this is because they can manipulate the survey attributes to increase response rates (tailor the survey); on the other it is because certain types of organisations engender greater trust and are therefore more likely to have better response rates (Dillman et al. 2009). In the ASCS, where each CASSR in England manages the survey process for their locality, there is the possibility of divergence between CASSRs in survey attributes and therefore response rates. Evidence from the ASCS and performance surveys that have a similar design to the ASCS show that there are fairly large differences in response rates between organisations (Zaslavsky et al. 2002, Health and Social Care Information Centre 2013, Health and Social Care Information Centre 2014). There is, however, little research that looks at the reasons for observed differences in response rates across organisations in the context of performance surveys. Only one study by Zaslavsky et al. (2002) of the CAHPS survey of health plan reporting units has looked at this question, albeit not in a significant detail. They found that differences in the quality of contact information were one important factor in determining differences in response rates between health plan reporting organisations. This study did not consider the possibility that differences may be explained by the way the organisations manage and run the survey and this possibility has not been examined in the ASCS.

Although CASSRs follow national guidance there are two reasons to expect processes to vary between CASSRs. First, CASSRs have a degree of discretion over how the survey is implemented. For example, they can choose to add questions, change the appearance, include incentives and alter various other attributes that have been shown to have an effect on response rates to postal surveys (Edwards et al. 2002). Second, CASSRs vary in their attitudes towards the survey, which is a mandatory requirement and is sometimes seen as an inconvenience (Heath et al. 2015). CASSRs' attitudes towards the survey may influence their choices over how to implement the survey; certainly there is anecdotal evidence that some choose (intentionally or unintentionally) to flout the rules, by, for example, not sending out reminders. A focus of this research will therefore be on exploring whether differences in response rates between CASSRs can be explained by differences in the way CASSRs manage and run the survey.

The characteristics of the survey population

The population studied in performance surveys is frequently dominated by disadvantaged groups whose characteristics may make it difficult for them to participate in surveys. By definition, ASC users have a disproportionately high burden of illness and disability, compared to the general population (Qureshi and Rowlands 2004). Furthermore, due to the extent of rationing of ASC, the majority of the population is likely to be severely disabled, have at least one (if not several) long-term health conditions with associated functional impairment, and there is also likely to be a high prevalence of cognitive impairment (Fernandez et al. 2013b). Since both cognitive and functional impairments limit a person's ability to undertake ADLs, it seems highly likely that these characteristics will also have an effect on response propensity. However, due to the small numbers of people with serious cognitive and functional impairments in the general population, the role of health and mental ability or cognition in determining response propensity is understudied and under-theorised (Hendershot et al. 2003, Hauser 2005, Goyder et al. 2006).

There is evidence that poor health (Paganini-Hill et al. 1993, Strayer et al. 1993, Cohen and Duffy 2002, Perneger et al. 2005, Hutchings et al. 2012, Peters et al. 2014b), as well as visual impairments (Rahi et al. 2004) and proximity to death (Hébert et al. 1996, Kauppi et al. 2005) negatively affect response to postal surveys. Yet, aside from a study by Hébert et al. (1996) which found that nonrespondents to a postal survey were more cognitively impaired and more disabled than respondents, there is very little evidence about the impact of disabling conditions, such as cognitive impairment, physical impairment, sensory impairments, intellectual disability, mental health problems and substance misuse problems on survey participation.

Like other surveys, the ASCS contains very few indicators that could be used to explore the effect of health and disability on response propensity. The ASCS, however, does categorise service users according to broad categories of disability types, which can be used as indicators of the effects of different types of disabling conditions on response propensity. It also has information about service use, which itself reflects the person's underlying level of disability, as more severely disabled respondents will have more services and more intense forms of service provision, such as residential care. I explore the effects of these factors on response propensity in this research. Saliency of the survey to those sampled is recognised as an important factors affecting participation (Dillman et al. 2009). Saliency is likely to vary according to the type of care package received, as

services vary in the amount of contact the person has, and the amount and type of care or help that is delivered. Some services such as equipment involve a one-off contact and little or no caregiving, whereas others, such as home care, Direct Payments and care homes, are characterised by high levels of contact and in most cases significant amounts of highly personal care being delivered. Those receiving the most intense services are more likely to see the survey as worthwhile. In the case of Direct Payments, another factor may be at play. Since Direct Payments are a cash payment given in lieu of services, such recipients may see the ASCS questionnaire as a way for the CASSR to check on their ability to manage their own care, and they may therefore be less likely to complete it. This may be particularly the case if trust in the CASSR is low. It therefore seems important to explore the effect of receipt of different types of services on response propensity to the ASCS.

There may be some variation in the effects of health conditions and disability on response propensity across CASSRs, as they could influence response propensity by implementing strategies that reduce the costs of responding by making it more convenient or easier for them to respond. For example, alternative modes of survey administration, such as face-to-face interviews, are more appropriate for visually impaired and physically disabled people, and alternative formats, such as Easy Read are more appropriate for people with intellectual disabilities (Sheldon et al. 2007, Beadle-Brown et al. 2012). All CASSRs send an Easy Read version directly to people with intellectual disabilities, but responses to the ASCS suggest that some CASSRs use other strategies to improve response rates from hard-to-reach populations. In this research I explore the extent to which CASSRs use different approaches and the effect of doing so on response propensity.

Studies of the nonresponse to general population surveys have generally been more concerned with socio-demographic variables, like age, sex and ethnicity, as predictors of response propensity. As Groves et al. (1992) argue, socio-demographic variables have their influence through intervening variables. For example, it is argued that people from minorities and those who are foreign-born may feel more disconnected from the government than non-minorities or natives and therefore may be less likely to participate in a government-sponsored study (Axinn et al. 2011). Much research has found that the young and the very old less likely to respond to surveys (Herzog and Rodgers 1988, Kaldenberg et al. 1994, Zaslavsky et al. 2002, Elliott et al. 2005). It is argued that these effects are due to intervening variables, such disability for the very old, and for people of working age a lack of time due to the multiple demands of family and employment. The demands of family and employment are likely to be less important for determining

response propensity to the ASCS as ASC clients are much less likely to be in employment and, in many cases, have dependent children (Berthoud 2006). Equally in a population such as the ASCS, where all sample members have high levels of social care need any gradient in disability severity with age is likely to be less steep. It is, nevertheless, of interest to explore the role of these factors in influencing response propensity to the ASCS given the role that surveys such as the ASCS have in demonstrating whether services are provided equally well to minority and vulnerable groups.

The social environment

A number of studies have investigated the role of the social environment, usually in the context of social disorganisation theory. Across England there are wide variations in the social environment in terms of urbanicity, population density and social cohesion / disorganisation. All of these factors have been found to be important in determining response rates, with more urban, dense, deprived and disorganised places having lower response rates (House and Wolf 1978, Goyder et al. 1992, Groves et al. 1992, Couper and Groves 1996, Johnson et al. 2006). It is argued that these variables have their effects by reducing helping behaviour and trust in institutions and others (see e.g. Couper and Groves 1996), and that strategies to increase trust, such as the use of incentives or efforts to increase social capital, will moderate the effects of the social environmental variables. I therefore investigate the role of the social environment in influencing response to the ASCS.

In summary, I propose that response to performance surveys will be a function of four factors: survey attributes ($\boldsymbol{S}$), individual characteristics ($\boldsymbol{X}$), the social environment ($\boldsymbol{N}$) and CASSR characteristics ($\boldsymbol{C}$). The role of the CASSR is important because of the possibility that CASSRs manage and implement the survey differently in their localities. The model for response propensity can be written as,

$$\mathrm{Y}_{ij} = f(\boldsymbol{X}_{ij}, \boldsymbol{S}_j, \boldsymbol{C}_j, \boldsymbol{N}_{ij}, \varepsilon_{ij}) \qquad (4)$$

where Y_{ij} denotes the response status, for person i in CASSR j, and ε_{ij} is the random error.

The set of variables included within each category of factors will vary according to the survey. For the ASCS, I have argued that the relevant set of individual characteristics,

$\boldsymbol{X}_{ij}$, includes health conditions, disability severity, service receipt and other socio-demographic factors. The relevant survey attributes are defined by the choices made by CASSRs and I examine how these attributes vary between CASSRs in Chapter 4, as part of the exploratory analysis associated with modelling response propensity. The organisational characteristics, $\boldsymbol{C}_j$, discussed here as relevant for the ASCS, i.e. data quality and attitudes of the organisation, are likely to be important across all surveys that are run by multiple organisations. The aspects of the social environment, $\boldsymbol{N}_{ij}$, as predicted by social disorganisation theory are also likely to be appropriate for most surveys.

Concluding remarks

The aim of this chapter was to set out conceptual frameworks for exploring, first, social care outcomes for the purpose of case-mix adjustment and, secondly, survey participation in the empirical chapters of this thesis. I have argued that social care outcomes can be understood through the theory of production relations, as adapted to the social care context by Davies and Knapp (1981). Using this framework, I identified five factors that are exogenous influences over QoL-outcomes and are therefore likely to confound the assessment of the impact of services on QoL-outcomes. These are personal characteristics and motivations, underlying health and disabling conditions, immediate physical environment, additional resources arising from social and economic capital and reporting-related factors. In order not to confound the estimation of the performance of organisations, it is therefore necessary to control for the effects of these case-mix factors. However, this is not straightforward in the ASC context. As I have shown, these factors also influence the quantity and type of care package given to an individual. As such, statistical modelling to control for case-mix may underestimate the impact of services on QoL-outcomes and overestimate the impact of case-mix factors, potentially biasing comparisons of organisational performance. I have suggested that a way of overcoming this problem is to introduce an indicator of the quantity and type of care package into the case-mix estimation. This approach, may itself suffer from endogeneity bias due to the difficulty of observing all the needs-related characteristics that are important in determining the quantity and type of care package. In the next chapter I will discuss how regression modelling techniques can be used to address these theoretical concerns and how I apply these techniques to the statistical modelling of social care outcomes in the ASCS datatset Chapters 6 and 7.

With respect to survey participation, the conceptual framework proposed here draws heavily on Dillman's (2009) social exchange theory, particularly as it is elaborated by Goyder (2006). I have, however, adapted this framework to fit the characteristics of the ASCS, since these theories, and the empirical evidence, demonstrate that the types of factors that predict participation will be specific to individual surveys. On this basis I identified four sets of factors that are likely to predict participation in the ASCS: survey attributes, individual characteristics, the social environment and CASSR characteristics. The role of CASSR characteristics is particularly important for the ASCS because of the much greater role CASSRs have in the design and management of surveys in their locality than is ordinarily the case for national surveys. I test the hypotheses developed from this conceptual framework empirically in Chapter 4. In the next chapter, I describe how I use the conceptual model to inform the analysis around the effect of nonresponse on performance assessment. Although the conceptual framework as it is discussed here is tailored to the features of the ASCS, the framework itself should be useful for exploring survey participation to performance surveys more generally, given how many of the features of the ASCS are shared across performance surveys.

Chapter 3
Tools for the Empirical Analysis

Introduction

The activity of collecting performance information to understand and evaluate the performance of organisations is sometimes referred to as "organisational profiling" (Shahian and Normand 2008, Ash et al. 2013). In earlier chapters, I described how the observational nature of data from profiling studies complicates the evaluation of organisational performance. In particular, differences in case-mix between organisations and, where survey data is used, differences in nonresponse rates, may confound inferences about performance. In the absence of an experimental design, statistical analysis is a key tool for addressing the effects of confounding influences and providing a fairer basis for understanding and evaluating CASSR performance (Bird et al. 2005). Statistical techniques therefore form the bedrock of the methods for this thesis.

The aim of this chapter is to review the statistical techniques that have been used to address nonresponse and case-mix and consider how they can be applied in this thesis, given the limitations of the data. My main concern in this review is therefore to consider how the methods can be adapted to analysing survey-based indicators in the ASC context. The chapter is organised into three sections. In the first section, I set out the methods for addressing differences in case-mix between CASSRs. I focus on the case where the indicators are PROMs because the empirical analysis uses the SCRQoL PI. In section two, I review the analytical approaches that can be used to adjust indicators for nonresponse and explore the characteristics that influence survey participation. In the third section, I set out the methods I use to explore the impact of adjusting for case-mix and nonresponse on performance assessment and address the question of whether adjustment is necessary.

Addressing differences in case-mix between CASSRs

The most common method used for profiling organisations is risk-adjustment (Iezzoni 2013). Risk-adjustment is a fairly loose term that is used to describe a collection of methods for facilitating meaningful performance assessment from observational data. It usually consists of two stages: first, a risk-adjustment model is developed, which statistically accounts for variations in case-mix characteristics on outcomes, so controlling for their confounding effects on outcomes; secondly, the results from the risk-adjustment model are used to

indirectly standardise the observed outcomes for each CASSR. My aim here is to set out how risk-adjustment can be applied to analysing the performance of CASSRs using the ASCOF PROM, known as the SCRQoL indicator that is derived from the ASCS.

Profiling studies raise a number of methodological challenges that make it difficult to determine performance, notably the observational nature of the data and lack of experimental design. It is therefore useful to start by describing risk-adjustment in the context of the theory of causal inference (Shahian and Normand 2008). Given the additional challenge in the ASC context of cross-sectional data, I then consider the types of adaptations to the risk-adjustment model that are required to reduce potential bias in inferences about the relative performance of organisations, the potential to apply these approaches to the ASCS data, and how these solutions compare to alternative approaches found in the literature where longitudinal data are available. Finally I discuss how the indirect standardisation method can be used to generate an adjusted SCRQoL indicator.

Risk-adjustment and causal inference

All studies of causality suffer from the "fundamental evaluation problem" (Holland 1986) that only one "potential outcome" can be experienced at any one time (Rubin 1974). Yet to determine causality it is not enough to know how an intervention affects participants, i.e. the actual outcome, we also need to know what would have happened to them if they had not received the intervention, i.e. the counterfactual outcome. In the performance assessment context, a method is needed to estimate what hypothetically would have happened to the participant had he/she experienced care from a different CASSR. Estimation of this counterfactual outcome is the main motivation for risk-adjustment (Shahian and Normand 2008).

Most profiling studies rely on regression modelling to estimate the counterfactual outcome. The regression model, often referred to as the 'risk-adjustment model', provides an estimate of the relationship between individual characteristics of service users and observed outcomes. It therefore accounts for observed differences in the populations of different CASSRs that may influence outcomes. This method has less internal validity than randomised experimental designs, which balance populations on observed *and* unobserved characteristics. However, as Shahian and Normand (2008) note, such randomisation would not be possible. Other methods for estimating counterfactuals, such as matching and stratification are also impractical in the profiling scenario because of the number of organisations and the need to account for a large number of observed characteristics. Risk-

adjustment is the only practical method, but the validity of the estimated counterfactual outcomes from the statistical models is limited by the ability to control for all relevant factors that may affect outcomes (Deeks et al. 2003, Lilford et al. 2004, Nicholl 2007, Iezzoni 2013, Clarke et al. 2015).

The aim of risk-adjustment is to understand how the performance of a CASSR compares to what would have been expected (i.e. the counterfactual) (Shahian and Normand 2008, Nicholl et al. 2013). To achieve this, the expected outcome is estimated as the outcome the CASSR would have achieved for its population, given its actual distribution of service users, but applying the effects of user characteristics on outcomes estimated from the reference population risk-adjustment model. The observed outcome for a CASSR is then compared to the expected outcome for the CASSR. It is assumed that differences in the relationship between expected and observed outcomes across CASSRs reflect differences in the performance of CASSRs.[1]

Risk-adjustment models for outcomes

For risk-adjustment models the general approach is to estimate a reduced form of the structural model for the production of social care outcomes (see Chapter 2, equation 2), where the term capturing service effectiveness, $f(\boldsymbol{x_i}, \boldsymbol{q_i}; \boldsymbol{\beta})$, is omitted. This can be represented as follows:

$$y_i = \alpha + \sum_{k=1}^{K} \beta_k z_{i,k} + \varepsilon_i, \tag{5}$$

where y_i is the observed outcome for individual, i, $z_{i,k}$ are a set of k risk factors for values of $k = (1, \dots, K)$, and ε_i is the independently distributed error term. Risk factors vary according to the outcome indicator, and are the set of factors that influence the likely success of the service in achieving the desired outcome but are outside its direct control. In the context of the social care PROM, relevant risk factors are the five categories of quasi-inputs identified in Chapter 2, i.e. (i) underlying health and disabling conditions, (ii) personal characteristics and motivations, (iii) additional resources arising from social and economic capital, (iv) the immediate physical environment, and (v) reporting-related factors. All of these groups of

[1] Where the characteristics of service users are quite different across CASSRs, some caution is needed when interpreting the indirectly standardised outcomes in this way (Shahian and Normand 2008, Nicholl et al. 2013). I reflect on these concerns in Chapter 9.

variables are expected to directly affect outcomes for individuals, but are not, I argued in Chapter 2, significantly affected by ASC services. Including these sets of factors in the model will adjust for their effect on outcomes, such that the predicted outcome for any individual from the model, $\hat{y}_i$, is conditional on their value.

For multi-attribute PROMs like the ASCOF SCRQoL indicator, the model in equation (5) can be estimated by OLS regression. Provided we can assume that the model coefficients are estimated without bias, then the predicted outcome, $\hat{y}_i$, should be an unbiased estimate of the counterfactual and the average error, $\bar{\varepsilon}_i$, should be an unbiased estimate of the effectiveness of ASC. For performance comparisons, where the interest is in differences between CASSRs, a slightly lower bar is required. As long as any bias in coefficients is constant across CASSRs then differences in the average error between CASSRs should be an unbiased estimate of relative differences in effectiveness. This assumption may not hold for PIs in general due to lack of independence of the observations. Specifically in the case of the SCRQoL PI it may not hold due to the complex relationships between the risk-adjustors and other variables that are unobserved in the risk-adjustment model (as described in Chapter 2) that confound the relationship between risk-adjustors and outcomes. I discuss these in turn and consider how to modify the risk-adjustment model to improve the estimation.

Observations are not independent

A key assumption for unbiased estimation is that observations are independent. This assumption is unlikely to be met in most instances where risk-adjustment is used for performance assessment, since it is premised on the idea that people residing within the same CASSR are affected by an unmeasured common factor, namely quality, and that this factor varies across CASSRs. Indeed, as Li et al. (2009, p. 84) remark, the OLS model seems to contradict the very "spirit of outcomes comparisons". As well as being at odds with the intentions of performance assessment, ignoring the clustering of individuals has been shown, in certain circumstances, to lead to biased coefficients and consequently biased indicators (DeLong et al. 1997, Normand et al. 1997, Li et al. 2009).[2]

There are two alternative approaches that can be used to appropriately reflect the clustering of observations within authorities – fixed-effects (FE) and random-effects (RE) models. These models have been used fairly extensively across the health and education

[2] It also will lead to incorrect standard errors (SEs) for model parameters, but the correct interpretation of SEs is less important here, except to the extent that it leads to the inclusion of variables that are not important predictors of outcomes.

sectors for performance comparison (Goldstein and Spiegelhalter 1996, Normand et al. 1997, Elliott et al. 2001, Austin et al. 2003, Huang et al. 2005, Arling et al. 2007, Mukamel et al. 2008, Li et al. 2009, Clarke et al. 2015, Nuttall et al. 2015). The models capture organisational-level clustering, through the inclusion of an additional parameter, ξ_j, which captures the effect for organisation, j. Using the same notation as above, the model is given by,

$$y_{ij} = \alpha + \sum_{k=1}^{K} \beta_k z_{ij,k} + \xi_j + \varepsilon_{ij} \quad (6)$$

The exact specification of the organisational term, ξ_j, depends on whether it is treated as a fixed or random effect. The RE model assumes that organisational effectiveness, ξ_j, arises from a normally-distributed population, i.e. $\xi_j \sim N(0, \sigma_u^2)$, while the FE model, makes no distributional assumptions for ξ_j – they can simply be specified as dummy variables for each organisation (Greene 2012, Clarke et al. 2015). The organisational term can be interpreted as an estimate of organisational effectiveness and the estimates for individual organisations can be obtained directly from the model. In this way, this approach deviates conceptually from risk-adjustment using OLS. The ability to derive estimates of the organisational effect directly is a major attraction of these models (particularly the RE model) (Goldstein and Spiegelhalter 1996, Normand et al. 1997), although, as I will discuss below, there are difficulties with using these estimates directly as PIs. Both the RE and FE models can, however, also be used within the classical risk-adjustment framework to obtain an estimate of the counterfactual outcome, $\hat{y}_i$, by setting the CASSR effectiveness term to zero, i.e. ξ_j=0 (Li et al. 2009).

Although the RE and FE models are more complex than OLS models, it is argued that explicitly modelling the clustering of observations within CASSRs means they have greater face validity for performance assessment (Li et al. 2009). Whether decisions about models matter for inferences regarding the relative performance of organisations is a moot point. Researchers make strong theoretical arguments about the need for RE or FE models, but the effects of such choices on inferences about performance vary depending on the study and on whether the estimates of organisational effects are used directly as PIs (Greenfield et al. 2002, Huang et al. 2005, Glance et al. 2006, Arling et al. 2007, Mukamel et al. 2008, Li et al. 2009, Eijkenaar and van Vliet 2014). In light of this, Li et al. (2009) argue that the importance of

modelling choices is an empirical question, which is likely to be affected by the particular characteristics of the data. For this reason it seems important to explore these models as alternatives to OLS in this particular study.

Important variables are unobserved

As the discussion in Chapter 2 highlighted, a number of the variables that are considered here as risk-adjustors also determine the technologies and quantity of technologies received as part of an individual's care package. Indeed a correlation between various risk-adjustors, for example functional ability, and the quantity of care provided to service users has been illustrated in previous research in social care (Davies et al. 2000b, Fernández 2005, Forder and Caiels 2011b, Forder et al. 2014). A consequence then of omitting a term capturing the quantity of service provided, i.e. $\boldsymbol{x}_i$ in equation 2 (see Chapter 2), is that the parameter estimates for the risk-adjustors, $z_{ij,k}$, will be biased. Specifically, the coefficients for the risk factors will be upwardly-biased, as they will capture some of the (positive) service effect (Davies et al. 2000b, Fernández 2005, Forder and Caiels 2011b, Forder et al. 2014). This in turn will lead to bias in the calculation of the expected outcome on which the PIs are based.

In the context of performance comparisons, this bias due to the confounding effect of service quantity is only problematic if the relationship between the risk-adjustors and service quantity varies across CASSRs. Should this relationship be fairly constant across organisations, e.g. because social services departments use similar algorithms for determining the intensity and nature of the care package received, then the risk-adjustment model may still provide a reasonable estimate of differences in effectiveness between organisations, even if it does not provide a good estimate of the average effectiveness of social care services. Differences in eligibility policies between CASSRs and in what local markets can offer, however, raises doubts over the plausibility of this assumption. It is therefore important to include an indicator of service quantity in the risk-adjustment model to control for its confounding effect on the relationship between the risk-adjustors and social care outcomes. Such a model would more closely resemble a production function (Fernandez et al. 2013a, Forder et al. 2016), so it is useful to look to this literature to explore how such a model could be specified in the context of ASC performance assessment.

It is important to recognise that the use of a production function to improve the estimation of effectiveness, is necessary in this application because of the cross-sectional nature of the data and the fact that most sample members receive community- or home-based

services, such as home care where the intensity of the service varies according to the social care needs of the individual. Regarding the latter point, the quantity of care provided should not exert a strong confounding effect for residents of nursing and care homes, as the service is more homogeneous. For residents of nursing and care homes, therefore, the risk-adjustment model should provide a relatively unbiased estimate of the expected outcome in the absence of services, assuming it has been possible to control for key risk factors. Regarding the former point, were longitudinal data available, in which data on outcomes and user characteristics were collected at regular intervals, an alternative approach would be feasible. Notably, it would be possible to use the scheme outlined by Smith and Street (2013) for interventions that aim to arrest or slow the decline in health status or maintain as far as possible the person's quality of life.

The approach suggested by Smith and Street (2013) draws on the fact that the longitudinal data makes it possible to chart the health status (or more appropriately here QoL status) pathway of individuals receiving care. An indicator can be constructed based on the change in QoL status between two adjacent data collection points. The inclusion of prior QoL status in the indicator controls for differences in the population of users cared for by organisations, by removing the effect of variations in 'baseline' characteristics. Since there is still the potential for user characteristics to confound differences in outcomes, a risk-adjustment model is needed to address selection bias. The risk-adjustment model estimates the expected change in QoL status, given service user characteristics at baseline (where baseline is understood as the previous time point) that would be considered to be risk factors for a change in QoL status. Examples of such risk factors might be multiple co-morbidities, age, type of health condition and functional ability. This approach has been applied to nursing homes quality indicators, such as the change in functional ability and change in mobility indicators, in the USA where data is collected on all residents on a quarterly basis (Arling et al. 2007, RTI International 2017). The main problem with this approach is that it does not provide a good estimate of the effectiveness of care, since prior QoL status (and the risk-adjustors) is not likely to be a good indicator of the counterfactual pathway (Smith and Street 2013). The counterfactual QoL status pathway is better estimated as a function of factors such as physical, cognitive and social impairment and environmental factors of the type identified as quasi-inputs in Chapter 2 (Forder et al. 2014). This is the main benefit of the production function approach over an approach based on longitudinal panel data.

Applying production functions to performance comparisons

A production function is used by economists to express the technical relationship between the factors of production (i.e. inputs such as labour and capital) and the product of the productive process. Despite the difficulties involved in estimating production functions empirically even where there is a high degree of technological determinacy in the relationship between inputs and the products of production (Intriligator et al. 1996), researchers have successfully managed to estimate production functions for social care (Davies et al. 2000b, Fernández 2005, Forder and Caiels 2011b, Forder et al. 2014, Forder et al. 2016). This thesis builds on this empirical work to specify the functional form of the production function and applies it to organisational profiling.

The main challenge for estimating production functions with the ASCS dataset are the data limitations. The only measure of service quantity in the ASCS dataset is a single variable that captures the 'personal budget' allocated to the user. The personal budget is designed to cover the costs of purchasing services to meet the service users' assessed needs. It can be thought of as a cost-weighted utilisation measure where the quantity of each of the services received are summed together using pounds sterling as the common unit. Although it should be possible to estimate a simplified production function with this variable, the personal budget is both inconsistently and very poorly recorded, with a large number of missing values. As well as reflecting the quantity and mix of services received as part of the care package, the personal budget also reflects differences in the price paid for the services across different parts of the country. These problems with the recording of the personal budget variable mean that I can only apply the production function method to a sub-population of the ASCS dataset and am limited in the statistical methods that I can use. The effect of the data limitations on the empirical modelling are discussed further in Chapter 7 where I use a production function to model social care outcomes, but here I reflect on the constraints of the data for the statistical methods.

To estimate a production function it is necessary to make some assumptions about the functional form of the production relationship. I draw on the work of Forder et al. (2014) and Fernández (2005) and use a functional form based around an expanded generalised linear production function with similarity to the work of Lau (1974) and Diewert (1971). This form allows for the existence of only one input (i.e. the cost-weighted utilisation indicator of service quantity), for diminishing returns to factor at higher levels of input (Knapp 1978b, Knapp 1979, Davies et al. 2000b, Fernández 2005), and can accommodate the presence of exogenous risk factors in the production relationship (Davies and Knapp 1981, Knapp 1984,

Davies 1985). The generalised form of the model for the production of social care outcomes is therefore given as,

$$y_i = \alpha + \gamma x_i + \delta \ln(x_i) + \sum_{k=1}^{K} \beta_k z_{i,k} + \sum_{k=1}^{K} \vartheta x_i . z_{i,k} + \sum_{k=1}^{K} \theta \ln(x_i) . z_{i,k} + \varepsilon_i \qquad (7)$$

where y_i is the observed outcome and x_i is the personal budget capturing service quantity. The logarithmic term allows for variations in returns to factor[3]. As in equation (5), quasi-inputs are captured through the term, z_i, with the direct effect on outcomes captured by β_k. The interaction terms, $x_i . z_{i,k}$ and $\ln(x_i) . z_{i,k}$, capture the mediating effect of quasi-inputs on outcomes and can be interpreted as differences in the marginal productivity of services for different groups of users (Knapp 1984, Fernández 2005). The independently-distributed error is given by ε_i.

From the perspective of face validity, the production function model is an improvement over the risk-adjustment model. It attempts to disentangle the effect of services on outcomes from the effect of user characteristics on outcomes, so addressing the bias in the estimation of the coefficients for risk factors in equation (5). The coefficient on the service quantity term, x_i, provides an estimate of the effectiveness of care and the risk factors, $z_{i,k}$, control for the effect of observable factors on SCRQoL status. These risk factors are for the most part needs-related characteristics of the user that we expect to be strongly related to both SCRQoL and to the care package provided (Forder et al. 2014). The difficulty is that we cannot observe all the relevant needs-related characteristics that are available to the assessor when carrying out the needs assessment that determines the care package. This leads to potential endogeneity bias in the estimation of the coefficient for the service quantity term, x_i. Put another way, the service quantity term depends on the outcome of the needs assessment, which depends on both observed and unobserved needs-related characteristics. As Forder et al. (2014) show OLS estimation will lead to biased estimates of the service quantity effect because changes in service quantity are associated with changes in both SCRQoL status and the error.

Instrumental variables (IV) estimation is the usual solution for endogeneity (Cameron and Trivedi 2005). The idea of IV estimation is to find a variable (or variables) – the instrument(s) – that induces variation in service quantity but does not directly affect the

[3] I in fact try a variety of alternative forms for this relationship as I describe in Chapter 7.

SCRQoL status of the person (except through its effect on service quantity). As Forder et al. (2014) argue a potential instrument for service quantity in home-based social care services is the CASSR in which the individual resides. CASSRs operate different eligibility policies, which are likely to induce variation in service quantity, but such policies should not have a direct effect on individuals' SCRQoL except through their effect on service quantity. Forder et al. (2014) successfully use this IV strategy to estimate the effectiveness of home care. In a later study Forder et al. (2016) refine the instrumentation strategy and apply a spatially-lagged service use variable as the instrument for service quantity. This variable is calculated as the average service use of other people living in the same area as the case in question. These strategies cannot be applied to the ASCS data because the personal budget variable also captures variations in prices across CASSRs. As instruments in the ASCS data these variables therefore induce variation in the personal budget variable due to both differences in eligibility policies (desirable) and differences in policies regarding price paid for a unit of service (undesirable). Policies regarding price are also likely to have a direct effect on social care outcomes, meaning any variable defined at the CASSR level will fail to meet a key condition of being an instrument. Without an appropriate instrument it is not possible to apply IV estimation to the ASCS dataset.

Returning to the production function model, as a consequence of the inclusion of the service effect in the model specification, the error has a slightly different interpretation compared to the risk-adjustment model. To the extent that quality is not a function of the mix of technologies and intensity of provision, as captured by the personal budget variable, its effect will be relegated to the error term. Differences between CASSRs in their average error will therefore reflect differences in quality and *not* effectiveness, as is the case in the risk-adjustment model. It is possible to estimate a predicted outcome, $\hat{y}_i$, and therefore counterfactual outcome, that is more analogous to the predicted outcome from the risk-adjustment model, by setting the care package effect to zero, i.e. $x_i = 0$. This approach has some analogy with the methods used to explore equity in the utilisation of health care, where there is a need to adjust for non-need factors that may confound the relationship between need and utilisation. To generate an estimate of need-expected utilisation where models include non-need factors these factors are set to their mean or some other value to neutralise their effect (O'Donnell et al. 2008).

Indirect standardisation using expected outcomes from regression

There are two main approaches to standardising outcomes: direct and indirect standardisation. The method of direct standardisation, in which the observed outcome is standardised to the case-mix of the reference population, is generally infeasible for performance assessment studies because the sample size of organisations is too small and the method cannot accommodate continuous variables. Consequently for performance assessment studies, results from regressions are used to indirectly standardise PIs (Ash et al. 2013). For this reason I apply this indirect standardisation to the SCRQoL, although I reflect on this decision in Chapter 9.

Indirect standardisation involves comparing the outcomes for the specific group of users in a CASSR with their expected outcomes had they been treated by an average CASSR in the reference population. There are two different ways of comparing the observed, y_{ij}, to the expected outcome, $\hat{y}_{ij}$, estimated from the regression models specified above. First, as a ratio of $y_{ij}/\hat{y}_{ij}$, and secondly, as the difference $y_{ij} - \hat{y}_{ij}$. Where the risk-adjustment model is estimated by OLS, the difference is simply the average error term, e_{ij}, from the model, hence hereafter I refer to the latter method as the 'error method'.

The relationship between the observed and expected outcome is sometimes referred to as the 'adjustment factor'. It conveys the extent to which performance is above or below what is expected. For both methods the adjustment factor is usually calculated at the individual level and then averaged across all individuals in the CASSR to generate an average adjustment factor for each CASSR. This average adjustment factor can be converted to the same scale as the original PI. This is achieved by multiplying the adjustment factor by the national average indicator score for the ratio method as in equation (8), and for the error method by adding the national average indicator score to the adjustment factor, as in equation (9).

$$PI_j = National\ average \times \frac{1}{N_j}\sum_{i=1}^{N}\frac{y_{ij}}{\hat{y}_{ij}} \tag{8}$$

$$PI_j = National\ average + \frac{1}{N_j}\sum_{i=1}^{N}\left(y_{ij} - \hat{y}_{ij}\right) \tag{9}$$

Neither method for estimating the adjustment factor is clearly superior (Ash et al. 2013). The ratio method has the benefit of giving more weight to deviations from expected outcomes for cases over the lowest part of the scale, but it has two drawbacks that favour the error method. First, since all values must be strictly greater than zero, it is not appropriate for scales with zeros and negative values, which is commonly the case where PIs are based on PROMs. To be used in this context a constant must be applied to scores, the size of which will affect the ratio. Second, because it gives more weight to deviations from expected performance at the lower extreme of the scale it can produce extreme values that may distort PIs. The error method is by contrast more conservative, since it gives equal weight to differences in observed and expected outcome scores over different parts of the distribution. This means that organisations do not receive greater recognition for 'raising' someone with a very low expected score up two points, than they do for 'raising' someone with an average expected score up two points.

Evidence about the effect these methods have on inferences about performance is thin. However, concerns have been raised about using the ratio method to risk-adjust the NHS surgical PROMs because of some extreme values (Ara et al. 2013, NHS England Analytical Team 2013). Consequently, the error method is now the preferred approach for NHS surgical PROMs in England (NHS England Analytical Team 2013). Ara et al. (2013) also suggest a method for calculating a more conservative ratio adjustment factor. They suggest estimating the expected outcome for each CASSR (rather than each individual) on the basis of CASSR-average scores for each risk factor as follows,

$$PI_j = National\ average \times \frac{\sum_{i=1}^{N} y_{ij}}{N_j . \hat{y}_j} \tag{10}$$

A criticism of the indirect standardisation approach outlined above is that it assumes that the deviation from average is entirely due to differences in performance. Since there is in fact a stochastic element to performance, this assumption is problematic. The smaller the numbers of cases each organisational score is based on, the larger the random element of the error. Not accounting for this uncertainty in PI scores when comparing organisations, can lead to incorrect identification of which organisations are outliers (both type I and II errors) and over time there is likely to be regression-to-the-mean. This concern over appropriately reflecting the uncertainty in PIs drove much of the early research into the use of RE models

in this context (Aitkin and Longford 1986, Goldstein and Spiegelhalter 1996, Goldstein 1997, Normand et al. 1997).

The attraction of RE models is that it is possible to produce estimates of organisational effectiveness even if organisations have very small samples. Where sample sizes are small, estimates obtained from FE models will have a large variance and, due to sampling variability, be unreliable (Li et al. 2009, Clark et al. 2010, Clarke et al. 2015). Since the RE model assumes that the organisational effects, ξ_j, are drawn from a normal distribution, a shrinkage factor that is inversely proportional to the precision of the estimate can be applied to each organisational estimate, in such a way that organisations with fewer observations show more shrinkage towards the grand mean than those based on a large number of observations. The shrunken residuals provide indicators for all organisations that appropriately address uncertainty and regression-to-the-mean bias is reduced (Aitkin and Longford 1986, Goldstein and Spiegelhalter 1996, Goldstein 1997, Normand et al. 1997, Austin et al. 2003).

Shrunken residuals consistently produce different results compared to PIs generated using the indirect standardisation method (Li et al. 2009). There is also evidence from simulation studies that the shrunken residual approach is a less sensitive approach producing more false positives, particularly where shrinkage to the mean is high (Austin et al. 2003, Clark et al. 2010, Ash et al. 2013). Shrinkage is highest where the sample size is small or where the level one variance is large and the level two variance is small, both of which seem to be fairly common in performance assessment studies (Hannan et al. 2005, Glance et al. 2006, Li et al. 2009). Some researchers argue that by construction shrunken residuals produce biased PIs (Mukamel et al. 2008, Li et al. 2010). From a practical perspective they are not very attractive either, since it is necessary to re-estimate the model for every new dataset (Nuttall et al. 2015). Since it is highly unlikely they would be adopted, for this reason I do not consider using shrunken residuals for the ASCOF PIs.

In the empirical analysis in chapter 8 I use only the indirect standardisation method for case-mix adjusting the ASCOF SCRQoL PI. Because of questions about the appropriateness of the error and ratio methods for calculating the adjustment factor, I consider both of these methods. I also consider the more conservative ratio method proposed by Ara et al. (2013). I now turn to discuss the methods I will use for addressing nonresponse to the ASCS.

Addressing nonresponse to social care surveys

Nonresponse leads to missing data, which presents two problems for interpreting PIs. The loss of data affects the precision of survey estimates and, at least in some cases, it can affect the representativeness of the sample, resulting in biased estimates. This will affect the validity of inferences about performance based on the PIs. This thesis has two aims with respect to the analysis of nonresponse: first to assess who is missing from performance surveys and how response rates can be improved; secondly, to establish whether nonresponse has an effect on performance assessment and consider what analytical approaches can be used to address differences in samples due to nonresponse. With respect to the second aim, various methods are available to address the effects of nonresponse on survey estimates, some of which can also be used to address the first aim. I review these methods focusing on their suitability for addressing nonresponse to performance surveys, in general, and to the ASCS specifically. I begin by setting out the different types of missingness and considering each approach for addressing nonresponse with respect to the assumptions it makes about the missingness mechanism. I then critically review the studies that have assessed the effects of nonresponse in performance surveys and suggest an appropriate approach for addressing nonresponse to the ASCS.

Types of missingness and approaches to addressing nonresponse

Complete case analysis (CCA), in which all cases with missing values for the variable (and covariates) of interest are dropped, is the most common approach to dealing with missing data in applied research (Horton and Kleinman 2007). The problem with CCA is that it makes the very strong assumption that the reasons for nonresponse are completely independent of both the observed and unobserved data, or, in the language of the nonresponse literature, the missingness mechanism is assumed to be missing completely at random (MCAR) (Rubin 1976, Little and Rubin 2002). The MCAR assumption may be plausible where all missingness is a result of, for example, data input error. Since nonresponse tends to be theorised as the result of particular behaviours, the MCAR assumption is untenable and requires testing for its plausibility. More plausible assumptions for missingness are (i) that the probability of missingness depends only the values of the observed data, a pattern of missing data said to be missing at random (MAR), or that (ii) the probability of missingness additionally depends on the values of the unobserved data, a pattern of missingness known as not missing at random (NMAR) (Rubin 1976, Little and Rubin 2002).

Where data are in fact MAR or NMAR, CCA will theoretically produce biased estimates. More principled techniques for handling missing data include multiple imputation (MI) and inverse probability weighting (IPW)[4] (e.g. Rubin 1987, Little and Vartivarian 2003, Horton and Kleinman 2007, Seaman and White 2013). In MI the observed data (from the survey, sampling frame or other sources) is used to model the values of the missing data. The missing values are then replaced with values generated randomly from the imputation model to create a complete dataset, consisting of the observed and imputed values. This procedure is repeated, to produce multiple 'imputed' datasets. Analysis is conducted on each imputed dataset and estimates are combined using Rubin's rules (1987). The purpose of MI is not to re-create the individual missing values as close as possible to the true ones but to handle missing data in such a way that it results in valid statistical inference (Rubin 1987, Rubin 1996). Where IPW is used to correct for nonresponse, complete cases are weighted by the inverse of their probability of being a respondent. A missingness model, often referred to as a 'response propensity model', is developed to predict the probability that a sample member is a respondent given observed characteristics (Little and Vartivarian 2003). The aim of IPW is to rebalance the respondent sample to better reflect the characteristics of the whole (observed and unobserved) sample. Using IPW researchers can then draw inferences from their analyses that are valid for the whole sample.

Estimates derived from the use of techniques such as MI and IPW are often described as 'adjusted', since they assume that data are MAR and produce revised estimates under that assumption. Should the missingness mechanism be in fact NMAR estimates may still be biased. In spite of this, MAR-based methods have gained widespread support among statisticians due to the principled way in which they address missing data (Molenberghs 2007). As an editorial to the *Journal of the Royal Statistical Society* concludes,

> it is time to place CC [complete case] analysis and simple imputation methods… in the Museum of Statistical Science, and to consider instead ignorable likelihood analysis, ignorable Bayesian analysis, MI [multiple imputation] and weighted estimation equations as the standard. In other words, one should enter the building at the MAR floor and not the MCAR level. (Molenberghs 2007, p. 863).

This strongly supports the application of MAR methods for addressing missingness to the analysis of survey-based PIs, where there is nonresponse.

---

[4] IPW is one of many different methods for weighting data for nonresponse (for a review see Kalton and Flores-Cervantes 2003). I consider only IPW here as the other methods are less suitable for auxiliary information with continuous predictors and cannot be applied to analyse the factors influencing response propensity as IPW can.

Analysing performance surveys under conditions of nonresponse

Performance surveys generally have high levels of nonresponse, particularly of *unit* nonresponse, which refers to those sample members who fail to respond to the survey request. This is because most performance surveys are postal surveys, and such surveys have higher rates of unit nonresponse than face-to-face or telephone interviews (Dillman et al. 2009). Of those who do respond, a proportion will also have missing data to certain questions. This type of nonresponse is often referred to as *item* nonresponse. For most performance surveys item nonresponse is generally much lower than unit nonresponse (Elliott et al. 2005, Roland et al. 2009, Klein et al. 2011, Health and Social Care Information Centre 2013, Health and Social Care Information Centre 2014). Nevertheless, even if there are low levels of item nonresponse for individual questions, where PIs are composed of multiple items or where PIs are risk-adjusted the fraction of cases missing from the PI could become quite substantial. Given what is known about the effect of nonresponse on PI estimates it is important to understand whether nonresponse affects inferences about performance.

Relatively few studies, however, have addressed this question directly in the context of performance surveys. A number of studies have taken a first step, and looked at whether nonrespondents differ from respondents. Much of this research has been conducted around the NHS PROMs and patient survey programme, as well as the US CAHPS patient survey programme. In general these studies find that socio-demographic (age, sex, ethnicity, living arrangements) and socio-economic factors (deprivation) are associated with nonresponse, although patterns are not always consistent across different surveys (Elliott et al. 2005, Roland et al. 2009, Klein et al. 2011, Hutchings et al. 2012, Peters et al. 2014a). Evidence from the PROMs programme suggests that nonrespondents and respondents to the post-intervention PROMs questionnaires differ on their baseline PROMs score, with nonrespondents having on average worse baseline health status (Hutchings et al. 2012, Peters et al. 2014a). This finding is consistent with other research that has found a relationship between nonresponse and poor health and quality of life (Paganini-Hill et al. 1993, Strayer et al. 1993, Cohen and Duffy 2002, Perneger et al. 2005). Given these findings concerns have been raised about the potential for nonresponse to lead to bias in PIs.

A small number of studies have attempted to look at whether nonresponse affects inferences about performance, focusing specifically on the question of bias. Various strategies are employed including following up nonrespondents with telephone interviews (Lasek et al. 1997), exploring (partial) correlations between response rates and PIs for

organisations (Roland et al. 2009), comparing raw PI scores to IPW-adjusted scores (Elliott et al. 2005) and correlating weights (derived from a response propensity model) with PI scores (Elliott et al. 2009). While these studies find differences between nonrespondents and respondents, they do not find that nonresponse significantly biases the PIs. For instance, Elliott et al. (2005) report that although IPW adjustment would reduce the bias, it would also reduce the precision of the estimates, particularly where estimates are risk-adjusted. These studies have in general concluded that nonresponse adjustments are unnecessary as they would needlessly reduce the precision of PIs (Elliott et al. 2005, Elliott et al. 2009, Roland et al. 2009).

A recent study by Gomes et al. (2016), which focuses on the effect of nonresponse on both precision and bias, however, reaches a different conclusion from the preceding studies. Gomes et al. (2016) use MI to explore the effects of missing data on inferences about performance from NHS PROMs data for hip replacements. Similarly to Elliot et al. (2005) they compare risk-adjusted PROMs scores estimated under CCA to MI-adjusted, risk-adjusted PROMs scores. They find that the loss of data arising from nonresponse affects conclusions about performance; specifically they find that more outliers are identified after adjusting for nonresponse. By decomposing the effect of MI-adjustment, they are able to show that the majority of the effect is due to the increased precision of MI-adjusted estimates not to bias reduction. The choice of MI as the method for nonresponse adjustment is an important contributing factor to Gomes et al.'s conclusions: MI is a more efficient method than IPW in part because it replaces missing values with plausible values and restores the dataset to its original size (Seaman et al. 2012, Seaman and White 2013). Gomes et al. argue that the PROMs PIs should be adjusted for nonresponse because of the positive effect it has on the precision of estimates with consequences for performance assessment.

MI as a method for exploring the effects of missingness on PIs has usually been confined to administrative data (Kirkham 2008, Gale et al. 2011) and it is unlikely to be an appropriate method for most performance surveys. The PROMs data are unusual in that they can be linked to extensive hospital records and the baseline questionnaire provides detailed information about those patients lost to follow-up. Most performance surveys are cross-sectional designs with no follow-up; the unit nonresponse rates are high and frequently over 50 per cent; and there is usually very little data available about the characteristics of nonrespondents. In such circumstances, it is likely to be difficult to specify an imputation model for MI. Where the imputation model is poorly-specified and nonresponse rates are

high (above 50 per cent), the potential for bias is so great that MI is generally considered unusable (McKee et al. 1999, Seaman and White 2013).

IPW is often preferred in circumstances where unit nonresponse is high precisely because of the difficulties inherent in using MI with such data (Seaman and White 2013). IPW methods are also likely to be more acceptable to a range of users since the missingness model is more transparent, its fit to the data can be tested, and it is easier to explain to non-specialist end-users of PIs. Yet IPW is less efficient than MI and where there are multiple PIs, multiple weights are required (assuming that rates of item missingness vary) which makes the method less practical. Where item nonresponse is present with high rates of unit nonresponse, Seaman and White (2013, p. 285) suggest that "it may be worth combining IPW and MI, imputing missing values in individuals with almost complete data and using IPW to adjust for the exclusion of individuals with more missing data". As Seaman et al. (2012) show in a different article, this approach, which they call IPW/MI, provides less biased results than MI/MI (denoting the fact that MI is used to address both unit and item nonresponse) and more efficient results than IPW/IPW (similarly where IPW is used for both unit and item nonresponse). Since the MI procedure results in a complete dataset for respondents, it makes IPW more practical as only one set of weights is required. Despite its potential, the IPW/MI method has not been applied to the analysis of performance surveys.

In this thesis I explore IPW/MI as an approach for analysing the effects of nonresponse on inferences about performance using the ASCOF indicators. This analysis is conducted in Chapter 5. A benefit of the IPW/MI approach is that the missingness model used to generate the weights for IPW can also be used to explain who is missing from the ASCS and the role that CASSRs have to play in affecting response rates. I explore this question in Chapter 4 and use the missingness or 'response propensity' model for this purpose. The specification of the response propensity model is grounded in the theoretical model developed in Chapter 2. The intention is that the results from the response propensity model should be useful for developing a more tailored survey design to increase response rates to the ASCS (Groves et al. 2000, Dillman et al. 2009). The results from this model also feed into the analysis in Chapter 5, as discussed.

Assessing the effects of adjusting for case-mix and nonresponse on performance assessment

An aim of this thesis is to understand whether survey-based PIs should be adjusted for case-mix and for nonresponse. There are many factors that policymakers need to consider in this

decision, including the validity of PIs, their acceptability and the various other characteristics identified in Chapter 1. Since this question cannot be answered directly through empirical analysis, the purpose of the analysis is to inform decisions about whether or not to adjust for nonresponse and for case-mix. To do this I observe the effect of adjustment on performance assessment and compare the results from the adjusted PIs to those from the raw PIs. Since adjustment methods generally introduce additional complexity into the calculation of PI scores for individual organisations, statistical adjustments are likely to be resisted unless they can be shown to substantially improve the validity of the results. Where there are clear effects on performance assessment, then there is a case for adjustment. In this section I outline the methods that I will use to inform judgements about the extent and importance of the effect of nonresponse and of case-mix adjustment on performance assessment.

Performance assessment through tabular and graphical display and implications for studying the effects of adjustment

There are many different ways to use PIs for performance assessment, which stems from the fact that PIs rarely have clear interpretations. What is a good, or conversely a bad, score is difficult to assess without a meaningful benchmark or information about how other organisations measure up against the same indicators. In this respect, "the success of performance indicator packages relies heavily on comparative data" (Smith 1990, p. 56). Performance assessment tends to be carried out by presenting PI scores, often graphically, for a collection of organisations to enable end-users to draw conclusions about the performance of their organisation *relative* to that of others. There is a sizeable literature that focuses on the best method for presenting PI scores, where the best method is the one that enables end-users with little statistical knowledge to easily and correctly interpret the data. Here I review the graphical and tabular methods used by policymakers to assess performance and consider what the choice of method means for performance assessment.

Over the years various methods have been used to present the results of organisational profiling studies. In the UK, the earliest efforts simply ranked organisations in order of PI scores, in 'league tables'. This approach is still popular and is used by newspapers to this day to rank schools and universities, as well as by international organisations to rank countries on their health care systems and so on. A variation on this theme, whereby the distribution of organisations is divided into (usually) quintiles and organisations are ranked according to the quintile ("banding") within which they fall, was used by the Audit Commission and associated regulatory bodies during the early days of the New Labour performance regime

(see e.g. Department of Health 2003, Commission for Social Care Inspection 2004). These approaches are widely seen as unfair, since they take no account of sampling error and random fluctuations in performance (Goldstein and Spiegelhalter 1996, Lilford et al. 2004, Spiegelhalter 2005a).

Methods that reflect the uncertainty in PIs are generally preferred and in this respect two methods have been regularly used. The first is the 'caterpillar plot', where PIs are plotted for each organisation along with (usually 95 per cent) confidence intervals, based on the variation within the data for the organisation. An example of a caterpillar plot based on the ASCS 2010/11 data is shown in Figure 4a. Although this method provides a visual display of the uncertainty in estimates, it is open to misinterpretation on two fronts. First, the ordering of organisations by their estimated scores implies a ranking of organisations that is spurious given the often wide confidence intervals (Spiegelhalter 2005a, Mohammed and Deeks 2008), and second, non-overlapping intervals are commonly and erroneously interpreted as evidence for a difference between organisations. The confidence interval provides a measure of the reliability of the estimate; a statistical test for the difference between two means uses the pooled standard error, which would produce a narrower interval[5] (Goldstein and Healy 1995).

As a result of these concerns around caterpillar plots, a second method, known as the 'funnel plot', is nowadays generally preferred (Mohammed and Deeks 2008). The funnel plot is inspired by Shewart's work on the control of variation (Mohammed et al. 2001). Organisations' scores on PIs are plotted against some measure of the precision of this estimate, usually the sample size (volume of cases) on which the estimate is based. Limits are set to distinguish cases that are "in control" from those that are "out of control". In control units are assumed to be subject to "common-cause variation", while out of control units are assumed to be subject to "special cause variability" (Spiegelhalter 2005a). By convention, limits are set at two or three standard deviations (SDs) from some target value (often the population average). An example of a funnel plot based on the ASCS 2010/11 data is shown in Figure 4b. These charts are perceived as better than caterpillar plots because they avoid spurious rankings and they illustrate clearly the greater expected variability in small samples. This method is currently used to identify outlying hospitals within the NHS PROMs programme and is also used by the regulator (Department of Health 2011b).

[5] This would need further adjustment for multiple comparisons.

Hospitals lying outside the three SD control limit are given an "alarm" status and those lying between the two and three SD control limits are given an "alert" status.

When discussing the effects of adjustment on performance assessment, a language is needed to describe the differences between the outlier status under the raw PI and the outlier status under the adjusted PI. It is common to treat the adjusted PI as the true situation and to therefore speak of movements from an in control state to an out of control state as a false negative and movements from an out of control state to an in control state as a false positive. Although this is a corruption of the idea of false negatives and positives, since we cannot know whether the adjusted PI represents the true situation (or even takes us closer to the true situation), it is useful for assessing the importance of adjustment. The concepts of false negatives and positives can be extended to calculate the type I and type II error rates, which provide a sense of the significance of the number of changes. I use these concepts in this thesis for this purpose, but to acknowledge the corrupted usage, I enclose them in inverted commas.

The form of presentation chosen – caterpillar or funnel plot – has consequences for performance assessment. Where caterpillar plots are used, attention is focused on the ranking of organisations and on the variability of individual PI scores, as shown by the confidence intervals. Where funnel plots are used, attention is instead focused on those organisations that are outliers because they are plotted outside of the funnelling control limits. The effects of precision are also highlighted by the position of organisations on the x-axis. In ASC, PIs still tend to be presented using caterpillar plots (Department of Health 2014), so it is important to understand the effects of adjustment on ranking in this context. I also assess the impact of adjustment on the identification of outlying organisations.

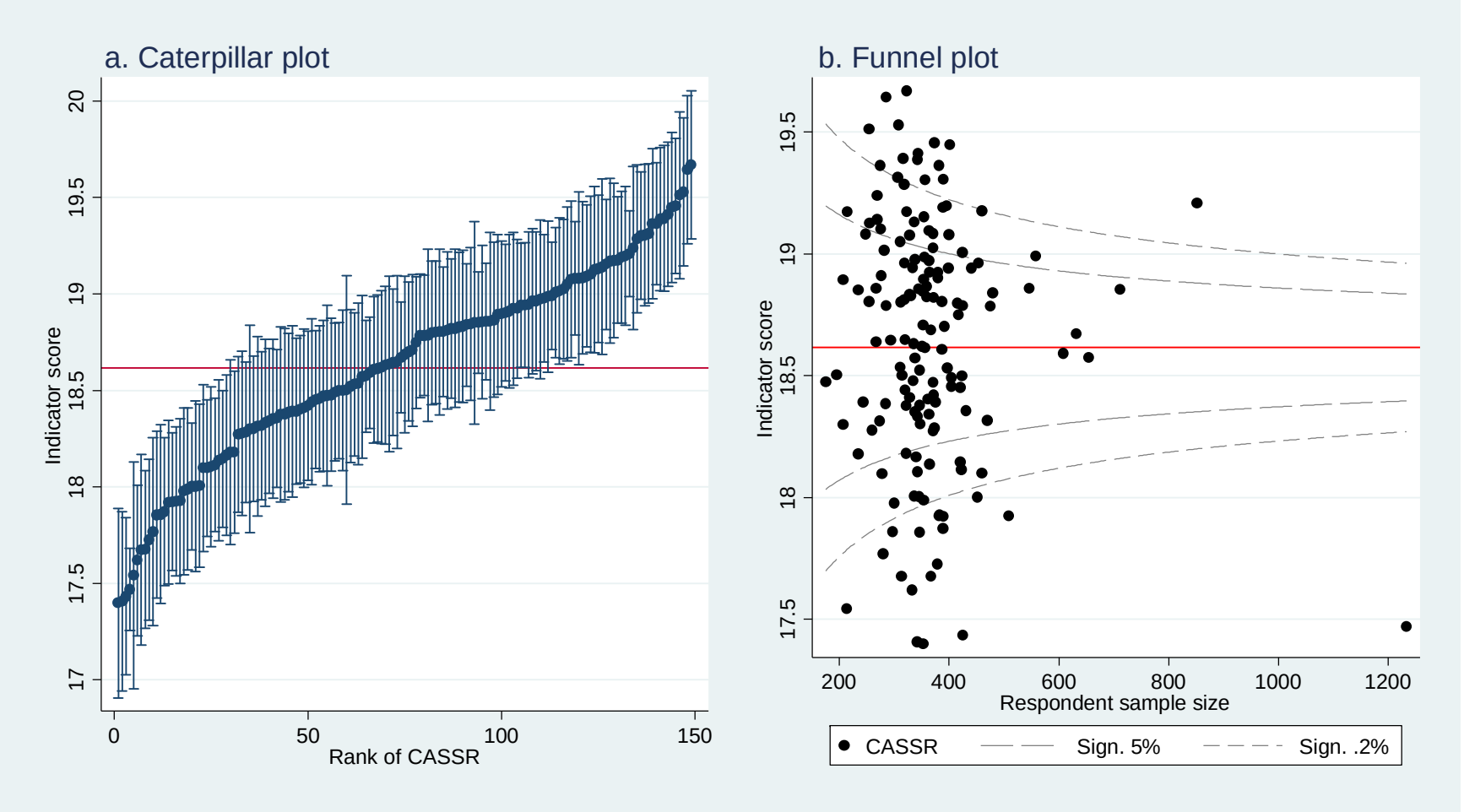

Figure 4: Caterpillar and funnel plots showing distribution of ASCOF SCRQoL PI scores for CASSRs, ASCS 2010-11

Decomposing the effect of adjustment: mean outcome, volume and variability effects

The methods proposed here aim to account for differences in the respondent samples between CASSRs arising from differences in the case-mix of the population served and differences due to nonresponse. An implicit assumption in much of the organisational profiling literature is that any differences in inferences about performance between the adjusted and raw PIs is due to removal of bias. This is, however, only part of the story. While adjustment can reduce bias, which I refer to following Gomes et al. (2016) as a "mean outcome effect", it can also affect performance assessment through changes to the sample size ("volume effect") or through changes to the variability of estimates ("variability effect"). I adapt Gomes et al.'s discussion of the effects of nonresponse adjustment on performance assessment to discuss the effects of adjustment for both nonresponse and case-mix. Like Gomes et al., I illustrate these ideas using funnel plots.

Considering nonresponse first, suppose point A in Figure 5 represents a CASSR with a low response rate. Its observed mean SCRQoL score (outcome) and respondent sample size (volume) are such that it is designated to be 'in control'. If all service users had responded to the survey, however, and the outcomes of the unobserved users were similar to the observed users, then the CASSR would be located at point B, which is no longer in control. This is labelled as a 'volume effect' in Figure 5. Under MI, where the sample size increases this will always have the effect of shifting the data points to the right. No volume effect will be observed using IPW methods, unless the effective sample size is used to plot the position on the x-axis. Since the effective sample size is a fairer measure of the precision of the PI (Deeks et al. 2005), it is appropriate to use this instead of the achieved sample size for IPW. Since weighting is usually less efficient, the effective sample size tends to be smaller than the achieved sample size. In such cases the volume effect will therefore be in the opposing direction, i.e. a shift to the left.

Since the MCAR assumption (that the outcomes of unobserved service users are similar to the outcomes of the observed service users) is unlikely to hold, the effect of nonresponse is also likely to be associated with a shift along the y-axis. Assuming missingness is MAR, the direction of the shift will depend on the pattern of nonresponse for individual CASSRs. In Figure 5, the CASSR would move from point B to C if the service users for whom data are missing have relatively better outcomes than those for whom data are observed. This would move the CASSR back into control status. Conversely the CASSR would move from point B to D, even further out of control if the unobserved sample had relatively worse outcomes than the observed sample. These effects

are identified as 'mean outcome effects' in Figure 5. There is also potentially a variability effect from adjustment, whereby changes in the variance of total population affect the position of the control limits. If the variability of the sample after adjustment is greater than in the unadjusted sample, then the control limits will be wider, so decreasing the likelihood of the CASSR being found to be out of control (and vice-versa).

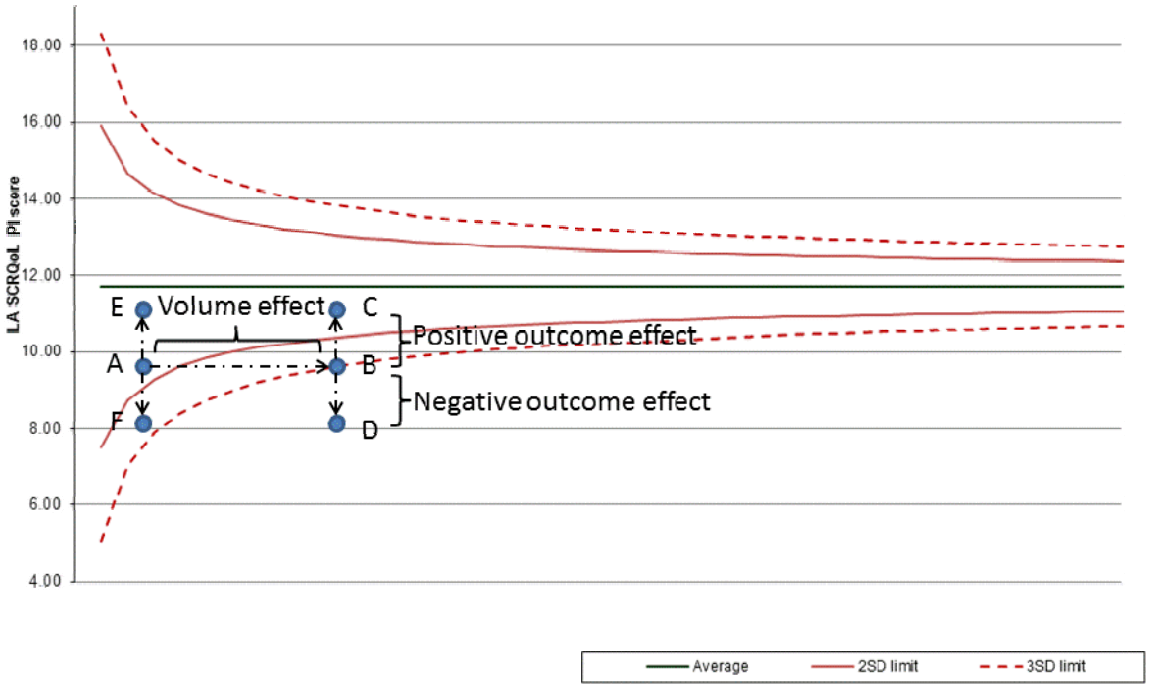

Figure 5: The implications of missing data for assessing performance using funnel plots

Adapted from Gomes et al. (2016)

Turning now to the effect of adjusting for case-mix on performance assessment, consider instead that point B in Figure 5 represents the raw PI score for a CASSR. After risk-adjustment, the CASSR would move from point B to C if the CASSR has a relatively unfavourable profile and from point B to D if it had a relatively more favourable profile. However, it is important to recognise that there may also be a volume effect resulting from risk-adjustment. The adjusted PIs are specified on the basis of calculating an adjustment factor for every sample member, defined in terms of the relationship between the observed outcome and the expected outcome, as discussed above. The adjustment factor can therefore only be calculated for cases with observations for all the risk adjustors and the outcome indicator. In most instances, under CCA, this will mean the adjusted PI is based

on fewer cases than the unadjusted PI. In Figure 5, the CASSR would therefore shift leftwards to point A after adjustment. Mean outcome effects resulting from adjustment are shown by the movements upwards from point A to E and downwards from A to F.

Although I have illustrated these effects with funnel plots, similar mean outcome effects will be observed using caterpillar plots. The volume effects, however, would not be as obvious since any changes in sample size would work through into changes in the confidence interval limits, which would also be affected by any changes in the variability of the sample post-adjustment. In caterpillar plots, therefore, the variability and volume effects are conflated making them a less useful tool for understanding the effects of nonresponse and case-mix adjustment upon performance assessment. Since the volume effect is likely to be significant where large amounts of data are missing, this is a limitation of using caterpillar plots to investigate the effects of adjustment on performance assessment. For this reason I make use of funnel plots to analyse the effects of adjustment on inferences about performance throughout this thesis, despite the fact that this method is not currently used within ASC.

Concluding remarks

In this chapter I have set out in broad terms the methods that I will employ in this thesis to address the effects of nonresponse and the effects of variations in the case-mix of CASSRs on performance assessment. For nonresponse, I have proposed using an IPW/MI method. For case-mix, I have proposed using risk-adjustment techniques, but have suggested a number of adaptations to the standard risk-adjustment model in order to address the clustering of observations within CASSRs and the dependency between resource inputs and need-related characteristics. To address clustering I have proposed using random and fixed effects models to estimate SCRQoL outcomes. I compare these estimation methods to OLS in Chapter 6. To address the dependency between resource inputs and need-related characteristics, I have suggested using a production function model. I will compare this model to a risk-adjustment model in Chapter 7.

The purpose of carrying out adjustment for case-mix and for nonresponse is twofold. First, one aim of the thesis is to determine which methods to use to adjust the PIs, should adjustment be necessary. All of the methods proposed can be considered as alternatives for adjusting the PIs. In the empirical chapters, I compare and contrast these methods to determine which is preferable. Secondly, I have proposed comparing the adjusted and unadjusted PIs in order to determine whether nonresponse and case-mix

differences between CASSRs have an impact on performance assessment. There is a degree of circularity in this strategy that deserves some reflection.

This strategy effectively equates the adjusted PI with true performance, which lacking any objective measure of true performance, is unavoidable. Nevertheless, it means careful attention should be paid to the validity of the adjusted PIs to assess whether there is good reason to believe that they are better estimates of true performance than the unadjusted PIs. The concepts of validity outlined in the first chapter are critical in this regard. I use them to structure my discussion of the empirical research in Chapters 5 and 8, examining whether adjustments for nonresponse and case-mix, respectively, have an effect on performance assessment. The lack of a measure of true performance also raises questions about how to assess the importance of differences between the unadjusted and adjusted PIs. There is no straightforward solution to this question, and I have suggested using the language of false positives and negatives to inform policy in this regard. I acknowledge, however, that stretches the concepts beyond their intended usage.

So far, I have set out the rationale and context for this thesis, and the theoretical frameworks and methods that will guide the empirical analyses. The empirical chapters form the core contribution of this thesis and it is to these I now turn. I begin with an analysis that examines who is missing from performance surveys and seeks to understand how response rates can be improved.

Chapter 4

Who is Missing from Survey-Based Performance Indicators and What Strategies Improve Response Rates? Analysis of the Factors Influencing Participation in the Adult Social Care Survey

Abstract

Objective: To determine the characteristics of service users that predict nonresponse to the ASCS and explore the factors related to survey delivery that may improve response rates.

Design: Auxiliary data on all those sampled as part of the English 2010-11 ASCS and paradata from the same survey were analysed using a multilevel multinomial logistic (MNL) model to understand the effects of individual- and CASSR-level observed factors on response propensity to the ASCS. Unobserved CASSR-level factors were controlled for through random-effects estimation.

Participants: The full sample of 150,672 cases, of which 61,026 are respondents and the remainder are nonrespondents. Nonrespondents comprise 16,294 cases who returned blank questionnaire and 73,352 cases who failed to return the questionnaire.

Outcome measures: Survey response which has three outcomes: respondent (0), returned a blank form (1), no response (2).

Results: Individual-level and CASSR-level factors were both important in determining response outcomes to the ASCS, but overall variation explained was low. The variables with the largest effects were age, where increasing age was associated with a greater probability of responding although this trend reversed for the oldest old; client group, where mental health clients were less likely to respond compared to people with physical disabilities, who were less likely to respond than people with learning disabilities; and the CASSR not chasing nonrespondents. The type of services received and social disorganisation variables, such as level of deprivation, had smaller effects. Unobserved CASSR-level factors also explained variation in the blank form and no response outcomes.

Conclusions: Younger age groups, mental health users and nursing home residents are underrepresented in the ASCS. The pattern of results suggests that disability severity may be important in determining nonresponse and future work should directly explore this effect. Chasing nonrespondents is important to improve response rates and should be enforced.

Introduction

Missing data is a problem for all research. The loss of data affects the precision of estimates and, where data is missing systematically, estimates may also be biased. In surveys missing data are primarily a result of nonresponse. Because survey nonresponse is a behaviour it is generally assumed that those who do not respond are systematically different from those who do respond, i.e. data are not missing completely at random. From a technical perspective precision and bias are the important issues affecting the validity of PIs in the context of nonresponse; for many users of the data it is low response rates and the suggestion that certain groups of people are missing from the sample that are a cause for concern since these indicate that the sample is unrepresentative of the population. If we are to improve response rates to performance surveys and address concerns about the representativeness of survey-based PIs, it is important to first understand who is missing from PIs and secondly how response rates can be improved. These two questions are the focus of the present chapter.

Public sector performance surveys differ in three principle ways from the types of general population surveys that are the staple of the nonresponse literature. First, they tend to be postal surveys. Secondly, since public services are often distributed on the basis of need, the populations surveyed are frequently dominated by marginalised and 'hard-to-reach' groups. Thirdly, the organisations being assessed are frequently also in charge of running and managing the surveys. While it is well-known that postal surveys have lower response rates than surveys conducted using other modes of administration (Dillman et al. 2009), there is little evidence about the importance of the other two aspects for nonresponse. Since these two aspects of performance surveys – individual characteristics and differences in how surveys are managed – are theoretically important determinants of survey participation (see Chapter 2), it is important to understand how these factors influence a person's likelihood of being a respondent to improve participation in performance surveys.

For this analysis I use the 2010-11 ASCS data. The analysis is novel in that exploits the natural variation across CASSRs to understand how individual characteristics and decisions regarding the management and design of the surveys by CASSRs affect response rates. While many studies have looked at the role of individual and interviewer characteristics on response propensity using this method (for a review see e.g. Groves et al. 1992), no studies have examined in any detail the effect of organisational factors on response propensity where the surveys are managed by multiple sites. I use multilevel modelling to explore the influence of variables at both the individual level and the CASSR level on survey participation. This method for modelling has not been applied widely in the literature and

where it has most studies have examined the role of interviewer characteristics alongside characteristics of the sample members (Johnson et al. 2006, Durrant and Steele 2009). The selection of variables for the modelling is guided by the literature reviewed in Chapter 2 to develop a conceptual framework for the analysis (see Figure 3). This framework is based on the modified version of social exchange theory proposed by Goyder (2006) that accommodates multiple theoretical influences over the participation decision. Previous empirical research has rarely tested the role of multiple theories or multiple influences on survey participation (Goyder et al. 2006, Durrant and Steele 2009). In addition to providing evidence regarding participation decisions in performance surveys, and specifically the ASCS, this study will also contribute to furthering our understanding of survey participation.

With respect to this thesis, the aim of this analysis is twofold: to help those administering performance assessment surveys to develop principled strategies for minimising nonresponse, and to improve our understanding of who is missing from performance surveys. The chapter is organised as follows. First I discuss the dataset for analysis. I then present the statistical model and describe the estimation approach used, the variables used to test the theoretical model in the ASC context, the modelling strategy and how I manage missingness within the auxiliary data. This is followed by the results. In the discussion I focus on the main findings and limitations. The wider implications are discussed in Chapter 9.

Data

In this analysis, I use the auxiliary data and paradata from the English 2010-11 ASCS. As explained in Chapter 1, the auxiliary data is captured from CASSR records and is available for both respondents and nonrespondents. The ASCS paradata captures information about the process of delivering and managing the survey in each CASSR. Although there is detailed guidance for the ASCS, CASSRs do have some discretion over how to implement various aspects of the survey process largely to allow them to improve response rates from hard-to-reach groups. As a result, a number of survey design and management features are mandatory, while others can be optionally introduced by CASSRs, as set out in Table 1.

In the 2010-11 wave of the ASCS, of the 150,672 questionnaires sent out[32], 73,278 were not returned (48.7 per cent), 16,294 were returned blank (10.8 per cent), and 61,100 were returned completed or partially completed (40.5 per cent). However, 74 questionnaires

[32] This excludes the eight questionnaires sent to people who should not have been included in the survey.

could not be linked back to the CASSR client records as the questionnaire identification number was missing, either because it had been removed by the service user (five cases) or through administrative error (in total 69 cases across two councils). These cases are treated as nonrespondents in this analysis (and subsequent analyses) as they lack any auxiliary data, leaving a total respondent population of 61,026 and a nonrespondent population of 73,352. Although this solution is not ideal, the number of affected cases is tiny in relation to the size of the sample (<0.1 per cent).[33]

[33] I also re-estimated the models excluding the affected CASSRs, but this did not alter the results.

Table 1: Survey design and management features and their status in the guidance

Survey design feature	Status in guidance
Incentives	Allowed but not recommended
Delivery mode (e.g. telephone, face-to-face, interpreter, supported completion)	Main mode required to be postal to avoid response bias. Alternative modes allowed and encouraged for chasing nonrespondents and engaging hard-to-reach groups
Different format or version of questionnaire (e.g. braille, audiotape, translated, large font)	Allowed and encouraged to engage hard-to-reach groups
Chasing nonrespondents	Alternative modes and formats/versions of the questionnaire allowed. Required to send at least one reminder and allowed to send a maximum of two reminders. Required to send reminders to whole nonrespondent population
Sampling frame	Required to exclude clients who lacked capacity to consent in line with ethics committee requirements. Allowed to remove people who had recently taken part in surveys. Allowed to remove people who had requested not to take part in future surveys. Allowed to exclude people in active dispute with the CASSR. No other exclusions allowed.
Supplementary sample	Not recommended in the guidance but allowed for CASSRs that failed to meet target for respondents[34]
Date questionnaires were sent out	Recommended mid to late January in order to complete fieldwork before the census date (end March), deviation discouraged but allowed
Addition of questions	Allowed, but neither encouraged nor discouraged
Modification of questions	Not allowed, standard questionnaire must be used

Adapted from NHS Information Centre (2010)

Statistical modelling

I use a multilevel multinomial logistic (MNL) model to understand the effects of different factors on response propensity to the ASCS. The multilevel model computes separate intercepts for each CASSR and estimates between-CASSR variance directly with the model parameter, $u_{ij}^{(s)}$. This allows me to test the hypothesis that CASSRs have an effect on response propensity. It also appropriately accounts for the clustering of responses within CASSRs in the standard errors, which is important to ensure correct inferences (Rabe-Hesketh and Skrondal 2008). I use a MNL model, rather than separate binary logistic

[34] Standard ASCS practice for determining sample size is that councils should aim to achieve a margin of error of no more than ± 5 per cent for each question (NHS Information Centre 2010).

models, to evaluate the effect of the various explanatory variables on each of the two nonresponse outcomes simultaneously. This also allows me to include and test for correlation between the unobserved CASSR influences on the different types of nonresponse.

The response variable for the ASCS data, which I denote y_{ij} for individual i in CASSR j, has three outcomes: respondent (0), returned a blank form (1), no response (2). The latter two outcomes (blank form and no response) are both nonrespondents. Although not a neat fit, those returning a blank form can be considered to be a mixture of 'refusals' and people who are 'unable to respond', since, in line with ethics committee requirements, the questionnaire cover letter instructed survey recipients to return a blank form if they could not or did not wish to participate in the survey. Those nonrespondents who do not return the form, are likely to be a mixture of non-contacts, refusals and people who are unable to respond, as well as people who run out of time or forgot to respond. Although the overlapping mix of reasons for being in these two outcome groups may mean that the factors predicting membership are not very different, I retain this distinction because it seems likely that different conditions, characteristics and motivations lead people to choose to return a blank form rather than not return a form. I test for the distinctiveness of the groups by testing for the equality of coefficients across the simultaneous equations, and find the MNL model to be appropriate.

As Durrant and Steele (2009) set out, the probability of y_{ij} having one of the three response outcomes is given by $\pi_{ij}^{(s)} = \Pr(y_{ij} = s)$, for s =0, 1, 2. If I take respondent (0) as the base category, the multilevel MNL model can be written,

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(0)}}\right) = \boldsymbol{\beta}_1^{(s)}\mathbf{x}_{ij}^{(s)} + \boldsymbol{\beta}_2^{(s)}\mathbf{s}_j^{(s)} + \boldsymbol{\beta}_3^{(s)}\mathbf{c}_j^{(s)} + \boldsymbol{\beta}_4^{(s)}\mathbf{n}_j^{(s)} + u_{ij}^{(s)}, \qquad s = 1,2 \qquad (11)$$

where $\mathbf{x}_{ij}^{(s)}$ is a vector of individual characteristics, $\mathbf{s}_j^{(s)}$is a vector of survey attributes, $\mathbf{c}_j^{(s)}$ is a vector of CASSR characteristics, $\mathbf{n}_j^{(s)}$is a vector of social environment variables, the $\boldsymbol{\beta}^{(s)}$s are vectors of coefficients and $u_{ij}^{(s)}$ is a random effect representing the unobserved CASSR-level characteristics. The statistical model in (11) is the counterpart to the theoretical model in equation (4) (see Chapter 2). It is a random-intercept model, in which the effect of the CASSR, j, is to increase the log-odds of returning a blank form or being a nonrespondent versus being a respondent by the amount $u_j^{(s)}$.

Equation (11) above consists of two simultaneous equations. The first ($s = 1$) models the logarithm of the ratio of the probability of returning a blank form to that of being a respondent as a function of covariate individual and CASSR effects; the second ($s = 2$) models the logarithm of the ratio of the probability of providing no response to that of being a respondent as a function of covariate individual and CASSR effects. Although it is possible to include different covariates as predictors of the different outcomes, lacking a clear rationale for selecting different covariates I do not attempt this here; rather I include the same covariates in both models and compare the effects of covariates on both types of nonresponse.

The CASSR random effects are assumed to follow a bivariate normal distribution, and are similarly specific to the outcome. Thus, $u_j = (u_j^{(1)}, u_j^{(2)}) \sim N(0, \Omega)$ where,

$$\Omega = \begin{pmatrix} \sigma^{2(1)} & \\ \sigma^{(12)} & \sigma^{1(2)} \end{pmatrix}.$$

The two variance parameters $\sigma^{2(1)}$ and $\sigma^{2(2)}$ are the residual between-CASSR variances in the log-odds of returning a blank form versus being a respondent and the log-odds of providing no response versus being a respondent, respectively. They can be interpreted as the effect of unobserved CASSR effects on the two nonresponse outcomes. The parameter $\sigma^{(12)}$ is the covariance between the unobserved CASSR influences on the probabilities of returning a blank form and providing no response. A positive covariance indicates that CASSRs with high rates of blank form nonrespondents have high rates of no response nonrespondents.

To help with interpreting the model (and assess the fit of the model), predicted probabilities for each outcome can be calculated. As the log-odds of an outcome compared to itself is zero the effects of any independent variables are also zero for the base category. The simultaneous equations can be solved, by rearranging model (11) such that the probability of observing outcome s given $\mathbf{x}$ with the reference category given by r, is:

$$\pi_{ij}^{(s)} = \frac{\exp(\beta^{(s)}\mathbf{x}_{ij}^{(s)} + u_{ij}^{(s)})}{1 + \sum_{r=1}^{2} \exp(\beta^{(r)}\mathbf{x}_{ij}^{(r)} + u_{ij}^{(r)})}, \qquad s = 1,2,$$
$$\pi_{ij}^{(0)} = 1 - \pi_{ij}^{(1)} - \pi_{ij}^{(2)} \tag{12}$$

In addition, marginal effects can be calculated by taking the derivative of the probability with respect to a given predictor $\mathbf{x}$. Predicted probabilities and marginal effects are all calculated in Stata 14 using the margins command.

The final models are estimated in Stata 14 using either of mean-and-variance or mode-and-curvature adaptive Gauss-Hermite quadrature, depending upon which method led to convergence. Due to difficulties attaining convergence using Stata's starting values, I first estimated the multilevel MNL models in the MLWIN software and passed these estimates to Stata as starting values[35]. I performed sensitivity checks of the converged estimates to the quadrature approximation, by increasing the number of integration points and comparing the estimates obtained with different numbers of integration points. The blank form random effect, the constants, some of the derived and CASSR-level variables were all sensitive to the quadrature approximation, but increasing the number of integration points to 20 provided estimates that had adequate stability.

Explanatory variables

The choice of explanatory variables to be tested is driven by the hypotheses developed from the conceptual model in Chapter 2, and by the availability of data in CASSR records. I also supplement the ASCS data with area-level data drawn from a variety of freely-available sources to explore the effect of the social environment variables on response propensity. The variables tested for inclusion in the model are set out in Table 2. The distribution of the auxiliary data used in this analysis is shown in Table 3; the distribution of the area-level variables in Table 4 and the distribution of the paradata in Table 5. For this analysis, only the individual characteristics are measured at the individual level; survey attributes, organisational characteristics and the characteristics of the social environment are all measured at the CASSR level.

[35] The models were estimated in MLWIN using Markov chain Monte Carlo methods (Browne 2015). Non-informative priors were assumed for all parameters and estimates were captured after 150,000 chains with a burn-in of 1,500, using approximate quasi-likelihood estimates (Goldstein 2003, Durrant and Steele 2009) as starting values. All MLWIN models were run using the runmlwin (Leckie and Charlton 2013) command from Stata 12.

Table 2: Variables tested in the response propensity models

Variable	Rationale
Service receipt • Nursing home resident • 'personal care' home resident • Home care • Direct Payments • Low-level services • Equipment	• Service receipt reflects the underlying disability of service users and may also be related to the saliency of the survey. People receiving high-contact, high-frequency and high-care services (e.g. Direct Payment, home care, nursing and care home residents), are likely to see the questionnaire as more salient than users of low-level services (defined as people receiving one of meals, day care or 'other', e.g. transport, services), and one-off services (e.g. equipment users). • Direct Payments users are differentiated from home care users as they may consider the survey as a way of monitoring their ability to manage their own care, and therefore be more reticent to respond. • Residents of care home (nursing or personal care) will be more severely disabled than other service users.
Socio-demographic • Age group • Ethnicity • Gender	• Effect is likely to be due to intervening variables, such as disability. • I expect young and very old to be less likely to respond, but these effects may be weak due to lower levels of employment and weak correlation between disability severity and age in this population. • People with white ethnic background are likely to have higher response rates.
Client group • Physical and sensory disability (PSD)[a] • Learning disability (LD) • Mental health (MH) • Substance misuse (SM)	• These are proxy indicators for health conditions and disability, which are likely to influence survey participation.
Count of services[b]	• This is a proxy for disability severity, as more disabled people are likely to receive more services.
Number of auxiliary items fully observed for a given CASSR[c]	• This is a proxy for the quality of the CASSR client records, which may affect whether people receive the survey.
Social environment • Population density (people/hectare)[d] • Authority type[d] • Indices of multiple deprivation average score (IMD score)[e]	• IMD score is a proxy for social cohesion • Authority type is a proxy for urbanicity • More urban, dense and less social cohesive places are likely to have lower response rates.

Variable	Rationale
Survey design variables • added questions • used incentives • did not chase nonrespondents • did not report using any strategies to engage hard-to-reach groups • Proportion of people removed due to lacking capacity	• Theory suggests that aspects that increase the benefits of participation, e.g. incentives, and reduce the costs, e.g. length of questionnaire, will increase response rates. • Chasing nonrespondents and removing people who lack the capacity to respond should increase response rates. • Strategies to engage hard-to-reach groups are likely to reduce the costs of participation for ill and disabled people, so will increase response rates.

Legend: [a] This group also includes a small number of 'vulnerable people'; [b] This indicator is derived from counting the number of services received (maximum 10, censored at 6); [c] This indicator is derived from counting the number of fully observed auxiliary items (across all requested variables not just those tested in the analysis models) for a given CASSR (maximum 24); [d] From ONS (Office for National Statistics 2001). Authority type is recorded as one of four categories: London, Metropolitan area (not London), Unitary Authority (urbanized area, but not large like the Metropolitan areas), County Council (largely rural authority with some smaller towns or cities); [e] Measures deprivation in seven areas (income, employment, health, education, housing and services, living environment and crime) and is suggested by the authors as a measure of community cohesion (McLennan et al. 2011).

Modelling strategy

The aim of the statistical analysis is to develop a statistical model to explore the factors influencing survey participation. I use a hierarchical modelling approach to make it easier to understand the interactions between the different types of variables. First, I estimate the null model (model 0), which contains no independent variables or random effects. This model is extended to include the CASSR random effects (model 1). I then introduce the variables in blocks. Model 2 includes individual-level covariates, model 3 includes the CASSR-level variables covering CASSR characteristics, the social environment and survey design variables.

The residual between-CASSR variances and covariances provide insight into the effect of unobserved CASSR effects on the probabilities of returning a blank form and providing no response. The concept of the variance partition coefficient (VPC), in which the residual variation in response outcomes is apportioned between the different analytical levels, is useful in this context since it provides insight into "the level 'at which the action lies'" (Browne et al. 2005, p. 600). Here I use the 'latent variable' method[36] to provide an

[36] In this method, I assume that the propensity to respond is continuous but given the timeframes of the survey and for other reasons, what is observed is only the trichotomous outcome – respond, respond with blank form or did not respond. Since I am using the MNL regression model, the underlying continuous variable comes from a logistic distribution, with a variance of $\pi^2/3$, and this constant is substituted into the VPC formula for the level

approximation of the proportion of the total residual variation in response outcomes that is explained by between-CASSR differences (Goldstein et al. 2002). The comparison between the null model and model 1 is a test of the significance of the random effect. The comparison between models 1 and 2, and 2 and 3 helps to understand the effect of the inclusion of covariates on between-CASSR variation in response propensity, albeit the order of introduction of covariates affects the interpretation of these results. Since adjusting for individual-level factors may reduce between-CASSR variance if people with a low propensity to respond are clustered within certain CASSRs, and adjusting for area/CASSR-level factors will reduce between-CASSR variance, such comparisons provide an indication of how much between-CASSR variation is unexplained variation.

At all stages of the modelling, because of the complexity of the models, I only retain those variables that are significant at the five per cent level or interact with other variables. As there are potentially a large number of interactions I only test those that are substantively interesting. No interaction effects and very few CASSR-level main effects were retained in the final models due to difficulties attaining convergence and a lack of stability in the estimates of these variables.

Missing data

The auxiliary data are subject to item nonresponse due largely to differences in procedures and systems for recording client data within CASSRs. Out of the 150,672 cases, only 123,805 (82 per cent) have full information for all the explanatory variables tested and only 128 of the 149 authorities (86 per cent) are represented in complete case analysis. While it is possible to recover the missing auxiliary data on the characteristics of the full sample, software limitations mean it is not possible to estimate the multilevel MNL models on imputed datasets[37]. Therefore in similar style to Durrant and Steele (2009), I estimate the models on the complete cases.

---

one variance. The Taylor series approximation and simulation methods are more accurate, but neither is straightforward to implement in standard software. Since the latent variable method provides reasonable approximations where the underlying probabilities are not close to zero or one (Goldstein et al. 2002, Browne et al. 2005), it is probably accurate enough in determining the VPC for the nonresponse category but may be less accurate for the blank form category.

[37] The gsem procedure is not compatible with mi estimate, which is necessary to combine the output from each of *m* estimations using Rubin's (1987) rules.

Results

Descriptive analysis

The individual-level explanatory variables tested in the models are shown in Table 3, by response status. When crossed with response status several of the groups are quite small (e.g. non-white ethnic groups) and in such cases response levels were collapsed in the regression analysis as indicated in the table. I ran chi-squared tests and analysis of variance, as appropriate. All of the variables have a significant association ($p<.001$) with response rate. It is clear from the row percentages, however, that response rates differ across the different variables. For example, the very young, people with mental health problems, nursing home residents and people receiving low-level services have a low response rate, whereas people with learning disabilities have a high response rate. There also seem to be differences across nonrespondent groups, with people receiving low-level services having a particularly high rate of return of blank forms, while the youngest age groups have much higher numbers of non-respondents. Some groups, for example people with mental health problems and nursing home residents, have high rates of both returning blank forms and not returning a form.

The CASSR-level variables capturing CASSR characteristics and the social environment that were tested for inclusion in the model are summarised in Table 4, along with their association with response, blank form and nonresponse rates. For the continuous variables, the Pearson correlation coefficient is reported, which is significant at the five per cent level for both deprivation and population density. Although the data quality indicator (count of auxiliary variables) is not significant at the five per cent level, the relationship with response rate and blank form rates borders on significance at the ten per cent level, with $p=0.12$ and $p=0.13$ respectively. For authority type, the F-test is reported. Below for each authority type there are the rates (with standard deviations) for each outcome. There is very little difference in blank form rates between authority types, but there do appear to be some differences in response rates and to a lesser extent in nonresponse rates. Differences in response rates are in fact significant at the ten per cent level, although nonresponse rates are not significant ($p=0.16$).

Table 3: Distribution of individual characteristics within the sample and across response outcomes

Variable	Category	Respondent (n=61,026)	Blank form (n=16,294)	Form not returned (n=73,352)	Total (n=150,672)
		N (row %) / Mean (SD)	N (row %) / Mean (SD)	N (row %) / Mean (SD)	N (col. %) / Mean (SD)
Gender	Female	38,842 (41.0)	10,640 (11.2)	45,151 (47.7)	94,633 (62.8)
	Missing	41	2	28	71
Age group	18-34	4,095 (31.3)	1,093 (8.4)	7,876 (60.3)	13,064 (8.7)
	35-44	3,892 (35.4)	963 (8.8)	6,128 (55.8)	10,983 (7.3)
	45-54	5,516 (39.2)	1,190 (8.5)	7,351 (52.3)	14,057 (9.3)
	55-64	6,390 (42.7)	1,330 (8.9)	7,261 (48.5)	14,981 (10.0)
	65-74	8,189 (43.0)	2,020 (10.6)	8,850 (46.4)	19,059 (12.7)
	75-84	14,923 (42.1)	4,189 (11.8)	16,361 (46.1)	35,473 (23.6)
	85 and over	17,948 (41.8)	5,488 (12.8)	19,462 (45.4)	42,898 (28.5)
	Missing	73	21	63	157
Ethnicity	White	55,002 (41.3)	14,391 (10.8)	63,842 (47.9)	133,235 (90.0)
	Missing	860	314	1,379	2,553
Primary client group	PSD[a]	44,957 (42.5)	11,922 (11.3)	48,806 (46.2)	105,685 (70.4)
	MH[b]	6,945 (28.6)	3,163 (13.0)	14,191 (58.4)	24,299 (16.2)
	LD	8,876 (44.3)	1,137 (5.7)	10,043 (50.1)	20,056 (13.4)
	Missing	248	72	312	632
Personal care only home	Resident	9,769 (42.2)	2,195 (9.5)	11,203 (48.4)	23,167 (15.4)
	Missing	26	2	36	64
Nursing home	Resident	2,619 (33.9)	1,038 (13.4)	4,075 (52.7)	7,732 (5.2)
	Missing	533	162	595	1,290
Home care	Recipient	21,848 (43.9)	4,467 (9.0)	23,504 (47.2)	49,819 (33.4)
	Missing	557	157	830	1,544
Direct Payments	Recipient	5,392 (43.8)	955 (7.8)	5,967 (48.5)	12,314 (8.4)
	Missing	1,345	411	1,732	3,488
Low level services[c]	Recipient	8,961 (34.6)	3,748 (14.5)	13,179 (50.9)	25,888 (17.9)
	Missing	1,921	539	3,196	5,656

Variable	Category	Respondent (n=61,026) N (row %) / Mean (SD)	Blank form (n=16,294) N (row %) / Mean (SD)	Form not returned (n=73,352) N (row %) / Mean (SD)	Total (n=150,672) N (col. %) / Mean (SD)
Equipment	Recipient	13,412 (43.4)	3,665 (11.8)	13,859 (44.8)	30,936 (21.3)
	Missing	2,043	451	3,049	5,543
Count of services[d]	Mean (SD)	1.45 (0.77)	1.39 (0.75)	1.39 (0.76)	1.41 (0.77)
	Missing	8,867	2,118	11,631	22,616

Legend: [a]also includes vulnerable people client group, [b]also includes substance misuse users, [c] includes people in receipt of meals, day centres and other services; [d]censored at 6.

Table 4: Distributional statistics for CASSR characteristics and social environment variables (n=149)†, and association with response outcome rates

Variable	Category	Mean (SD) / N (%)	Association with…		
			R rate r / F R rate (SD)	BF rate r / F BF rate (SD)	NR rate r / F NR rate (SD)
Count of fully observed auxiliary variables		18.63 (2.94)	0.128	-0.126	0.010
Deprivation		23.23 (8.56)	-0.205*	-0.162*	0.251**
Population density		23.84 (26.52)	-0.162*	-0.200*	0.243**
Authority type			2.12	0.40	1.75
	London	31 (20.8)	0.39 (0.05)	0.10 (0.04)	0.51 (0.07)
	Metropolitan	36 (24.2)	0.43 (0.07)	0.10 (0.09)	0.47 (0.12)
	Unitary	48 (32.2)	0.44 (0.11)	0.11 (0.10)	0.45 (0.16)
	County	34 (22.8)	0.41 (0.08)	0.09 (0.11)	0.51 (0.13)

Legend: R rate, response rate; BF rate, blank form rate; NR rate, nonrespondent rate; †except for density where n=148; * p<.05; ** p<.01; *** p<.001; F-test reported for authority type; Pearson correlation coefficient reported for count of fully observed auxiliary variables, deprivation, and population density

Table 5 summarises the degree of variation in the paradata that was collected by CASSRs about the management and conduct of the survey. A number of CASSRs took advantage of opportunities to modify the survey and some deviated from the guidance, e,g. by not chasing nonrespondents (with any form of reminder). In general, however, fairly small numbers of CASSRs reported using alternative methods to chase nonrespondents, engage with minority groups or add more questions to the survey. Differences between CASSRs in the percentage of people removed due to a lack of capacity to respond, which takes a range of values from zero to 58 per cent, is likely explained by differences in the ability of CASSRs to execute this task with the same degree of thoroughness. This was a requirement of the ethics committee, and CASSRs were in general unprepared to conduct this onerous task.

Table 5: Extent of variation in survey management and deviation from the guidance across CASSRs (n=149), and association with response outcome rates

Type of variation	Mean (SD) / N (%)	Association with… R rate r / F R rate (SD)	BF rate r / F BF rate (SD)	NR rate r / F NR rate (SD)
Used incentives		0.07	0.53	0.50
Yes	12 (8%)	0.42 (0.06)	0.12 (0.18)	0.46 (0.16)
No	137 (92%)	0.42 (0.08)	0.1 (0.08)	0.49 (0.12)
How nonrespondents were chased				
Post		7.48***	5.45*	12.61***
Yes	127 (85%)	0.42 (0.08)	0.11 (0.1)	0.47 (0.12)
No	22 (15%)	0.38 (0.07)	0.06 (0.03)	0.57 (0.07)
Email		1.19	0.02	0.36
Yes	5 (3%)	0.46 (0.21)	0.09 (0.05)	0.45 (0.25)
No	144 (97%)	0.42 (0.07)	0.10 (0.09)	0.49 (0.12)
Phone		2.18	0.85	2.68
Yes	17 (11%)	0.39 (0.07)	0.08 (0.05)	0.53 (0.10)
No	132 (89%)	0.42 (0.08)	0.10 (0.09)	0.48 (0.13)
Interview		1.13	1.31	0.02
Yes	5 (3%)	0.45 (0.05)	0.05 (0.02)	0.49 (0.05)
No	144 (97%)	0.42 (0.08)	0.10 (0.09)	0.48 (0.13)
no chasing took place†		8.04**	2.74	9.51**
Yes	16 (11%)	0.37 (0.07)	0.06 (0.03)	0.57 (0.07)
No	133 (89%)	0.42 (0.08)	0.10 (0.09)	0.47 (0.12)
How engaged minority groups				
IC translations		0.02	1.48	0.63
Yes	33 (22%)	0.42 (0.07)	0.08 (0.03)	0.50 (0.08)
No	116 (78%)	0.42 (0.08)	0.10 (0.10)	0.48 (0.13)
local translations		0.03	5.34*	3.25
Yes	8 (5%)	0.42 (0.04)	0.17 (0.20)	0.41 (0.17)
No	141 (95%)	0.42 (0.08)	0.09 (0.08)	0.49 (0.12)
interpreter via phone		2.86	0.00	1.15
Yes	21 (14%)	0.39 (0.06)	0.10 (0.14)	0.51 (0.14)
No	128 (86%)	0.42 (0.08)	0.10 (0.08)	0.48 (0.12)
interpreter face-to-face interview		0.20	0.99	0.20
Yes	12 (8%)	0.43 (0.10)	0.07 (0.04)	0.5 (0.11)
No	137 (92%)	0.42 (0.08)	0.10 (0.09)	0.48 (0.12)
friend/family provide interpretation		2.12	0.40	1.75
Yes	30 (20%)	0.42 (0.10)	0.09 (0.04)	0.49 (0.13)
No	119 (80%)	0.42 (0.07)	0.1 (0.10)	0.48 (0.12)
no engagement†		0.35	1.10	0.15
Yes	78 (52%)	0.41 (0.07)	0.11 (0.10)	0.48 (0.13)
No	71 (48%)	0.42 (0.09)	0.09 (0.08)	0.49 (0.12)

Type of variation	Mean (SD) / N (%)	Association with… R rate r / F R rate (SD)	BF rate r / F BF rate (SD)	NR rate r / F NR rate (SD)
Per cent of sample removed due to lacking capacity‡	11.45 (13.4)	0.050	-0.126	-0.059
Added questions		0.45	0.14	0.03
Yes	27 (18%)	0.43 (0.06)	0.09 (0.12)	0.48 (0.12)
No	122 (82%)	0.42 (0.08)	0.1 (0.08)	0.49 (0.12)

Legend: †These categories were generated from responses to the questions and comments. The question was not directly asked of CASSRs; ‡Three CASSRs did not provide a response to this question; * p<.05; ** p<.01; *** p<.001; F-test reported for incentives, nonrespondent chasing methods, engagement of minority groups, and addition of questions; Pearson correlation coefficient reported for per cent of sample removed due to lacking capacity.

Table 5 also shows the association between the indicators and response, blank form and nonresponse rates for CASSRs. For the majority of the variables, the F-statistic from an analysis of variance test is reported, and appearing below for each of the binary variables, the rates (with standard deviations) for each outcome type. Very few variables show significant differences in response outcome rates at the five per cent level. It is likely that this is a consequence of small numbers of CASSRs reporting using the alternative survey attributes. For the percentage of the sample removed due to lacking capacity, the Pearson correlation coefficient with each type of response outcome rate is reported, but it is not significant for any of the response outcomes. The negative relationship with blank form rates, however, is close to significance at the ten per cent level (p=0.13).

In addition to these changes and deviations from the guidance, CASSRs differed from each other in the management of the survey in a number of other ways. Some CASSRs also modified the formatting of questions[38] (2, 1%), surveyed a second supplementary sample in order to meet the target sample (6, 4%), used telephone interviews as the preferred mode of delivery (1, 1%), sent out questionnaires in batches rather than on one day (1, 1%), and chased a sample of nonrespondents rather than all nonrespondents (2, 1%). As well as excluding people who lacked the capacity to respond, some CASSRs reported that they had excluded mental health clients from the sample due to problems generating a sampling frame for this group (3, 2%), and excluded those people who had recently participated in a survey run by the council (13, 9%).

[38] One authority modified the yellow highlighting to make it specific to its situation and the other authority modified the formatting of the numbers slightly for its form recognition software. These are not major alterations.

Interestingly a number of CASSRs (10, 7%) reported that they had carried out their own surveys or consultations very close to or over the period of the ASCS, which suggests that survey fatigue may be a problem. Partly to avoid clashes with other activities, CASSRs also differed in when they sent out their questionnaires. While most started in late January (38, 26%), early February (73, 49%) and late February (39, 26%), some started in early January (4, 3%) and others went into early March (22, 15%) and late March (7, 5%). Although it would have been interesting to explore many of these variables in more detail, since this information was volunteered rather than requested, I could not be certain that other CASSRs had not implemented these or very similar features into their surveys. This underlines the importance of including a CASSR indicator in the models to capture the effect of these unobserved variables.

CASSR random effects

Table 6 shows the results of the random effects covariance matrix for each of the model specifications, with associated correlations and VPC. In addition the table shows the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), which balance model fit with model complexity. A comparison of the AIC and BIC statistics between model 1 and the null model (AIC difference: 236,287-219,138=17,149; BIC difference: 236,307-219,187= 17,119), and the likelihood ratio test ($\chi^2(3)$=17,154.6, p<0.001), suggests that there is between-CASSR variation in response propensity. The highly significant (p<.001) negative random effect correlation means that CASSRs with high return of blank forms have low levels of nonrespondents and vice-versa, suggesting that there is to some extent a trade-off between the two nonresponse categories. In model 1, the VPC is higher for the blank form than the nonrespondent outcome, at 17 and 15 per cent respectively, but both VPCs are relatively high, indicating that a substantial proportion of the variation in response outcomes is due to differences between CASSRs. Importantly, the correlated random effects model is significantly better than a model with shared effects (likelihood ratio test: $\chi^2(2)$= 14,138.4, p<0.001; AIC difference: 233,273-219,138=14,134; BIC difference: 233,302-219,187=14,115), with much larger estimates for between CASSR variation (shared random effect variance = 0.120, VPC=3.5%, not shown).

The addition of individual-level variables to the model (model 2), leads to a large reduction in AIC and BIC statistics (AIC difference: 219,138-215,204=3,934; BIC difference: 219,187-215,584=3,603). The likelihood ratio test is also highly significant ($\chi^2(34)$= 4,002.1, p<0.001) suggesting that the model with covariates is an improvement over

model 1 without covariates. The inclusion of individual-level variables also reduces the variance of the blank form random effect, but not the nonrespondent random effect. The conditional VPC for the blank form outcome is also reduced from 17 to 16 per cent. This suggests that at least some of the differences between CASSRs in the blank form outcome are explained by differences in the characteristics of individuals within the sample. The correlation between the two random effects slightly decreases after inclusion of individual-level covariates, suggesting that some of the correlation is driven by similarities in the characteristics of the individuals.

After inclusion of the CASSR-level covariates, the BIC (215,584-215,619=-36) and AIC (215,204-215,201=3) statistics suggest that the model with CASSR-level covariates is not an improvement over the model with only individual-level covariates. However, a likelihood ratio test suggests that model 3 is an improvement over model 2, at least at the five per cent significance level ($\chi^2(4)=11.44$, $p=0.022$). Addition of the CASSR-level variables also leads to a reduction in the random effects variances and the VPCs, suggesting that they may explain some of the differences between CASSRs beyond what is explained by individual characteristics. The reduction in the covariance and correlation between the two random effects also suggests that the CASSR-level variables make a unique contribution to explaining the correlation between CASSRs in the rate of return of blank forms and non-return of forms. For these reasons and despite the contrary indications from the AIC and BIC statistics the final model is 3.

Table 6: Estimates† (with SEs) of the between-CASSR variance-covariance matrix, under alternative model specifications

Parameter	Model 1 (n=123,805)	Model 2 (n=123,805)	Model 3 (n=123,805)
Var ($u_j^{(BF)}$)	0.677 (0.096)	0.633 (0.094)	0.61 (0.09)
Var ($u_j^{(NR)}$)	0.586 (0.086)	0.591 (0.086)	0.542 (0.08)
Covariance	-0.201 (0.053)	-0.178 (0.058)	-0.143 (0.041)
Correlation	-0.319	-0.291	-0.249
VPC $^{(BF)}$	0.171	0.161	0.156
VPC $^{(NR)}$	0.151	0.152	0.141
AIC	219,138	215,204	215,201
BIC	219,187	215,584	215,619
Log-Likelihood	-109,564	-107,563	-107,558

Legend: BF blank form outcome, NR nonrespondent outcome, † using mode-curvature adaptive Gauss–Hermite quadrature for model 2 and 3 and mean-variance adaptive Gauss-Hermite quadrature for model 1. Model 1 contains only random effects for the CASSR. Model 2 extends Model 1 and includes individual-level independent variables. Model 3 extends Model 2 and also includes CASSR-level independent variables.

Factors affecting response propensity

The final model (model 3) from the hierarchical block procedure is shown in Table 7. All individual-level variables except gender are significant in the model. By contrast very few CASSR-level variables are found to be significant, probably due to small numbers. Despite including a fairly large number of predictors, the percentage of predictions correctly classified is fairly low at 55 per cent. The final column of Table 6 suggests that the MNL model is preferable to a binary logistic regression of response-non response. The estimates for the effect of the covariates on returning a blank form or not returning a form (compared to being a respondent) vary significantly for the majority of the covariates. Only the estimates for the MH client group (as opposed to PSD), being aged between 35 and 44 years (as opposed to between 18 and 34), and low-level services are not significantly different from each other.

Table 7: Multilevel MNL regression of response propensity (model 3, n=123,805)

	Blank form†		Nonrespondent†		B^{BF}-B^{NR}=0
	RRR	SE	RRR	SE	χ^2
Fixed part					
MH[a]	1.595***	0.052	1.611***	0.032	0.1
LD[a]	0.522***	0.025	0.760***	0.018	59.8***
Age: 35-44[b]	0.837**	0.053	0.761***	0.025	2.3
Age: 45-54[b]	0.785***	0.047	0.631***	0.020	13.2***
Age: 55-64[b]	0.772***	0.046	0.536***	0.017	37.8***
Age: 65-74[b]	0.983	0.057	0.508***	0.016	131.4***
Age: 75-84[b]	1.131*	0.063	0.507***	0.015	212.7***
Age: 85 and over[b]	1.329***	0.074	0.514***	0.016	295.0***
White	1.122**	0.046	0.810***	0.019	59.2***
Count of service types	0.790***	0.041	0.897***	0.027	6.9**
Count of service types – squared	1.047***	0.012	1.027***	0.007	
Nursing Home	1.235***	0.067	1.342***	0.046	2.5***
Residential Home	0.786***	0.034	0.928**	0.023	14.1***
Home Care	0.652***	0.026	0.956*	0.022	92.0***
Low-level services	1.074	0.045	1.072**	0.027	0.0
Direct Payment	0.655***	0.035	0.872***	0.025	28.0***
Equipment	1.090*	0.041	0.950*	0.021	13.0***
IMD deprivation score	0.987	0.008	1.016*	0.008	5.3*
No chase	0.661	0.157	1.688*	0.366	6.9**
Constant	0.305***	0.069	1.601*	0.324	
Model fit statistics					
AIC				215,201	
BIC				215,619	
% correctly classified				55.1%	

Legend: †Base category: Respondent; [a] Base category: Physically disabled; [b] Base category: 18-34 years; * p<.05; ** p<.01; *** p<.001

The relative risk ratios convey the effect of each of the variables on the odds of either returning a blank form compared to returning a completed form, or not returning a form compared to returning a completed form. This can be useful for interpreting the model, but here I am more interested in understanding the effects of the variables on the probability of a sample member being a respondent, returning a blank form or not returning a form. I have therefore estimated the average marginal effects for each of the model variables, as shown in Table 8. The average marginal effects for the continuous variables, count of services and deprivation score, are more difficult to interpret since they capture the average instantaneous rate of change. The effect of continuous variables is more easily understood when the predicted probabilities and the marginal effects are estimated at different values of the continuous variable, and the resultant graphs interpreted together, since the marginal effects point to the region over which the effect of the variable on the probability of the outcome changes both significantly and most rapidly. Predicted probabilities and marginal effects at representative values are illustrated separately in Figure 6 and Figure 7, respectively, for the number of services received and similarly in Figure 8 and Figure 9 for the deprivation score.

The average marginal effects of the variables are in general larger for the respondent and nonrespondent outcomes than the blank form outcome, although this difference in the size of the effects is likely to be due to the much lower observed (and predicted) probability of returning a blank form[39]. For all outcomes, the variables with the largest effects are age group (particularly the comparison between those age groups over 55 and the age group 18 to 34), client group (particularly between the MH and PSD groups) and not chasing nonrespondents. For the last of these, the standard error is very wide because of the small number of CASSRs that did not chase respondents; consequently there is more uncertainty over the effect of this variable. In general the different type of services received and being white do not have a large (and in some cases significant) effect on the probability of any of the response outcomes. The effect of the number of services received and the deprivation score is more variable, but in some cases the predicted probability of the outcome can be different for people at different ends of these scales.

Considering the respondent outcome first, the discrete variable with the largest effect is age. Age has a positive effect on the probability of being a respondent, which increases until the very oldest age groups at which point the effect of age starts to decrease (although it

[39] It is a property of the logistic function that it is easier to change probabilities for cases at the margin, where the probability is close to 0.5 than those with lower or higher probabilities.

is still positive). Thus the greatest effect of age is seen for people aged 55 to 84, with people in these age groups having on average a 14 per cent increased probability of being a respondent, compared to someone in the 18 to 34 age group. A strong effect is also observed for people living in a CASSR that did not chase nonrespondents, as these people are nearly ten per cent less likely to be a respondent. Similarly, MH clients are ten per cent less likely to be a respondent than PSD clients. LD clients are more likely to be a respondent than PSD clients, but the effect is smaller at around eight per cent.

Smaller positive effects are observed for being white, and receiving one of the frequent, high contact and high care service forms, i.e. Direct Payments and home care, which increase the probability of being a respondent by between two and five per cent. As expected receiving low-level services has a negative effect on the probability of being a respondent, but this is small at slightly under two per cent. Interestingly, nursing home residents are more than five per cent less likely to be a respondent than other service users, but people in (personal care) residential homes are relatively more likely to be respondents. This suggests either differences in the characteristics of people living in these two types of institutional settings, differences in the success of removing people who lack capacity, or differences in the way that these two settings engage with the survey and encourage or support their residents to participate.

Table 8: Average marginal effects for multilevel MNL model covariates on each of the response outcomes

	Respondent		Blank form		Nonrespondent	
	AME	SE	AME	SE	AME	SE
MH[a]	-0.111***	0.005	0.016***	0.003	0.095***	0.005
LD[a]	0.079***	0.006	-0.031***	0.003	-0.048***	0.006
Age: 35-44[b]	0.059***	0.007	0.000	0.003	-0.059***	0.007
Age: 45-54[b]	0.101***	0.007	0.002	0.003	-0.102***	0.007
Age: 55-64[b]	0.137***	0.007	0.006	0.003	-0.142***	0.007
Age: 65-74[b]	0.140***	0.007	0.022***	0.003	-0.162***	0.007
Age: 75-84[b]	0.135***	0.007	0.033***	0.004	-0.168***	0.007
Age: 85 and over[b]	0.126***	0.007	0.046***	0.004	-0.171***	0.007
White	0.040***	0.005	0.015***	0.003	-0.055***	0.006
Count of service types	0.010**	0.003	-0.007***	0.002	-0.004	0.003
Nursing Home	-0.066***	0.007	0.004	0.004	0.063***	0.008
Residential Home	0.024***	0.006	-0.014***	0.003	-0.010	0.006
Home Care	0.024***	0.005	-0.027***	0.003	0.004	0.005
Low-level services	-0.017**	0.006	0.003	0.003	0.014*	0.006
Direct Payment	0.042***	0.007	-0.022***	0.003	-0.020**	0.007
Equipment	0.007	0.005	0.008**	0.003	-0.016**	0.005
IMD deprivation score	-0.003	0.002	-0.002*	0.001	0.004*	0.002
No chase	-0.099*	0.044	-0.041**	0.014	0.139**	0.053

Legend: AME average marginal effect; [a] Base category: PSD; [b] Base category: 18-34 years; * p<.05; ** p<.01; *** p<.001

The average marginal effect of the number of services received on the probability of being a respondent is fairly small and people receiving zero services only have about a six per cent higher probability of responding than people receiving six services (see Figure 6a). Nevertheless, the relationship between the count of services and the probability of being a respondent is as expected if this variable is acting as an indicator of disability severity; as the number of services received increases beyond two, the probability of responding starts to decrease, at an increasingly rapid rate. By contrast the marginal effect of deprivation is not significant, either on average or at any value of the deprivation score (see Figure 9a), indicating that deprivation does not have a significant effect on the probability of being a respondent.

Many of the covariates have an effect on the probability of returning a blank form, but none have a strong effect – probably due to the shape of the logit function. The variable capturing whether CASSRs chased nonrespondents, has one of the largest effects on the probability of returning a blank form, with sample members living in CASSRs that did not chase nonrespondents having on average a roughly five per cent lower probability of returning a blank form than those who live in another CASSR. The large SE associated with this variable, however, means that the significance of this effect is lower than that of the effect of being an LD client (compared to being a PSD client), which only decreases the probability of returning a blank form by on average three per cent. Being an MH client (compared to PSD) has a small positive effect of less than two per cent.

Two other variables have comparatively large effects on the probability of returning a blank form. Being in the 85 and over age group (compared to being in the 18 to 34 age group), increases an individual's probability of returning a blank form by around 4.5 per cent. The two non-institutional, frequent, high-contact and high-care service forms, Direct Payments and home care, both have a negative effect on the probability of returning a blank form of slightly over two per cent. The marginal effect of the count of the number of services received on the probability of returning a blank form is also fairly small and only significant where people are receiving between zero and two services (see Figure 7b). Over this range the effect is negative, such that people receiving zero services have a slightly higher probability of returning a blank form than people who receive two services (see Figure 6b). All other effects are not significant or also very small at less than two per cent. This includes the effect of living in a deprived area, where the marginal effect is just significant at high levels of deprivation (see Figure 8b) and indicates that those people living in higher

deprivation areas have a lower probability of returning a blank form than people living in lower deprivation areas (see Figure 9b).

For the nonrespondent outcome age has an important effect. Compared to people in the 18 to 34 age group, people in the age groups over 65 are more than 15 per cent less likely to be nonrespondents. The magnitude of the effect decreases for younger age groups. Not chasing nonrespondents also has a powerful effect on the probability of being a nonrespondent, with people living in CASSRs that did not chase nonrespondents being nearly 15 per cent more likely not to respond to the survey request. Deprivation has an effect on the probability of being a nonrespondent, with a significant marginal effect across the range of deprivation scores (see Figure 9c). As the deprivation score increases, the probability of being a nonrespondent increases, such that people living in high deprivation areas are nearly 20 per cent more likely to be nonrespondents than people living in low deprivation areas (see Figure 8c).

Compared to PSD clients, MH clients are nearly ten per cent more likely to be a nonrespondent. The effect of being an LD client compared to being a PSD client on the probability of being a nonrespondent is smaller and is negative. A similar sized effect is observed for ethnicity, with white people being just over five per cent less likely to not respond to the survey request than people of non-white ethnic backgrounds.

The types of services received have very small effects of less than two per cent, on the probability of being a nonrespondent. People receiving low-level services are slightly more likely than those not receiving such services to be nonrespondents and people receiving either equipment or Direct Payments are slightly less likely to be nonrespondents than people not receiving such services. Being a nursing home resident had a large effect on the probability of being a nonrespondent, being around six per cent more likely to be nonrespondents compared to non-residents. The effects of other services, including receipt of residential care homes, home care and the count of services received, on being a nonrespondent are not significant. For the number of services received, although the marginal effects appear to be significant over the range of receipt of three to six services (see Figure 7c), the margin of error around the predicted probabilities is so great that there is only a slight upward trend in the likelihood of being a nonrespondent as the number of services received increases from three to more services (see Figure 6c).

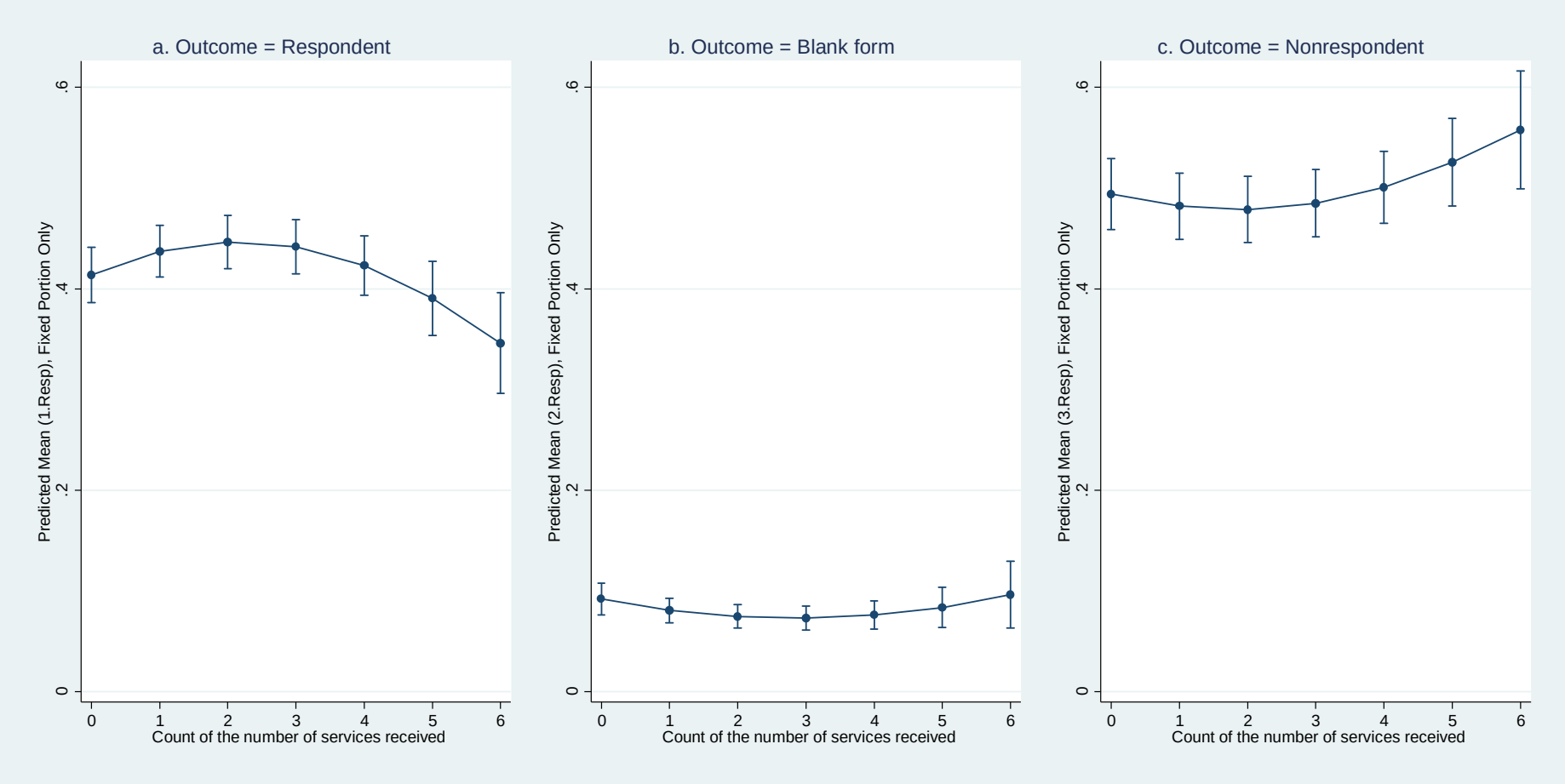

Figure 6: Predicted probabilities of response outcomes at representative values for the number of services received, with 95% confidence intervals

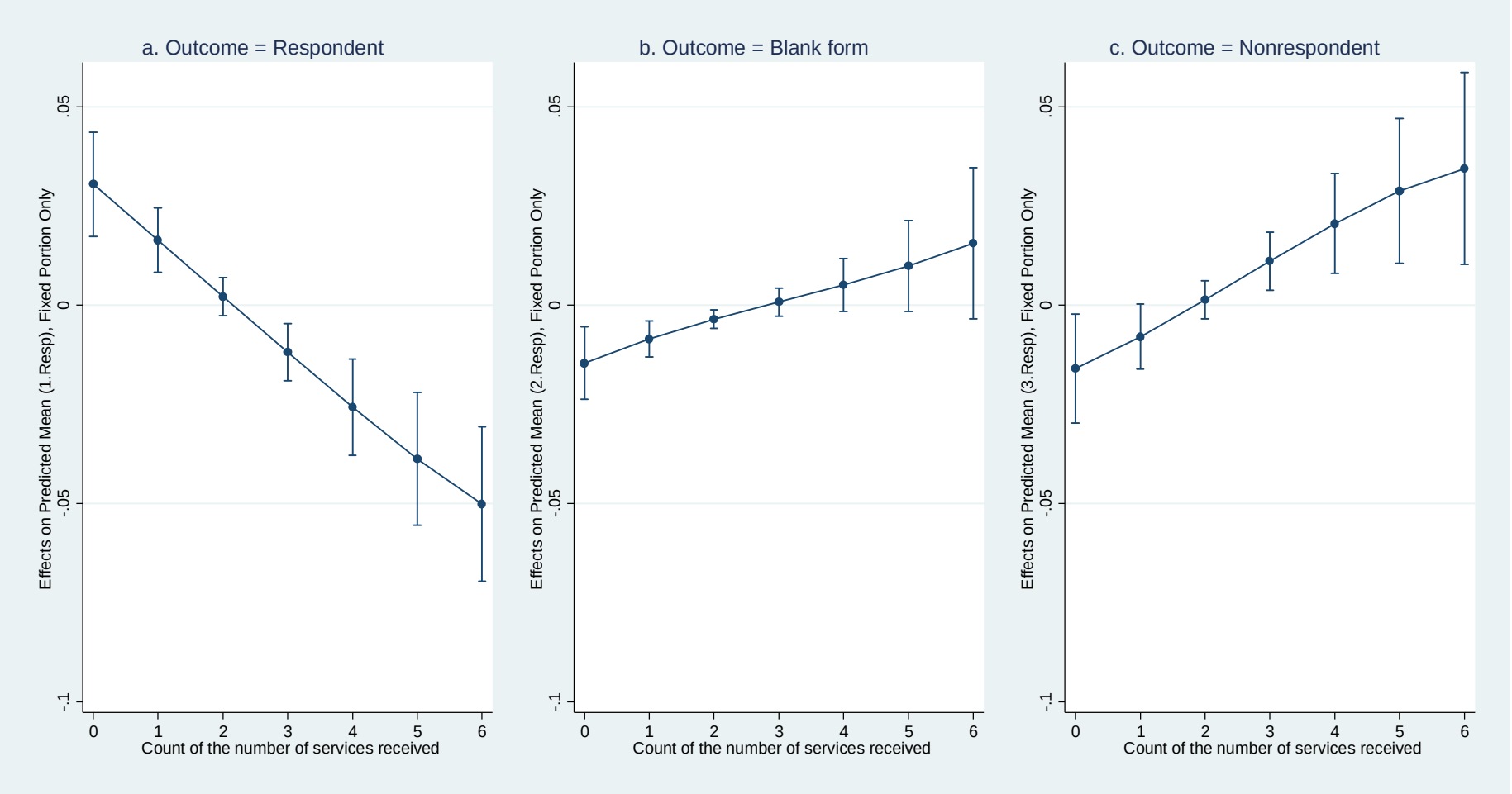

Figure 7: Marginal effects of number of services received at representative values, with 95% confidence intervals

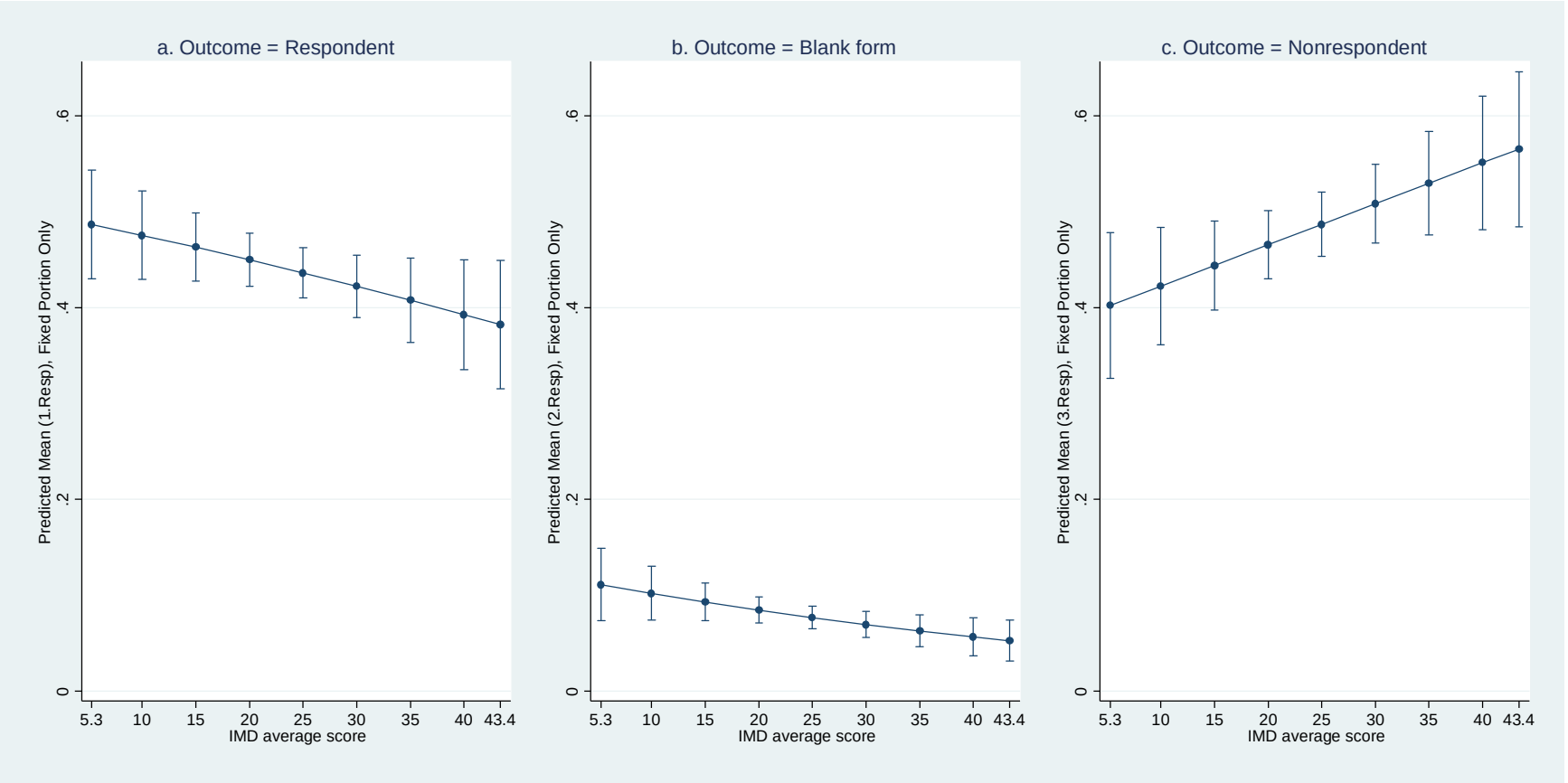

Figure 8: Predicted probabilities of response outcomes at representative values for the average IMD score, with 95% confidence intervals

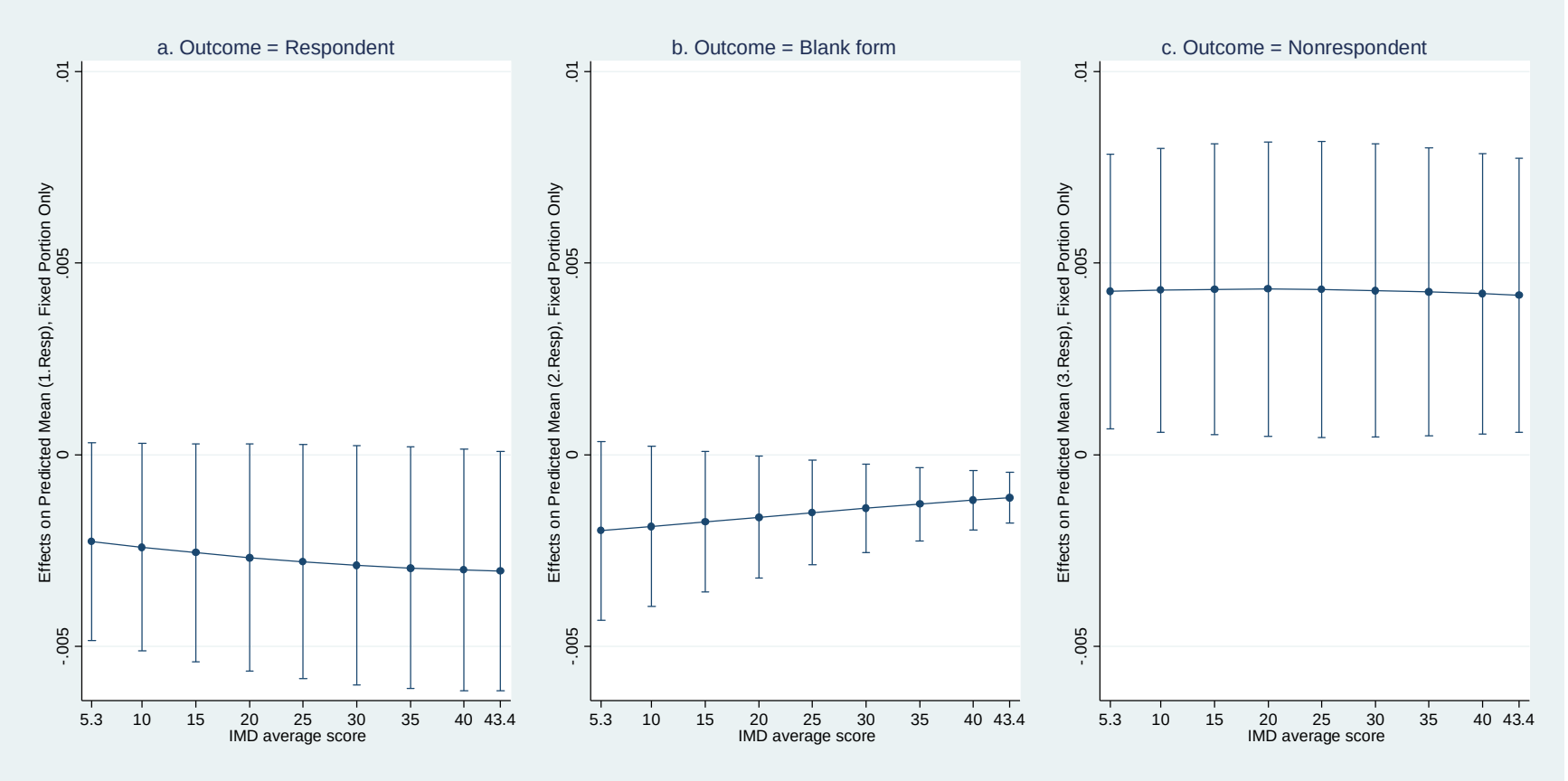

Figure 9: Marginal effects of the average IMD score at representative values, with 95% confidence intervals

Discussion

The purpose of this analysis was to understand who is missing from performance surveys and to explore whether there are any strategies that could be used to improve response rates. These questions are important since nonresponse is likely to affect the perceived representativeness of the survey and potentially the validity of inferences about performance. I have used the ASCS to explore these questions, exploiting the variation in the way the survey is managed by different CASSRs to provide insight into the methods that might increase response rates.

The modelling is set within the modified social exchange framework proposed by Goyder (2006). This framework allowed me to explore the extent to which different theories and types of factors may explain nonresponse to the ASCS, such as individual characteristics, social disorganisation, survey attributes and CASSR characteristics. I found that all of these types of factors were important in explaining participation to the ASCS. I used a multilevel modelling approach to test the hypothesis that variations between CASSRs in response rates can be explained by both differences in individuals' characteristics and CASSR-level factors, including those related to the organisation and the area. This method has also allowed me to account for unobserved CASSR-level variation in participation rates, which was important as a significant amount of variation was explained by unobserved CASSR-level factors.

I discuss these results in more detail below and outline some of the limitations of the analysis. I argue that the combination of the modelling approach and the social exchange framework were important for understanding participation in the ASCS. The implications of the findings for other studies and policy are discussed in Chapter 9.

Differentiating response outcomes and the role of the CASSR in determining participation

The analysis supported the proposition that different processes are involved in deciding whether to return a blank form or not return the questionnaire. In general coefficients for the explanatory variables in the model differed across the simultaneous equations, and some of the variables that were associated with an increased likelihood of someone returning a blank form, were also associated with a decreased likelihood of not returning a questionnaire (e.g. white ethnic background, and chasing nonrespondents). This study shows that nonrespondents to postal surveys can be differentiated and that where it is possible to differentiate nonrespondents this improves models of survey participation.

It also proved to be important to differentiate between the two nonresponse outcomes in modelling the CASSR random effects. Since CASSRs run the survey within their area and have some discretion over its management, I expected them to have a significant impact on response propensity. However, where the CASSR clustering variable was shared between the simultaneous equations – the shared random-effects model – a very small percentage of the variation in response rates was explained by between-CASSR differences. Where a correlated random-effects model was used, in which separate random effects for each of the nonresponse outcomes are estimated, between CASSR differences accounted for a much larger percentage of the variation in the nonresponse outcomes. The correlated random-effects model outperformed the shared random-effects model on model fit indices and produced results that were consistent with my expectations. There are according to this analysis, important differences between CASSRs in rates of nonresponse and return of blank forms.

Factors explaining survey participation

I found that multiple theories explain response to the ASCS. The most important factors with the greatest effects on participation were age, client group and chasing nonrespondents. The strong effect of age on participation is consistent with previous research (Herzog and Rodgers 1988, Kaldenberg et al. 1994, Elliott et al. 2005, Peters et al. 2014b). I had not expected such a strong effect for the ASC population as factors such as employment and disability that are often posited as mediating variables for the observed age effects should be less important in this sample, due to low levels of employment and high levels of disability. Since it was the youngest age groups (people between the ages of 18 to 34) who were the least likely to respond and the older age groups that were most likely to respond (people between the ages of 55 to 84) an alternative explanation may be related to the lower rates of political participation found in younger age groups (Jowell and Park 1998, Henn et al. 2002).

Although the type of services received had a less important effect on participation, the effect of nursing and care home residents is important to consider given these groups are excluded from the sample due to the instruction to remove people who lack capacity. I found that patterns of participation for residents of nursing homes and personal care only homes were in opposing directions. Thus residents of nursing homes were less likely to respond, and residents of personal-care only homes were more likely to respond. Although the costs of participation are likely to be higher for care home residents than people in

private household because they are more disabled, other factors are likely to intervene in care homes. One explanation for this finding is that the costs of participation are higher for nursing home residents as they are more severely disabled and have a higher prevalence of dementia than residents of other care homes (Darton et al. 2006, Matthews et al. 2016). Alternatively, it may be the case that staff in nursing homes are more aggressive gatekeepers, perhaps perceiving the survey of limited usefulness due to the disability level of residents. If the attitude of staff is important this may also have affected the ability of CASSRs to successfully remove people who lacked the capacity to respond to the survey from the nursing home population. This would inflate nonresponse rates from nursing homes compared to other groups. Follow-up research with care homes would help to assess the plausibility of these explanations.

There was some indication from the results that disability severity reduces response rates. While people aged 85 and over were significantly more likely to respond than people aged between 18 and 34, the effect was not as large as that observed for the 55 to 84 age groups. This is consistent with a number of other studies which have found that the oldest old are less likely to respond to postal surveys (Sheldon et al. 2007). The model indicates that the reason being aged 85 and over has a smaller effect on the probability of being a respondent was due to the increased likelihood that people in this age group will return a blank form. A possible explanation for this finding is that these people are returning a blank form because they are unable to respond and want to indicate their poor health, so as to prevent further contacts. This would fit with findings from other studies that have shown older nonrespondents to be less healthy than older respondents (Paganini-Hill et al. 1993, Strayer et al. 1993). I also found a fairly weak but negative association between the number of services received and the likelihood of being a respondent. These findings, combined with the lower rates of participation among nursing home residents, suggest that disability severity may be important in determining survey participation and that the most severely disabled may be underrepresented in the current ASCS. This finding deserves further exploration.

The fairly small effects if deprivation are fairly consistent with social disorganisation theory and other literature, which have shown that after controlling for household variables area-level effects are less important (Couper and Groves 1996, Groves 2006, Johnson et al. 2006). In contrast to previous research, however, I did not find significant effects from urbanicity and population density (Couper and Groves 1996,

Groves 2006). It is argued that variables such as deprivation have their effects by reducing helping behaviour and trust in institutions and others (see e.g. Couper and Groves 1996).

Turning to the factors that may increase response rates, not chasing nonrespondents had a relatively large positive effect on the probability of being a nonrespondent and a corresponding fairly large negative effect on the probability of being a respondent or returning a blank form. This finding is consistent with findings from experimental studies (Edwards et al. 2002) and with research by Hébert et al. (1996), which found reminders to be highly effective among older people over 75. This provides a clear direction for a strategy to increase response rates. No other survey attributes were significant in the models; nor was the indicator of CASSR data quality.

The strong effect of client group on participation, especially the finding that LD clients had the greatest likelihood of responding and MH clients the least likelihood of responding may point to another way to improve response rates. There are two possible explanations for the high participation rates among LD clients. First, all LD clients were sent an Easy Read version of the questionnaire, which would have substantially lowered the costs of responding for this group. Second, LD clients report much higher rates of help to complete the questionnaire, which may mediate the observed relationship (The Information Centre for Health and Social Care 2012b). For MH clients there are also two explanations for the low response rates. MH clients tend to receive care through multi-disciplinary teams based within the health services. Consequently they may fail to see the relevance of a social services survey. Alternatively, low response rates may be due to poor data quality for this client group. Many CASSRs do not have direct access to case records, and rely on health services (who have no vested interest in the survey) to provide contact information and other details. Indeed several CASSRs reported that they had excluded all MH clients as they could not get the required information from health colleagues. These explanations deserve more detailed research, because of the implications for improving response rates.

Another finding that may point to ways to increase response rates is the effect of ethnicity. I found a small effect positive effect on the likelihood of participation of being white, however, other US and UK studies have found much larger differences in rates between ethnic groups (Elliott et al. 2005, Sheldon et al. 2007, Klein et al. 2011, Peters et al. 2014b). It has been posited that disengagement with government and 'official' institutions is a reason for low response rates by black and minority ethics groups in surveys (Groves and Couper 1998), but in their review of the evidence, Sheldon et al.

(2007) find little evidence to support such a theory of disengagement. They argue that, at least for the English national patient surveys, theories of acculturation, literacy and language – all of which raise the costs of participation – may be better explanations of low response rates. Certainly there is ample evidence that language and literacy skills affect social care users ability to access services and achieve good care outcomes (Gill et al. 2014, Willis et al. 2015, Blake et al. 2016, Yeung et al. 2016). Since ethnicity is a fairly noisy measure of literacy skills, it seems plausible that poor literacy skills explain the fairly small effect of ethnicity observed here for both return of blank forms and completed forms, since both responses require understanding of the questionnaire and its instructions. Targeted use of translated versions may therefore help to improve response rates among the non-white population.

Limitations of the analysis and directions for future research

Overall the wealth of available auxiliary and paradata has meant that I have been able to conduct a much more detailed analysis of the factors influencing nonresponse than is frequently possible with surveys. However, the predictive ability of the model was low. One possible reading of this finding is that nonresponse is largely a quasi-random process (Smith 2011). Another would be that key variables are unobserved. In this analysis I lacked measures for theoretically important variables such as availability of someone to help the person complete the survey (Renfroe et al. 2002), and had fairly noisy proxy indicators for proficiency in English (ethnicity), disability type (client group) and severity (count of services). Since it is not possible to say whether, or to what extent, better (or any) measures of these variables would improve the models of response propensity, the question of whether nonresponse to the ASCS is largely a quasi-random process remains open.

The modelling also provided little insight into the reason for differences between CASSRs in nonresponse outcomes. Neither individual-level nor CASSR-level variables contributed greatly to explaining CASSR-level variations in response rates, although CASSR-level variables were slightly more important. There is some suggestion from this modelling that it may be possible with more data points to decompose the CASSR-level variation between organisational and area characteristics, as the analysis has shown that both of these groups of factors contribute to between-CASSR variation. Much between-CASSR variation is unexplained by this study. This is an important area for future

research given the possibility that much of the between-CASSR variation in response rates is explained by the behaviour of CASSRs.

There is also a need to be cautious about treating the lack of significance of most of the social environment, data quality and survey attributes variables as evidence that they are not important in predicting response outcomes. First, there are not many observations at the CASSR level and the binary variables that describe survey attributes have in most cases very skewed distributions. This makes it difficult for effects to be significant. Second, the exploratory analysis found that the social environment variables and many of the survey attributes have significant relationships with response, blank form and nonresponse rates that are in the expected direction. For example, the indicator of data quality is positively correlated with response rates; CASSRs that used incentives have lower nonresponse rates; and some of the alternative methods for chasing nonrespondents seemed to be associated with higher response rates. A study with more data points would have had greater power for detecting differences. This would be a useful direction for future research.

A third reason for caution is that due to the variety of ways in which CASSRs implemented the various strategies to increase response rates, many of the survey attribute variables are not particularly sensitive measures, which will dampen any association observed. For example, although all CASSRs that used incentives had a prize draw, the prize varied between CASSRs. Research has shown that the amount of money and the type of incentive are instrumental in determining the effectiveness of the incentive (Zaslavsky et al. 2001, Edwards et al. 2002, Dillman et al. 2009) and that incentives are differentially effective for subgroups of the population (Lynn 2012). Similarly, evidence around the effect of questionnaire length on response rates suggests a nonlinear relationship (Iglesias et al. 2001, Edwards et al. 2002), so the dummy variable capturing whether additional questions were added is probably not sensitive enough for assessing the effect of questionnaire length. A similar argument could be made for the variables that capture engagement strategies for hard-to-reach groups.

Following on from this point, the observational design limits the potential to draw conclusions about the factors that may improve response rates because the effects from the modelling can only be interpreted as associations. Studies that explore the effect of altering survey attributes usually use experimental designs to manipulate conditions and ensure random selection into ‘control’ and ‘treatment’ groups (Edwards et al. 2002). In observational studies selection into groups is not random and, as was the case here, often

the treatments vary in their application. This makes it much more difficult to estimate an effect since treatments are confounded with other, often unobservable, factors. In my view, the sample size was the more important limitation for this analysis. Nevertheless, it is certainly the case that to fully understand the effect of survey attributes, such as incentives and strategies to engage hard-to-reach groups, experimental studies would yield much more conclusive results. In the absence of resources and the political will to implement such a design, observational studies such as this can contribute to furthering our understanding of the factors that influence survey participation.

A further limitation of the modelling was that I was not able to appropriately address missingness in the auxiliary data. Investigation suggested that it may have been possible to estimate multilevel MNL models on multiply-imputed datasets in different software, for example MLWIN, but this would have come at the expense of model interpretation as the Bayesian estimation methods used by the software make it much more complicated to estimate marginal effects and adjusted predicted probabilities. It is of note that Durrant and Steele (2009) who apply the same model to their data in MLWIN report only the adjusted predictions at means and do not attempt to impute missing values. Interestingly they report using the Bayesian approach to overcome model convergence problems experienced by researchers in the past (Durrant and Steele 2009). Given the problems I experienced with convergence of the models using frequentist methods, it may be that Bayesian methods are more appropriate for these types of models.

A further aspect that requires some reflection is the recent changes to the survey methods and the impact this may have on the study conclusions. Since 2014-15, there have been some changes to the eligible population of the ASCS and the survey auxiliary data that have been driven by changes in the way data is collected and reported about social care users by CASSRs. The most significant change is that in the new data collection, the types of services people receive (i.e. home care, day care and so on) are no longer collected. Instead, social care users are defined as either receiving short-term or long-term support services, with additional information captured about accommodation status, support setting and delivery mechanism. Since 2014-15, only service users receiving ‘long-term support’ are eligible for the survey, which means people previously included because they received short-term residential care, home care re-ablement or other short-term support to maximise independence, equipment and adaptations are not included in the more recent ASCSs.

Since many of these groups of people had lower response rates, it is reasonable to expect response rates to the ASCS to increase as a result of these changes. In fact response rates have continued to decline from 41 per cent in 2010-11 to just under 36 per cent in the most recent survey (NHS Digital 2016b). There are many potential explanations for this decline in response rates that are consistent with the findings presented here and the interpretation that the survey was less salient to these groups, including the general phenomenon of declining participation in surveys (de Leeuw and de Heer 2002), survey fatigue, which may be more acute in authorities with smaller user populations, and the continued rationing of services due to financial pressure, which has led to increased targeting of services at the most severely disabled whom this analysis suggests are less likely to participate (Fernandez et al. 2013b, Burchardt et al. 2015). It would be of interest to unpick these different explanations. The more detailed information that is collected about primary support reason under the new data collection framework may be useful in this respect. These data would also enable a more detailed exploration of the role of disability and health in determining nonresponse than has been possible here. Such developments would support better targeting of alternative formats, and a more 'tailored' approach to the survey that would see improvements in response rates (Dillman et al. 2009). These are all possible avenues for future research.

Conclusion

This analysis has shown that survey participation is a complex process, with multiple factors intervening in different ways. This lends support to Goyder's (2006) proposal to integrate the theories of nonresponse into a single framework. One purpose of this analysis was to identify who was missing from the ASCS. In this analysis I have found low rates of response among MH clients, working age groups (particularly 18 to 34 year olds) and nursing home residents, which I suggest is due to the saliency of the survey, data quality problems for the MH group and higher barriers to participation for the other groups. To a lesser extent, low-level service users, people in more deprived neighbourhoods and black and minority ethnic groups are also underrepresented at present, possibly as a consequence of the salience of the survey, higher barriers to participation and social disorganiastion. There is also an indication that people with more severe disabilities are underrepresented and there are questions over the representativeness of the survey for care home residents more generally due to exclusion criteria. Should any (or all) of factors be associated with ASCOF indicator scores, there is the potential for the PIs to be biased. This would be a

strong reason to find strategies to improve response rates and the survey's representativeness. I turn to these questions in the next chapter.

Even if there is little evidence for bias, it may be considered politically important to address response rates among the groups with lower response rates for reasons of face validity and to ensure the precision of PIs. A second aim of this study was to identify strategies that could be used to increase response rates. In this respect the study was less successful. There is fairly good evidence that chasing nonrespondents will help increase response rates, but this strategy is not selective. It may increase response rates from all groups, which would not help to address the problems of underrepresentation. Based on the finding of higher responses rates from LD clients, I have suggested that targeting groups with adapted versions of the questionnaire may be a good way of increasing response rates. This strategy could be applied to various groups, such as people whose first language is not English and people with visual impairments. However, to follow this through CASSRs would need to improve their data systems. Other strategies for targeting specific groups should be investigated. Importantly, much of the between-CASSR variation in nonresponse rates was unexplained in this analysis. Further research to understand the causes of between-CASSR variations in response rates will be important for improving response rates to future surveys and for ensuring that differences in PI scores reflect real differences in performance rather than unobserved differences in respondent populations due to variations in survey design and implementation. I reflect further on these implications for research and policy in Chapter 9.

Chapter 5

What is the Effect of Nonresponse on Inferences about Performance and Does the Method Used to Address Differences in Samples Due to Nonresponse Matter? An Analysis into the Extent and Effects of Unit and Item Nonresponse on the ASCOF Indicators

Abstract

Objective: To explore whether nonresponse to performance surveys affects the interpretation of indicator scores.

Method: Data from the English 2010-11 ASCS are used to explore the effect of nonresponse on five performance indicators based on these data. To assess the effect of nonresponse I compare indicator estimates generated under the missing completely at random (MCAR) and missing-at-random (MAR) assumptions. I use inverse propensity weighting (IPW) and multiple imputation (MI) to correct for unit and item nonresponse, respectively, under MAR conditions. Differences in the estimates obtained under the two missingness assumptions indicate that nonresponse has an effect on the interpretation of PI scores. The impact of differences is assessed through examining changes in rank position and outlier status.

Results: After adjusting for nonresponse indicator scores are on average more negative, suggesting that nonrespondents are more likely to be dissatisfied with their care or have poorer quality of life. Absolute differences in adjusted and unadjusted indicator estimates for CASSRs are small, with the majority being less than one standard error. In general the impact of nonresponse had little effect on the identification of outliers or rank position of CASSRs, with high correlations between adjusted and unadjusted indicators. There are some exceptions, particularly for the quality of life (SCRQoL) indicator where the effect of IPW/MI adjustment was precision-enhancing. This indicator had the highest rate of item nonresponse and after IPW/MI adjustment a small number of CASSRs experienced large changes in rank position and changed outlier status.

Conclusions: There is little evidence to recommend adjustment for nonresponse for the majority of the indicators considered here, since it reduces precision and observed bias is small. IPW/MI adjustment for nonresponse could be considered for the SCRQoL indicator, since adjustment is precision-enhancing and observed bias has more impact on inferences about the relative performance of CASSRs. The effect of adjustment depends on the method used to generate estimates under the MAR assumption, with MI methods being more

important where item nonresponse is higher. We cannot conclude that there is no appreciable nonresponse bias, since the methods used here only adjust for bias due to observed factors.

Introduction

Missingness, due to nonresponse, in the data used to generate PIs raises questions about the validity of inferences regarding the performance of organisations drawn from the affected PIs. The loss of data affects the precision of survey estimates, which will affect the certainty there is over the estimates and make it more difficult to identify differences in performance. Nonresponse also affects the representativeness of the sample, which can result in biased estimates. Bias will only occur where missingness is systematic, or not missing completely at random (MCAR). MCAR is a strong assumption and is unlikely to be realistic for most datasets. Where the MCAR assumption is not plausible there will most likely be some degree of bias in PIs, which will increase the probability of falsely identifying poor or very good performance (false positive) and incorrectly identifying exceptional (poor or good) performance as ordinary (false negative). There are, however, a number of strategies for addressing nonresponse which assume that the missingness mechanism depends on the values of the observed data, i.e. it is missing-at-random (MAR). It is possible to use these strategies to generate 'adjusted PIs'. As a consequence of the statistical assumptions underlying the nonresponse adjustment, the PIs will be less transparent and more complicated to understand. This may mean people are less likely to use them despite their greater validity. A key question then for policymakers is: does nonresponse to performance surveys affect the interpretation of PI scores?

In this chapter I assess the effect of nonresponse on inferences about performance drawn using the ASCOF survey-based PIs. I use an approach that has been applied in previous studies, which involves comparing PI estimates generated under the MCAR and MAR assumptions (see e.g. Elliott et al. 2005, Höfler et al. 2005, Groves 2006). If there are differences in the estimates obtained under the two missingness assumptions, this indicates that nonresponse has an effect on the interpretation of PI scores. To generate estimates under MAR assumptions, I use a combined inverse propensity weighting (IPW)/multiple imputation (MI) approach, with analogy to the method proposed by Seaman et al. (2012). As I explain in Chapter 3, this method has not been used to adjust performance survey data for nonresponse before, but given that such data usually have high levels of *unit* nonresponse and limited auxiliary information about nonrespondents, this method is likely to be more efficient than an IPW approach (IPW/IPW) that addresses both unit and item nonresponse. It is also likely to produce less biased estimates than a MI approach (MI/MI) that addresses both unit and item nonresponse (Seaman et al. 2012). The aim is that this analysis should contribute to the

debate in the wider literature around whether there is a need to adjust survey-based PIs for the precision-reducing and potentially biasing effects of nonresponse and how best to do so.

The chapter is organised as follows. I start by discussing the ASCS data, focusing on the types and extent of nonresponse in the sample. Next, I discuss the empirical strategy, concentrating on how I implement the IPW/MI approach to adjust for item and unit nonresponse within the ASCS data. I then present the results in two parts. First, I assess the effect of nonresponse on the ASCOF indicators, by quantifying the loss of precision and the possible extent of bias. Secondly, I look at the effect that adjusting for nonresponse has on performance assessment. In the discussion I focus on the key findings and limitations of the study. The implications for policy and research are discussed in Chapter 9.

Data

For this analysis I use the 2010-11 ASCS to explore the effect of nonresponse on the five ASCOF 'outcome' indicators that are drawn from the ASCS. The distributional statistics for these indicators for the whole sample are shown in Table 9.

Table 9: Distributional and missingness statistics for the ASCOF PIs, 2010-11 ASCS

	SCRQoL	Satisfaction	Control	Safety	Information
Obs.	54,350	57,929	59,478	59,499	42,884
Mean	18.62	0.62	0.75	0.93	0.74
SD	3.93	0.49	0.43	0.26	0.44
Max	24	1	1	1	1
Min	0	0	0	0	0
% Missing	10.9%	5.1%	2.5%	2.5%	5.6%
% N/A	n/a	n/a	n/a	n/a	24.1%

Nonresponse to the ASCS

The ASCS has two types of nonresponse: *item nonresponse* to the individual questions or variables and *unit nonresponse* where the sample member has not responded to the survey request. As discussed in Chapter 4, the unit nonresponse is high but fairly typical for a postal survey, with a response rate of 41 per cent. Item nonresponse rates vary by question in the ASCS, and are generally fairly low at less than five per cent throughout the questionnaire (see Table 9 for the ASCOF indicators and Table 10 for the questionnaire items). They are higher where respondents were allowed to select a 'not applicable' option, as is the case for the

information PI; or where there is dependency between responses to questions, as is the case for the questions about assistance to complete the questionnaire; or where indicators are composed of multiple items, as is the case for SCRQoL indicator and the ADL scales.

Table 10: Extent of missingness to the questionnaire items, 2010-11 ASCS

Variable	Obs.	Response options	Percentage missing
Self-perceived health	59,243	1 to 5	2.92
EQ-5D: Pain	58,509	1 to 3	4.12
EQ-5D: Anxiety/depression	57,843	1 to 3	5.22
ADL: get around indoors	58,480	1 to 3	4.17
ADL: get in/out bed/chair	58,645	1 to 3	3.90
ADL: feed self	58,707	1 to 3	3.80
ADL: wash all over in bath/shower	58,773	1 to 3	3.69
ADL: get dressed/undressed	58,679	1 to 3	3.85
ADL: use WC/toilet	58,710	1 to 3	3.80
ADL: wash face and hands	58,865	1 to 3	3.54
IADL: deal with finances/paperwork	58,344	1 to 3	4.39
Count of ADLs with difficulty	56,092	0 to 7	8.09
Count of ADLs can't manage without help	56,092	0 to 7	8.09
Self-perceived home design	58,481	1 to 4	4.17
Regular practical help	57,291	3 categories	6.12
Make additional payments	56,135	3 categories	8.01
Assistance to answer the questionnaire	57,004	3 categories	6.59
Type of assistance	57,004	4 categories	6.59
Source of assistance	56,690	3 categories	7.11

The auxiliary data is also subject to item nonresponse. This is due largely to differences in procedures and systems for recording client data within CASSRs, which meant that some CASSRs were not able to provide all of the requested variables and the quality of some of the data is variable. As Table 11 shows, for the ASCS 2010-11 dataset, the majority of the variables are at least 90 per cent observed. Problems with recording are confined to a small number of variables, specifically the religion, sexual identity, secondary client group and budget data. Variables with high proportions of missing data or that were missing for a large proportion of the CASSRs were not considered in this analysis.

Table 11: Extent of missingness to the auxiliary data across the CASSRs and within the overall sample, 2010-11 ASCS

Variable	Number of CASSRs			Sample completion rate
	Fully observed	Partially observed	Completely missing	
Sex	135	14	0	100.0%
Age group	127	22	0	99.9%
Ethnicity	17	132	0	98.3%
Sexual identity	0	21	128	1.1%
Religion	4	128	17	52.2%
Primary client group	118	31	0	99.6%
Secondary client group	49	44	56	42.8%
Residential care	144	5	0	100.0%
Nursing home	142	7	0	99.1%
Home care	140	9	0	99.0%
Day care	140	9	0	98.3%
Meals	137	6	6	94.6%
Short-term residential care	138	10	1	97.5%
Direct Payments	141	8	0	97.7%
Personal Budgets	133	9	7	92.2%
Professional support	135	7	7	93.8%
Equipment	139	7	3	96.3%
Other services	137	9	3	95.6%
Total budget	13	70	66	41.0%
Other funding streams in budget	48	12	89	39.4%

Empirical strategy

I study the effect of nonresponse by comparing the distribution of scores on the ASCOF PIs under MCAR and MAR assumptions. Indicator estimates under MCAR are produced using complete case analysis (CCA), the MAR approach is the IPW/MI method. I also have two mixed MCAR-MAR methods, referred to as IPW/CCA and CCA/MI, of which the former assumes that item nonresponse is MCAR and unit nonresponse is MAR, and the latter assumes that item nonresponse is MAR and unit nonresponse is MCAR. One difficulty with using IPW with the ASCS data is the extent of missingness in the auxiliary data, as outlined above. Using CCA for the response propensity model would result in weights only for those

cases with full information, i.e. for 125,753 cases out of 150,672 (83 per cent). Nineteen authorities (87 per cent)[40] would also be excluded. As well as being an impractical method for correcting for nonresponse due to the loss of CASSRs, this practice is inefficient and could itself introduce bias should the cases excluded from the propensity models (because they have partially observed data) be different from those with fully observed data. If I am to use the IPW/MI method I also need to address missingness within the auxiliary data.

I use MI to recover the missing auxiliary data, but I carry out imputation of the auxiliary data separately from imputation of the respondent sample, because it is undesirable to impute outcomes for the nonrespondent sample. This is not ideal, as the variability in the weights from imputation cannot be reflected in the analysis. Nevertheless, it is better than assuming unit nonresponse is MCAR and the result is a set of weights that can be applied to the whole sample. I do, however, return to this limitation in the discussion. I now describe the three procedures in more detail.

Procedure for multiple imputation of the respondent sample

Analysis suggested that the MCAR assumption for the respondent data was not plausible (Little's (1988) MCAR test: $\chi^2(102)=1262$, $p<.001$)[41]. I use a chained equations MI approach to impute the missing values for the respondent data (van Buuren 2007, White et al. 2011). This method (compared to the multivariate normal imputation model) allows greater flexibility in the specification of imputation equations for the missing variables – important for these data given the mix of nominal, ordinal and continuous variables. Rather than impute the SCRQoL score, I impute the eight ordinal items in the SCRQoL measure individually. This is partly because two of the SCRQoL items are themselves ASCOF indicators, but it is also easier to reproduce the distribution of the ordinal items than the skewed SCRQoL distribution. A limitation of the chained equations method is that it cannot easily accommodate the clustered structure of the data. To overcome this, I include a set of dummy variables for the CASSR in the imputation models (except those using mlogit due to difficulties with estimation) to capture CASSR effects.

To ensure the imputation is valid, all variables in the analysis model are included in the imputation model. Although the analysis model here is the mean PI score, as I use these

[40] One CASSR has only three cases with full information and this CASSR is excluded in the complete case analysis are there are too few data points to estimate the performance indicator. Excluding this CASSR brings the total number of cases with complete information to 125,750 cases, and the total number of CASSRs excluded to 20.

[41] Little's MCAR test was estimated using the user-written mcartest routine in Stata (Li 2013).

data for subsequent chapters, the imputation model is designed to be congenial with the analysis models in Chapter 6. To strengthen the plausibility of the MAR assumption, I also included all variables that were available in the dataset and could be shown empirically to be associated with both missingness and the outcomes of interest, e.g. ethnicity. Due to the size of the dataset and number of variables with missing data, to make the imputation computationally feasible for some imputation models the variables were more limited (see Appendix 1 for more details).

In total 20 imputations were carried out, after a burn-in period of 20 iterations. I ran a series of tests to assess the necessary length of the burn-in period for the chain to converge to a stationary distribution and to explore the sufficiency of the number of imputations (White et al. 2011) – for details see Appendix 2. After imputation, I combined the individual SCRQoL components into an overall SCRQoL score. Estimates of means and variance for each of the ASCOF indicators were obtained for each imputed dataset, with and without weights (see below for description of how these are generated). The estimates obtained from each analysis were combined using Rubin's rules (1987), which take account of the variation within and between imputations.

Procedure for generating weights to adjust for unit nonresponse

Weights are applied such that missing or incomplete units in the sample are ignored and the complete data are inflated to reflect the probability of response for that unit. I generate the weights in several steps. First, I develop a statistical model to predict response propensity. This model is a simplified version of the model in Chapter 4. I then use the predicted probabilities from this model to develop weights to rebalance the respondent sample for nonresponse. Given the extent of missingness in the auxiliary data, these steps are carried out on an MI sample, to produce a weight for all cases in the dataset (i.e. 61,026 weights). The steps are described below.

Step 1: statistical model for predicting response propensity

A correlation of CASSR-level response rates against CASSR-level PI means, suggested that cluster-specific nonignorable (CSNI) nonresponse may be a problem for some PIs (see Table 59, Appendix 3). CSNI refers to a situation where cluster-level nonresponse depends on the cluster means of the survey variables of interest. Although cluster membership is fully observed for nonrespondents, missingness is not MAR because the cluster means are unobserved random effects (Yuan and Little 2007). Therefore to ensure weights were not

biased, I follow Skinner and D'Arrigo's (2011) advice and use a MNL model to predict response propensity to the ASCS with CASSRs entered into the model as fixed effects[42].

The ASCS response variable, which I denote y_{ij}, for individual i in CASSR j, has three outcomes: respondent (0), blank form (1), no response (2). The probability of y_{ij} having one of the three response outcomes is given by $\pi_{ij}^{(s)} = \Pr(y_{ij} = s)$, for s =0, 1, 2. If I take respondent (0) as the base category, the MNL model can be written

$$\log\left(\frac{\pi_{ij}^{(s)}}{\pi_{ij}^{(0)}}\right) = \boldsymbol{\beta}_1^{(s)}\mathbf{x}_{ij}^{(s)} + +\boldsymbol{\beta}_4^{(s)}\mathbf{d}_j^{(s)}, \qquad s = 1,2 \qquad (13)$$

where $\mathbf{x}_{ij}^{(s)}$ is a vector of individual characteristics, $\mathbf{d}_j^{(s)}$ is a set of dummy variables representing CASSRs, and the $\boldsymbol{\beta}^{(s)}$s are vectors of coefficients. The choice of covariates is informed by the empirical model in Chapter 4. Since CASSRs are included within the model as a set of dummy variables, I do not include any CASSR-level variables. Further checks were carried out to see if any variables omitted from the final model in Chapter 3 (see Table 3) were important in predicting any of the PIs, but no further variables were included in the model[43]. The final model is shown in Appendix 4, estimated by CCA and MI.

Step 2: generating inverse propensity weights

The model in (13) above consists of two simultaneous equations. The first ($s = 1$) models the logarithm of the ratio of the probability of returning a blank form to that of being a respondent as a function of covariate individual and CASSR effects; the second ($s = 2$) models the logarithm of the ratio of the probability of providing no response to that of being a

[42] When constructing inverse probability weights, Skinner and D'Arrigo (2011), caution against using a random-effects model to account for the clustering of cases, as it produces biased weighted estimates. If the MAR assumption is plausible they suggest using the marginal model, as this does not produce biased estimates. Importantly they note that the MAR assumption may not hold where cases are clustered, as non-response may be CSNI. In this situation they also find the marginal model to produce biased weighted estimates and instead propose a conditional logistic regression or fixed effects regression, which they demonstrate produces unbiased weighted estimates. A disadvantage of the conditional logistic regression approach is that it is limited to binary outcome variables (the ASCS has two nonrespondent groups), so it is not useful for this analysis. The fixed effects approach is more promising and as Skinner and D'Arrigo (2011, p. 26) note it "performed similarly to the conditional maximum likelihood estimator and it may be that in practice it will often provide a reasonable proxy to this estimator, while not requiring such strong model assumptions nor so much computation".

[43] In selecting covariates there is a trade-off between bias reduction and increasing the variance of the weight (Höfler et al. 2005). The best covariates are i) those with the highest predictive value, so predicted propensity scores have a low predictive standard deviation for each individual and the model fits the data (Höfler et al. 2005), and ii) those that they are predictive of survey outcomes, in this case the PIs (Little and Vartivarian 2003).

respondent as a function of covariate effects. Predicted probabilities for computing response propensity weights, can be calculated for each outcome as follows. As the log-odds of an outcome compared to itself is zero the effects of any independent variables are also zero for the base category. The simultaneous equations can be solved, by rearranging equation (13) such that the probability of observing outcome s given $\mathbf{x}$ with the reference category given by r, is

$$\pi_i^{(s)} = \frac{\exp(\beta^{(s)}\mathbf{x}_i^{(s)})}{1 + \sum_{r=1}^{2}\exp(\beta^{(r)}\mathbf{x}_i^{(r)})}, \qquad s = 1,2 \qquad (14)$$

$$\pi_i^{(0)} = 1 - \pi_i^{(1)} - \pi_i^{(2)}$$

The predicted probabilities thus obtained are then used to compute the weights. The coefficient of variation was very similar across the models and there were no large outlier values. (See Appendix 5, Table 61 and Table 62 for the distribution of the probabilities and weight, respectively.) For this reason I have simply used the inverse of the propensity score and not trimmed the weights or used adjustment cells (Little 1986, Kalton and Flores-Cervantes 2003). The inverse propensity weights are calibrated by multiplying by a post-stratification adjustment to ensure the population totals for each CASSR remain the same after weighting.

Step 3: treatment of missing auxiliary data

The auxiliary data used to model response propensity is subject to missingness, so I use MI to impute the missing information. To reflect the distributional characteristics of the variables, I use chained equations MI (van Buuren 2007, White et al. 2011). A limitation of this method, already mentioned, is that it does not easily accommodate the clustered structure of the data. Problems arise where I am imputing a missing value within a CASSR-level variable. In such cases, the imputation model treats the variable as if it were an individual-level variable. Consequently, imputed cases are likely to vary within the CASSR. Since this then affects how these variables are treated within the analysis model, this implies that analysis and imputation models are not congenial. I therefore take steps to limit excessive variability (see Appendix 6 for discussion), but after imputation there is a small amount of variability within CASSRs.

Due to the size of the dataset and number of variables with missing data a pragmatic approach had to be taken to the specification of the imputation equations to make the imputation computationally feasible. In general, the imputation equations for each variable include all of the variables used in the analysis model. Where necessary, however, variables were omitted where the predictive significance of the variable was found to be less than 0.15. Derived variables could not be included as their derivation could not be specified within the software, which may produce some bias in results for these variables. To strengthen the plausibility of the MAR assumption, where possible any variables not included in the response propensity model that nevertheless were strongly correlated with the variables to be imputed were also included. Since receipt of some services is entirely missing for some CASSRs, I also use additional CASSR-level data on service receipt and social care expenditure to allow for shifts in service receipt by CASSR (The Information Centre for Health and Social Care 2012a, The Information Centre for Health and Social Care 2012c). Variables with more than 50 per cent of the cases missing (i.e. secondary client group, religion and budget) were not considered for imputation (McKee et al. 1999, Rubin 2003). (See Appendix 6 for details regarding the imputation models.)

In total 20 imputations were carried out, as 17 per cent of cases had incomplete data (White et al. 2011), after a burn-in period of 20 iterations. Tests for the appropriateness of the burn-in period for achieving convergence of the chain to a stationary distribution and number of imputations are described in Appendix 7. After imputation all derived variables were generated. I then estimated the response propensity models and the predicted response probabilities. I use Rubin's rules to combine the predicted probabilities from each dataset and generate a weight (White et al. 2011) that can be applied to the PIs calculated on the multiply-imputed respondent dataset. All imputations and analysis were carried out in Stata 13.

Method for comparing the effects of adjusting for nonresponse on performance assessment

To assess the effect of adjusting PIs on inferences about the performance of organisations, I look at whether there are differences between the unadjusted PIs and adjusted PIs, in terms of (i) the ranking of CASSRs and (ii) the identification of outlying organisations. I compare the ranking of CASSRs attained under the MAR assumption to the ranking of CASSRs under the MCAR assumption, using caterpillar plots and a variety of correlation statistics. These include Pearson's and Spearman's rank correlation coefficients, as well as Kendall's tau statistic. The tau statistic is calculated based on the degree of concordance and discordance

in rankings between pairs of CASSRs. A Kendall's τ of one indicates that organisations are ranked in exactly the same order, both before and after adjustment. The tau coefficient therefore makes it possible to quantify the proportion of CASSR pairs that change order after the adjustment, using the formula $100*(1-\tau)/2$ (Johnson et al. 2010).

To examine the effect of nonresponse on the identification of outliers, I use a funnel plot, as described in Chapter 3 and decompose the effect of nonresponse on performance assessment into the mean outcome, volume and variability components. I calculate the number of outliers under the different methods for addressing missingness in the data, specifying whether cases are 'in control', high or low 'alerts', or high or low 'alarms' depending on where they lie in relation to the funnel plot control limits (see Chapter 3 for discussion). I also look at the number of 'movements' between the different outlier states. To assess the significance of the changes in the number of movements as a consequence of nonresponse adjustment, I draw on the concept of 'false negatives' and 'false positives', as discussed in Chapter 3. The assumption of this analysis is that the data are truly MAR, the likelihood of which I reflect on in the discussion.

Results

Assessing the effect of nonresponse on PI scores

In this section I provide estimates of the extent of bias due to item and unit nonresponse for each of the ASCOF indicators. Several different estimates are presented in Table 12 and Table 13 to enable interpretation of both the significance and meaningfulness of the bias. A point estimate for the amount of bias is presented which is calculated by subtracting the indicator estimated under the MAR assumption from the indicator estimated under the MCAR assumption. To appreciate the magnitude of the bias I also present the absolute bias and the absolute bias as a percentage of the scale, since the latter provides a way of comparing the relative magnitude of the bias across the different indicators. I also present the absolute bias as a percentage of the SE estimated under the MCAR assumption, to provide an indication of the statistical significance of the bias. Changes in variability can be understood by examining the SE difference (MAR – MCAR) as a percentage of the SE estimated under the MCAR assumption. Since bias and changes in variability may vary across CASSRs, it is important to consider their distribution among CASSRs. Therefore, I report the number of CASSRs where the bias is greater than one percentage point of the given scale, the number of CASSRs where the bias is negative, and the number of CASSRs where the bias is greater than 196 per cent of the indicator's SE (corresponding to the 95% confidence interval)

estimated under the MCAR assumption, and the number of CASSRs with increases in variability.

In general, at both the overall population level and CASSR level the estimates for the ASCOF indicators derived under MAR assumptions are very similar to those derived under the MCAR assumption. The mean difference observed at the population level, as well as for CASSRs is in all cases less than one percentage point of the scale, with slightly greater shifts being observed after IPW/CCA compared to CCA/MI. The different levels of unit and item nonresponse means it is difficult to compare the IPW/CCA and CCA/MI results directly, but it seems likely that IPW/CCA has more effect than CCA/MI, potentially as a consequence of the higher rates of unit compared to item nonresponse. This is illustrated in Figure 10 and Figure 11, which show how the magnitude of the estimated bias increases as nonresponse rates increase. The fitted lines show that the gradient of the line is steeper for the relationship between item nonresponse rate and the MAR-MCAR difference, than for unit nonresponse rate. This suggests that the effect of MI is minimised by the fairly low levels of item nonresponse observed in the dataset.

Table 12: Estimates of the extent of bias due to item and unit nonresponse for the SCRQoL and satisfaction indicators

Treatment of missingness (n for MAR)	SCRQoL indicator[a]			Satisfaction indicator[a]		
	CCA/MI (n=61,026)	IPW/CCA (n=54,350)	IPW/MI (n=61,026)	CCA/MI (n=61,026)	IPW/CCA (n=57,929)	IPW/MI (n=61,026)
Difference (adjusted – raw)						
Total sample	-0.031	-0.060	-0.093	-0.002	-0.002	-0.004
Mean	-0.035	-0.063	-0.102	-0.002	-0.002	-0.005
Maximum	0.137	0.111	0.153	0.005	0.017	0.012
Minimum	-0.300	-0.340	-0.481	-0.016	-0.028	-0.032
Freq more +ve (% sample)	45 (30%)	23 (15%)	25 (17%)	36 (24%)	51 (34%)	33 (22%)
Absolute Difference (|adjusted – raw|)						
Total sample	0.031	0.060	0.093	0.002	0.002	0.004
Mean	0.068	0.070	0.120	0.003	0.004	0.006
Maximum	0.300	0.340	0.481	0.016	0.028	0.032
Minimum	0.000	0.000	0.000	0.000	0.000	0.000
Absolute difference as a percentage of the scale						
Total sample	0.1%	0.2%	0.4%	0.2%	0.2%	0.4%
Mean	0.3%	0.3%	0.5%	0.3%	0.4%	0.6%
Maximum	1.2%	1.4%	2.0%	1.6%	2.8%	3.2%
Minimum	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Freq >1% pt (% sample)	2 (1%)	4 (3%)	13 (9%)	3 (2%)	14 (9%)	35 (23%)
Absolute difference as a percentage of raw SE						
Total sample	194%	374%	583%	116%	113%	233%
Mean	34%	35%	60%	13%	18%	26%
Maximum	147%	169%	236%	65%	118%	136%
Minimum	0%	0%	0%	0%	0%	0%
Freq > 196% (% sample)	0 (0%)	0 (0%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)

Treatment of missingness (n for MAR)	SCRQoL indicator[a]			Satisfaction indicator[a]		
	CCA/MI (n=61,026)	IPW/CCA (n=54,350)	IPW/MI (n=61,026)	CCA/MI (n=61,026)	IPW/CCA (n=57,929)	IPW/MI (n=61,026)
Change in SE as a percentage of raw SE (adjusted-raw)						
Total sample	-5.0%	5%	-0.6%	0.1%	2.1%	2.0%
Mean	-5%	5%	0%	-0.3%	2.1%	2%
Maximum	1%	23%	20%	9.4%	8.6%	11%
Minimum	-13%	-1%	-13%	-2.8%	-0.2%	-2%
Freq SE incr. (% sample)	5 (3%)	145 (97%)	62 (42%)	42 (28%)	148 (99%)	128 (86%)

Key: [a] number of cases under MCAR assumption: SCRQoL=54,350, Satisfaction=57,929

Table 13: Estimates of the extent of bias due to item and unit nonresponse for the control, safety and access to information indicators

Treatment of missingness (n for MAR)	Control indicator[a]			Safety indicator[a]			Information indicator[a]		
	CCA/MI (n=61,026)	IPW/CCA (n=59,478)	IPW/MI (n=61,026)	CCA/MI (n=61,026)	IPW/CCA (n=59,499)	IPW/MI (n=61,026)	CCA/MI[b] (n=45,313)	IPW/CCA (n=42,884)	IPW/MI[b] (n=45,313)
Difference (raw – adjusted)									
Total sample	0.000	-0.003	-0.003	0.000	-0.004	-0.004	0.001	-0.003	-0.003
Mean	0.000	-0.003	-0.003	0.000	-0.004	-0.004	0.000	-0.004	-0.003
Maximum	0.005	0.010	0.012	0.003	0.003	0.003	0.009	0.026	0.023
Minimum	-0.009	-0.020	-0.020	-0.005	-0.023	-0.022	-0.013	-0.019	-0.021
Freq more +ve (% sample)	83 (56%)	43 (29%)	45 (30%)	83 (56%)	23 (15%)	25 (17%)	84 (56%)	37 (25%)	40 (27%)
Absolute Difference (|raw – adjusted|)									
Total sample	0.000	0.003	0.003	0.000	0.004	0.004	0.001	0.003	0.003
Mean	0.002	0.005	0.005	0.001	0.004	0.005	0.003	0.005	0.005
Maximum	0.009	0.020	0.020	0.005	0.023	0.022	0.013	0.026	0.023
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Absolute difference as a percentage of the scale									
Total sample	0.0%	0.3%	0.3%	0.0%	0.4%	0.4%	0.1%	0.3%	0.3%
Mean	0.2%	0.5%	0.5%	0.1%	0.4%	0.5%	0.3%	0.5%	0.5%
Maximum	0.9%	2.0%	2.0%	0.5%	2.3%	2.2%	1.3%	2.6%	2.3%
Minimum	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Freq >1% pt (% sample)	0 (0%)	15 (10%)	16 (11%)	0 (0%)	15 (10%)	14 (9%)	1 (1%)	16 (11%)	21 (14%)
Absolute difference as a percentage of raw SE									
Total sample	11%	170%	161%	1%	376%	374%	35%	171%	138%
Mean	8%	22%	24%	7%	34%	36%	11%	20%	21%
Maximum	41%	104%	102%	55%	172%	161%	45%	91%	88%
Minimum	0%	0%	0%	0%	0%	0%	0%	0%	0%
Freq > 196% (% sample)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

Treatment of missingness (n for MAR)	Control indicator[a]			Safety indicator[a]			Information indicator[a]		
	CCA/MI (n=61,026)	**IPW/CCA (n=59,478)**	**IPW/MI (n=61,026)**	**CCA/MI (n=61,026)**	**IPW/CCA (n=59,499)**	**IPW/MI (n=61,026)**	**CCA/MI[b] (n=45,313)**	**IPW/CCA (n=42,884)**	**IPW/MI[b] (n=45,313)**
Change in SE as a percentage of raw SE (adjusted-raw)									
Total sample	-0.1%	2.7%	2.5%	0.5%	7.5%	7.8%	-0.5%	3.0%	2.4%
Mean	-0.4%	2.7%	2%	0.1%	7.2%	7%	-0.3%	3.0%	3%
Maximum	2.5%	12.1%	16%	19.5%	37.3%	37%	5.0%	10.4%	13%
Minimum	-3.5%	-1.5%	-2%	-3.0%	-3.8%	-4%	-2.4%	-3.9%	-4%
Freq SE incr. (% sample)	36 (24%)	137 (92%)	125 (84%)	61 (41%)	134 (90%)	129 (87%)	45 (30%)	140 (94%)	136 (91%)

Key: [a] number of cases under MCAR assumption: Control=59,478, Safety=59,499, Information=42,884; [b] Sample varies between 45,313 and 45,396, as imputations differ in how many responses are imputed as 'don't know' response options.

The effect of adjusting for item and unit nonresponse using the IPW/MI approach is in general to make the ASCOF estimates more negative, implying that it is on average people who have worse quality of life and are less satisfied with their care who do not respond or do not fully complete the questionnaires. Notably, however, for the control, safety and information indicators, CCA/MI adjustment has on average no effect or, in the case of the information indicator, a slightly positive effect. The effect of adjustment also varies across the CASSRs. For all indicators there are some CASSRs that have more positive estimates after nonresponse adjustment. This is illustrated more clearly in the graphs in Figure 12, which show the distribution of bias in CASSR indicator scores, or shifts in indicator estimates, following adjustment for unit and item nonresponse for each of the indicators, as a proportion of the standard error.

The extent of bias estimated as a result of adjusting for both item and unit nonresponse differs by indicator as illustrated in Figure 12. Where adjustment for item nonresponse has a perceptible effect on average scores (i.e. for the SCRQoL and satisfaction indicators), the effect of adjusting for both item and unit nonresponse is to shift the distribution of indicators scores further to the left. Interestingly this does not always result in more CASSRs with negative scores. Examining movements for individual CASSRs, it seems that in some cases item and unit nonresponse have opposing biasing effects on scores. This means that for some CASSRs the effects of adjusting for both item and unit nonresponse is to some extent to cancel each other out.

Adjustments for nonresponse using IPW/CCA increase the variance of PIs. Standard errors increase after weighting for at least 90 per cent of CASSRs for all indicators. By comparison CCA/MI tends to have a negative effect on the standard errors for the majority of CASSRs and for the whole sample for most PIs. Only for the safety indicator does CCA/MI appear to lead to larger standard errors for a substantial number of CASSRs, with some fairly large decreases in the precision of indicator estimates. The combined IPW/MI effect for most PIs follows the IPW/CCA effect and is, on average and for the majority of CASSRs, precision-reducing. The only indicator for which this is not the case is the SCRQoL indicator. Although some CASSRs experience reductions in the precision of this PI, on average, and for the majority of CASSRs, the effect is precision-enhancing. While the effects of MI and IPW on estimated bias are fairly similar and broadly all in the same direction, the procedures have very different effects on precision.

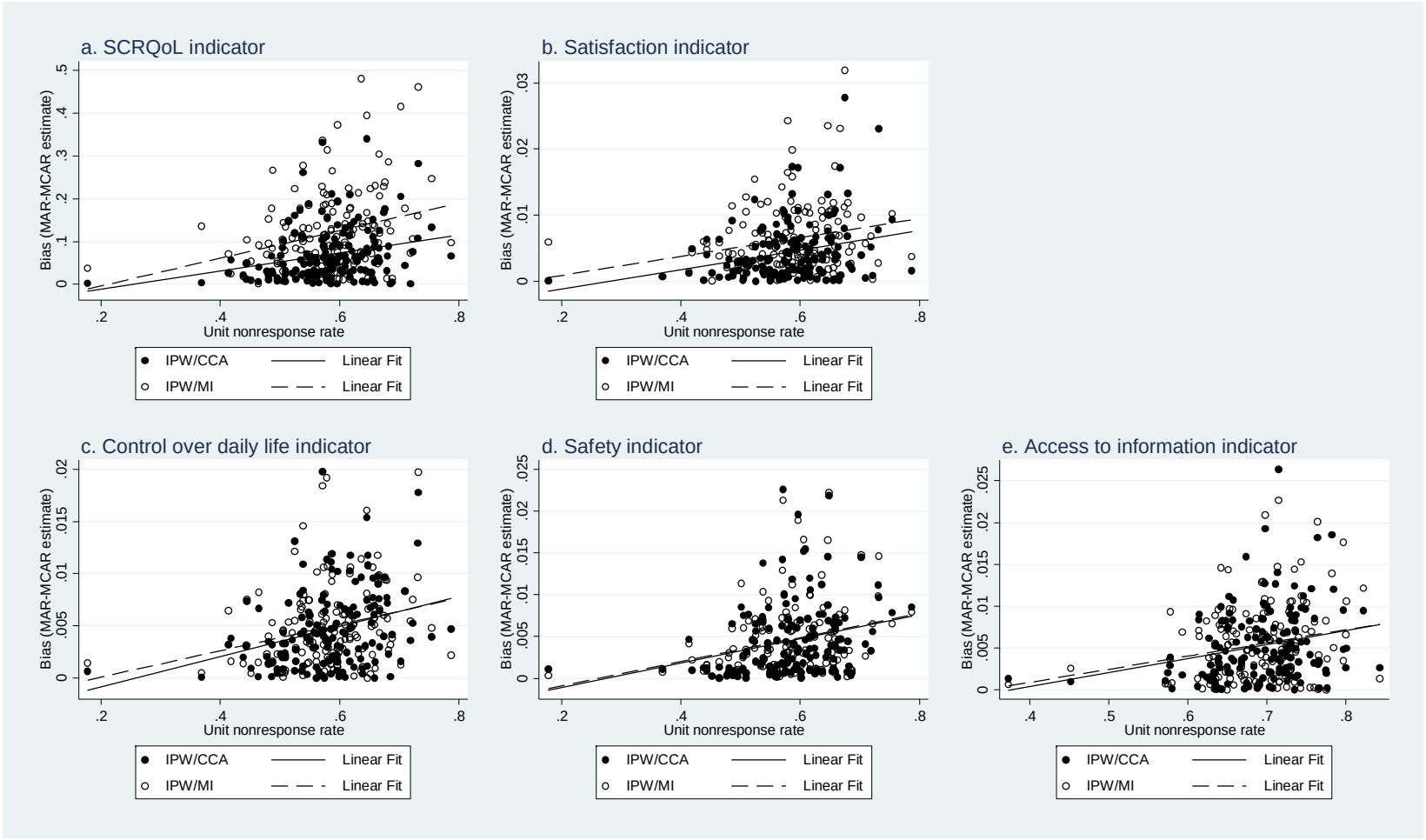

Figure 10: Relationship between bias (MAR-MCAR) and CASSR unit nonresponse rate for all indicators

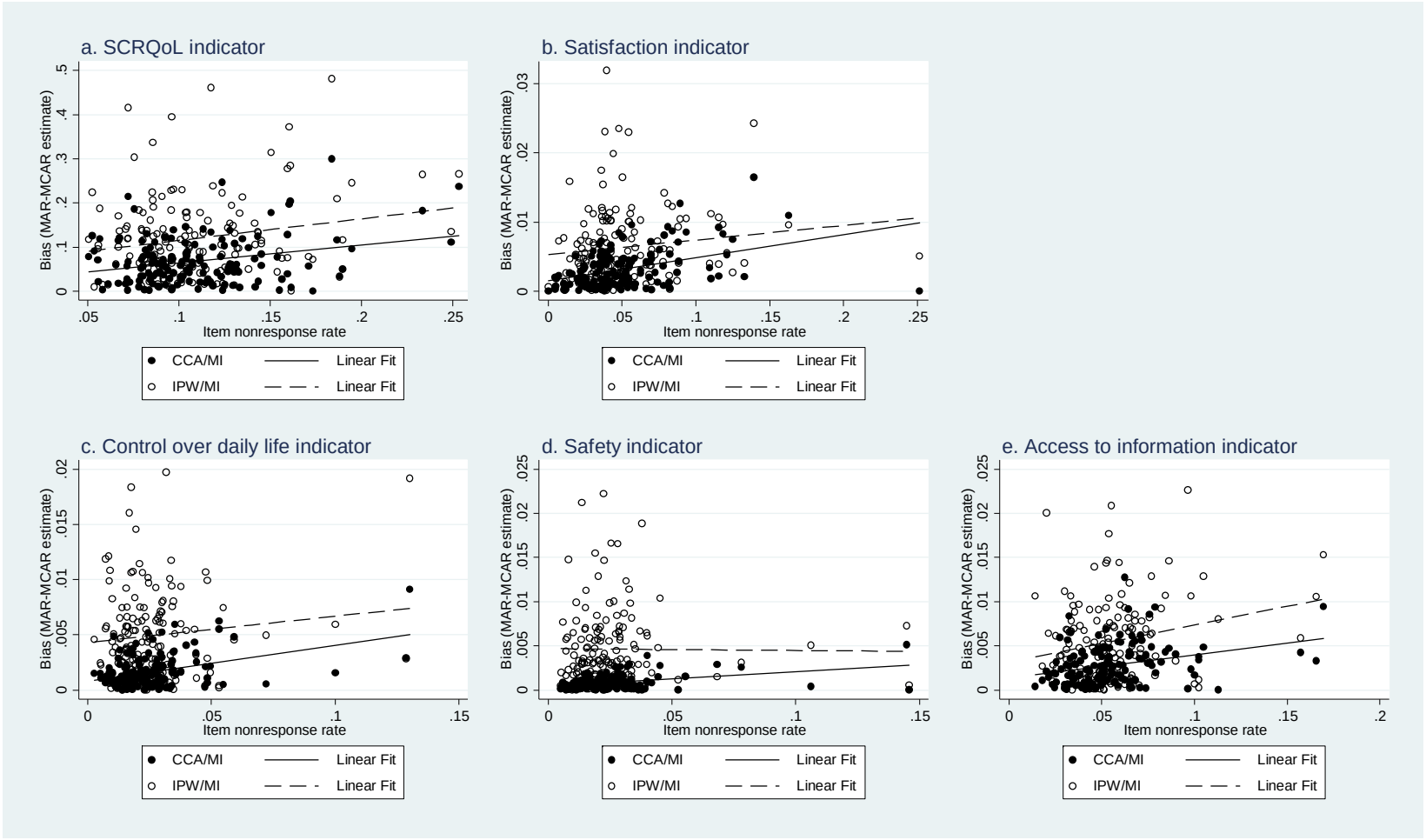

Figure 11: Relationship between bias (MAR-MCAR) and CASSR item nonresponse rate for all indicators

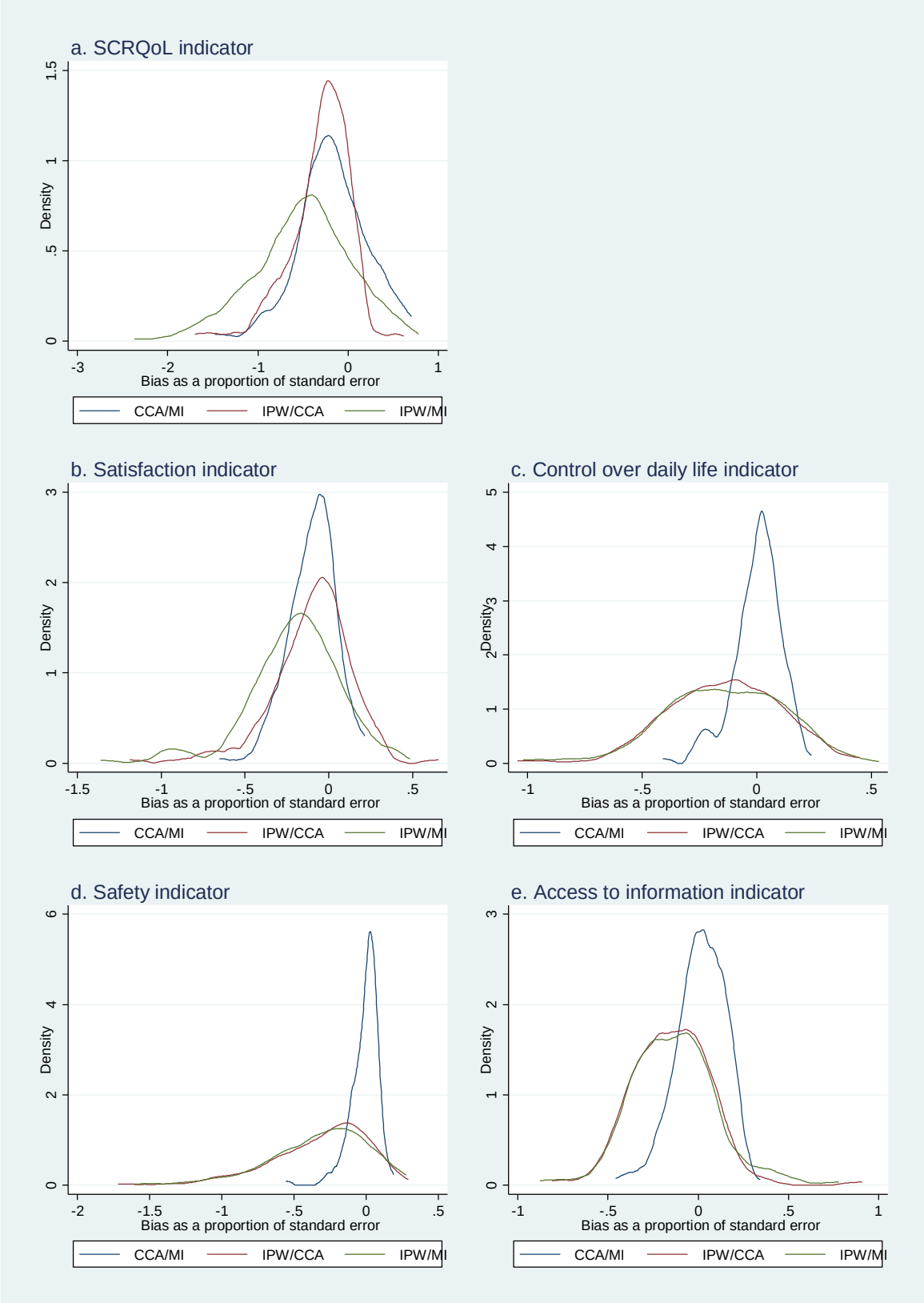

Figure 12: Distribution of bias (MAR-MCAR) in PI scores as a proportion of SE

The magnitude of all shifts in the indicator scores as a result of nonresponse is small. However, at the overall population level the sample size is much larger and estimates more precise, meaning that for many indicators the estimated bias is greater than the width of the 95 per cent confidence interval (shift >196% of the SE). At the CASSR level the picture is very different, with only one CASSR moving more than 1.96 standard errors, and then only for the SCRQoL indicator following adjustment for both item and unit nonresponse. As Figure 12 illustrates for each indicator and for the overwhelming majority of CASSRs movements are less than one standard error.

Assessing the effect of nonresponse on performance assessment

In this section I assess the effect of nonresponse to the ASCS on performance assessment. I first look at the effect of nonresponse on the rankings of CASSRs within caterpillar plots, and consider the extent to which nonresponse may influencing rankings by looking at the number and magnitude of changes in rank ordering using a range of correlation statistics. Following this, I present the indicators scores for CASSRs using funnel plots and examine how nonresponse may be affecting the identification of outliers.

1. Caterpillar plots and rankings

The graphs in Figure 13 to Figure 17 illustrate the effect of adjusting for item and unit nonresponse on the distribution of indicators scores for CASSRs using caterpillar plots, with approximate 95% confidence intervals. From visual inspection of the plots, without labels for individual CASSRs it is difficult to see much difference in the plots after adjustment for nonresponse (plots b, c and d) compared to under CCA (plot a). There is a perceptible increase in variability after weighting for unit nonresponse, as seen by wider-looking confidence intervals for some of the CASSRs, but the general distribution of the indicator scores looks fairly similar.

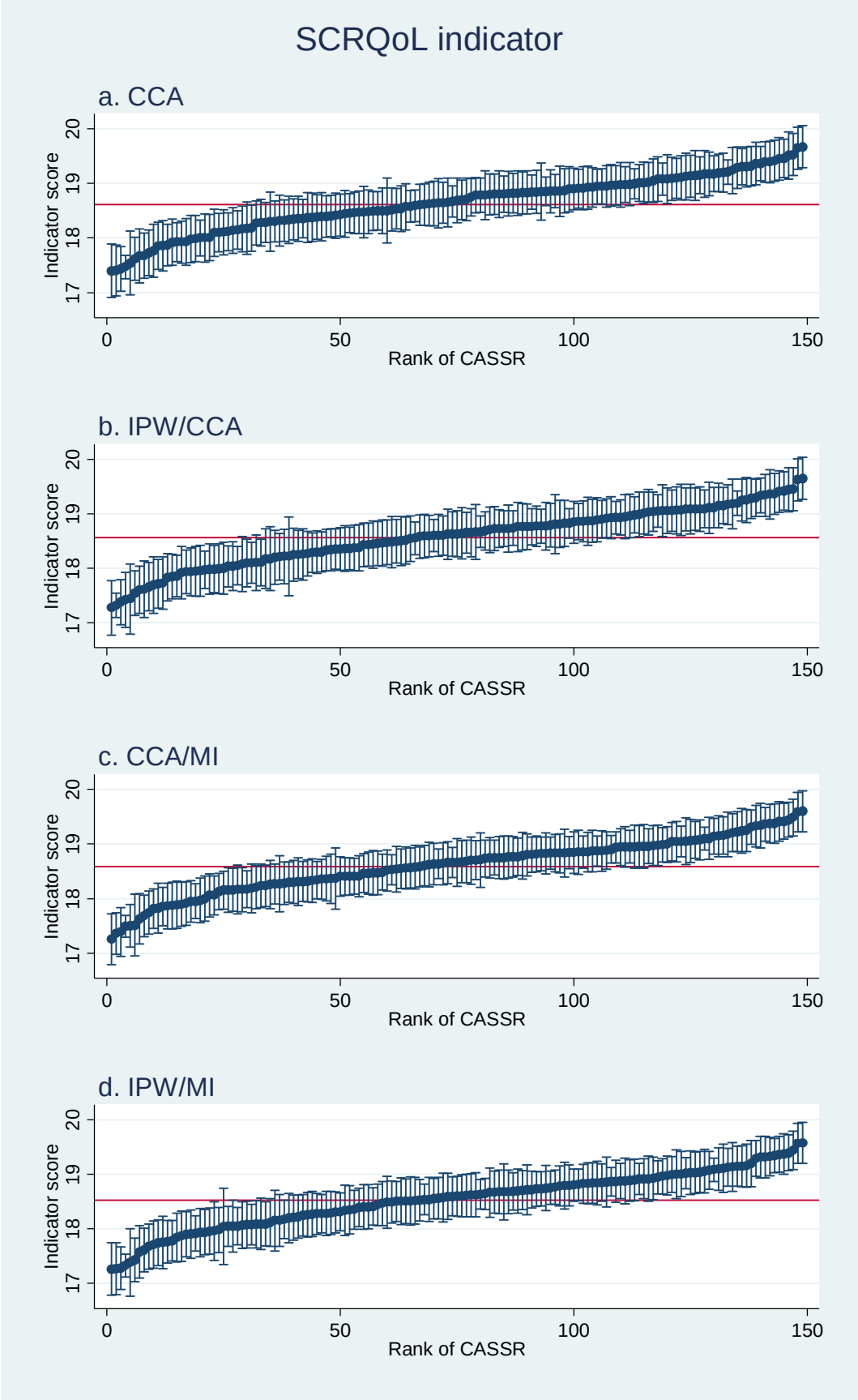

Figure 13: Caterpillar plots of the SCRQoL indicator, with approximate 95% confidence intervals, under different nonresponse adjustments

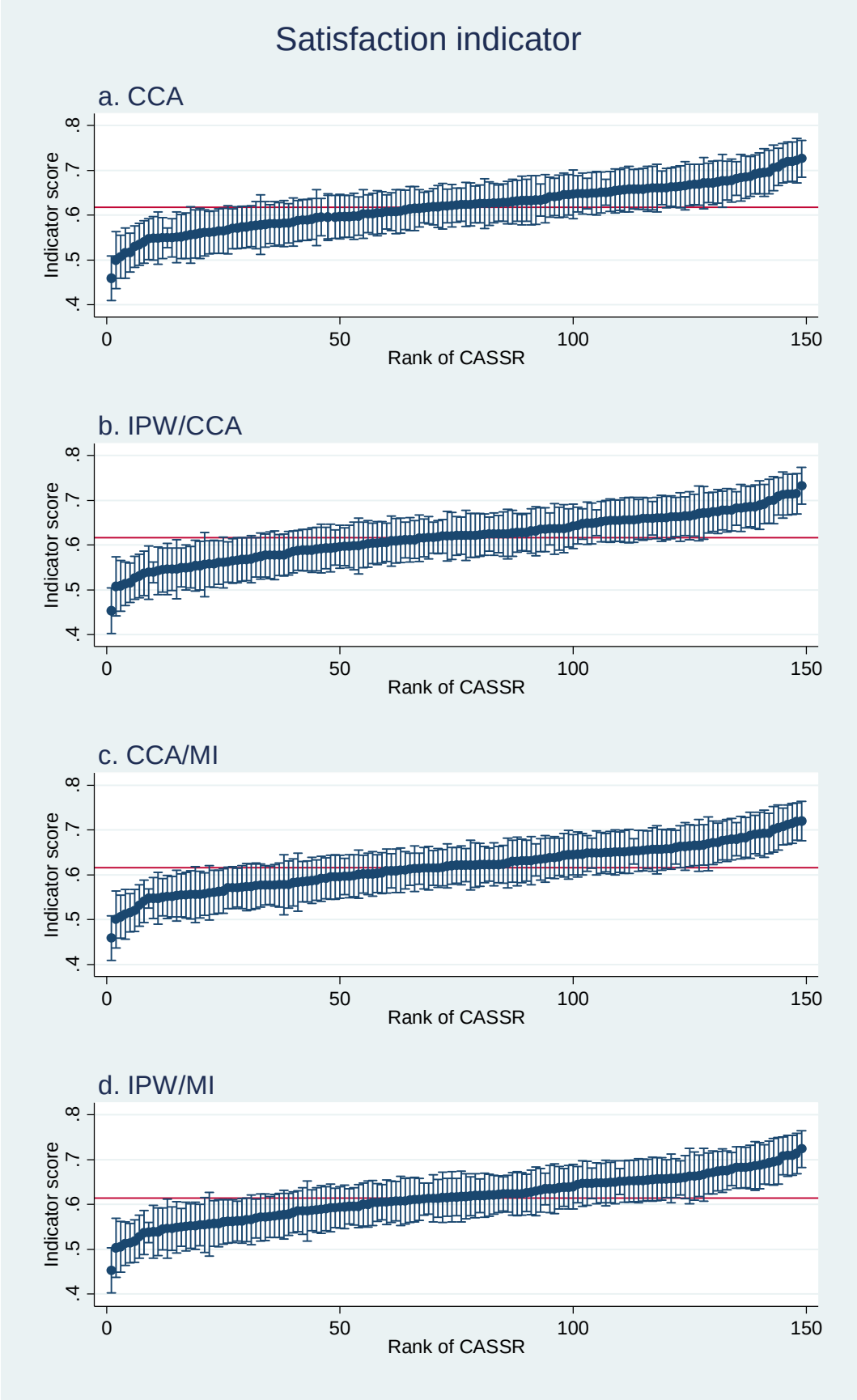

Figure 14: Caterpillar plots of the satisfaction indicator, with approximate 95% confidence intervals, under different nonresponse adjustments

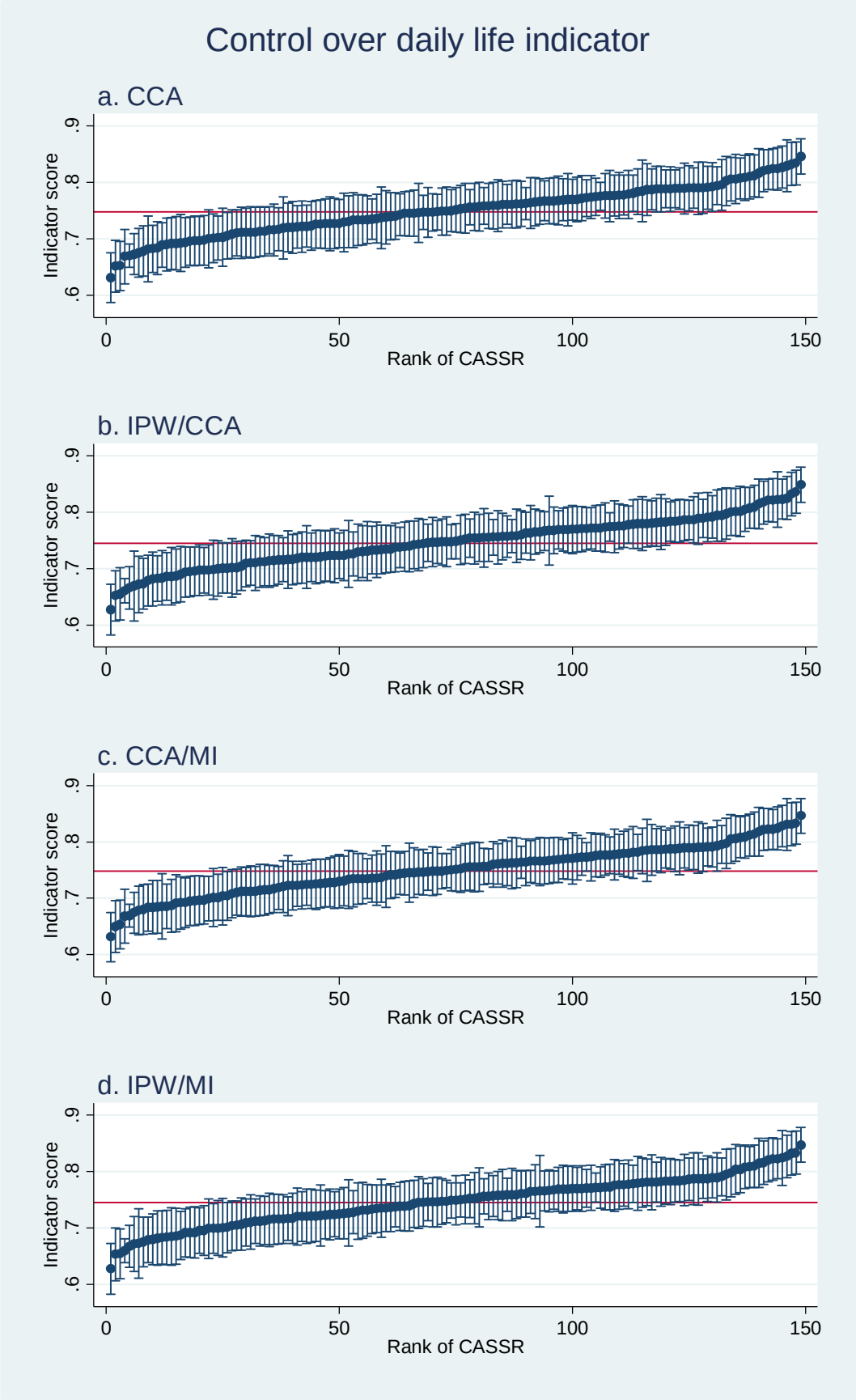

Figure 15: Caterpillar plots of the control over daily life indicator, with approximate 95% confidence intervals, under different nonresponse adjustments

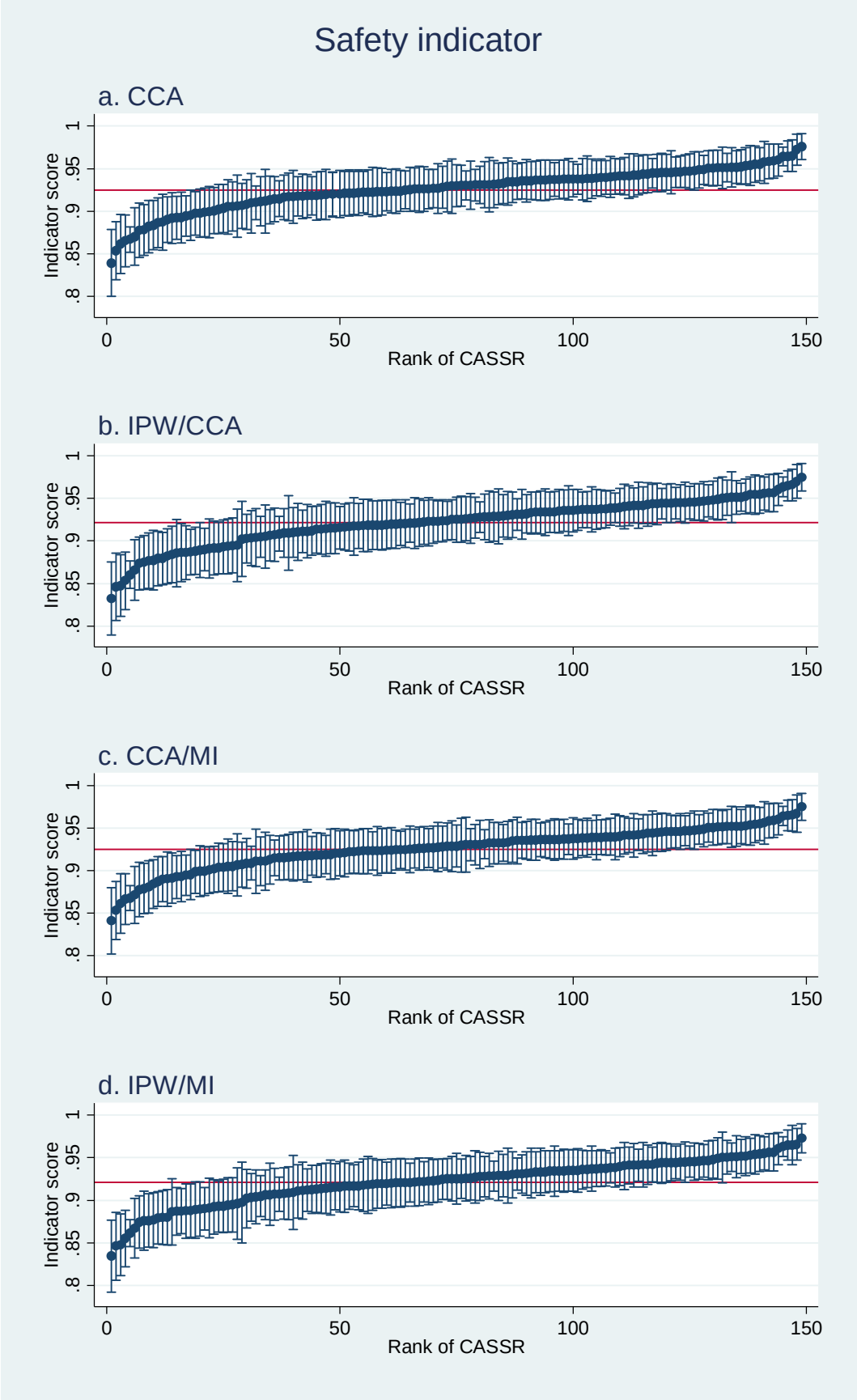

Figure 16: Caterpillar plots of the safety indicator, with approximate 95% confidence intervals, under different nonresponse adjustments

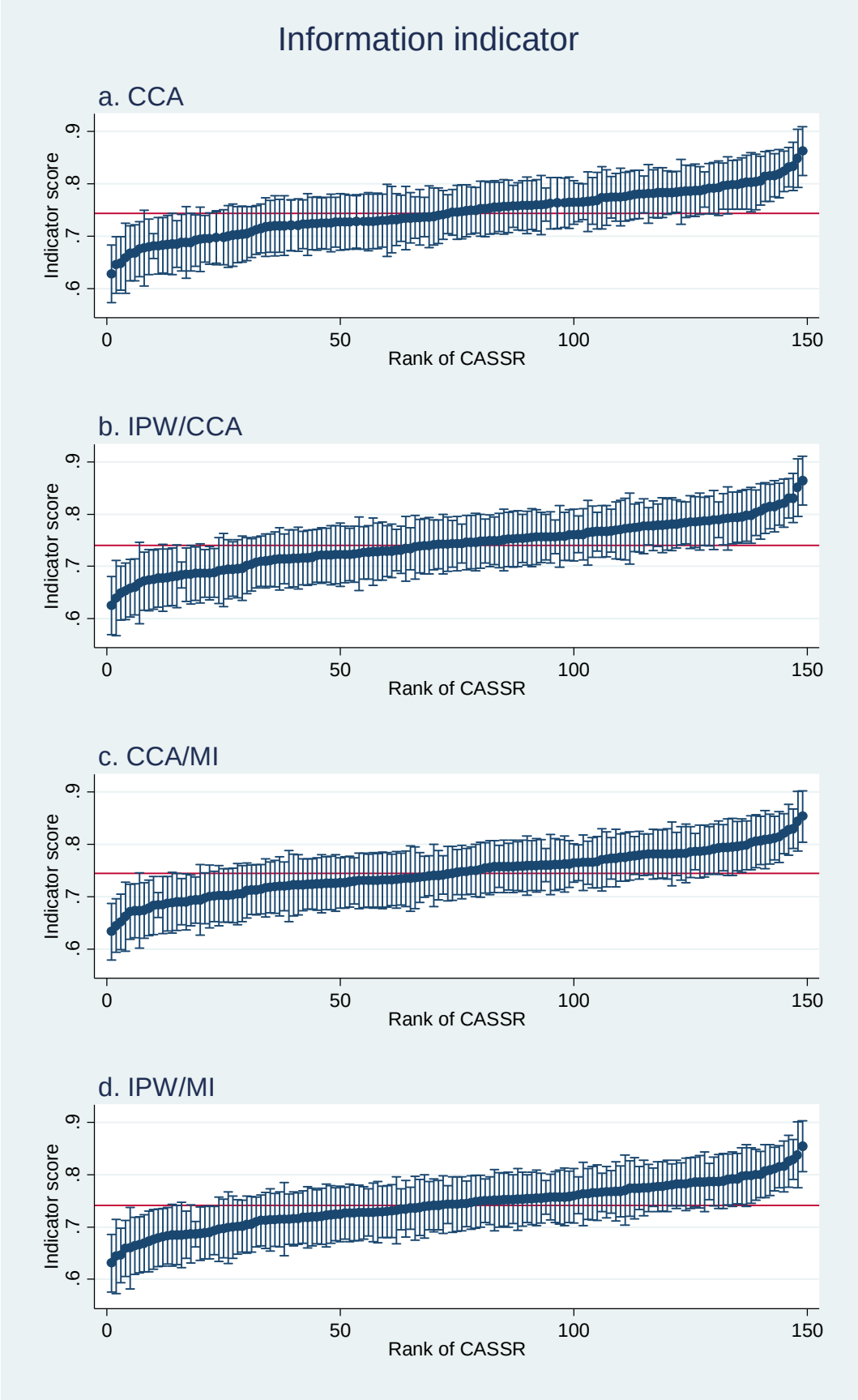

Figure 17: Caterpillar plots of the information indicator, with approximate 95% confidence intervals, under different nonresponse adjustments

Since the caterpillar plots imply a ranking of CASSRs it is useful to see how similar the ordering of CASSRs is following adjustment for unit and item nonresponse. Table 14 summarises the similarity in ranking of CASSRs pre and post adjustment using a range of correlation statistics. The Pearson and Spearman's rank correlation coefficients are in all cases, for all indicators, always above 0.9, which indicates near perfect correlation. Kendall's tau statistic is always above 0.85 and indicates that fewer than ten per cent of the CASSR pairs change order. In all cases the percentage changing order is highest for the combined IPW/MI adjustment, and in general it is following adjustment for unit nonresponse that the most pairs change order. The exception is for the SCRQoL indicator where more pairs of CASSRs change order after CCA/MI adjustment compared to after IPW/CCA adjustment. This may reflect the higher rates of item nonresponse observed for this indicator.

Table 14: Correlation statistics between indicators estimated under MCAR and MAR assumptions

Indicator	Missing data adjustments	Pearson's R^2 (p-value)	Rho (p-value)	Tau (p-value)	% pairs change order
SCRQoL PI	IPW/CCA	0.991	0.991	0.928	3.6%
	CCA/MI	0.988	0.983	0.897	5.1%
	IPW/MI	0.976	0.971	0.860	7.0%
Satisfaction PI	IPW/CCA	0.994	0.993	0.936	3.2%
	CCA/MI	0.998	0.997	0.959	2.0%
	IPW/MI	0.991	0.990	0.920	4.0%
Control PI	IPW/CCA	0.993	0.992	0.927	3.6%
	CCA/MI	0.999	0.998	0.970	1.5%
	IPW/MI	0.992	0.991	0.924	3.8%
Safety PI	IPW/CCA	0.987	0.982	0.899	5.1%
	CCA/MI	0.999	0.998	0.971	1.5%
	IPW/MI	0.986	0.981	0.892	5.4%
Information PI	IPW/CCA	0.992	0.989	0.922	3.9%
	CCA/MI	0.997	0.995	0.954	2.3%
	IPW/MI	0.990	0.987	0.911	4.5%

Although Table 14 shows that correlations are very high between the PIs estimated under different assumptions about the missingness mechanism, for individual CASSRs some of the changes in rank can be quite large. This is illustrated in the histograms (Figure 18 to Figure 22), which show the difference between the rank position of CASSRs relative to each

other before and after adjustment for nonresponse using the three different methods. Positive changes in rank position indicate that after adjustment for nonresponse the rank position of the CASSR improves, and vice-versa for negative changes. For the SCRQoL indicator, CCA/MI and IPW/CCA adjustment has a similar effect on the rank position of CASSRs. After adjustment a small number of CASSRs move over 20 positions, and a larger number move over 10 positions. The graphs demonstrate how the combined effect of IPW/MI increases the number of changes in rank, and also leads to more and larger changes in the rank position of CASSRs, with one CASSR dropping 45 positions on the SCRQOL indicator.

In contrast for the satisfaction, control over daily life, safety and access to information indicators, weighting for unit nonresponse has a larger effect than MI on rank positions, causing more and larger movements in rank. Graphs showing the combined effect of adjustment for unit and item nonresponse for these indicators look very similar to those generated following IPW/CCA adjustment. The movements after IPW/MI adjustment are in general smaller for these four indicators than those for the SCRQoL indicator, with the maximum movement being a drop of 24 positions for the satisfaction indicator, 22 positions for the control over daily life indicator, 42 positions for safety, and an increase of 34 positions for the access to information indicator.

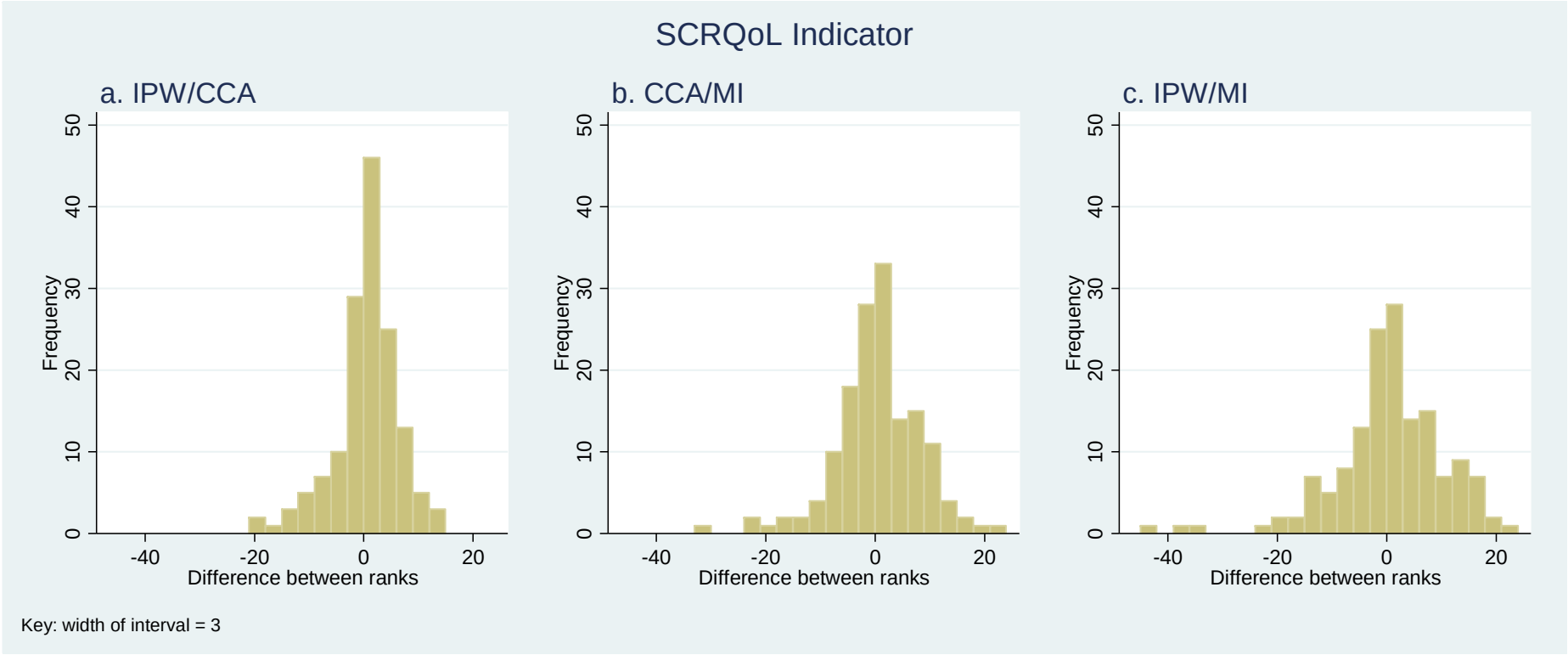

Figure 18: Distribution of CASSR changes in rank (MAR-MCAR) for the ASCOF SCRQoL indicator

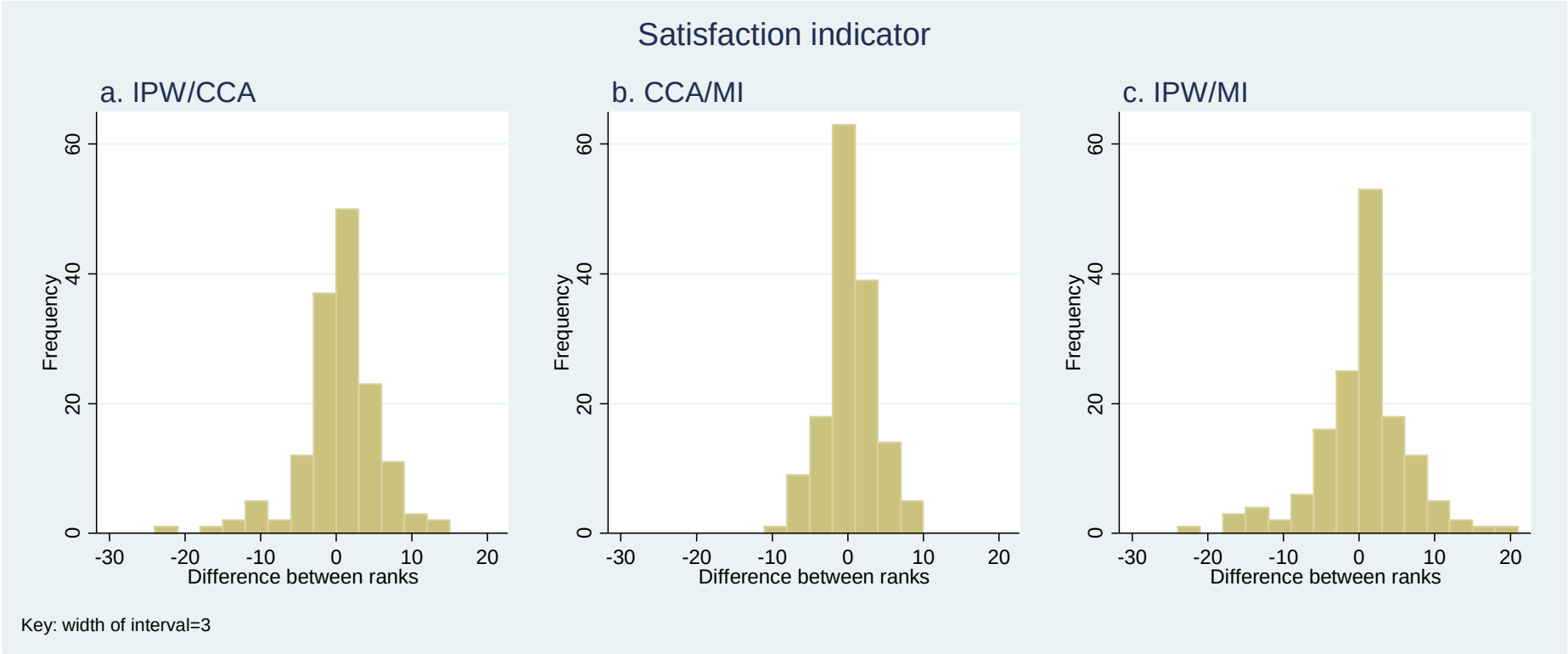

Figure 19: Distribution of CASSR changes in rank (MAR-MCAR) for the satisfaction indicator

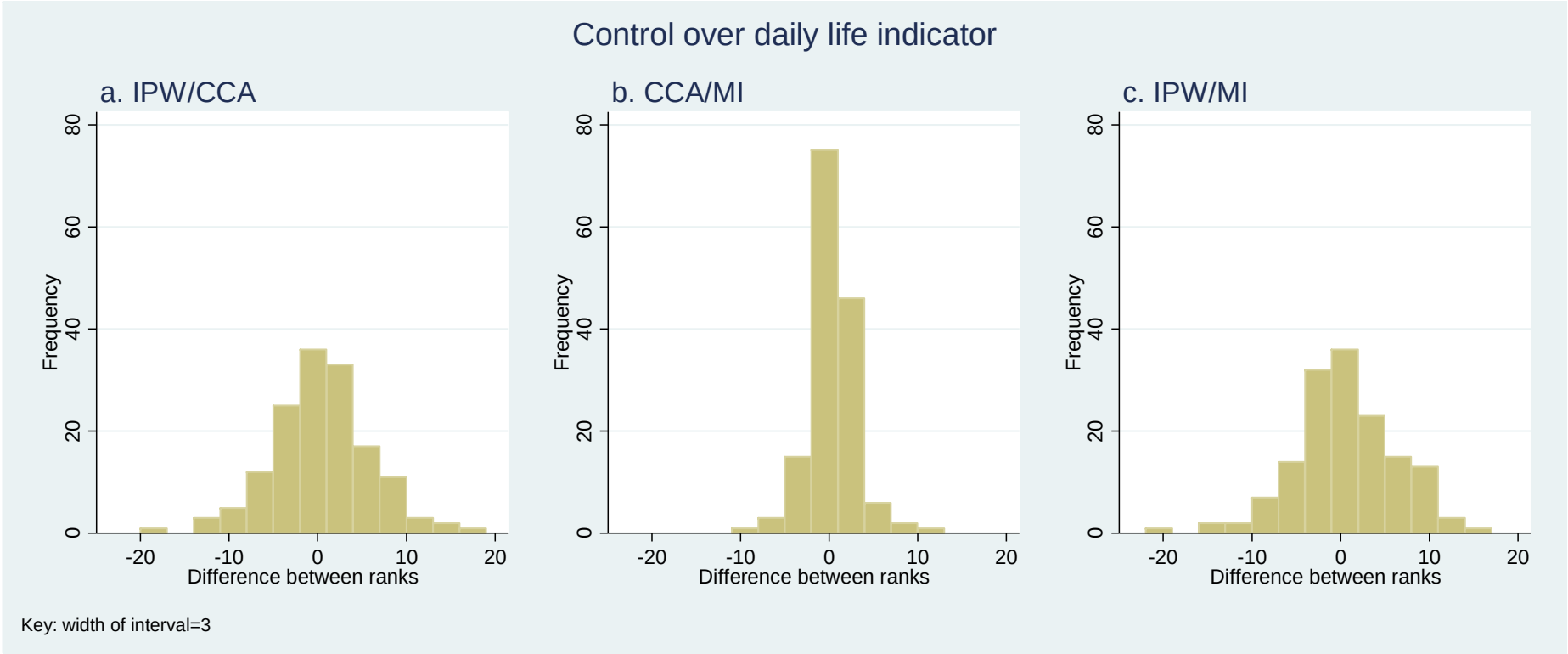

Figure 20: Distribution of CASSR changes in rank (MAR-MCAR) for the control over daily life indicator

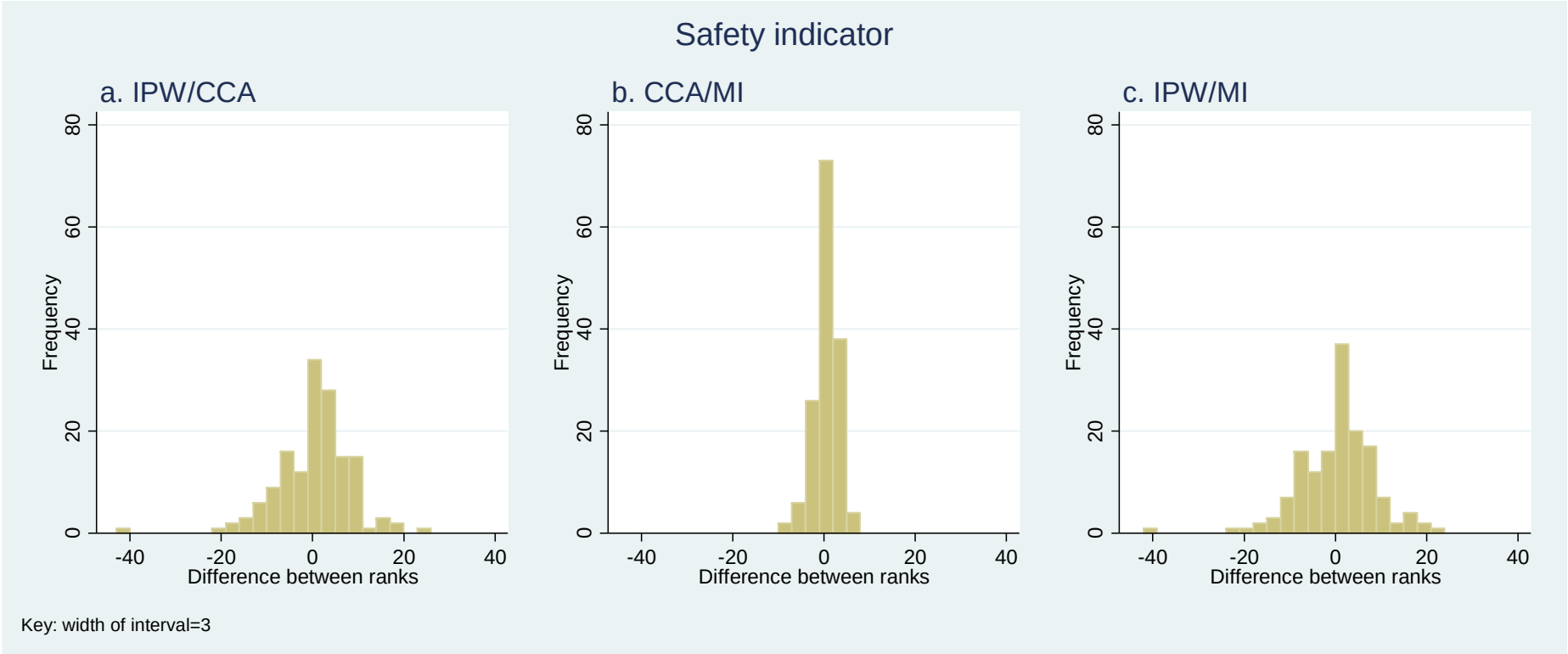

Figure 21: Distribution of CASSR changes in rank (MAR-MCAR) for the safety indicator

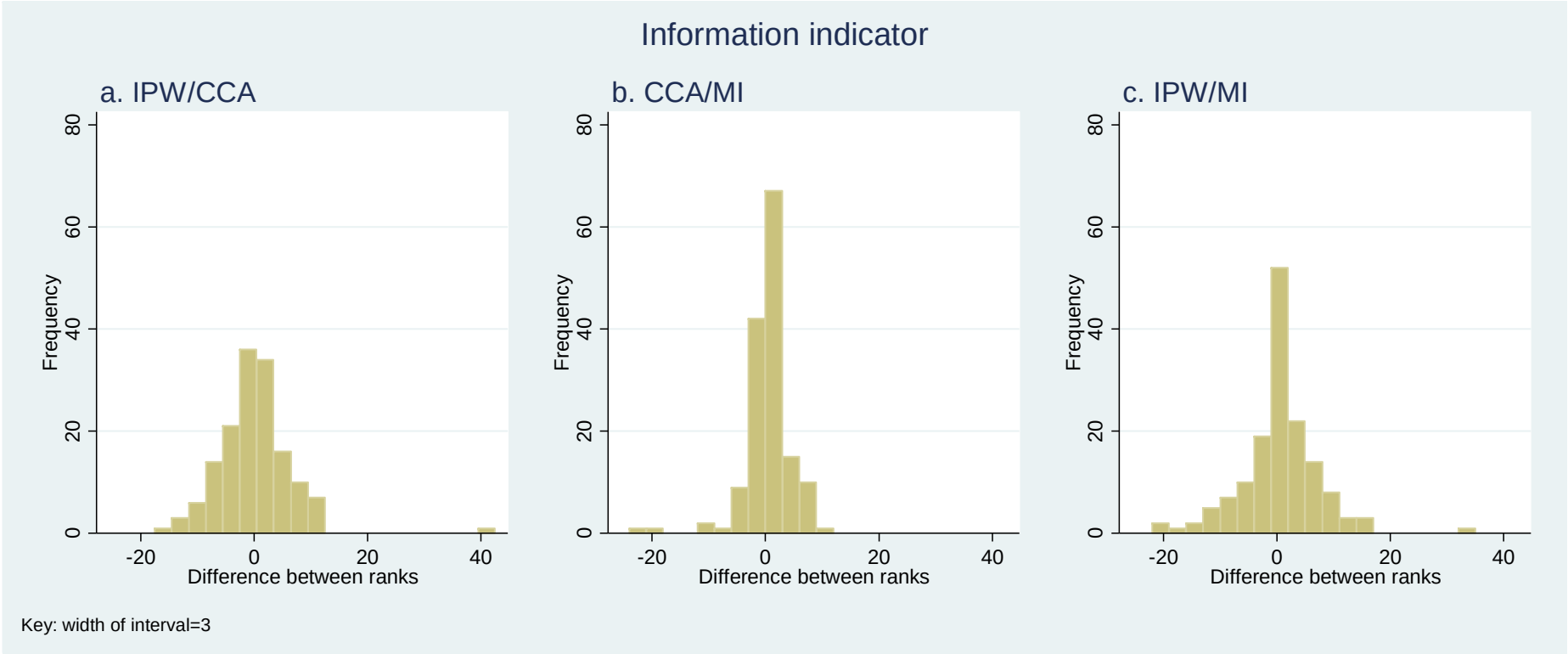

Figure 22: Distribution of CASSR changes in rank (MAR-MCAR) for the access to information indicator

2. Funnel plots and the identification of outliers

Figure 23 to Figure 27 illustrate the effect of adjusting for item and unit nonresponse on the distribution of CASSR scores using funnel plots. The top and bottom ten CASSRs under CCA are marked in green and blue, respectively, so it is possible to follow the movements of these CASSRs after nonresponse adjustment. Before considering the effects of nonresponse adjustment, it is worth noting two features of the charts. First, the vast majority of organisations have similar (complete case and respondent) sample sizes, at around 350 to 400 people, but a small number have much larger sample sizes. This reflects decisions taken within the CASSRs about how to interpret the national guidance and differences in the assumptions they made about expected response rates. While a few CASSRs seem to have potentially overestimated (or been overly cautious in estimating) the rate of nonresponse, the CASSR with the largest respondent sample did not sample from their population, but sent out questionnaires to the entire eligible adult social care population. The funnel plots illustrate strikingly the greater accuracy attained with larger sample sizes – information which is lost in the caterpillar plots.

Second, for the SCRQoL indicator and the satisfaction indicator a large proportion of CASSRs are 'out of control', meaning that they fall beyond the 95% and 99.8% control limits. This can be seen more clearly by inspection of the results in Table 15, which shows the frequency and percentage of CASSRs identified as 'in control' and out of control, as represented by the alert or alarm status. In the case of the SCRQoL indicator, nearly half the CASSRs fall into one of the positive or negative alert statuses and around 20 per cent into one of the negative or positive alarm statuses. For satisfaction, fewer CASSRs are 'out of control'; nevertheless, nearly 40 per cent fall into the (positive or negative) alert status and 15 per cent fall into one of the negative or positive alarm status. Checks for 'overdispersion' (see Table 69, Appendix 8), however, did not indicate that the process is 'out of control' so I have not made any adjustments to control limits (Spiegelhalter 2005b).

Inspection of the plots shows that there is very little difference in the width of the control limits after nonresponse adjustment. Although nonresponse adjustment does increase the variability of estimates, it is not a large enough effect to have a perceptible impact on the width of the control limits. Equally, as might be expected given the results already presented, there is little difference to be seen between the distribution of points on the three funnel plots that include adjustment for item and unit nonresponse ('b', 'c' and 'd' in the charts) and the funnel plot ('a' in the charts) that makes no adjustment for nonresponse. There are a few small movements in the top and bottom ten, sometimes driven by CCA/MI adjustment and

sometimes by IPW/CCA adjustment, depending on the indicator. These seem to be most significant for the SCRQoL and satisfaction PIs.

Table 15 shows that adjustment for either of, or both, item and unit nonresponse does not have a large effect on the number of outliers identified. In most instances there are only a few changes in the number of outliers following adjustment for either or both of item and unit nonresponse. There are also no clear patterns to the effects, with adjustment sometimes leading to the identification of more outliers (e.g. satisfaction PI) and sometimes fewer (e.g. control PI).

Analysis at this aggregated level, however, masks the extent of movements into and out of control status. These movements are summarised in Table 16 for all the indicators. Taking as an example the SCRQoL indicator, after IPW/MI adjustment, only one more CASSR is out of control but in fact three move to alarm from alert status, nine move to alert status from in control, and a further five move to alert from alarm status. There are fewer movements for the other indicators and it is important to note that the overwhelming majority are between adjacent categories. Nevertheless, without adjustment for nonresponse a proportion of CASSRs would be falsely identified as either out of control or in control for each indicator, with the SCRQoL indicator having the highest rates of false negatives and positives. For the SCRQoL and control PIs the 'type one error rate' (assuming the adjusted PI is the 'true' score) is over five per cent after IPW/MI adjustment.

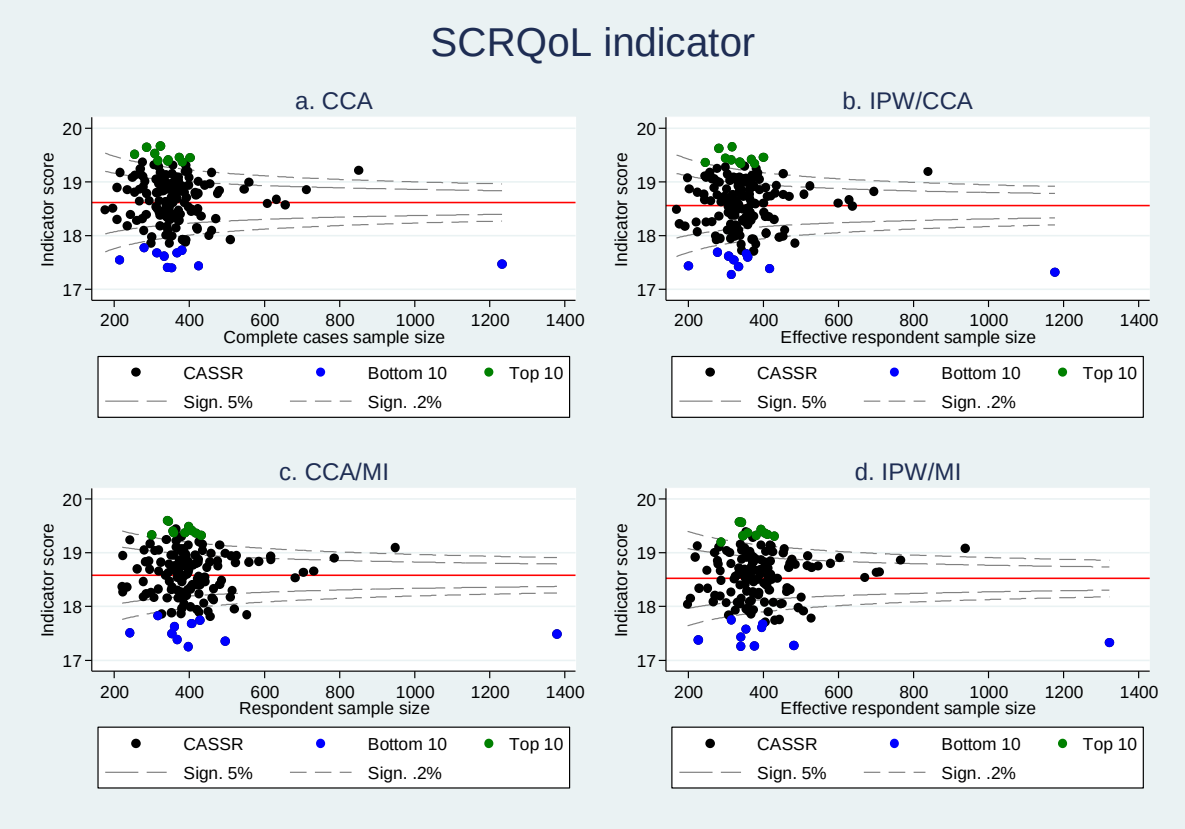

Figure 23: Funnel plots of CASSR scores on the SCRQoL indicator under different nonresponse adjustments

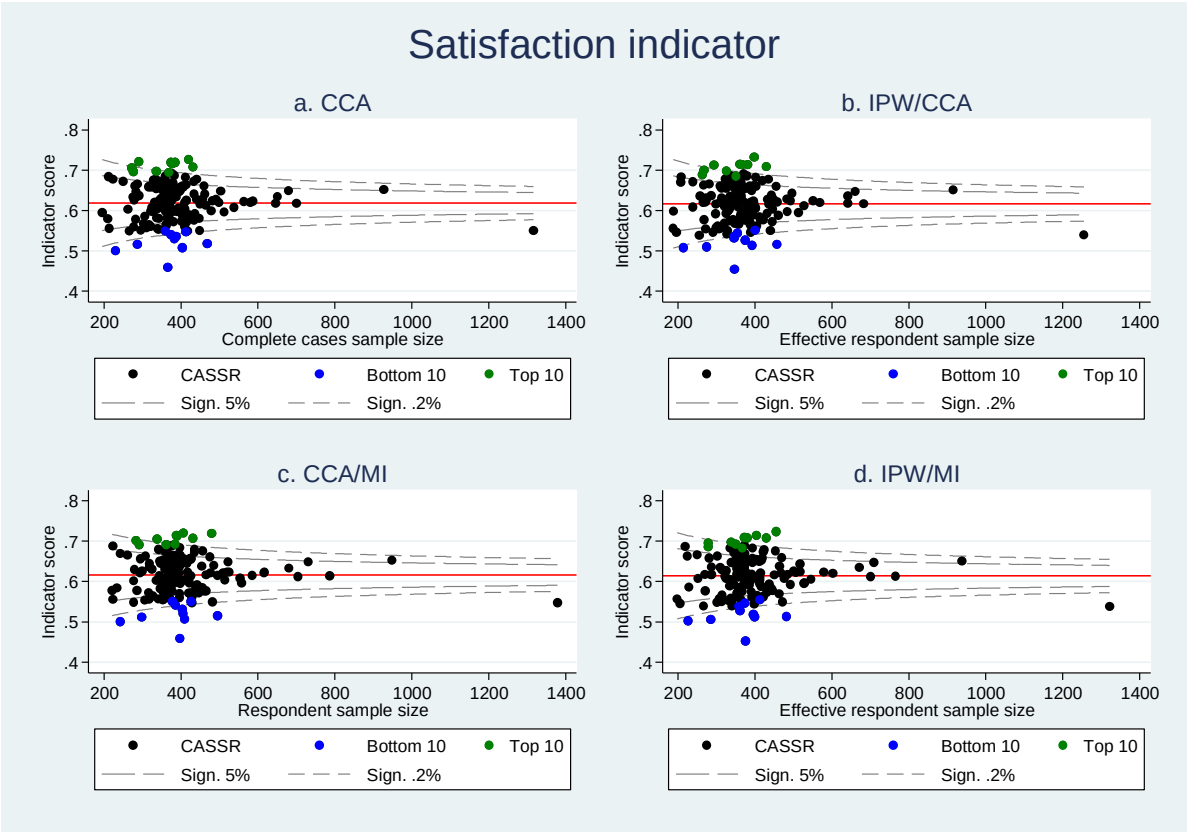

Figure 24: Funnel plots of CASSR scores on the satisfaction indicator under different nonresponse adjustments

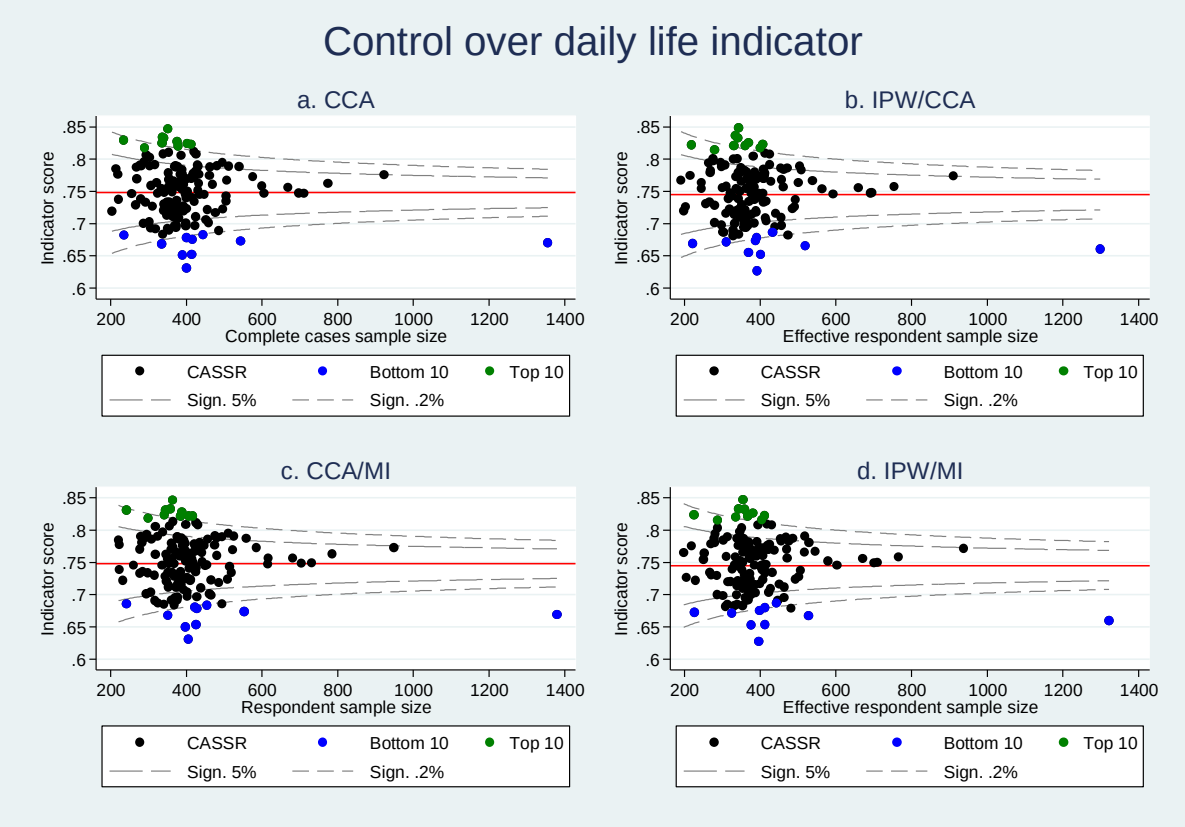

Figure 25: Funnel plots of CASSR scores on the control over daily life indicator under different nonresponse adjustments

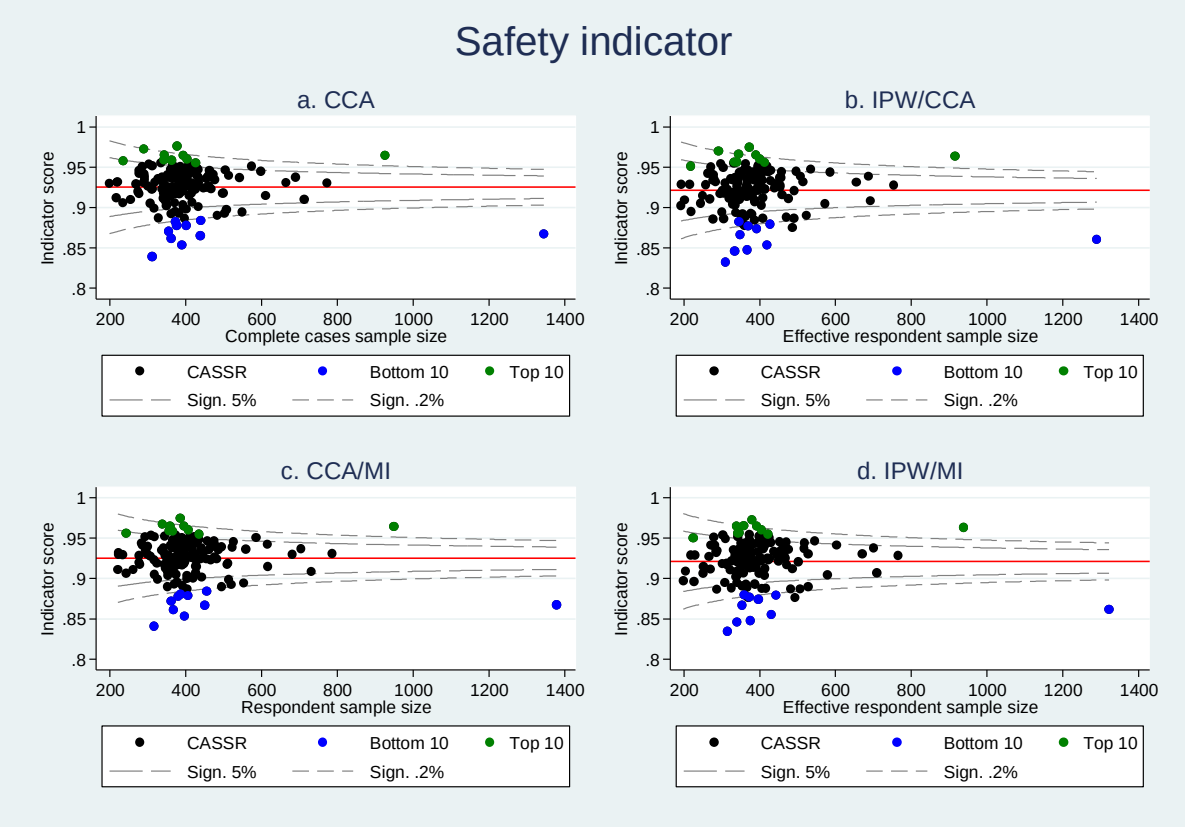

Figure 26: Funnel plots of CASSR scores on the safety indicator under different nonresponse adjustments

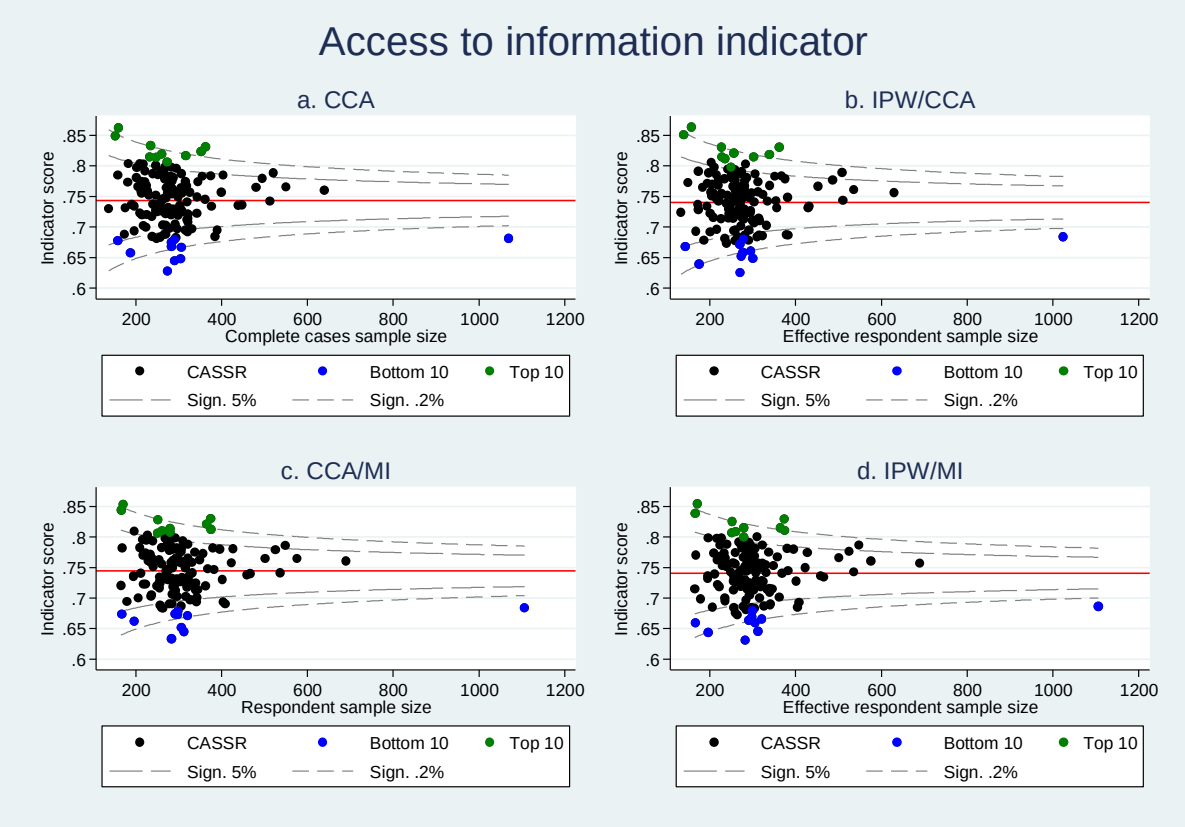

Figure 27: Funnel plots of CASSR scores on the access to information indicator under different nonresponse adjustments

Table 15: Number of CASSRs identified as outliers for each ASCOF indicator, using different adjustments for missingness (n=149)

PI	Missing data adjustment	Outlier status				
		High alarm y>-3 SD	High alert y>2 SD	'Normal' range	Low alert y>-2 SD	Low alarm y>-3 SD
SCRQoL PI	CCA/CCA	15 (10%)	34 (23%)	52 (35%)	30 (20%)	18 (12%)
	IPW/CCA	15 (10%)	35 (23%)	51 (34%)	32 (21%)	16 (11%)
	CCA/MI	15 (10%)	36 (24%)	50 (34%)	29 (19%)	19 (13%)
	IPW/MI	14 (9%)	34 (23%)	51 (34%)	33 (22%)	17 (11%)
Satisfaction PI	CCA/CCA	6 (4%)	23 (15%)	86 (58%)	25 (17%)	9 (6%)
	IPW/CCA	6 (4%)	22 (15%)	87 (58%)	26 (17%)	8 (5%)
	CCA/MI	6 (4%)	25 (17%)	87 (58%)	23 (15%)	8 (5%)
	IPW/MI	6 (4%)	23 (15%)	82 (55%)	30 (20%)	8 (5%)
Control PI	CCA/CCA	8 (5%)	24 (16%)	84 (56%)	24 (16%)	9 (6%)
	IPW/CCA	8 (5%)	24 (16%)	87 (58%)	23 (15%)	7 (5%)
	CCA/MI	8 (5%)	23 (15%)	84 (56%)	24 (16%)	10 (7%)
	IPW/MI	8 (5%)	20 (13%)	91 (61%)	23 (15%)	7 (5%)
Safety PI	CCA/CCA	2 (1%)	15 (10%)	101 (68%)	21 (14%)	10 (7%)
	IPW/CCA	3 (2%)	19 (13%)	92 (62%)	25 (17%)	10 (7%)
	CCA/MI	2 (1%)	15 (10%)	103 (69%)	19 (13%)	10 (7%)
	IPW/MI	4 (3%)	17 (11%)	92 (62%)	25 (17%)	11 (7%)
Information PI	CCA/CCA	4 (3%)	17 (11%)	105 (70%)	19 (13%)	4 (3%)
	IPW/CCA	4 (3%)	14 (9%)	103 (69%)	22 (15%)	6 (4%)
	CCA/MI	3 (2%)	17 (11%)	106 (71%)	19 (13%)	4 (3%)
	IPW/MI	3 (2%)	13 (9%)	106 (71%)	22 (15%)	5 (3%)

Table 16: Movements into and out of control status following adjustment for item and unit nonresponse

Adjustment	IPW/MI					CCA/MI					IPW/CCA				
Indicator	**SCRQoL**	**Satisfaction**	**Control**	**Safety**	**Information**	**SCRQoL**	**Satisfaction**	**Control**	**Safety**	**Information**	**SCRQoL**	**Satisfaction**	**Control**	**Safety**	**Information**
Movements into high 'alarm' status from															
high alert	1	0	0	2	0	1	0	0	0	0	1	0	0	1	0
in control	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Movements into low 'alarm' status from															
low alert	2	0	1	2	1	3	0	1	0	0	1	0	1	1	2
in control	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Movements into high 'alert' status from															
in control	5	1	0	3	0	5	2	0	1	1	3	0	2	5	0
high alarm	2	0	0	0	1	1	0	0	0	1	1	0	0	0	0
Movements into low 'alert' status from															
in control	4	5	1	4	3	1	0	0	0	1	2	3	0	4	3
low alarm	3	1	3	1	0	2	1	0	0	0	3	1	3	1	0
Movements into 'control' from															
low alert	1	0	2	0	0	2	2	0	2	1	0	2	1	0	0
high alert	5	1	4	1	4	3	0	1	1	1	2	1	2	1	3
'False positives'	6	1	6	1	4	5	2	1	3	2	2	3	3	1	3
'False negatives'	9	6	1	7	3	6	2	0	1	2	5	3	2	9	3
'Type I error rate'	7.3	1.0	5.7	0.9	3.5	6.0	2.0	1.0	2.6	1.8	2.4	3.0	2.9	1.0	2.7
'Type II error rate'	15.3	11.5	2.6	18.9	9.1	10.3	4.3	0.0	2.9	5.7	8.2	6.4	4.7	22.0	8.8

Note: only movements that are observed are listed above.

Since there is very little change in the width of the control limits, the status changes and shifts that do occur are due to either a volume or 'mean' outcome effect. Table 17 shows the aggregate changes in the number of outliers after MI and Table 18 shows the number of individual moves into and out of control statuses due to the MI volume and mean outcome effect. In general the effect of the increased number of cases is to increase the number of outliers, and in the case of SCRQoL indicator which has the lowest item response rate, there is a noticeable volume effect. For the satisfaction, control over daily life and safety indicators which have very high item response rates there is very little volume effect. Since the mean outcome effect is negative for the SCRQoL and satisfaction indicators the overall effect of imputation is a tendency to 'pull' high scorers closer to the average, so cancelling out some of the volume effect among those CASSRs in the upper part of the distribution.

The opposite effect is observed for IPW. This is illustrated in Table 19, which shows the aggregate changes in the number of outliers, and Table 20, which shows the number of individual moves into and out of control statuses due to the volume and mean outcome effect. There is a small volume effect for all indicators as a result of the decreased effective sample size. In some cases the mean outcome effect enhances the volume effect (e.g. control PI); in other cases it works in opposition to the volume effect (e.g. safety PI). Funnel plots for the decomposition of the mean outcome and volume effects are shown in Appendix 9.

Table 17: Decomposition of the effect of adjusting for item nonresponse between volume and mean outcomes effects on outlier status, for each ASCOF indicator (n=149)

	Missing data adjustments	Outlier status				
		High score y>3 SD	High score y>2 SD	'Normal' range	Low score y>-2 SD	Low score y>-3 SD
SCRQoL PI	CCA	15 (10%)	34 (23%)	52 (35%)	30 (20%)	18 (12%)
	Volume effect	18 (12%)	37 (25%)	45 (30%)	30 (20%)	19 (13%)
	Mean and volume	15 (10%)	36 (24%)	50 (34%)	29 (19%)	19 (13%)
Satisfaction PI	CCA	6 (4%)	23 (15%)	86 (58%)	25 (17%)	9 (6%)
	Volume effect	6 (4%)	24 (16%)	85 (57%)	25 (17%)	9 (6%)
	Mean and volume	6 (4%)	25 (17%)	87 (58%)	23 (15%)	8 (5%)
Control PI	CCA	8 (5%)	24 (16%)	84 (56%)	24 (16%)	9 (6%)
	Volume effect	8 (5%)	25 (17%)	83 (56%)	24 (16%)	9 (6%)
	Mean and volume	8 (5%)	23 (15%)	84 (56%)	24 (16%)	10 (7%)
Safety PI	CCA	2 (1%)	15 (10%)	101 (68%)	21 (14%)	10 (7%)
	Volume effect	3 (2%)	16 (11%)	99 (66%)	21 (14%)	10 (7%)
	Mean and volume	2 (1%)	15 (10%)	103 (69%)	19 (13%)	10 (7%)
Information PI	CCA	4 (3%)	17 (11%)	105 (70%)	19 (13%)	4 (3%)
	Volume effect	6 (4%)	17 (11%)	100 (67%)	21 (14%)	5 (3%)
	Mean and volume	3 (2%)	17 (11%)	106 (71%)	19 (13%)	4 (3%)

Table 18: Summary of movements into and out of 'out of control' status following MI, showing volume and mean outcome effects

Missing data adjustment	MI, volume effect					MI, mean outcome and volume effect				
Indicator	**SCRQoL**	**Satisfaction**	**Control**	**Safety**	**Information**	**SCRQoL**	**Satisfaction**	**Control**	**Safety**	**Information**
Movements into high 'alarm' status from										
high alert	3	0	0	1	2	1	0	0	0	0
in control	0	0	0	0	0	0	0	0	0	0
Movements into low 'alarm' status from										
low alert	1	0	0	0	1	3	0	1	0	0
in control	0	0	0	0	0	0	0	0	0	0
Movements into high 'alert' status from										
in control	3	1	1	1	0	5	2	0	1	1
high alarm	0	0	0	0	0	1	0	0	0	1
Movements into low 'alert' status from										
in control	0	0	0	0	2	1	0	0	0	1
low alarm	0	0	0	0	0	2	1	0	0	0
Movements into 'control' from										
low alert	0	0	0	0	0	2	2	0	2	1
high alert	0	0	0	0	0	3	0	1	1	1

Table 19: Decomposition of the effect of adjusting for unit nonresponse between volume and mean outcomes effects on outlier status (n=149)

	Missing data adjustments	Outlier status				
		High score y>3 SD	High score y>2 SD	'Normal' range	Low score y>-2 SD	Low score y>-3 SD
SCRQoL PI	CCA	15 (10%)	34 (23%)	52 (35%)	30 (20%)	18 (12%)
	Volume effect	14 (9%)	34 (23%)	53 (36%)	30 (20%)	18 (12%)
	Mean and volume	15 (10%)	35 (23%)	51 (34%)	32 (21%)	16 (11%)
Satisfaction PI	CCA	6 (4%)	23 (15%)	86 (58%)	25 (17%)	9 (6%)
	Volume effect	6 (4%)	21 (14%)	90 (60%)	24 (16%)	8 (5%)
	Mean and volume	6 (4%)	22 (15%)	87 (58%)	26 (17%)	8 (5%)
Control PI	CCA	8 (5%)	24 (16%)	84 (56%)	24 (16%)	9 (6%)
	Volume effect	8 (5%)	23 (15%)	85 (57%)	24 (16%)	9 (6%)
	Mean and volume	8 (5%)	24 (16%)	87 (58%)	23 (15%)	7 (5%)
Safety PI	CCA	2 (1%)	15 (10%)	101 (68%)	21 (14%)	10 (7%)
	Volume effect	2 (1%)	14 (9%)	104 (70%)	20 (13%)	9 (6%)
	Mean and volume	3 (2%)	19 (13%)	92 (62%)	25 (17%)	10 (7%)
Information PI	CCA	4 (3%)	17 (11%)	105 (70%)	19 (13%)	4 (3%)
	Volume effect	4 (3%)	16 (11%)	107 (72%)	18 (12%)	4 (3%)
	Mean and volume	4 (3%)	14 (9%)	103 (69%)	22 (15%)	6 (4%)

Table 20: Summary of movements into and out of 'out of control' status following IPW, showing volume and mean outcome effects

Missing data adjustment	IPW, volume effect					IPW, mean outcome and volume effect				
Indicator	**SCRQoL**	**Satisfaction**	**Control**	**Safety**	**Information**	**SCRQoL**	**Satisfaction**	**Control**	**Safety**	**Information**
Movements into high 'alarm' status from										
high alert	0	0	0	0	0	1	0	0	1	0
in control	0	0	0	0	0	0	0	0	0	0
Movements into low 'alarm' status from										
low alert	0	0	0	0	0	1	0	1	1	2
in control	0	0	0	0	0	0	0	0	0	0
Movements into high 'alert' status from										
in control	0	0	0	0	0	3	0	2	5	0
high alarm	1	0	0	0	0	1	0	0	0	0
Movements into low 'alert' status from										
in control	0	0	0	0	0	2	3	0	4	3
low alarm	0	1	0	1	0	3	1	3	1	0
Movements into 'control' from										
low alert	0	1	0	1	1	0	2	1	0	0
high alert	0	2	1	1	1	2	1	2	1	3

Discussion

The primary goal of this analysis was to assess whether nonresponse had an effect on inferences about performance. I have argued that this question is important because the high levels of nonresponse to the survey lead to negative perceptions about the validity of the results. In this analysis I used the ASCOF indicators derived from the ASCS, to explore the impact of nonresponse on inferences about performance. The approach I used compared PI estimates obtained under the assumption that both item and unit nonresponse is MCAR (CCA/CCA), to those obtained under several more appropriate assumptions about the missingness mechanism for item and unit nonresponse, i.e. that

- unit nonresponse is MAR and item nonresponse is MCAR (IPW/CCA),
- unit nonresponse is MCAR and item nonresponse is MAR (CCA/MI),
- unit and item nonresponse are both MAR (IPW/MI).

I examined how adjusting for nonresponse using these three approaches influences performance assessment, through its effect on the ranking of CASSRs and the identification of outlying CASSRs. One aim of comparing these methods was to allow me to assess whether the method used to address differences in samples due to nonresponse matters for performance assessment.

This analysis has shown that there are small differences in the value and precision of ASCOF indicator scores after adjustment for nonresponse. While adjustment has little impact on conclusions about the rank position of CASSRs and outlier status for most indicators, there are some exceptions. Using the IPW/MI method, nonresponse adjustment has more impact for the SCRQoL indicator and is in general precision-enhancing. In what follows I discuss the findings in more detail and consider how they might inform the decision about whether or not to adjust the ASCOF PIs for nonresponse. I consider the implications for policy and research of this analysis in Chapter 9.

The effect of nonresponse on ASCOF indicators

In general I found that adjusting for nonresponse leads on average to decreases in indicator scores, implying that it is more often than not people with worse quality of life, and those who are less satisfied who choose not to respond to the survey and individual questions. It is not straightforward, however, to judge whether the extent of bias due to nonresponse is large enough to warrant ASCOF indicators being adjusted for nonresponse. I have

examined the absolute degree of nonresponse bias, effects on precision and the effect on the ranking and outlier status of CASSRs to inform decisions concerning this question.

When examining the extent of nonresponse bias the main problem is determining what a meaningful difference in indicator scores is and considering that alongside assessments of statistical significance. Here I have used two arbitrary criteria for both of these concepts: for the former, I equated a meaningful difference with a change exceeding one percentage point of a given indicator scale; and for the latter, I considered a significant change as one that means the new estimate falls outside of the 95 per cent confidence interval for the raw indicator. The important point to note is that using these criteria, I reach different conclusions about the number of CASSRs affected by nonresponse, with the meaningful difference criterion producing more affected CASSRs than the statistical significance criterion. Furthermore, when I consider the effect of adjustment on the national-level estimates, differences of less than half a percentage point of the given scale are found to be statistically significant due to the greater degrees of freedom and therefore precision of the estimates.

On purely statistical grounds these results would imply that it would be important to adjust for nonresponse for producing national-level, but not local-level, statistics. This interpretation, however, ignores the fact that there are fewer degrees of freedom at the local level, so the statistical criterion used may be considered too restrictive. While I think it is sensible to use 95 per cent confidence intervals because of their use in national statistics, the charts in Figure 12 show how decisions about the number of statistically significant differences would change using different statistical cut-off points. Decisions about when to apply adjustment for nonresponse could instead be guided by judgements about whether differences between adjusted and unadjusted PI scores exceed meaningful differences in ASCOF indicator scores. The value of one per cent that I have used may be too permissive and larger differences may be more meaningful from a policy point of view. Indeed the margin of error required for survey estimates is five per cent (NHS Information Centre 2010). There is, however, no evidence to guide decisions using this criterion of 'meaningful difference' and without evidence it is difficult to judge what is meaningful. This is an issue that applies particularly to performance surveys where effects are averaged across very different user groups and service types – a point noted by other researchers (Valderas et al. 2012, Fernandez et al. 2013a). From a policy perspective, it is perhaps more useful to look at whether adjustment for nonresponse changes inferences about performance.

One way in which adjustment for nonresponse would change inferences about performance is through its effect on the precision of estimates. Indeed, nonresponse adjustment was rejected for PIs from the GP patient survey precisely on the basis that it would needlessly reduce precision (Roland et al. 2009). It is therefore important that the IPW/MI approach adjusts for nonresponse bias without substantial loss of precision and in some cases is precision-enhancing. As Seaman et al. (2012) found, the IPW/MI approach leads to gains in precision compared to the IPW/CCA approach through the introduction of MI to address item nonresponse. The analysis presented here suggests that the gains from the IPW/MI approach in terms of precision are related to the extent of item nonresponse. More than half the CASSRs have more precise SCRQoL estimates after IPW/MI adjustment, but only 80 to 90 per cent have more precise estimates for the other ASCOF indicators. The difference in item nonresponse between these indicators is appreciable, with overall item response rates of slightly under 90 per cent for the SCRQoL indicator and 95 per cent or over for the other indicators. The gains in precision from adjustment for the SCRQoL indicator combined with the correction of bias provide an argument for considering IPW/MI adjustment.

Another way to consider whether nonresponse adjustment changes inferences about performance is to consider whether it affects the ranking of CASSRs and which CASSRs are identified as outliers. A small proportion of CASSRs – in general less than ten per cent – have changes to their rank position and outlier status post adjustment. In the majority of cases, changes in rank involve moving only a few places up or down and changes in outlier status involve moving between adjacent categories. There are, however, a handful of CASSRs that have large changes in rank. The majority of movements are between in control and out of control status, meaning that not adjusting for nonresponse would lead to the wrong conclusion about outlier status for between five and ten per cent of CASSRs depending upon the indicator. Most often the errors are ‘false negatives’, meaning that CASSRs are erroneously identified as ‘in control’ when in fact they are ‘out of control’. For most CASSRs, the ‘false negative’ and ‘false positive’ rates are probably within acceptable bounds, but the ‘false positive rate’ for the SCRQoL (and possibly control) indicator seems too high at over five per cent. Altogether this creates an argument for adjusting the SCRQoL indicator for nonresponse.

Limitations of the analysis and directions for future research

The validity of the conclusions depends on the extent to which the assumptions regarding the missingness mechanism are valid, the extent to which the imputation and response propensity models are correctly specified, and the extent to which the procedures are appropriately applied. I address each of these considerations in turn.

The MCAR assumption is not tenable given the evidence of a relationship between the observed characteristics of individuals and unit nonresponse (see Chapter 4), and the relationship between the characteristics of individuals and item nonresponse. On the basis of these relationships, the MAR assumption seems more plausible. For the MI process the plausibility of the MAR assumption is strengthened by the fact that I included a large number of covariates in the imputation model. However, the true missing data mechanism is unknown, and it is possible that unobserved factors also influence item nonresponse. An extension to the analysis presented here therefore would be to test whether the findings are robust to alternative assumptions about reasons for missingness, e.g. through sensitivity analysis or explicit modelling of the NMAR mechanism (Carpenter et al. 2007, Kirkham 2008, Carpenter and Kenward 2013, Gomes et al. 2014). Given the low levels of item nonresponse observed in this dataset, such detailed testing of assumptions seemed unnecessary in this instance.

Concerning the specification of the models, assuming the data are MAR then the imputation model for the outcome variables seems fairly robust: I included a large number of covariates and, leaving aside the issue of the weights, it was straightforward to develop a congenial imputation and analysis model given the interest in only the PI estimate. It seems less plausible that the response propensity model is correctly specified as there are several factors that seemed theoretically important to include, but for which I was unable to find good measures (see discussion in Chapter 4). These include availability of informal care, disability type and severity, and proficiency in English. Although I have some indicators of English proficiency, disability type and severity (ethnicity, client group and number of service and type of services received) the measures are not very sensitive and are certainly not optimal. There is good evidence that these factors are important predictors of quality of life and satisfaction with care (see discussion in Chapter 2), implying that they are important omissions from the response propensity model (Little and Vartivarian 2003). It may therefore be appropriate to stratify the weights by forming adjustment cells, since stratification places less emphasis on the correct specification of the regression model (Little 1986). Seaman and White (2013) also suggest a number of

strategies for developing a response propensity model for IPW, which I did not follow in this application as it was not clear how they could be extended from the binomial to the MNL situation. An area for future research would be to find ways of extending the checks suggestions by Seaman and White (2013) to the MNL situation.

The most significant limitation of this analysis is the appropriateness of the strategy I used to address the problem of missingness in the auxiliary data. I used MI to recover the missing auxiliary data. Such an approach has been used in several studies where propensity models have been used to balance the characteristics of treatment and control groups (Song et al. 2001, Mattei 2009, Qu and Lipkovich 2009, Mitra and Reiter 2011), but not in the context of IPW for nonresponse except in a very recent paper by Seaman and White (2014). The approach I took produced two overlapping imputed datasets that could not easily be combined. One contained the imputed auxiliary data and was used to estimate the weights and the second contained the respondent dataset with the outcome indicators. There are two consequences of this strategy. First, it meant that the missing auxiliary covariates were imputed on the basis that they were not related to the outcome indicator, which could introduce a degree of bias into the PI estimates (Petticrew and Roberts 2003). Secondly, it meant that the variability introduced into the weights from the MI procedure was not reflected in the analysis.

On examining the question of how to address missingness in the auxiliary data in the context of IPW, Seaman and White (2014) suggest an approach that produces two non-overlapping imputed datasets that can be combined. Their method involves imputing the respondent sample separately from the nonrespondent sample, and then combining these two datasets. With such data, it is possible to develop a different set of weights for each imputed dataset and therefore reflect the additional variability due to MI. It is likely that the way I have implemented MI to recover the missing auxiliary data has introduced some bias into the results. It is also clear that a consequence of my method is that I underestimate the uncertainty around results for CASSRs. Both of these factors will affect the conclusions about the relative efficiency of the IPW/CCA and IPW/MI approach compared to CCA/MI, with both approaches likely to be more precision-reducing than I find here. The effect on the estimation of bias due to nonresponse is unclear without re-estimation of the PIs using the Seaman and White (2014) approach.

Concluding remarks

This analysis finds some effect of nonresponse on ASCOF indicators. The effects are in general fairly modest and for most indicators, there does not seem to be a strong case for making adjustments for nonresponse. There is, however, a stronger case for adjusting the SCRQoL indicator for nonresponse. Notably the case is only strong where the effects of nonresponse are examined through the lens of the IPW/MI approach and to a lesser extent the CCA/MI approach. Where unit and item nonresponse are corrected for using IPW/MI, there appears to be more bias in the SCRQoL indicator from nonresponse, adjustment for nonresponse is on average precision-enhancing and the false negative rate is over five per cent. The limitations of the modelling mean there is some doubt over the effect of IPW/MI and IPW/CCA on performance assessment. Further research using the methods suggested by Seaman and White (2014) for implementing IPW in the context of missing auxiliary data is recommended.

Importantly, and with implications for other studies as I discuss in Chapter 9, the findings from this analysis suggest that the method chosen to adjust for nonresponse is likely to influence conclusions about the extent to which nonresponse affects inferences about performance. The higher level of item nonresponse to the SCRQoL indicator compared to the other ASCOF indicators also suggests adjustment may be more important where item nonresponse is higher. This may have implications for case-mix adjusted PIs where rates of item nonresponse will be higher due to additional loss of data from the case-mix variables. At the time of conducting this analysis the ASCOF PIs were not adjusted for case-mix, so I could not examine this question. In Chapters 6 and 7, however, I consider a range of methods for adjusting the ASCOF PIs for case-mix, focusing on the SCRQoL indicator. In Chapter 8, I look at the effect of adjustment for case-mix on performance assessment and alongside this I consider the effect that item nonresponse has on performance assessment.

Chapter 6

What is the Most Appropriate Method for Modelling Adult Social Care Outcomes? Part One: An Investigation into the Effect of the Choice of Model on the Estimation of Outcomes Using Risk-Adjustment

Abstract

Objective: To develop a risk-adjustment model for the self-reported social care outcome indicator and compare the effect on model fit of modelling decisions around the choice of regression model and risk adjustors.

Design: Cross-sectional survey data that captured self-reported SCRQoL and a variety of risk-adjustors were analysed using (i) OLS regression, (ii) fractional response models, and (iii) random and fixed-effects regression. Two different specifications were estimated: one with a set of theoretically-chosen but statistically significant risk-adjustors, and a second more parsimonious statistically-driven specification. Separate models were estimated for different sub-groups of the population. Model fit statistics were compared to assess the effect of modelling decisions.

Participants: This analysis uses only respondents from the physically and sensorily disabled (PSD) client group. This includes 29,786 people aged 65 and over living in private households, 7,447 people aged between 18 and 64 living in private households, and 7,580 adults aged 18 and over resident in care homes.

Outcome measures: The SCRQoL indicator, based on the ASCOT-SCT4 measure.

Results: The amount of variation explained by the risk-adjustors varied across the sub-groups and by specification, with R^2 ranging from 0.30 to 0.43 for the theoretically-driven specification and from 0.14 to 0.42 for the statistically driven specification. For the two private household sub-groups the specification was less important than for the care home sub-group where the statistically-driven method produced a poor model. The choice of regression model made little difference to the model fit.

Conclusions: Risk-adjustors explain a sizeable amount of the variation in the SCRQoL indicator. The choice of regression model was not important but the choice of risk-adjustors did influence the model fit.

Introduction

For survey-based PIs to be valid performance measures, differences in scores between organisations should reflect differences in performance. The design of performance surveys, however, means that differences in performance are only one of many explanations for differences in PI scores. One approach to improve the interpretability of PIs and account for the confounding influence of other factors is risk-adjustment (Iezzoni 2013). In risk adjustment the aim is to estimate empirically the expected outcome for individuals, on the basis of a set of 'risk factors' that predict outcomes but are unrelated to performance (Ash et al. 2013). The relationship between the expected and observed outcome for an individual can then be used as the basis for an adjusted PI. This approach has been used extensively to adjust indicators in a number of performance datasets (Iezzoni 2013), and has been used to develop adjusted indicators for long-term care (Berlowitz and Intrator 2013). Although, the technique does not adequately address the bias in estimates caused by the correlation between service intensity, outcomes and characteristics of users (see Chapter 3), it is straightforward to implement with the ASCS data.

A concern when developing adjusted PIs is that the resultant indicators are intelligible and for this to be the case the methods used need to be transparent and ideally as simple as possible (Mukamel 1997, Iezzoni 2013). There is, however, always a trade-off between simplicity and validity. More simple models tend to make strong assumptions that may not be tenable; whereas models that relax these assumptions may be more valid but are generally more complex and harder to explain to non-specialists in statistical analysis. The validity of risk-adjustment models is a key focus of the literature, where the concern has been to establish whether the choice of risk-adjustors and the assumptions inherent in the choice of statistical model are valid for the data. The research base suggests that modelling decisions can have implications for the validity of adjusted indicators and inferences regarding the relative performance of organisations, but the importance of different choices seems to vary depending on the study (Greenfield et al. 2002, Huang et al. 2005, Glance et al. 2006, Arling et al. 2007, Mukamel et al. 2008, Li et al. 2009, Eijkenaar and van Vliet 2014). In light of this, Li et al. (2009) argue that the importance of modelling choices is an empirical question, which is likely to be affected by the particular characteristics of the data. Before settling on a more complicated model, there is therefore value in exploring whether the improvements in face validity translate to (i) improvements in the validity of the model understood in terms of model fit, and (ii) differences in PI scores and inferences about performance. Where there is

no obvious gain from using a more complicated model and it cannot be shown to have an effect on inferences about performance, the simpler model may be preferable.

In this chapter, I am concerned with the first of these aspects, i.e. whether statistical models that have greater face validity substantially improve the fit of the model. I explore the second of these questions in Chapter 8. As my focus is the ASCOF SCRQoL indicator, which is a multi-item scale, the base model is estimated by OLS. I compare the OLS to:

- a quasi-likelihood estimator for a Beta distribution based mean-variance model that is similar to maximizing a Bernoulli likelihood function. This model allows for negatively-skewed distributions and restricts predictions to the correct range (Basu and Manca 2012); and
- a random-effects (RE) and fixed-effects (FE) regression model. These both relax the assumption of the independence of observations, which is undermined by the hierarchical organisation of the data, with individuals clustered within CASSRs (Goldstein and Spiegelhalter 1996, Normand et al. 1997, Austin et al. 2003, Clarke et al. 2015).

The choice of risk-adjustors for the modelling is guided by the modified POW framework (Davies and Knapp 1981, Knapp 1984) set out in Chapter 2, where the risk-adjustors are referred to as quasi-inputs. I use two different approaches to retain theoretically-relevant variables in the models: one based on theoretical value and statistical significance, and a more parsimonious but statistically-driven approach based on the extent to which the variables explain variation in outcomes and vary across CASSRs (Zaslavsky et al. 2001).

In the rest of this chapter, I first describe the data and the statistical models used, focusing on the specification of the models and other aspects of the empirical approach, including the comparison of models and how I address missing data. I then present the results, providing details of the sample and comparing the models results and predictions. In the discussion I consider whether the modelling choices examined here are important for understanding heterogeneity in ASC outcomes. The implications for policy and future research are discussed in Chapter 9.

Data and statistical analysis

The dataset for this analysis is the 2010-11 ASCS. In this analysis I make use of the questionnaire data as well as the auxiliary data from the records systems of CASSRs. The ASCS sample is highly diverse, comprising many sub-groups of ASC users. The groups of

users vary in a number of important ways that means it is not possible in some circumstances, and not sensible in others, to analyse the dataset as a whole, due to differences in the relevance of variables as the risk adjustors for the various sub-groups. For example, the suitability of the design of care home environment for caring can reasonably be considered an aspect of quality, but in private households these aspects are to a large extent beyond the immediate control of the service. CASSRs may provide small adaptations, and over longer timescales they can provide Disabled Facilities Grants or alternative (perhaps supported) accommodation, but these options may not always be possible and some service users may consider them undesirable. A consequence is that while it is important to control for design of housing for people in private households it is inappropriate to do so for the care home population. Similar types of arguments can be made for other variables (see Appendix 10 for a discussion).

The approach I take here is to split the dataset into subsets defined according to the categories of primary client group (physically and sensorily disabled/vulnerable person, mental health problem, substance misuse problem or learning disability), age (65 and over and under 65), and care setting (residential versus non-residential). This provides groups in which the relationship between risk-adjustors and outcomes is similar. Splitting the sample in this way, however, produces sixteen sub-groups many of which are very small (see Table 21). To make the number of comparisons manageable, I restrict this analysis to the physically and sensorily disabled (PSD) client group and combine the over and under 65s in the care home setting. Since I look at three sub-groups it means I can assess whether different modelling assumptions are more or less appropriate for different segments of the ASC population. Together these sub-groups cover over 70 per cent of the respondent sample.

Table 21: Numbers within each sub-group in ASCS 2010-11 respondent sample

Setting	Primary client group	Age group: Over 65	Age group: 18 to 64	Total
Private household	Physical and sensory disability (PSD)†	29,786	7,447	37,233
	Mental health (MH)	1,924	3,010	4,934
	Learning disability (LD)	423	5,454	5,877
	Substance misuse (SM)	18	82	100
Care home	Physical and sensory disability (PSD)†	6,843	657	7,500
	Mental health (MH)	1,246	557	1,803
	Learning disability (LD)	406	2,237	2,643
	Substance misuse (SM)	15	21	36

Legend: † PSD group also include a small number of 'vulnerable people'

Statistical modelling

Risk-adjustment models are based on a reduced form of the conceptual model presented in equation (2) (see Chapter 2), which includes only the risk factors to predict outcomes, i.e.

$$y_i = \alpha + \boldsymbol{\beta_1 z_i^0} + \boldsymbol{\beta_2 z_i^1} + \boldsymbol{\beta_3 z_i^2} + \boldsymbol{\beta_4 z_i^3} + \boldsymbol{\beta_5 z_i^4} + e_i, \quad (15)$$

where y_i is the outcome measure for individual, i, α is a constant term, $\boldsymbol{z_i^0}$ is a vector of variables capturing personal characteristics and motivations, $\boldsymbol{z_i^1}$ is a vector of variables representing underlying health and disabling conditions, $\boldsymbol{z_i^2}$ is a vector of variables representing the immediate physical environment, $\boldsymbol{z_i^3}$ is a vector of variables capturing additional resources arising from social and economic capital, $\boldsymbol{z_i^4}$ is a vector of variables capturing reporting-related factors related to the postal method of survey administration, the $\boldsymbol{\beta}$s, represent the direct effect of these factors on the observed outcome, and e_i is the independently-distributed error term. The assumption of risk-adjustment models is that by accounting of the confounding effects of the case-mix variables, $\boldsymbol{z}$, the remaining residual outcome differences, e_i, are related to service effectiveness.

I use a statistical analysis to estimate the reduced form equation (15). The dependent variable in this analysis is the SCRQoL indicator and the base model is estimated by OLS regression. The negatively-skewed and bounded nature of the SCRQoL indicator, however is potentially a problem for OLS estimation. OLS can be sensitive to the extreme outliers in the tails of the distribution and can yield out of range predictions, both of which are problematic in the context of risk-adjustment as they could yield inconsistent estimates of covariate

effects (Pullenayegum et al. 2010, Basu and Manca 2012). For this reason I also estimate the equation (15), using a model based on the Beta distribution, which accommodates negatively skewed variables and restricts predictions to the correct range.[44] As Basu and Manca (2012) explain, modelling the full distributional assumptions of the Beta distribution is restrictive as the likelihood function of Beta cannot support values of zero or one for the outcome variable – a problem for these data given the proportion of cases who have a maximum SCRQoL score. The quasi-likelihood approach is able to overcome this limitation. Since the quasi-score equation generated using the mean-variance relationship of the Beta distribution is identical to maximizing a Bernoulli likelihood function with an additional overdispersion parameter. Such models are often referred to as fractional response (FR) models in the literature (Papke and Wooldridge 1996). I implement the FR models in Stata using the glm routine with a logit link function and a binomial distribution (Papke and Wooldridge 1996, Baum 2008, Basu and Manca 2012).

For OLS estimates to be unbiased, observations are assumed to be independent, but this is unlikely to be the case where indicators are used for performance assessment for the reasons discussed in Chapter 3. I consider two alternative approaches to appropriately reflect the clustering of observations within authorities, FE and RE models, which have been used extensively for risk-adjustment (Goldstein and Spiegelhalter 1996, Normand et al. 1997, Austin et al. 2003, Huang et al. 2005, Li et al. 2009, Clarke et al. 2015). The statistical model for the RE and FE estimation is similar to that in (15), but with an additional parameter, u_j, which captures the CASSR-level clustering, as follows,

$$y_{ij} = \alpha + \boldsymbol{\beta_2 z_{ij}^0} + \boldsymbol{\beta_3 z_{ij}^1} + \boldsymbol{\beta_3 z_{ij}^2} + \boldsymbol{\beta_4 z_{ij}^3} + \boldsymbol{\beta_5 z_{ij}^4} + e_{ij} + u_j, \qquad (16)$$

where the additional subscript j denotes the CASSR. In the RE regression, u_j are random effects that are assumed to follow a normal distribution, i.e. $u_j \sim N(0, \sigma_u^2)$. Although this makes the model more flexible it requires an additional assumption that the errors, u_j, are also assumed to be uncorrelated with the risk factors $\boldsymbol{z_{ij}^0, z_{ij}^1, z_{ij}^2, z_{ij}^3, z_{ij}^4}$ (DeLong et al. 1997, Clarke et al. 2015). By contrast the FE model makes no assumptions about the distribution of

[44] There is a growing field of study exploring models based on alternative distributions and assumptions to better represent the bounded and skewed distribution of multi-attribute QoL measures. Researchers have investigated Tobit models (Austin et al. 2000, Austin 2002b, Austin 2002a), censored least absolute deviations (CLAD) models (Austin 2002b, Huang et al. 2008), two-part models (TPMs) (Austin 2002b, Huang et al. 2008), latent class models (LCMs) (Huang et al. 2008), and models based on a beta distribution (Basu and Manca 2012), including 'FR models' (Papke and Wooldridge 1996).

u_j and is able to deal with situations where the CASSR effectiveness estimate, u_j, correlates with the risk factors (Greene 2012). All models are estimated in Stata 14.

Selection of risk adjustors

The models above specify five groups of quasi-inputs. The ASCS dataset contains a number of variables that can be used as proxies for these factors, as follows:

- for personal characteristics and motivations, the only available variables are sex (MALE) and age group (AGE_GP);
- for underlying health and disabling conditions, available variables are self-perceived health (SPH), the EQ-5D anxiety/depression domain (ANXDEP), the EQ-5D pain domain (PAIN), activities of daily living (ADLs) and instrumental activities of daily living (IADLs), vulnerable person client group (VP);
- for additional resources arising from social and economic capital, there is receipt of regular practical help (PH) and whether the respondent purchased additional private help (PRIVATE)[45];
- for environmental characteristics, variables include self-perceived design of home (SPHD); and
- for reporting-related factors there are the following variables: source and type of help given to complete questionnaire (ASSIST and PROXY).

These variables are described more fully in Table 22.

[45] This variable can also be considered a reporting-related factor as it may affect the individual's perceptions of the quality or value of formal provision.

Table 22: Description of all the theoretically-relevant risk adjustor variables

Variable description (label)	Response categories (label)	Rationale
Personal characteristics		
Age (AGE_GP)	18-24 (1), 25-34 (2), 35-44 (3), 45-54 (4), 55-64 (5), 65-74 (6), 75-84 (7), 85 and over (8)	Advanced age is associated with disability, but age is also included due to its descriptive importance and potential relationship with motivations.
Sex (MALE)	male (1), female (0)	Sex is included due to its descriptive importance and potential relationship with motivations.
Vulnerable person (VP)	Physical disability (0), vulnerable person (1)	This is included to test whether there are differences between the client groups in outcomes.
Underlying health and disabling conditions		
Count of the ADLs[a] an individual has at least difficulty completing (ADLDIFF) or can't complete (ADLCANT)	Scale from 0 through 7, where, for ADLDIFF, 7 means the person has difficulties with (or can't do) all the ADLs, and for ADLCANT, 7 means the person can't do any of the ADLs without help	This variable measures the functional ability of the individual. The greater the number of ADLs the person fails or has difficulty completing the more severely the person is disabled.
IADL Management of finances and paperwork (FINANCES)	can do easily by myself (0), have difficulty doing this by myself (1), can't do this by myself (2)	This IADL is included separately in the model, as a predictor of early stages of dementia (Barberger-Gateau et al. 1993, De Lepeleire et al. 2004, Sikkes et al. 2011).
EQ-5D anxiety/depression item (ANXDEP)	None (0), moderate (1), extreme (2)	This is included as an indicator of underlying mental health state.
EQ-5D pain/discomfort item (PAIN)	None (0), moderate (1), extreme (2)	This is included as an indicator of the underlying health state of the person.
Self-perceived health (SPH)	Very good (0), good (1), fair (2), bad(3), very bad (4)	This is included as an indicator of the underlying health state of the person.
Resources		
Receipt of regular practical help (PH)	From person… inside home (PH, in home), outside home (PH, out home), no help (PH, none) [d]	This is included as an indicator of receipt of informal care.
Buy additional care privately, or pay more to top-up care (PRIV)	Yes, user's own money (PRIV, own), yes, family member money (PRIV, fam), none (PRIV, none) [d]	This captures whether there is an additional privately-purchased care input.

Variable description (label)	Response categories (label)	Rationale
Immediate environment		
Self-perceived design of home (SPHD) [b]	Meets needs very well (0), meets most needs (1), meets some needs (2), totally inappropriate for need (3)	This is an indicator of the extent to which the immediate home environment is more or less disabling, either compounding or ameliorating the person's difficulties related to their functional disability.
Reporting-related variables		
Whether help was given to complete the questionnaire	Completed by self (SELFCOMP)[d], proxy assistance (PROXYASSTD), proxy (PROXY)	This captures potential response bias due to the person having help to complete the questionnaire or the questionnaire being completed by a proxy respondent (Elliott et al. 2008).
Source of help to complete questionnaire [c]	From… care worker (ASSIST, cw), someone inside the home (ASSIST, in home), someone outside the home (ASSIST, out home)	This captures potentially different degrees and directions of response bias due to differences in who helps the person to complete the questionnaire (Elliott et al. 2008).
Type of help given to complete the questionnaire [d]	Help to… read (ASSIST, read), translate (ASSIST, translate), write (ASSIST, write), talked through answers (ASSIST, discuss)	This captures potential differences in response bias due to the type of help given to the person to complete the questionnaire.

Notes: [a] Composed of Katz (1963) ADLs (ability to get around indoors, get in/out of bed/chair, feed self, use WC/toilet, wash face and hands, wash all over using bath/shower, get dressed/undressed), except the continence ADL as it is not possible to ask about continence in the context of a self-completion questionnaire for ethical reasons. Response scale for all: can do easily by myself, have difficulty doing this by myself, can't do this by myself. [b] Not included in care home models as too closely associated with structural quality. [c] These variables cannot be included in the model if PROXYASSTD is included. All options can be included, as base category is SELFCOMP. [d] The base category for nominal variables. For ordinal and binary variables the base category is indicated by a zero.

One approach to risk-adjustment would be to include all theoretically-relevant variables outlined in Table 22, regardless of their empirical relationship with outcomes (Murtaugh et al. 2007). Although this approach would maximise the variation explained, it would also result in models with many covariates that are difficult to interpret and are therefore unappealing for policymakers. Instead I use a theoretically-driven approach that I refer to as the 'significant variables' (SV) approach. For this method, I select a set of variables (including theoretically-relevant interactions) for each outcome indicator based on the ability of each variable to predict the outcome and the interpretability of the relationship with the outcome indicator (Li et al. 2009). I screen each variable for inclusion in the model and reject those with a p-value for the B-coefficient of <.1. Although it is common to use stepwise regression methods for this stage (see e.g. O'Malley et al. 2005, Eselius et al. 2008), I test each variable manually, to ensure that the final model makes sense and does not suffer from problems of collinearity. The aim is to maximise the variation explained, but have a model that is theoretically-plausible and provides a less complex explanation of the data, making it more intelligible to policymakers and other users of the adjusted PIs (Murtaugh et al. 2007, Iezzoni 2013).

The second data-reducing method that I use draws on the argument that to be a relevant risk-adjustor a variable must have good predictive power and vary in distribution across CASSRs. Zaslavsky et al. (2001) refer to this concept as 'explanatory power', hence I refer to this method as the 'explanatory power' (EP) approach. Zaslavsky et al. develop an index of EP, by combining measures of predictive power and heterogeneity (the variability in distribution across CASSRs). They quantify predictive power as the improvement in model fit (R^2) attributable to the given variable and assess it by regressing each outcome indicator on the entire set of adjustor variables, excluding the adjustment variable of interest. Predictive power is the difference between the R^2 for this model compared to the R^2 for a model including the entire set of adjustor variables. Dummy variables for CASSRs are also included in the regressions so the resulting coefficients are estimates of the within-CASSR effects. The heterogeneity factor is a variance ratio of the adjustor's between-CASSR variance divided by its within-CASSR variance. The variance components for a given adjustment variable can be estimated from a regression of the adjustment variable on the rest of the adjustment variables, including random effects for CASSRs. The EP of each variable is found by multiplying the heterogeneity factor by the predictive power and scaling by a factor of 1000. Following previous studies (Zaslavsky et al. 2001, O'Malley et al. 2005, Eselius et al. 2008), I use a minimum EP of one for a variable to be included, although I relax

this criterion slightly for the care home sub-group, as the sample size is not very large. The results of this procedure are discussed in Appendix 11.

For both approaches, I initially treat all ordinal variables as continuous. This approach assumes interval scale properties and is taken for convenience. I test this assumption by replacing each ordinal variable with a set of dummies and examining the improvement in model fit over the baseline model (where all ordinal variables are treated as continuous) using an F-test. I retain the baseline specification, unless the dummy variable specification significantly improves the model R^2 (*p*-value <.01). Similarly I explore whether the assumption of linearity for the ADL measure is justified by examining partial residual plots and replacing the ADL measure with an appropriate transformation (e.g. square root, log and so on). I also tested for theoretically-plausible interaction effects. In all cases the baseline specification is retained unless the transformed variables or interaction terms improve model fit ($p<.01$), as determined through an F-test or likelihood ratio test.

Because of indications that the OLS models are misspecified due to the failure of Ramsey's (1969) Reset test, I tested a derived need variable in addition to the non-linear Beta-based fractional regression. This variable is the sum of the items SPHD, ADLDIFF, PAIN and ANXDEP, standardised so their means were zero and standard deviations one. However, I did not retain this variable as it was not significant and did not improve the specification tests.

Comparing statistical models

I compare the models using the concepts of construct validity, convergent validity and predictive validity. Construct validity is understood as finding that the "effects of risk factors on outcome are estimated in the expected direction", convergent validity as finding that the "effects of risk factors on outcome show close agreement when estimated by alternative models", and predictive validity as finding that the model predicts the actual outcome well (Li et al. 2009, p. 89). To explore construct and convergent validity, I examine the estimated model coefficients to both check that the effects are in the expected direction and that the estimates are fairly similar across the models. I also explore whether the coefficients for the adjustor variables change in the expected direction after the addition of the CASSR effect, i.e. that where higher values of the adjustor predict more services the coefficients becomes more negative and vice-versa. To explore predictive validity, I am most interested in the amount of variation explained by the models, for which estimates of R^2 and RMSE are the most

important measures[46]. Given that the rationale for comparing multiple models is to explore whether the various models provide a better fit to the data than the OLS, I also look at a range of other goodness of fit statistics.

Relevant goodness of fit statistics vary according to the models tested. Basu and Manca (2012) and Manning et al. (2005), suggest the following tests to assess whether the distributional assumptions of FR improve the prediction: a modified Hosmer and Lemeshow (2000) test[47], Pregibon's (1980, 1981) link test for nonlinearity[48], Ramsey's (1969) Reset test[49], and the Pearson product moment correlation coefficient of raw-scale residuals against the fitted values from the model[50]. To assess whether the RE and FE models improve model fit over the OLS, I present the Akaike Information Criterion (AIC) (Akaike 1974) and Bayesian Information Criterion (BIC) (Schwarz 1978), which balance model fit with model complexity. I take differences of greater than ten as strong evidence of the superiority of the model with the smaller BIC or AIC value over the one with the larger value (Raftery 1995, Burnham and Anderson 2004). Additionally, for the RE model, I report the estimate of Rho (the intraclass correlation coefficient), which provides a measure of amount of variation explained by differences between CASSRs. A likelihood ratio test of the null hypothesis that the between-cluster variance is equal to zero, provides evidence concerning the existence of the CASSR random effect (Rabe-Hesketh and Skrondal 2008). Along with the Hausman test this provides useful information for evaluating the fit of the RE model (Greene 2012).

I also compare predictions from the models across the range of observed values for the outcome indicator. Predicted values are calculated as the expected value for each person within the sub-group, given the values of the covariates and including the CASSR effect,

---

[46] For the FR regression R^2 is estimated as the square of the correlation between the prediction and the observed value. For the FE model I report the adjusted R^2 from the version of the FE model that includes CASSRs as dummy variables. The estimation of R^2 in random-effects models is an area of active research, and there is no agreed upon method (Gelman and Pardoe 2006). Kreft and de Leeuw (1998) advise against using R^2 because it lacks a clear meaning. In all the models RMSE is calculated from the square root of the mean difference between the predicted and observed values, with the predicted values calculated as described in the text.

[47] This uses an F-test that the means of the raw scale residuals across all 10 of the deciles of the linear predictor are not significantly different from zero.

[48] To calculate this, the linear predictor and the square of the linear predictor are included as covariates in a second version of the model. If the coefficient on the square of the linear predictor is different from zero then this indicates that there is additional nonlinearity unaccounted for in the current specification.

[49] In a second model, containing all the covariates, higher-order powers (specifically the square, cube and fourth power) of the linear predictor are also included. If the joint test of the higher-order powers is different to zero this implies that there is additional nonlinearity in the specification.

[50] If this statistic (rho) is significantly different from zero then the model is providing a biased prediction of E(y|x) (Manning et al. 2005).

where relevant, i.e. $E[y_{ij}|z_{ij}, u_j]$. Error scores are calculated as the deviation of the predicted from the observed value, i.e. $y_{ij} - E[y_{ij}|z_{ij}, u_j]$.

Missing data

The questionnaire and auxiliary data are both subject to item nonresponse. To avoid problems of bias, I recover the missing data on the characteristics of the full sample using chained equations multiple imputation (van Buuren 2007, White et al. 2011). I use the multiply-imputed dataset described in Chapter 5 for this analysis. Appendices 1 and 2 provides more details of the imputation models, procedure and sufficiency of the number of imputations.

After imputation, I combine the individual SCRQoL components into an overall SCRQoL score and also generate all other multi-item variables required for the analysis, such as the ADL measure. All models are re-estimated on each of the imputed datasets and the estimates are combined using Rubin's rules (1987). Missing cases for the dependent variables are excluded from the estimation of these models as they would increase the variance without adding any further information. The results from these models estimated on the multiply-imputed dataset were not very different from those estimated using the complete cases. For the sake of clarity and parsimony, I therefore only present the results from the complete cases analysis. A comparison between the estimates from the complete case analysis and MI is given in Appendix 12.

Results

Sample descriptives

The analysis subsample consists of 44,744 cases out of the total respondent sample of 61,026 (around 73 per cent)[51]. Two-thirds of the subsample (29,786 cases), are in the sub-group comprising people aged 65 and over, with PSD, in private households (referred to hereafter as '65 and over sub-group'). The rest of the subsample is split fairly evenly between the other two groups, with 7,447 people in the sub-group of people aged between 18 and 64, with PSD, living in private households (referred to hereafter as '18 to 64 sub-group'), and 7,510 cases,

[51] A small fraction (1.5 per cent) of the sample was lost when stratifying into sub-groups for analysis, due to missingness of a key stratification variable (313 cases) and being sent a questionnaire inconsistent with the client group (585 cases). For example, people living in a care home or with a learning disability have been sent the standard questionnaire, or people without learning disability as a primary client have been sent the learning disability questionnaire.

in the sub-group of people with PSD who are resident in a care home (referred to hereafter as 'care home sub-group').

Table 23 shows the distributional statistics for the outcome and risk-adjustor variables for all sub-groups. All ordinal variables (including age group) are treated in these tables as continuous. For most of the variables, the fraction of missing cases is fairly small. Nevertheless, were all of the variables in Table 23 to be included in the adjustment models 7,330 cases from the 65 and over sub-group would be lost (25 per cent), 1,362 cases from the 18 to 64 sub-group would be lost (18 per cent) and 1,579 cases from the care home sub-group would be lost (21 per cent). Since the models use a selection of the variables and there is also missingness in the outcome variable, around 13 to 30 per cent of the data are missing from the analysis, depending on the model specification and the sub-group.

The final column of Table 23 contains the test statistic for whether there are differences in the means (proportions for the binary variables) of the three sub-groups. The F-statistic is reported for ordinal variables and the chi-squared statistic is reported for all binary variables. No test statistic is reported for age group, since two sub-groups are defined with respect to this variable. There are significant differences across the sub-groups for each one of the adjustor variables, except the proportion of vulnerable people which is constant across the groups. This provides some support for the decision to split the analysis by sub-group.

Table 23: Distributional statistics for the adjustor variables, all sub-groups

Variable	Care home (n=7,510)		18 to 64 (n=7,447)		65 and over (n=29,786)		Equality of means †
	N (% missing)	Mean (SE)	N (% missing)	Mean (SE)	N (% missing)	Mean (SE)	
MALE	7,505 (0.07)	0.29 (0.005)	7,445 (0.03)	0.4 (0.006)	29,778 (0.03)	0.3 (0.003)	344.22***
AGE_GP	7,500 (0.13)	7.21 (0.014)	7,447 (0)	4.08 (0.013)	29,786 (0)	7.22 (0.004)	n/a
VP	7510 (0)	0.03 (0.002)	7447 (0)	0.03 (0.002)	29786 (0)	0.03 (0.001)	1.58
ADLDIFF	7,043 (6.22)	4.14 (0.029)	6,950 (6.67)	3.68 (0.030)	26,832 (9.92)	2.91 (0.015)	857.0***
ADLCANT	7,043 (6.22)	2.79 (0.030)	6,950 (6.67)	1.48 (0.026)	26,832 (9.92)	1.15 (0.011)	1,907.2***
SPH	7,318 (2.56)	1.61 (0.01)	7,288 (2.14)	2.16 (0.012)	28,744 (3.5)	1.98 (0.005)	744.3***
PAIN	7,234 (3.68)	0.67 (0.007)	7,220 (3.05)	1.15 (0.008)	28,478 (4.39)	1.03 (0.004)	1,252.5***
ANXDEP	7,136 (4.98)	0.46 (0.007)	7,190 (3.45)	0.79 (0.008)	27,963 (6.12)	0.55 (0.004)	638.5***
FINANCES	7,230 (3.73)	1.68 (0.008)	7,190 (3.45)	1.03 (0.01)	28,255 (5.14)	1.08 (0.005)	1,568.5***
SPHD	7,282 (3.04)	0.4 (0.007)	7,187 (3.49)	0.99 (0.011)	28,322 (4.92)	0.69 (0.005)	1,097.5***
PH, in home	6,999 (6.8)	0.32 (0.006)	7,169 (3.73)	0.52 (0.006)	27,813 (6.62)	0.35 (0.003)	795.3***
PH, out home	6,999 (6.8)	0.55 (0.006)	7,169 (3.73)	0.37 (0.006)	27,813 (6.62)	0.58 (0.003)	986.3***
PH, none	6,999 (6.8)	0.26 (0.005)	7,169 (3.73)	0.22 (0.005)	27,813 (6.62)	0.15 (0.002)	521.2***
SELFCOMP	7,163 (4.62)	0.1 (0.003)	6,956 (6.59)	0.46 (0.006)	27,464 (7.8)	0.4 (0.003)	2,700.0***
PROXY	7,163 (4.62)	0.15 (0.004)	6,956 (6.59)	0.03 (0.002)	27,464 (7.8)	0.06 (0.001)	973.4***
ASSIST, read	7,163 (4.62)	0.55 (0.006)	6,956 (6.59)	0.31 (0.006)	27,464 (7.8)	0.32 (0.003)	1,300.0***
ASSIST, translate	7,163 (4.62)	0.09 (0.003)	6,956 (6.59)	0.08 (0.003)	27,464 (7.8)	0.06 (0.001)	69.4***
ASSIST, write	7,163 (4.62)	0.48 (0.006)	6,956 (6.59)	0.29 (0.005)	27,464 (7.8)	0.32 (0.003)	764.0***
ASSIST, discuss	7,163 (4.62)	0.33 (0.006)	6,956 (6.59)	0.22 (0.005)	27,464 (7.8)	0.24 (0.003)	276.4***
ASSIST, in home	7,131 (5.05)	0.04 (0.002)	6,931 (6.93)	0.22 (0.005)	27,370 (8.11)	0.16 (0.002)	1,000.0***
ASSIST, out home	7,131 (5.05)	0.41 (0.006)	6,931 (6.93)	0.16 (0.004)	27,370 (8.11)	0.31 (0.003)	1,100.0***
ASSIST, cw	7,131 (5.05)	0.3 (0.005)	6,931 (6.93)	0.11 (0.004)	27,370 (8.11)	0.05 (0.001)	3,700.0***
PRIV, own	6,860 (8.66)	0.16 (0.004)	7,070 (5.06)	0.25 (0.005)	27,119 (8.95)	0.41 (0.003)	1,900.0***
PRIV, fam	6,860 (8.66)	0.13 (0.004)	7,070 (5.06)	0.05 (0.003)	27,119 (8.95)	0.06 (0.001)	420.8***
PRIV, none	6,860 (8.66)	0.73 (0.005)	7,070 (5.06)	0.71 (0.005)	27,119 (8.95)	0.54 (0.003)	1,200.0***
SCRQoL‡	6,849 (8.80)	19.63 (0.043)	6,765 (9.16)	17.1 (0.054)	25,965 (12.83)	18.28 (0.023)	751.73***

Legend: * p<.1; ** p<.01; *** p<.001; † χ^2 test for nominal variables and anova F-test for ordinal variables and multi-item scales; ‡ median for care home sub-group=20, median for 18 to 64 sub-group=17, median for 65 and over sub-group=19.

Estimates for the risk-adjustment models

In this section I present the results for the risk-adjustment models, estimated using the four alternative regression procedures. The data in this section are presented by population sub-group and by model specification, i.e. the theoretically-driven, SV, or statistically-driven, EP, model. All of the OLS models failed tests for normality of the error terms and homoscedasticity (confirmed on visual inspection of the data). Likewise all of the FE models fail the modified Wald test for groupwise heteroscedasticity, as implemented via the xttest3 routine (Baum 2000, Greene 2012). All results for these models therefore use the Huber-White sandwich estimator to correct standard errors (Huber 1967b, White 1980).

Comparing the theoretically-driven, SV, models across the sub-groups, it is clear that a similar set of variables are relevant (Table 24, Table 26 and Table 28). Notably, the variables capturing underlying conditions (SPH, ADLDIFF, ANXDEP, FINANCE) are strong predictors of SCRQoL, as is the immediate environment variable (SPHD), and the resources arising from social capital variables (PH). Sociodemographic variables (MALE, AGE_GP), resources arising from economic capital (PRIV) and the reporting-related variables (ASSIST, PROXY) are less consistently included, save for the variable capturing assistance from a care worker (ASSIST, cw). Despite these similarities across the models in the variables included, R^2 is very different suggesting that these variables have different explanatory power within the models. This is confirmed to some extent by a comparison of the statistically-driven, EP, models (Table 25, Table 27 and Table 29). The variables retained are very similar for the two private household sub-groups (18 to 64 and 65 and over), differing only in the fact that ADL score is retained for the 65 and over sub-group. None of the resources or reporting-related variables are retained. By contrast for the care home sub-group, only two variables are retained, one from the underlying conditions category and one from the resources category. This suggests some differences across sub-groups in the ability of the variables to predict SCRQoL and in the extent to which the characteristics vary across CASSRs.

In general, all the models show good construct validity, regardless of regression method, variable selection or sub-group (Table 24 to Table 29). The effects are all in the expected direction, except in the FR regression for a few of the interaction terms which are mostly no longer significant. The differences between the FR and OLS model probably reflect the non-linear transformation achieved through the logit link function. The age group variable (AGE_GP) for the care home sub-group has a negative coefficient for all categories (although some are not significant) (see Table 28). Since the base group is adults aged 85

and over, this indicates that people younger than 85 have worse SCRQoL than people aged over 85, which is contrary to what we would expect if this variable is acting as an indicator of need. This relationship is present (and stronger) before adding in other covariates, suggesting it is not a consequence of collinearity. It is likely that age is capturing some other, perhaps reporting-related phenomenon. Indeed age is a likely to be a fairly poor indicator of need in a care home sample. The finding is consistent with evidence of a negative u-shaped relationship between age and life satisfaction, which some have suggested is related to differences in the criteria people use for evaluating their QoL (Easterlin 2006, Blanchflower and Oswald 2008).

It is not possible to compare coefficients from the FR regression with the OLS and other regressions directly, because (i) of the scaling applied to the outcome indicator in the FR regression, and (ii) the right-hand side variables in the FR regression are related to the outcome indicator by a non-linear 'link' function. Nevertheless, I can explore convergent validity by looking at the significance of the coefficients. In this respect the models have good convergent validity. The significance of the estimated coefficients differs in only a small number of cases, and all of these instances bar one are related to the interaction terms. The only example of the significance of an estimated coefficient that is not an interaction term changing between the regression models, is found for gender (MALE), in the theoretically-driven SV model for the care home sub-group (see Table 28), but this is a consequence of the p-value being close to the significance boundary.

For the OLS, RE and FE models, it is possible to directly compare the coefficients. In general, there is also good agreement between the models, providing evidence for convergent validity. The largest differences are found for the age group coefficients in the theoretically-driven SV model estimated on the care home sub-group, where several coefficients have differences between 0.1 to 0.25 scale points (>1 percentage point) (Table 28). The change in the coefficients for the risk-adjustors is mostly as would be expected if the (random or fixed) CASSR effects are capturing (a part of) the impact of services on SCRQoL. There are several variables, however, where the coefficients change in the unexpected direction. This is the case across the models for the anxiety/depression (ANXDEP), but is also a problem for some of the other variables (SPHD, SPH) for some of the models, where the coefficients are more positive (rather than negative) in the RE and FE models compared to the OLS model. This suggests that mean differences between CASSRs in anxiety/depression and other indicators are negatively related to SCRQoL.

Table 24: Estimates for OLS, FR, RE and FE regression of the SCRQoL indicator, SV models, 18 to 64 sub-group (n=5,856)

Variable	OLS *B* (Robust SE)	FR *B* (Robust SE)	RE *B* (SE)	FE *B* (Robust SE)
Fixed effects				
SPH: good [a]	-1.061*** (0.16)	-0.419*** (0.06)	-1.068*** (0.203)	-1.118*** (0.162)
SPH: fair [a]	-2.170*** (0.157)	-0.724*** (0.057)	-2.172*** (0.194)	-2.202*** (0.16)
SPH: bad [a]	-3.483*** (0.18)	-0.970*** (0.059)	-3.487*** (0.209)	-3.521*** (0.181)
SPH: very bad [a]	-3.964*** (0.242)	-1.033*** (0.066)	-3.962*** (0.242)	-3.974*** (0.256)
SPHD	-1.092*** (0.097)	-0.260*** (0.021)	-1.088*** (0.096)	-1.062*** (0.101)
ADLDIFF	-0.069* (0.031)	-0.021** (0.007)	-0.069* (0.032)	-0.072* (0.029)
ADLDIFF*SPHD	-0.091*** (0.021)	-0.009* (0.004)	-0.091*** (0.02)	-0.090*** (0.021)
FINANCE: diff [b]	-0.715*** (0.115)	-0.153*** (0.025)	-0.717*** (0.116)	-0.722*** (0.116)
FINANCE: can't [b]	-0.575*** (0.119)	-0.121*** (0.027)	-0.579*** (0.12)	-0.598*** (0.117)
ANXDEP	-1.665*** (0.077)	-0.348*** (0.016)	-1.663*** (0.073)	-1.661*** (0.073)
PH, in home	0.357* (0.151)	0.081* (0.037)	0.359* (0.163)	0.372* (0.155)
PH, out home	0.288** (0.096)	0.058** (0.021)	0.284** (0.096)	0.275** (0.101)
PH, in*ADLDIFF	0.093** (0.035)	0.016* (0.008)	0.091** (0.035)	0.082* (0.036)
MALE	-0.238** (0.089)	-0.052** (0.019)	-0.236** (0.089)	-0.226* (0.095)
ASSIST, cw	0.878*** (0.155)	0.204*** (0.037)	0.880*** (0.149)	0.878*** (0.171)
ASSIST, translate	0.320* (0.179)	0.071* (0.04)	0.326* (0.169)	0.313* (0.183)
ASSIST, write	0.221* (0.108)	0.043* (0.024)	0.216* (0.107)	0.191 (0.116)
Constant	22.299*** (0.183)	2.318*** (0.062)	22.310*** (0.221)	22.347*** (0.18)
Random effects				
σ_u			0.210 (0.083)	
σ_e			3.289 (0.031)	
Model statistics				
F-stat	259.98***	211.23***	191.33***	253.91***
R^2	0.431	0.431		0.447

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: can do easily by myself.

Table 25: Estimates for OLS, FR, RE and FE regression of the SCRQoL indicator, EP models, 18 to 64 sub-group (n=6,513)

Variable	OLS *B* (Robust SE)	FR *B* (Robust SE)	RE *B* (SE)	FE *B* (Robust SE)
Fixed effects				
SPH: good [a]	-1.187*** (0.151)	-0.451*** (0.058)	-1.192*** (0.197)	-1.244*** (0.146)
SPH: fair [a]	-2.425*** (0.148)	-0.771*** (0.056)	-2.429*** (0.188)	-2.472*** (0.151)
SPH: bad [a]	-3.817*** (0.168)	-1.038*** (0.057)	-3.822*** (0.201)	-3.872*** (0.173)
SPH: very bad [a]	-4.422*** (0.227)	-1.139*** (0.063)	-4.424*** (0.229)	-4.465*** (0.254)
SPHD	-1.291*** (0.091)	-0.339*** (0.022)	-1.286*** (0.088)	-1.248*** (0.087)
ANXDEP: mod [b]	-1.290*** (0.123)	-0.397*** (0.033)	-1.290*** (0.135)	-1.288*** (0.121)
ANXDEP: extreme [b]	-2.985*** (0.251)	-0.745*** (0.051)	-2.978*** (0.218)	-2.916*** (0.272)
SPHD* ANXDEP: mod [c]	-0.264* (0.114)	0.039 (0.026)	-0.264* (0.11)	-0.264* (0.11)
SPHD* ANXDEP: ext [c]	-0.573*** (0.168)	0.02 (0.034)	-0.576*** (0.139)	-0.618*** (0.17)
Constant	22.314*** (0.13)	2.370*** (0.055)	22.316*** (0.179)	22.318*** (0.13)
Random effects				
σ_u			0.183 (0.089)	
σ_e			3.345 (0.03)	
Model statistics				
F-stat	517.20***	415.93***	384.50***	475.57***
R^2	0.417	0.416		0.432

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: not anxious or depressed; [c] Base category: SPHD*ANXDEP: not anxious or depressed.

Table 26: Estimates for OLS, FR, RE and FE regression of SCRQoL indicator, SV models, 65 and over sub-group (n=20,881)

Variable	OLS	FR	RE	FE
	B (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects				
SPH: good [a]	-0.632*** (0.097)	-0.303*** (0.041)	-0.635*** (0.118)	-0.641*** (0.094)
SPH: fair [a]	-1.624*** (0.095)	-0.603*** (0.04)	-1.626*** (0.114)	-1.631*** (0.089)
SPH: bad [a]	-2.676*** (0.109)	-0.806*** (0.041)	-2.681*** (0.124)	-2.686*** (0.108)
SPH: very bad [a]	-3.165*** (0.15)	-0.864*** (0.045)	-3.161*** (0.147)	-3.157*** (0.15)
ANXDEP	-1.350*** (0.04)	-0.304*** (0.009)	-1.345*** (0.037)	-1.342*** (0.04)
SPHD	-1.397*** (0.031)	-0.305*** (0.007)	-1.391*** (0.028)	-1.383*** (0.03)
ADLDIFF	-0.297*** (0.013)	-0.068*** (0.003)	-0.294*** (0.012)	-0.293*** (0.013)
FINANCES	-0.188*** (0.029)	-0.050*** (0.007)	-0.182*** (0.029)	-0.178*** (0.03)
PH, in home	0.156* (0.075)	0.037* (0.02)	0.153* (0.08)	0.145* (0.087)
PH, out home	0.276*** (0.052)	0.063*** (0.013)	0.266*** (0.051)	0.258*** (0.056)
PH, in *ADLDIFF	0.154*** (0.018)	0.031*** (0.004)	0.153*** (0.018)	0.155*** (0.021)
MALE	-0.204*** (0.045)	-0.047*** (0.011)	-0.202*** (0.044)	-0.200*** (0.047)
PROXY	-0.854*** (0.104)	-0.178*** (0.022)	-0.860*** (0.095)	-0.863*** (0.113)
ASSIST, cw	1.039*** (0.103)	0.272*** (0.028)	1.004*** (0.1)	0.974*** (0.139)
ASSIST, out home	-0.323*** (0.06)	-0.073*** (0.014)	-0.331*** (0.06)	-0.336*** (0.064)
ASSIST, read	0.159** (0.051)	0.033** (0.012)	0.153** (0.05)	0.150** (0.055)
PRIV, own	-0.159*** (0.042)	-0.044*** (0.01)	-0.156*** (0.042)	-0.155*** (0.042)
Constant	22.42*** (0.099)	2.377*** (0.041)	22.41*** (0.122)	22.41*** (0.094)
Random effects				
σ_u			0.271 (0.029)	
σ_e			2.878 (0.014)	
Model statistics				
F-stat	742.34***	668.84***	597.86***	830.19***
R^2	0.391	0.391		0.401

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good.

Table 27: Estimates for OLS, FR, RE and FE regression of SCRQoL indicator, EP models, 65 and over sub-group (n=23,110)

Variable	OLS *B* (Robust SE)	FR *B* (Robust SE)	RE *B* (SE)	FE *B* (Robust SE)
Fixed effects				
SPH: good [a]	-0.734*** (0.091)	-0.321*** (0.039)	-0.737*** (0.113)	-0.743*** (0.089)
SPH: fair [a]	-1.730*** (0.089)	-0.609*** (0.038)	-1.733*** (0.11)	-1.740*** (0.088)
SPH: bad [a]	-2.837*** (0.104)	-0.828*** (0.039)	-2.841*** (0.119)	-2.847*** (0.106)
SPH: very bad [a]	-3.275*** (0.147)	-0.894*** (0.044)	-3.275*** (0.142)	-3.276*** (0.151)
SPHD: most [b]	-1.425*** (0.063)	-0.434*** (0.017)	-1.432*** (0.068)	-1.435*** (0.056)
SPHD: some [b]	-2.552*** (0.132)	-0.687*** (0.028)	-2.532*** (0.127)	-2.512*** (0.145)
SPHD: inappropriate [b]	-3.278*** (0.417)	-0.832*** (0.077)	-3.214*** (0.36)	-3.161*** (0.398)
ANXDEP: moderate [c]	-1.316*** (0.042)	-0.318*** (0.01)	-1.314*** (0.042)	-1.313*** (0.052)
ANXDEP: extreme [c]	-2.887*** (0.119)	-0.597*** (0.022)	-2.866*** (0.096)	-2.849*** (0.12)
ADLDIFF	-0.241*** (0.013)	-0.074*** (0.003)	-0.239*** (0.013)	-0.237*** (0.013)
ADLDIFF*SPHD: most [d]	-0.024 (0.018)	0.020*** (0.004)	-0.022 (0.018)	-0.02 (0.019)
ADLDIFF*SPHD: some [d]	-0.057* (0.03)	0.023*** (0.006)	-0.057* (0.028)	-0.056 (0.035)
ADLDIFF*SPHD: inappropriate [d]	-0.137 (0.09)	0.012 (0.016)	-0.144* (0.071)	-0.148 (0.09)
Constant	22.43*** (0.084)	2.430*** (0.038)	22.42*** (0.109)	22.42*** (0.084)
Random effects				
σ_u			0.261 (0.028)	
σ_e			2.924 (0.014)	
Model statistics				
F-stat	1,032.44***	899.523***	1,024.36***	1,034.73***
R^2	0.370	0.371		0.379

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: meets needs very well; [c] Base category: not anxious or depressed; [d] Base category: ADLDIFF*SPHD: meets needs very well.

Table 28: Estimates for OLS, FR, RE and FE regression of SCRQoL indicator, SV models, care home sub-group (n=5,710)

Variable	OLS	FR	RE	FE
	B (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects				
SPH	-0.820*** (0.055)	-0.230*** (0.016)	-0.819*** (0.052)	-0.821*** (0.054)
ANXDEP: moderate [a]	-1.289*** (0.09)	-0.366*** (0.025)	-1.286*** (0.088)	-1.276*** (0.086)
ANXDEP: extremely [a]	-3.052*** (0.28)	-0.680*** (0.057)	-3.057*** (0.216)	-3.076*** (0.29)
ADLDIFF^3	-0.004*** (3.59E-04)	-0.001*** (9.75E-05)	-0.005*** (2.41E-04)	-0.005*** (3.55E-04)
PH, in home	0.686*** (0.082)	0.207*** (0.026)	0.686*** (0.087)	0.684*** (0.094)
MALE	-0.326*** (0.094)	-0.086** (0.027)	-0.325*** (0.094)	-0.315** (0.107)
AGE_GP: 18-34 [b]	0.057 (0.344)	0.066 (0.124)	0.069 (0.4)	0.043 (0.347)
AGE_GP: 35-44 [b]	-1.595*** (0.406)	-0.445*** (0.102)	-1.564*** (0.368)	-1.374*** (0.401)
AGE_GP: 45-54 [b]	-0.103 (0.241)	-0.026 (0.078)	-0.086 (0.266)	0.001 (0.243)
AGE_GP: 55-64 [b]	-0.477* (0.217)	-0.148* (0.063)	-0.460* (0.218)	-0.354 (0.226)
AGE_GP: 65-74 [b]	-0.455** (0.15)	-0.134** (0.043)	-0.448** (0.147)	-0.405* (0.175)
AGE_GP: 75-84 [b]	-0.118 (0.097)	-0.023 (0.028)	-0.117 (0.097)	-0.111 (0.101)
PROXY	-1.051*** (0.186)	-0.235*** (0.05)	-1.044*** (0.171)	-1.006*** (0.194)
ASSIST, cw	0.720*** (0.145)	0.264*** (0.047)	0.705*** (0.148)	0.656*** (0.158)
ASSIST, out home	-0.530*** (0.142)	-0.153*** (0.042)	-0.530*** (0.138)	-0.527*** (0.154)
ASSIST, read	0.313** (0.103)	0.090** (0.029)	0.310** (0.1)	0.283* (0.112)
PRIV, own	-0.492*** (0.129)	-0.130*** (0.033)	-0.484*** (0.114)	-0.455** (0.138)
Constant	22.13*** (0.149)	2.247*** (0.047)	22.13*** (0.152)	22.14*** (0.147)
Random effects				
σ_u			0.204 (0.081)	
σ_e			3.018 (0.029)	
Model statistics				
F-stat	122.98***	127.64***	116.13***	110.65***
R^2	0.296	0.301		0.318

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: not anxious or depressed; [b] Base category: 85 and over age group.

Table 29: Estimates for OLS, FR, RE and FE regression of SCRQoL indicator, EP models, care home sub-group (n=6,344)

Variable	OLS *B* (Robust SE)	FR *B* (Robust SE)	RE *B* (SE)	FE *B* (Robust SE)
Fixed effects				
ANXDEP: moderate [a]	-2.218*** (0.114)	-0.588*** (0.029)	-2.190*** (0.105)	-2.157*** (0.107)
ANXDEP: extremely [a]	-5.027*** (0.346)	-1.147*** (0.065)	-4.988*** (0.257)	-4.932*** (0.38)
PH in home	0.741*** (0.097)	0.261*** (0.035)	0.752*** (0.113)	0.770*** (0.105)
PH in* ANXDEP: moderate [b]	0.540** (0.184)	0.052 (0.053)	0.511** (0.189)	0.471** (0.175)
PH in*ANXDEP: extremely [b]	1.784** (0.656)	0.234* (0.14)	1.758*** (0.495)	1.708* (0.668)
Constant	20.33*** (0.063)	1.712*** (0.02)	20.30*** (0.078)	20.30*** (0.049)
Random effects				
σ_u			0.447 (0.062)	
σ_e			3.293 (0.030)	
Model statistics				
F-stat	172.23***	187.632***	188.15***	195.50***
R^2	0.140	0.140		0.174

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: not anxious or depressed; [b] Base category: PH in*ANXDEP: not anxious or depressed.

Comparison of model fit and predictive validity

In this section I compare the predictive validity of the models. I first look at the comparison between the OLS and FR regressions for both outcome indicators, across all sub-groups. Then I compare the OLS, RE and FE regressions. These comparisons are presented separately because I use different indices of model fit, as previously described. After comparing model fit using standard indices, I examine the distribution of the predicted values over the range of the observed scale for each outcome indicator.

Goodness of fit statistics for the OLS and FR regression models are set out in Table 30. The R^2s are not very high but are fairly typical for models with these populations (Davies et al. 2000b, Glendinning et al. 2008a, Forder et al. 2016). For the two non-care home sub-groups, the OLS has a slightly higher R^2 (and lower RMSE) and also marginally outperforms the FR regression on the other goodness of fit tests. Neither model fits the data well; both have poor link and Ramsey reset tests. In the care home sub-group SV model, in which the SCRQoL indicator has a slightly more skewed and peaked distribution, the FR model slightly outperforms the OLS. It has a slightly higher R^2 (and lower RMSE), and passes the tests that are concerned with the distribution of the errors. There is, however, not much difference between the models; again both fail the link and Ramsey reset tests, suggesting some unaccounted for nonlinearity. The tests for the statistically-driven, EP, models for the care home sub-group suggest a slight preference for the OLS model, but the R^2 is very low for this specification compared to the other models

Table 30: Model fit statistics for OLS and FR regression

	SV model		EP model	
Estimators	**OLS**	**FR**	**OLS**	**FR**
18 to 64 sub-group				
Hosmer-Lemeshow, F-stat (p-value)	0.961 (0.475) †	2.186 (0.016)	1.027 (0.417) †	2.331 (0.01)
Pearson correlation, Rho (p-value)	0.000 (1.000)†	-0.005 (0.721)†	0.000 (1.000)†	-0.004 (0.757)†
Link test, t stat (p-value)	-0.017 (0.02)	-0.031 (<.001)	-0.002 (0.606) †	-0.032 (<.001)
Reset test, F-stat (p-value)	13.95 (<.001)	10.84 (<.001)	2.72 (0.043)	8.743 (<.001)
R^2	0.431	0.431	0.417	0.416
RMSE	3.296	3.294	3.350	3.354
65 and over sub-group				
Hosmer-Lemeshow, F-stat (p-value)	2.088 (0.022)	8.004 (<.001)	3.13 (0.001)	4.591 (<.001)
Pearson correlation, Rho (p-value)	0.000 (1.000)†	-0.009 (0.216)†	0.000 (1.000)†	-0.003 (0.656)†
Link test, t stat (p-value)	-0.016 (<.001)	-0.025 (<.001)	-0.013 (<.001)	-0.037 (<.001)
Reset test, F-stat (p-value)	19.00 (<.001)	42.80 (<.001)	23.55 (<.001)	24.64 (<.001)
R^2	0.391	0.389	0.370	0.371
RMSE	2.891	2.895	2.936	2.935
Care home sub-group				
Hosmer-Lemeshow, F-stat (p-value)	5.888 (<.001)	1.03 (0.415)†	0.881 (0.551)†	1.800 (0.055)†
Pearson correlation, Rho (p-value)	0.000 (1.000)†	-0.004 (0.753)†	0.000 (1.000)†	0.000 (1.000)†
Link test, t stat (p-value)	-0.042 (<.001)	-0.036 (<.001)	0.000 (1.000)†	-0.046 (<.001)
Reset test, F-stat (p-value)	17.66 (<.001)	3.923 (0.008)	9.26 (<.001)	---
R^2	0.296	0.301	0.140	0.140
RMSE	3.025	3.013	3.322	3.322

Legend: --- estimation not possible due to collinearity, † passes goodness of fit test at p<.05 level

A comparison between the OLS, RE and FE models reveals differences across the sub-groups in terms of the predictive validity of the different regression methods (Table 31). For the 18 to 64 sub-group the likelihood ratio test for rho is not significant ($p>0.05$)[52], suggesting that predictive validity is not improved by modelling the organisational effect as a random variable. The BIC and AIC statistics support this finding. The F-test for the inclusion of fixed effects for CASSRs is more ambiguous ($p>0.1$ for both models)[53], but the adjusted R^2, RMSE, AIC and BIC statistics all clearly favour the FE model over the OLS. The FE model may have better predictive validity than the OLS for this sub-group.

For the 65 and over sub-group, the predictive validity of the RE and FE models are both better than the OLS. The AIC and BIC statistics very strongly favour the RE and FE models. While the differences in adjusted R^2 and RMSE statistics are not large they are greater than those observed for the 18 to 64 age group. Importantly, the likelihood ratio test for rho is highly significant ($p<.001$)[54] and so is the F-test for the inclusion of fixed effects for CASSRs ($p<.001$)[55]. Although results from the FE model show fairly low levels of correlation between estimates of the FE and the other model coefficients (ranging from .057 to .064), a Hausman test indicates that the FE model is preferred ($p<.001$) for all the models[56].

The choice of covariates has the most notable effect on the results for the care home sub-group. For the statistically-driven, EP, model the AIC, BIC, adjusted R^2 and RMSE statistics very strongly favour the RE and FE models over the OLS; for the theoretically-driven SV model only the AIC and BIC statistics strongly favour the FE regression over the OLS. Supporting these findings, the likelihood ratio test for rho for the statistically-driven EP model is highly significant ($p<.001$) but not for the theoretically-driven SV model ($p>0.1$)[57], while the F-test for the inclusion of fixed effects for CASSRs is also highly significant for the statistically-driven EP model ($p<.001$), but only significant at a higher threshold for the theoretically-driven SV model ($p<.05$)[58]. A Hausman test indicates that for the care home statistically-driven EP model the RE model is preferred ($p>.1$)[59]. Since this specification contains only two variables plus their interaction, it seems likely that the

---

[52] SV model: $\chi^2(2)=2.09$, $p=0.074$; EP model: $\chi^2(2)=1.31$, $p=0.126$.
[53] SV model: $F(148,5690)=1.14$, $p=0.116$; EP model: $F(148,6355)=1.15$, $p=0.103$.
[54] SV model: $\chi^2(2)=29.33$, $p=0.032$; EP model: $\chi^2(2)=19.53$, $p=0.021$.
[55] SV model: $F(148,20715)=2.25$, $p<.001$; EP model: $F(148,22948)=2.24$, $p<.001$.
[56] SV model: $\chi^2(17)=55.80$, $p<.001$; model: $\chi^2(13)=43.32$, $p<.001$.
[57] SV model: $\chi^2(2)=2.09$, $p=.352$; EP model: $\chi^2(2)=29.49$, $p<.001$.
[58] SV model: $F(147,5545)=1.20$, $p=.049$; EP model: $F(147,6191)=1.77$, $p<.001$.
[59] EP model: $\chi^2(5)=9.21$, $p=0.101$.

statistically-driven approach to specification in the care home sub-group does not adjust for sufficient of the CASSR-level variation that is due to differences in the profile of service users.

Three further points are worth noting here. First, as would be expected, the theoretically-driven SV model which results in a specification with more risk-adjustors explains more of the variation in scores (R^2 is greater and RMSE is lower) than the statistically-driven and more parsimonious EP model across all the population sub-groups. Additionally, despite the increased model complexity, AIC and BIC statistics very strongly favour the SV over the EP models. Secondly, the predictive ability of the regressions differs across the sub-groups. The models estimated on the 18 to 64 sub-group having the highest R^2, followed by the models estimated on the 65 and over sub-group and finally followed by the models estimated on the care home sub-group, where the R^2 is much lower, but where fewer variables were available for inclusion in the models. This suggests that the indicators are better suited to capturing the social care needs of the 18 to 64 sub-group. Thirdly, conditional rho, that is the between-CASSR variation after accounting for risk-adjustors, was in all cases less than one per cent of the overall variation. Even prior to the inclusion of risk-adjustors, rho was very low across the sub-groups, explaining around two per cent of the total variation. This suggests that very little of the variation in SCRQoL can be directly attributed to CASSRs.

Table 31: Model fit statistics for OLS, RE and FE regressions

	SV model			EP model		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
18 to 64 sub-group						
Log-likelihood	-15,293	-15,292	-15,207	-17,115	-17,114	-17,029
F-stat	259.98***	191.33***	253.91***	517.20***	384.50***	475.57***
AIC	30,622	30,624	30,450	34,250	34,252	34,077
BIC	30,742	30,757	30,571	34,318	34,334	34,145
Adjusted R^2	0.429		0.431	0.416		0.418
RMSE	3.296	3.248	3.248	3.350	3.306	3.306
Conditional rho (SE)		0.004 (0.003)			0.003 (0.003)	
65 and over sub-group						
Log-likelihood	-51,794	-51,762	-51,627	-57,681	-57,650	-57,515
F-stat	742.34***	597.86***	830.19***	1,032.44***	1,024.36***	1,034.73***
AIC	103,624	103,564	103,291	115,389	115,331	115,058
BIC	103,767	103,723	103,434	115,502	115,460	115,171
Adjusted R^2	0.391		0.396	0.370		0.375
RMSE	2.891	2.868	2.868	2.936	2.915	2.915
Conditional rho (SE)		0.009 (0.002)			0.008 (0.002)	
Care home sub-group						
Log-likelihood	-14,422	-14,421	-14,332	-16,618	-16,603	-16,487
F-stat	122.98***	116.13***	110.65***	172.23***	188.15***	195.50***
AIC	28,880	28,882	28,701	33,247	33,222	32,986
BIC	29,000	29,015	28,821	33,288	33,276	33,026
Adjusted R^2	0.294		0.298	0.139		0.154
RMSE	3.025	2.978	2.978	3.322	3.254	3.254
Conditional rho (SE)		0.005 (0.004)			0.018 (0.005)	

Legend: *** p<.001

Figure 28 illustrates the distribution of the predictions from the regression models, over the range of values of the SCRQoL indicator. The distribution of predictions from the OLS, RE and FE models are very similar, and seem to coincide for the two non-care home sub-groups. Only in the case of the care home EP model, do the predicted values for the OLS have a slightly more peaked distribution than those from the RE and FE regressions. The FR regression, however, captures the skewed nature of the distribution slightly better than the OLS, particularly for the SV models. It also does not predict values beyond the maximum score, as the OLS, RE and FE regressions do. This comes at a cost of the predicted values having a more peaked distribution. The relatively poor performance of all the models in predicting the extremes of the distribution is shown more clearly in Table 84 to Table 86 in Appendix 13, which examine the mean error across five regions of the distribution. The mean error and mean absolute error are both largest for the categories that are farthest away from the mean of the distribution and the errors from the statistically-driven EP model are in all cases larger than those from the theoretically-driven SV model.

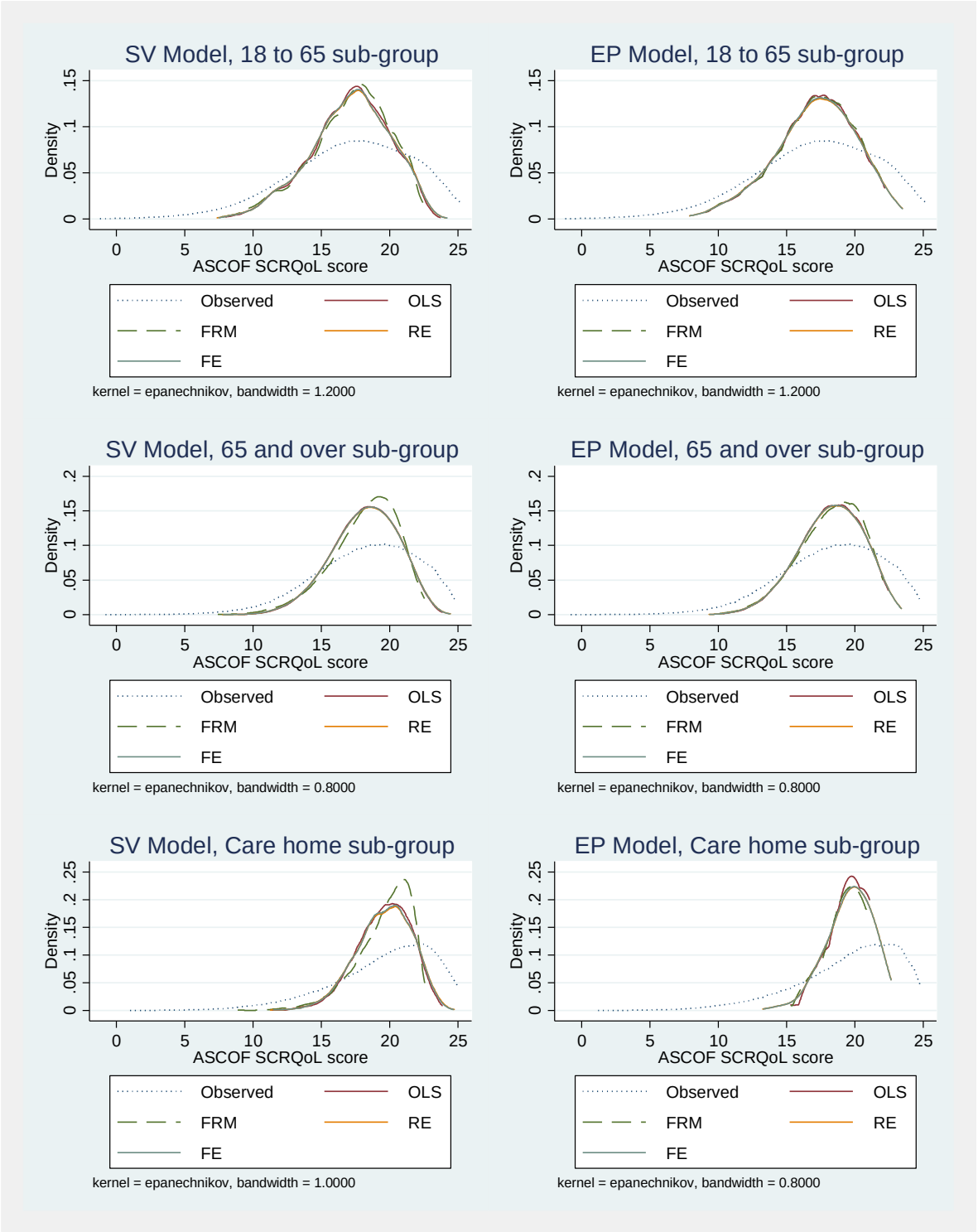

Figure 28: Distribution of observed and predicted scores from OLS, FR, RE and FE regressions

Discussion

This analysis sought to answer the question of whether modelling decisions matter for explaining non-performance-related heterogeneity in ASC outcomes. I have used data from the ASCS to explore this question, and have focused the modelling on the SCRQoL indicator from the ASCOF. I estimated the base model by OLS and used two different specifications for the risk-adjustors. I also considered three alternative estimators: (i) FR regression, which has distributional assumptions that are a better fit for the SCRQoL indicator, and (ii) a RE and (iii) a FE regression, which both address the clustering of the data by CASSR. Where models produce similar conclusions, I would expect model coefficients to be very similar (models will have high convergent validity) and for the predictive validity of models to be similar. In this analysis I found high convergent validity between the four regression estimators and only very marginal gains in predictive validity from the alternatives to the OLS. By contrast the specification of the risk-adjustors had a more noticeable effect on R^2, particularly for the care home sub-group where the statistically-driven approach seemed to provide a poor model. There were also some differences across sub-groups in the fit of the different regressions and in the predictors retained in the models. I discuss these findings in more detail below, and conclude by considering the implications for risk-adjusting the SCRQoL PI. The policy and wider implications of this analysis are considered in Chapter 9.

The effect of different modelling choices on the estimation of the SCRQoL outcome

The regression method used did not have a very large effect on the model estimates. There was very high convergent validity between the models, in terms of the significance, direction and value of the coefficients. There were in some cases some very small gains in predictive validity from alternative regression methods, but the small differences imply that the OLS provides unbiased results for these data.

The similarity in the coefficients obtained from the OLS, FE and RE models is likely to be a consequence of the fairly small amount of variation in SCRQoL scores explained by differences between CASSRs. Nevertheless, the small differences in coefficients between the RE, FE and OLS models were in general in the direction anticipated, assuming that the small CASSR effect is picking up systematic differences in the intensity of provision across areas. Since the risk factors are included as predictors of social care need, I would expect them to be strongly correlated with the intensity of provision, such that greater need would result in more intense services. Under such

circumstances inclusion of a CASSR effect will then lead to more negative coefficients. This is indeed the case for the ADL score variable, but is not the case for some of the variables capturing perceived health. The anxiety and depression variable consistently defied expectations and had a more positive coefficient in the FE and RE models. This finding suggests that within CASSRs worse average levels of anxiety and depression are associated with lower average SCRQoL scores. The correlation between health and social care outcomes implies that the health variables are behaving as outcome variables rather than need variables. I consider this relationship further in Chapter 7 and discuss the implications of this for case-mix adjustment of the PIs in Chapter 9, where I also discuss the implications of the small CASSR effect on SCRQoL for quality improvement policies.

I used the theoretical framework in Chapter 2 to select variables for inclusion in the models. The variable specification seemed to be important for predictive validity and more important than the choice of regression model, which concurs with other studies (Hannan et al. 1997, Huang et al. 2005, Mukamel et al. 2008, Li et al. 2009, Iezzoni 2013). For all sub-groups, the theoretically-driven (SV) methods which produced a specification with more risk adjustors was preferred to the statistically-driven and more parsimonious (EP) method, explaining more of the variation in SCRQoL and producing predictions with less error. For the care home sub-group the statistically-driven method produced a poorly-specified model with limited explanatory power. This suggests that theoretically-driven methods may be more appropriate for developing risk-adjustment models.

Although less useful for developing a final specification for risk adjustment, the statistically-driven EP models point to differences in the explanatory power of the risk-adjustors for the different sub-groups. This is because variables were selected based on their predictive power and the extent to which they varied between CASSRs across the sub-groups. For the private household sub-groups, the variables capturing underlying conditions and the immediate environment have the greatest explanatory power; whereas for the care home sub-group it is the resources variable capturing receipt of practical help from someone else in the care home and the anxiety and depression variable from the underlying conditions category that are most important. Differences in the importance of risk-adjustors for different sub-groups suggest that there is a benefit in analysing the data by sub-group.

Limitations of the modelling and directions for future research

Although the models had high convergent validity, it is, nevertheless, important to acknowledge that none of the regression models provided a good fit to the data. In all cases, goodness of fit statistics suggested problems with non-linearity and problems with predicting the extremes, particularly the tail of the distribution, even in the non-linear fractional response regression. Using the available variables, I attempted to improve the fit by introducing polynomial terms and derived needs variables, but none of these attempts was successful. Consequently there will be large errors for cases with poor SCRQoL.

The extent to which this is a problem depends on whether one believes that the SCRQoL of these people in the tail of the distribution could be radically improved, such that the predicted outcome is realistic. If the cases in the tail region are people who are receiving genuinely poor quality services then the predicted outcome may be fairly realistic. If, however, the underlying distribution (assuming all people were receiving optimal care) is truly skewed, then the expected outcome for cases in the tail region is likely to be unrealistic. It is possible to think of situations in which this latter scenario may be the case. For example, social care services are likely to be unable to compensate people who are in very poor health, perhaps close to death, for loss of functional ability. For this group of people the model will provide a poor and upwardly biased prediction of their SCRQoL (Ara et al. 2013). This limitation of the models has implications for the methods used for adjustment, which I discuss in Chapter 8, and for policy as I discuss in Chapter 9. It is also an area that deserves further research.

As with all modelling of this kind, I was limited by the available indicators in the dataset. It is possible that the misspecification is due to omitted variable bias. One possible source of this is the omitted indicator of service intensity. I explore whether the inclusion of a term for this variable improves the models in Chapter 7. Other unobserved variables that are potentially relevant are health conditions, which have been shown to be important for the estimation of SCRQoL in a recent study (Forder et al. 2016). Personal characteristics and motivations identified in Chapter 2 as relevant, such as self-care abilities are also unobserved. These have been shown to be related to health outcomes (Greene and Hibbard 2012) and theoretically should be important for social care outcomes given the high levels of co-production (Knapp 1984, Netten and Davies 1990). To the extent that these unobserved variables are also correlated with the observed variables, the estimates will be biased, with the direction of the bias depending on the direction of the correlation.

A further limitation was the fairly small numbers in two of the sub-groups. It is generally considered good practice in risk-adjustment to develop models on a sub-sample of the data and then use another sample to validate the models (Ash et al. 2013). This strategy guards against over-fitting of the model to the peculiarities of a particular dataset. It also means risk-adjustment models have more validity when applied to different samples, which is important given that generally the model will be applied to many future waves of the data to generate adjusted-PIs. This strategy, however, requires large numbers to implement. While it would have been possible for the 65 and over sub-group, the other sub-groups were too small to split the sample. A dataset combining multiple waves of the ASCS would probably have sample sizes large enough to split the data into a development and validation sample for the PSD sub-groups (and possibly other sub-groups). A combined dataset would also have the benefit of making models more robust to differences in the relationships between variables between survey waves. Future studies could explore re-estimating the models developed here on multiple waves of the ASCS in order to validate the models.

Concluding remarks

This analysis has shown that the five sets of quasi-inputs predicted by the modified POW framework are important for explaining heterogeneity in SCRQoL outcomes. Given the risk-adjustment models explain a sizeable amount of the variation in SCRQoL outcomes, risk-adjustment is likely to make a difference to PI scores and inferences about performance. It is less clear how important the different modelling choices explored here will be for performance assessment. The choice of regression model did not substantially alter predictions from the risk-adjustment models. Depending on the sub-group, there were small gains from using either the FR, FE or RE model over an OLS. The model specification made a more important difference to the predictive ability of the models, with models using the more statistically-driven specification that had fewer risk-adjustors having less predictive validity than models based on the more theoretically-driven specification that had more risk-adjustors. In this chapter I have examined the effect of modelling choices on model parameters, case-level predictions and model fit statistics. The real test of whether modelling choices matter for risk-adjustment is in whether they affect inferences about performance. I explore this question in Chapter 8.

I consider the policy and wider research implications of this analysis in Chapter 9, after I have explored the effect of modelling choices on performance assessment in

Chapter 8. It is important for subsequent discussions to highlight a few of the points raised in this discussion. First, for these data the analysis suggests that sub-group analysis was beneficial for developing risk-adjustment models, as there was some variation across the sub-groups in the distribution of the risk-adjustors and in their ability to predict SCRQoL. This has implications for the adjustment of PIs for non-performance-related sources of heterogeneity, which I pick up on in Chapter 8 and discuss in more detail in Chapter 9. Second, the specification problems identified for these models and the poor prediction of the extremes of the distribution are key areas for future research, because of the implications they have for risk-adjustment. To some extent it may be possible to improve on some of the analyses conducted here with recent versions of the ASCS. Changes to data collections in 2014-15 mean that the more recent surveys have more detailed information about health conditions that may help to address potential omitted variable bias. In the next chapter, I explore the possibility that bias arising from the correlation between unobserved service intensity and the risk-adjustors is a cause of the misspecification.

Chapter 7

What is the Most Appropriate Method for Modelling Adult Social Care Outcomes? Part Two: A Comparison between Production Function Models and Risk-Adjustment Models for Explaining Heterogeneity in Outcomes

Abstract

Objective: To compare the validity and model fit of a risk-adjustment model and a production function model for the self-reported social care outcome indicator.

Design: Cross-sectional survey data that captured self-reported SCRQoL, an indicator of the intensity of social care provision and a variety of risk-adjustors were analysed using a production function and a risk-adjustment modelling approach. Both approaches were estimated via (i) OLS regression, and (ii) random and fixed-effects regression. Two different specifications were estimated: one with a set of theoretically-chosen but statistically significant risk-adjustors, and a second more parsimonious statistically-driven specification. Model parameters and fit statistics were compared to assess the effect of modelling decisions.

Participants: This analysis uses the 4,201 respondents to the ASCS from the physically and sensorily disabled (PSD) client group aged between 18 and 64 living in private households with an available indicator of the intensity of social care provision.

Outcome measures: The SCRQoL indicator, based on the ASCOT-SCT4 measure.

Results: The predictive ability and model fit of the risk-adjustment and production function models were very similar. The theoretically-chosen specification produced a more valid production function model. Including indicators of functional ability in the specification seemed important for validity of the model. Including indicators of self-perceived health and self-perceived home design as covariates may underestimate the effect of social care services on SCRQoL.

Conclusions: From a statistical perspective there is no clear reason to prefer the production function to the risk-adjustment model for explaining heterogeneity in SCRQoL outcomes. The production function model, however, does have greater face validity and it seems likely that these results are biased by the fact that the budget variable available from the ASCS is a very noisy indicator of service intensity. The production function approach requires data that is difficult to collect, but it would seem unwise to abandon it on the basis of these findings.

Introduction

A key assumption for unbiased estimation in risk-adjustment models is that the effect of risk factors on outcomes is not confounded by unobserved factors. For certain complex interventions and populations, this assumption is unlikely to hold as the factors that are used to risk adjust outcomes are also likely to influence the type and quantity of the intervention received. Where there is variation in the care package provided to the sample and an indicator of the quantity of care provision is omitted from the estimation, risk-adjustment models are likely to be biased. The extent of the bias and its significance, in terms of its effect on inferences about the performance of organisations, is an empirical question. It is likely to depend on the nature of the population and the intervention. The extent to which the omission of the quantity of provision is a problem for risk-adjustment models can be explored by comparing the results from regression models estimated including and then excluding the effect of services. In other words, it can be explored by comparing the results estimated using a production function model to those estimated using a risk-adjustment model.

Social care interventions are exactly the type of interventions in which service quantity is likely to confound the relationship between risk factors and outcomes. As I explained in Chapters 2 and 3, social workers assess a person's social care need on the basis of factors such as their functional ability, availability of informal care and health conditions, which are independently related to their quality of life. They then allocate a care package (or in modern parlance 'personal budget') to the individual on the basis of this assessment, with the aim of matching the person's needs and preferences to the care package so as to maximise their outcomes. Care packages will vary according to the types of interventions included (e.g. day centre sessions, home care, meals services) and according to the frequency and intensity of each of the interventions. Those with the greatest assessed needs should receive the most frequent, most intense and therefore most expensive interventions. Since the variables used to assess need are the same as those used to risk-adjust outcomes, parameter estimates for the risk-adjustors are likely to be biased – they will be correlated with the type and level of intervention received, with the direction of the bias depending on the relationship between the outcome, services and each risk adjustor. Including the service effect in the model should help to reduce the bias due to confounding in the estimates for the risk adjustors (Knapp 1984, Davies et al. 2000b, Fernández and Knapp 2004, Fernández 2005, Forder and Caiels 2011b, Forder et al. 2014). Any differences in the estimates between the risk-adjustment and the production function models can be interpreted as bias due to the confounding effect of service quantity.

In this chapter, I compare a risk-adjustment and production function approach to statistically modelling self-reported outcome measures to understand which model is better for explaining heterogeneity in survey-based indicators. Motivating this analysis is the same trade-off that was behind the consideration of various alternative modelling choices for risk-adjustment in Chapter 6; that is the production function models have greater face validity, but since they are more complicated to estimate than the risk-adjustment models and therefore understand they may be less acceptable to practitioners and other end-users of PIs. They may also be less feasible as they have more data requirements than risk-adjustment models. I reflect on this concern only briefly in the discussion to this chapter, but give it fuller consideration in Chapter 9.

This chapter proceeds first with a description of the statistical analysis used to estimate the production function models and some discussion of the choice of indicator for service type and intensity. Since I use the same approach to estimating the risk-adjustment models as in Chapter 6, i.e. the same variable specification, regression methods, methods for model comparison and the treatment of missing data, these subjects are only mentioned briefly. I then present the results of the modelling, but begin first with a description of the estimation sample. As in Chapter 6, I compare the models in terms of their construct, convergent and predictive validity, following the work of Li et al. (2009). I end with a discussion of the results, in which I focus on the key findings and limitations of the analysis. The policy and wider implications are discussed in Chapter 9.

Data and statistical analysis

As in the previous chapter, the data for this analysis comes from the 2010-11 ASCS. I use the questionnaire as well as the auxiliary data from the records systems of CASSRs. I use a more restricted sample in this analysis compared to the previous chapter due to the problems with the budget variable, which captures the recorded budget allocated to the person on the basis of an assessment of their needs[60]. I use this variable as a proxy measure for resource inputs in the production function models, on the basis that the cost of the package is related to the type and intensity of provision. In effect the budget variable is a cost-weighted utilisation measure, capturing the intensity of all the (recorded) service inputs to the care package. It is

[60] Importantly I have not included 'and means' in this definition, since the intention was for councils to provide the budget based on needs, i.e. the gross budget, not the contribution paid by the council (or indeed user) on the basis of a further means test.

critical to the production function analysis, since it provides an estimate for x_i, the quantity of care provided.

It is, however, extremely poorly reported, being missing for approximately half the sample and entirely missing for a large number of CASSRs. It was also very inconsistently recorded across CASSRs, with the following problems: (i) omission of certain types of services from the overall budget (notably services for mental health clients), with variations in what is omitted across authorities; (ii) poor or inconsistent recording of budgets associated with certain types of services across authorities (especially equipment services); (iii) incomplete information about the budget, where the individual receives funding from other sources (e.g. the independent living fund or the NHS); and (iv) differences between authorities in whether both the user and authority contributions to the overall budget are included, combined with a lack of clarity over which portion of the budget (the user or authority paid portion) was recorded. Since it was not clear what assumptions CASSRs had made in reporting these data, and it was not possible to make any adjustments for these assumptions, I limited this analysis to the a subgroup of users for whom I have more confidence in the data. The subgroup is PSD users living in private households, from the 18 to 64 age group. User contributions are fairly low for adults under 65 and this client group is subject to fewer data problems (The Information Centre for Health and Social Care 2012c).

Due to missingness of the budget data, the sample for the 18 to 64 sub-group is much smaller than in the previous analysis presented in Chapter 6. For this reason I re-estimate all the models to ensure a consistent basis for comparison of the risk-adjustment and production function approaches. I do not repeat the description of the methods used to estimate the risk-adjustment models here and readers are referred to the 'Statistical modelling' section of Chapter 6 for a description of this analysis. For simplicity of presentation I do not consider the fractional response regression for this stage of the work, as it did not seem to have obviously better predictive validity than the other models, and its adds a further level of complexity to the presentation and comparison of results due to differences in the types of fit indices that can be used. All analysis is conducted in Stata 14.

Statistical model for the production function approach

The production function models are based on the conceptual model presented in Chapter 2, as specified in equation (7) in Chapter 3. The production function model attempts to disentangle the effect of intensity from the effect of service user characteristics on SCRQoL, so addressing bias in the estimation of the coefficients for risk-adjustors due to the

confounding effect of service quantity. The production function model also has the added benefit of providing a direct estimate of the effectiveness of services through the coefficient on the service quantity term (minus the quality effect which is relegated to the error term).

I conduct a statistical analysis to estimate the relationship between service quantity and the SCRQoL indicator in the presence of risk factors. The statistical model estimated by OLS is given by,

$$y_i = \alpha + \beta_1 \ln(x_i) + \boldsymbol{\beta_2 z_i^0} + \boldsymbol{\beta_3 z_i^1} + \boldsymbol{\beta_3 z_i^2} + \boldsymbol{\beta_4 z_i^3} + \boldsymbol{\beta_5 z_i^4} + e_i, \qquad (17)$$

where, as in Chapter 6, y_i is the observed outcome SCRQoL, $\ln(x_i)$ captures the effect of service quantity on outcomes through a logarithmic relationship, $\boldsymbol{z_i^0}$ is a vector of variables capturing personal characteristics and motivations, $\boldsymbol{z_i^1}$ is a vector of variables representing underlying health and disabling conditions, $\boldsymbol{z_i^2}$ is a vector of variables representing the immediate physical environment, $\boldsymbol{z_i^3}$ is a vector of variables capturing additional resources arising from social and economic capital, $\boldsymbol{z_i^4}$ is a vector of variables capturing reporting-related factors related to the postal method of survey administration, and the $\boldsymbol{\beta}$s, represent the direct effect of these risk factors on the observed outcome. The error term is given by ε_i.

Although I tested for the inclusion of user group marginal productivity terms, none of these interactions (between the risk factors and $\ln(x_i)$), were significant. For clarity, I therefore have not included them in equation (17). I also tested for alternative relationships for the service quantity term, including quadratic, cubic and square root terms, but these were not significant or not more strongly estimated than the log term so I use the more familiar log relationship. In addition, I tested for different methods for controlling for price differences between councils that may affect the relationship between the cost-weighted intensity variable and outcomes. The best model included the Area Cost Adjustment index (Department for Communities and Local Government 2010) for adults aged 18 to 64 as a set of dummy variable, but I also tested mean and median wages for the sector and experimented with these variables as divisors for the budget variable. It should be noted that while there is no direct measure of quality, as intensity is captured by the local cost of care package, to the extent that it is a factor of cost, it will be included in the effect of $\ln(x_i)$.

For OLS estimates to be unbiased, the assumption that observations are independent must hold. Since individuals are clustered within CASSRs, and it is expected that there is correlation in outcomes between individuals within a given CASSR due to a common quality

experience, the independence assumption is unlikely to hold. The CASSR-specific problems with the reporting of the budget variable may also contribute to a clustering in these relationships. As in Chapter 6, to address this I therefore estimate a RE and FE regression. The statistical model for the RE and FE estimation is similar to that in (17), but with an additional parameter, u_j, which captures the CASSR-level clustering, as follows,

$$y_{ij} = \alpha + \beta_1 \ln(x_i) + \boldsymbol{\beta_2 z_{ij}^0} + \boldsymbol{\beta_3 z_{ij}^1} + \boldsymbol{\beta_3 z_{ij}^2} + \boldsymbol{\beta_4 z_{ij}^3} + \boldsymbol{\beta_5 z_{ij}^4} + e_{ij} + u_j, \qquad (18)$$

where the additional subscript j denotes the CASSR. In the RE regression, u_j are random effects that are assumed to follow a normal distribution, i.e. $u_j \sim N(0, \sigma_u^2)$. They are also assumed to be uncorrelated with the risk factors $\boldsymbol{z_{ij}^0, z_{ij}^1, z_{ij}^2, z_{ij}^3, z_{ij}^4}$ and $\ln(x_i)$ (DeLong et al. 1997, Clarke et al. 2015). By contrast the FE model makes no assumptions about the distribution of u_j and is able to deal with situations where the organisational effect correlates with the risk factors and the service effect (Greene 2012), as is likely in the case of the production function model where council policies might be expected to explain variations in service intensity beyond that captured by differences in geographical economic conditions. The FE model excludes the area cost adjustment indicator since this is an LA-level variable.

I also explored the possibility of estimation by stochastic frontier analysis but tests showed that it was not appropriate for the SCRQoL indicator[61].

Selection of risk adjustors

The models above specify five groups of risk factors derived from the literature reviewed in Chapter 2, in addition to the service intensity effect. For consistency with the analysis in Chapter 6 I use the same two specifications for the risk adjustors (see Table 22). The first specification includes the list of theoretically-relevant indicators in Chapter 6that were found to be significant predictors of SCRQoL for this sub-group (SV model). Since the sample used here is smaller a number of these risk adjustors were no longer significant (p<.1), so I also include a simplified version of this specification (simplified SV model). The second

[61] Tests for negative skewness of the OLS residuals, as implemented with the Stata sktest routine (D'Agostino et al. 1990, Royston 1991), were not significant (skewness=-0.03, p=0.449); and plots show that the residuals appear normally distributed. The null hypothesis for the likelihood-ratio test that the efficiency error component, σ_u^2=0, gives X^2 values of 0.756 for the half-normal, 1.028 for the truncated normal, and 0.274 for the exponential model. Since all of these values are <1.642, these are not significant at the p<.1 level (Kodde and Palm 1986).

specification is statisticall-driven, and includes only those risk adjustors with high explanatory power (EP model).

Comparing statistical models

I compare the models using the concepts of construct validity, convergent validity and predictive validity from Li et al. (2009). To explore construct and convergent validity, I present the tables of estimated model coefficients to both check that the effects are in the expected direction and that the estimates are fairly similar across the models. I also explore whether the coefficients for the adjustor variables change in the expected direction after the addition of the organisational effect or the cost-weighted utilisation measure, i.e. that where higher values of the variable are expected to predict the use of more intensive services the coefficient becomes more negative and vice-versa). To explore predictive validity, I am most interested in the amount of variation explained by the models, for which estimates of R^2 and RMSE are the most important measures[62]. Given that the rationale for comparing multiple models is to explore whether the various models provide a better fit to the data than the OLS, I also look at a range of other goodness of fit statistics as per Chapter 6, i.e. the Akaike Information Criterion (AIC) (Akaike 1974) and Bayesian Information Criterion (BIC) (Schwarz 1978), which balance model fit with model complexity. I take differences of greater than ten as strong evidence of the superiority of the model with the smaller BIC or AIC value over the one with the larger value (Raftery 1995, Burnham and Anderson 2004). For the RE model, I report the estimate of Rho (the intraclass correlation coefficient), which provides a measure of amount of variation explained by the CASSR clustering variable for the RE model. Along with the Hausman test for fixed and random-effects this provides useful information for evaluating the fit of the RE model (Greene 2012).

I also compare predictions from the models across the range of observed values for the outcome indicator. Predicted values are calculated as the expected value for each person within the sub-group, given the values of the covariates and including the intensity and CASSR effect, where relevant, i.e. $E\left[y_{ij}|x_{ij}, z_{ij}, u_j\right]$. This differs from how the predicted

[62] For the FE model I report the adjusted R^2 from the version of the FE model that includes CASSRs as dummy variables. The estimation of R^2 in random-effects models is an area of active research, and there is no agreed upon method for estimating R^2 (Gelman and Pardoe 2006). Kreft and de Leeuw (1998) advise against using R^2 because it lacks a clear meaning. In all the models RMSE is calculated from the square root of the mean difference between the predicted and observed values. Predicted values are given by the expected value for each person within the sub-group, given the values of the covariates and including the CASSR effect, i.e. $E(y_{ij}|z_{ij}, u_j)$. Error scores are calculated as the deviation of the predicted from the observed value, i.e. $y_{ij} - E(y_{ij}|z_{ij}, u_j)$.

values are estimated to generate the performance indicators, considered in Chapter 8 where the service and organisational effects are set to zero. Error scores are calculated as the deviation of the predicted from the observed value, i.e. $y_{ij} - E[y_{ij}|x_{ij}, z_{ij}, u_j]$.

Missing data

The questionnaire and auxiliary data used as covariates in the adjustment models are both subject to item nonresponse. To avoid problems of bias, I attempt to recover the missing data on the characteristics of the full sample using chained equations multiple imputation (van Buuren 2007, White et al. 2011). I generate a separate multiply-imputed dataset for this analysis that excludes the 66 CASSRs where the budget data is completely missing, as in these cases there is no information on which to impute the budget data. Appendix 14 provides more details of the imputation models. Following guidance by White et al. (2011) 30 imputations were carried out, after a burn-in period of 30 iterations, which appeared sufficient for achieving convergence of the chain to a stationary distribution. (See Appendix 15 for details of the sufficiency of this number of imputations.)

After imputation, I combined the individual SCRQoL components into an overall SCRQoL score and also generated all other multi-item variables required for the analysis, such as the indicator of difficulties with ADLs. All models are re-estimated on each of the imputed datasets and the estimates obtained from each analysis are combined using Rubin's rules (1987). Missing cases for the dependent variables are excluded from the estimation of these models as they would increase the variance without adding any further information (McKee et al. 1999). The results from these models estimated on the MI dataset were not very different from those estimated using the complete cases. For the sake of clarity and parsimony, I therefore only present the results from the models estimated on the complete cases below. A comparison between the models estimated on the complete cases and the multiply-imputed dataset is given in Appendix 16.

Results

Sample descriptives

The subsample used in this analysis consists of data from only the 83 LAs (56 per cent of the total) that provided budget data. In total there are 4,201 respondents within the 18 to 64 subgroup from these CASSRs, but of these 4,201 respondents 24 per cent of cases are missing budget data (see Table 32), which is a fairly large loss of data.

Data is also missing for the risk adjustors as shown in Table 32. For most of the risk adjustors, the fraction of missing cases is fairly small and in all instances it is less than ten per cent. Combined with the loss of data for the budget variable, however, were all of the variables in Table 32 to be included in the adjustment models, a much larger fraction of the sample would be lost – a total of 1,604 cases (38 per cent). Since the models use a selection of the variables shown in Table 32, and there is also missingness associated with the SCRQoL indicator as shown, the actual loss of respondent data is between 33 to 40 percent depending on the covariates selected.

Table 32: Distributional statistics for the budget and risk-adjustor variables for the subsample analysis (n=4,201)

Variable	N (% missing)	Mean (SE)
MALE	4,199 (0.05)	0.41 (0.008)
AGE_GP	4,201 (0)	4.08 (0.017)
VP	4,201 (0)	0.04 (0.003)
ADLDIFF[a]	3,913 (6.86)	3.7 (0.04)
ADLCANT[b]	3,913 (6.86)	1.48 (0.034)
SPH	4,114 (2.07)	2.17 (0.016)
PAIN	4,066 (3.21)	1.15 (0.011)
ANXDEP	4,050 (3.59)	0.81 (0.011)
FINANCE	4,060 (3.36)	1.05 (0.013)
SPHD	4,070 (3.12)	1.03 (0.014)
PH, in home	4,054 (3.5)	0.5 (0.008)
PH, out home	4,054 (3.5)	0.37 (0.008)
PH, none	4,054 (3.5)	0.23 (0.007)
SELFCOMP	3,926 (6.55)	0.45 (0.008)
PROXY	3,926 (6.55)	0.02 (0.002)
ASSIST, read	3,926 (6.55)	0.32 (0.007)
ASSIST, translate	3,926 (6.55)	0.09 (0.004)
ASSIST, write	3,926 (6.55)	0.29 (0.007)
ASSIST, discuss	3,926 (6.55)	0.23 (0.007)
ASSIST, in home	3,908 (6.97)	0.23 (0.007)
ASSIST, out home	3,908 (6.97)	0.17 (0.006)
ASSIST, care worker	3,908 (6.97)	0.11 (0.005)
PRIV, own	3,993 (4.95)	0.24 (0.007)
PRIV, fam	3,993 (4.95)	0.06 (0.004)
PRIV, none	3,993 (4.95)	0.71 (0.007)
BUDGET (£s logged)[c]	3,208 (23.6)	8.37 (0.030)
SCRQoL[d]	3,828 (8.88)	17.02 (0.072)

Legend: [a] median=4, minimum=0, maximum=7; [b] median=0, minimum=0, maximum=7; [c] log to base e of annual budget for which median= 8.71, minimum=0, maximum=12.37; [d] median=17, minimum=0, maximum=24.

Estimates for the risk-adjustment and production function models

I now compare the coefficients from the models estimated by the risk-adjustment method against those estimated by the production function method, using OLS, RE and FE regression techniques. The data are presented by specification, i.e. whether the set of risk adjustors are the theoretically-driven specification (SV model and simplified SV model), , or whether they are statistically-driven (EP model). All of the OLS models failed the link test and Ramsey's (1969) Reset as well as tests for normality of the error terms and homoscedasticity (confirmed on visual inspection of the data). Likewise all of the FE models fail the modified Wald test for groupwise heteroscedasticity, as implemented via the Stata routine xttest3 (Baum 2000, Greene 2012). All results for these models therefore use the Huber-White sandwich estimator to correct standard errors (Huber 1967b, White 1980).

The theoretically-driven production function models seem to have good construct validity (Table 33 and Table 34, respectively). The logged budget variable has a positive sign and is significant, and all the risk adjustors have the expected signs and are consistent with the risk-adjustment model, suggesting good convergent validity. By contrast, the statistically driven production function models (EP model) have poor construct validity (see Table 35). The budget variable is not significant, although it has a positive sign. The lack of significance of the budget variable, seems to be explained by the exclusion of the ADL score (ADLDIFF) and the IADL measure of ability to undertake finances and paperwork (FINANCES). When these two variables are included in the models (as they are in theoretically-driven models) the budget variable becomes significant.

The hypothesised correlation between service quantity and the various risk adjustors is illustrated by comparing the coefficients of the risk adjustors in the production function and risk-adjustment models. As expected, the ADL score (ADLDIFF) and the IADL measure of ability to undertake finances and paperwork (FINANCES) have coefficients that are more negative in the production function models than in the risk-adjustment models, reflecting the positive relationship between need and service quantity. Equally, as expected, the indicator of informal care (PH, in home) has a coefficient that is more positive in the production function models than the risk-adjustment models, reflecting the fact that receipt of practical help is generally negatively associated with service quantity – people with practical help receive less services all else being equal. Unexpectedly, however, the health-related variables (ANXDEP and SPH) and self-perceived home design (SPHD) all have coefficients that are more positive in the production function models than in the risk-adjustment models.

Table 33: Estimates for SCRQoL indicator risk adjustment and production function models, SV model (n=2,516)

Variable	Risk-adjustment models			Production function models		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
Budget (ln)	n/a	n/a	n/a	0.098* (0.042)	0.098* (0.042)	0.094* (0.046)
Area cost adjustment	n/a	n/a	n/a	-2.853** (1.063)	-2.887** (1.097)	n/a
SPH: good [a]	-1.143*** (0.249)	-1.160*** (0.318)	-1.247*** (0.263)	-1.097*** (0.249)	-1.109*** (0.318)	-1.217*** (0.261)
SPH: fair [a]	-2.215*** (0.243)	-2.225*** (0.307)	-2.268*** (0.271)	-2.155*** (0.244)	-2.163*** (0.307)	-2.225*** (0.273)
SPH: bad [a]	-3.188*** (0.279)	-3.197*** (0.332)	-3.229*** (0.278)	-3.103*** (0.28)	-3.111*** (0.333)	-3.159*** (0.278)
SPH: very bad [a]	-3.981*** (0.369)	-3.989*** (0.377)	-4.003*** (0.404)	-3.860*** (0.372)	-3.869*** (0.378)	-3.923*** (0.415)
SPHD	-1.023*** (0.163)	-1.035*** (0.157)	-1.067*** (0.163)	-1.002*** (0.162)	-1.012*** (0.157)	-1.066*** (0.163)
ADLDIFF	-0.018 (0.048)	-0.02 (0.048)	-0.028 (0.043)	-0.038 (0.049)	-0.039 (0.049)	-0.047 (0.044)
ADLDIFF*SPHD	-0.102** (0.034)	-0.099** (0.032)	-0.089* (0.035)	-0.100** (0.034)	-0.098** (0.032)	-0.088* (0.035)
FINANCES: diff [b]	-0.622*** (0.182)	-0.625*** (0.184)	-0.647*** (0.177)	-0.631*** (0.181)	-0.632*** (0.184)	-0.651*** (0.178)
FINANCES: can't [b]	-0.742*** (0.183)	-0.751*** (0.186)	-0.805*** (0.181)	-0.794*** (0.184)	-0.796*** (0.186)	-0.835*** (0.181)
ANXDEP	-1.607*** (0.121)	-1.613*** (0.113)	-1.643*** (0.132)	-1.586*** (0.121)	-1.591*** (0.113)	-1.630*** (0.131)
PH in home	0.404* (0.241)	0.407 (0.263)	0.419 (0.26)	0.428* (0.241)	0.43 (0.263)	0.444* (0.261)
PH out home	0.273* (0.148)	0.266* (0.147)	0.236 (0.16)	0.24 (0.148)	0.24 (0.147)	0.239 (0.161)
PH in*ADLDIFF	0.07 (0.054)	0.065 (0.055)	0.05 (0.061)	0.068 (0.054)	0.066 (0.055)	0.054 (0.061)
MALE	-0.171 (0.138)	-0.168 (0.137)	-0.161 (0.163)	-0.16 (0.137)	-0.159 (0.137)	-0.159 (0.162)
ASSIST, care worker	0.689** (0.226)	0.680** (0.214)	0.646* (0.256)	0.639** (0.227)	0.634** (0.214)	0.605* (0.257)
ASSIST, translate	0.291 (0.255)	0.285 (0.24)	0.26 (0.277)	0.32 (0.254)	0.313 (0.24)	0.267 (0.277)
ASSIST, write down	0.105 (0.165)	0.11 (0.163)	0.129 (0.195)	0.076 (0.166)	0.08 (0.163)	0.101 (0.197)
Constant	22.20*** (0.284)	22.24*** (0.346)	22.38*** (0.264)	24.38*** (1.193)	24.45*** (1.255)	21.61*** (0.481)
Random effects						
σ_u		0.283 (0.120)			0.214 (0.142)	
σ_e		3.325 (0.048)			3.321 (0.047)	
Model statistics						
F-stat	96.18***	76.05***	149.04***	87.80***	68.68***	141.40***
R^{2} [c]	0.405		0.428	0.408		0.429

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: can do easily by myself; [c] R^2 for FE calculated from dummy variable model.

Table 34: Estimates for SCRQoL indicator risk adjustment and production function models, Simplified SV model (n=2,517)

Variable	Risk-adjustment models			Production function models		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
Budget (ln)	n/a	n/a	n/a	0.098* (0.042)	0.098* (0.042)	0.094* (0.046)
Area cost adjustment	n/a	n/a	n/a	-3.044** (1.057)	-3.087** (1.11)	n/a
ADLDIFF	-0.070* (0.03)	-0.071* (0.031)	-0.077* (0.03)	-0.090** (0.031)	-0.090** (0.032)	-0.093** (0.031)
FINANCE: diff [a]	-0.583** (0.179)	-0.587** (0.183)	-0.608*** (0.177)	-0.597*** (0.178)	-0.598** (0.183)	-0.616*** (0.178)
FINANCE: can't [a]	-0.592*** (0.166)	-0.606*** (0.173)	-0.672*** (0.164)	-0.654*** (0.168)	-0.659*** (0.174)	-0.710*** (0.164)
SPH	-0.991*** (0.079)	-0.990*** (0.077)	-0.980*** (0.085)	-0.963*** (0.08)	-0.964*** (0.077)	-0.960*** (0.088)
SPHD: most [b]	-1.646*** (0.154)	-1.654*** (0.163)	-1.670*** (0.151)	-1.613*** (0.154)	-1.623*** (0.163)	-1.662*** (0.151)
SPHD: some [b]	-3.012*** (0.2)	-3.012*** (0.195)	-2.981*** (0.182)	-2.932*** (0.201)	-2.940*** (0.196)	-2.963*** (0.18)
SPHD: inappropriate [b]	-4.316*** (0.373)	-4.315*** (0.289)	-4.297*** (0.426)	-4.240*** (0.374)	-4.248*** (0.289)	-4.288*** (0.423)
ANXDEP: moderate [c]	-1.402*** (0.147)	-1.412*** (0.155)	-1.438*** (0.153)	-1.396*** (0.147)	-1.403*** (0.154)	-1.431*** (0.153)
ANXDEP: extreme [c]	-3.306*** (0.273)	-3.318*** (0.236)	-3.371*** (0.294)	-3.251*** (0.274)	-3.263*** (0.237)	-3.340*** (0.293)
PH in home	0.613*** (0.143)	0.600*** (0.142)	0.557*** (0.15)	0.642*** (0.146)	0.635*** (0.144)	0.597*** (0.154)
PROXY	-1.002* (0.436)	-0.953* (0.454)	-0.742 (0.455)	-1.022* (0.44)	-0.988* (0.453)	-0.744 (0.452)
ASSIST, care worker	0.682** (0.218)	0.677** (0.208)	0.663** (0.246)	0.630** (0.219)	0.627** (0.209)	0.614* (0.248)
Constant	22.29*** (0.185)	22.33*** (0.207)	22.39*** (0.184)	24.67*** (1.166)	24.75*** (1.235)	21.61*** (0.448)
Random effects						
σ_u		0.310 (0.114)			0.237 (0.132)	
σ_e		3.331 (0.048)			3.328 (0.048)	
Model statistics						
F-stat	133.46***	106.86***	168.58***	116.66***	93.23***	156.33***
R^2 [d]	0.402		0.427	0.405		0.428

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: meets needs very well; [c] Base category: not anxious or depressed; [d] R^2 for FE calculated from dummy variable model.

Table 35: Estimates for SCRQoL indicator risk adjustment and production function models, EP model (n=2,819)

Variable	Risk-adjustment models			Production function models		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
Budget (ln)	n/a	n/a	n/a	0.045 (0.037)	0.046 (0.038)	0.051 (0.042)
Area cost adjustment	n/a	n/a	n/a	-3.053** (1.001)	-3.079** (1.011)	n/a
SPH: good [a]	-1.258*** (0.228)	-1.267*** (0.303)	-1.348*** (0.225)	-1.222*** (0.228)	-1.228*** (0.303)	-1.331*** (0.227)
SPH: fair [a]	-2.458*** (0.224)	-2.463*** (0.293)	-2.503*** (0.255)	-2.414*** (0.225)	-2.418*** (0.293)	-2.480*** (0.258)
SPH: bad [a]	-3.550*** (0.256)	-3.552*** (0.314)	-3.581*** (0.269)	-3.502*** (0.257)	-3.505*** (0.314)	-3.549*** (0.273)
SPH: very bad [a]	-4.411*** (0.343)	-4.418*** (0.352)	-4.462*** (0.384)	-4.355*** (0.344)	-4.361*** (0.353)	-4.439*** (0.389)
SPHD	-1.286*** (0.141)	-1.288*** (0.133)	-1.291*** (0.125)	-1.282*** (0.141)	-1.284*** (0.133)	-1.296*** (0.125)
ANXDEP: mod [b]	-1.173*** (0.189)	-1.183*** (0.207)	-1.232*** (0.191)	-1.195*** (0.189)	-1.199*** (0.206)	-1.236*** (0.191)
ANXDEP: extreme [b]	-2.659*** (0.399)	-2.673*** (0.336)	-2.737*** (0.452)	-2.665*** (0.401)	-2.671*** (0.336)	-2.728*** (0.453)
SPHD* ANXDEP: mod [c]	-0.203 (0.177)	-0.199 (0.167)	-0.166 (0.16)	-0.181 (0.178)	-0.179 (0.166)	-0.159 (0.159)
SPHD* ANXDEP: ext [c]	-0.637* (0.26)	-0.630** (0.21)	-0.607* (0.275)	-0.602* (0.26)	-0.601** (0.21)	-0.600* (0.275)
Constant	22.21*** (0.196)	22.23*** (0.275)	22.28*** (0.189)	25.00*** (1.106)	25.03*** (1.139)	21.83*** (0.438)
Random effects						
σ_u		0.241 (0.124)			0.157 (0.168)	
σ_e		3.369 (0.045)			3.367 (0.045)	
Model statistics						
F-stat	198.6***	155.45***	237.48***	165.84***	128.13***	213.77***
R^2 [d]	0.395		0.416	0.398		0.416

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: not anxious or depressed; [c] Base category: SPHD*ANXDEP: not anxious or depressed; [d] R^2 for FE calculated from dummy variable model.

Comparison of model fit and predictive validity

In this section I compare the predictive validity of the risk-adjustment and production function models estimated using different regression methods with different covariate selections. After comparing model fit using standard indices, I then look at the distribution of the predicted values over the range of the observed scale for the SCRQoL indicator.

Comparing goodness-of-fit statistics for the risk-adjustment and production function models provides some evidence that the production function models have slightly better predictive validity than the risk-adjustment models (see Table 36). The adjusted R^2, RMSE and AIC statistics are in most cases better for the production function method than the risk-adjustment. It should be noted though that while the AIC difference is usually greater than two (except for the statistically-driven EP models), it is never greater than ten, so there is still reason to continue considering the risk-adjustment models with higher AIC (Burnham and Anderson 2004). Indeed in all cases, the BIC, which penalises more for model complexity, is lower in the risk-adjustment models, suggesting that on this statistic the risk-adjustment models are to be preferred. Again, the BIC differences are small and no greater than 10, so there is no 'very strong' evidence for preferring the risk-adjustment models based on this criterion (Raftery 1995). Interestingly, the differences in the information criteria statistics decrease from OLS to RE to FE, suggesting that modelling the organisational effect may partially account for some of the service intensity effect. This is as would be expected, and is consistent with the smaller differences in adjusted R^2 between the risk-adjusted and production function FE regressions compared to the risk-adjusted and production function OLS regressions.

As I found in the analysis completed on the dataset that included all councils in Chapter 6, the likelihood ratio test for rho is not significant for the risk-adjustment models ($p>0.1$ for EP model and $p>.05$ for the other two models)[63], nor for the production function models[64]. This is consistent with the information criteria statistics, which are higher for the RE compared to the OLS regressions, suggesting that predictive validity is not improved by modelling the organisational effect as a random variable for this indicator. The situation with respect to the FE regression compared to the OLS is more ambiguous. While the F-test for the inclusion of CASSRs as fixed effects is not significant for the risk-adjustment ($p>0.1$ for

[63] SV model: $\chi^2(2)=1.920$, $p=0.080$; Simplified SV model: $\chi^2(2)=2.68$, $p=0.051$; EP model: $\chi^2(2)=1.230$, $p=0.134$.

[64] SV model: $\chi^2(2)=0.690$, $p=0.203$; Simplified SV model: $\chi^2(2)= 1.030$, $p=0.155$; EP model: $\chi^2(2)=0.250$, $p=0.310$.

the EP model and p>0.05 for the other models)[65] and production function (p>0.1 for the EP model and p>0.05 for the other models)[66] models, the adjusted R^2, RMSE, AIC and BIC statistics are all better for the FE model than the OLS. This suggests that the FE model may have better predictive validity than the OLS for this sub-group.

The theoretically-driven models (SV and simplified SV models) explain more of the variation in scores (R^2 is greater) than the statistically-driven model (EP model). There is little difference between the two versions of the theoretically-driven midels, although BIC statistics, which penalise model complexity more, suggest a preference for the simplified version.

Figure 29 illustrates the distribution of the predictions from the risk-adjustment and production function models, over the range of values of the SCRQoL indicator. All of the predictions produce very similar shaped distributions, suggesting that the inclusion of the intensity effect does not improve the modelling from a distributional point of view. This is perhaps not surprising given the distributional assumptions of all the models. The relatively poor performance of all the models in predicting the extremes of the distribution is also shown in Table 98 in Appendix 17. All predictions are also within range for this estimation sample, which was not the case for the larger sample used in Chapter 6.

[65] SV model: F(82,2416)=1.23, p=0.082; Simplified SV model: F(82,2422)=1.27, p=0.053; EP model: F(82,2727)=1.23, p=0.110.
[66] SV model: F(82,2415)=1.24, p=0.095; Simplified SV model: F(82,2421)=1.26, p=0.063; EP model: F(82,2726)=1.20, p=0.109.

Table 36: Model fit statistics for SCRQoL indicator risk adjustment and production function models

	Risk-adjustment models			Production function models		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
Theoretically-driven SV models						
Log-likelihood	-6,602	-6,601	-6,551	-6,595	-6,595	-6,548
F-stat	96.18***	76.05***	149.04***	87.80***	68.68***	141.40***
AIC	13,240	13,242	13,135	13,231	13,234	13,133
BIC	13,345	13,359	13,234	13,347	13,362	13,238
Adjusted R^2	0.401		0.405	0.403		0.406
RMSE	3.349	3.270	3.336	3.341	3.267	3.279
Conditional rho (SE)		0.007 (0.006)			0.004 (0.005)	
Theoretically-driven simplified SV models						
Log-likelihood	-6,611	-6,610	-6,558	-6,604	-6,603	-6,556
F-stat	133.46***	106.86***	168.58***	116.66***	93.23***	156.33***
AIC	13,248	13,249	13,140	13,238	13,241	13,138
BIC	13,324	13,337	13,210	13,325	13,340	13,214
Adjusted R^2	0.399		0.404	0.402		0.405
RMSE	3.346	3.276	3.276	3.336	3.274	3.273
Conditional rho (SE)		0.009 (0.006)			0.005 (0.006)	
Statistically-driven EP models						
Log-likelihood	-7,431	-7,431	-7,381	-7,426	-7,422	-7,381
F-stat	198.6***	155.45***	237.48***	165.84***	128.13***	213.77***
AIC	14,883	14,886	14,781	14,875	14,881	14,781
BIC	14,942	14,957	14,834	14,947	14,988	14,841
Adjusted R^2	0.395		0.416	0.395		0.397
RMSE	3.378	3.319	3.319	3.378	3.318	3.324
Conditional rho (SE)		0.005 (0.005)			0.002 (0.005)	

Legend: *** p<.001

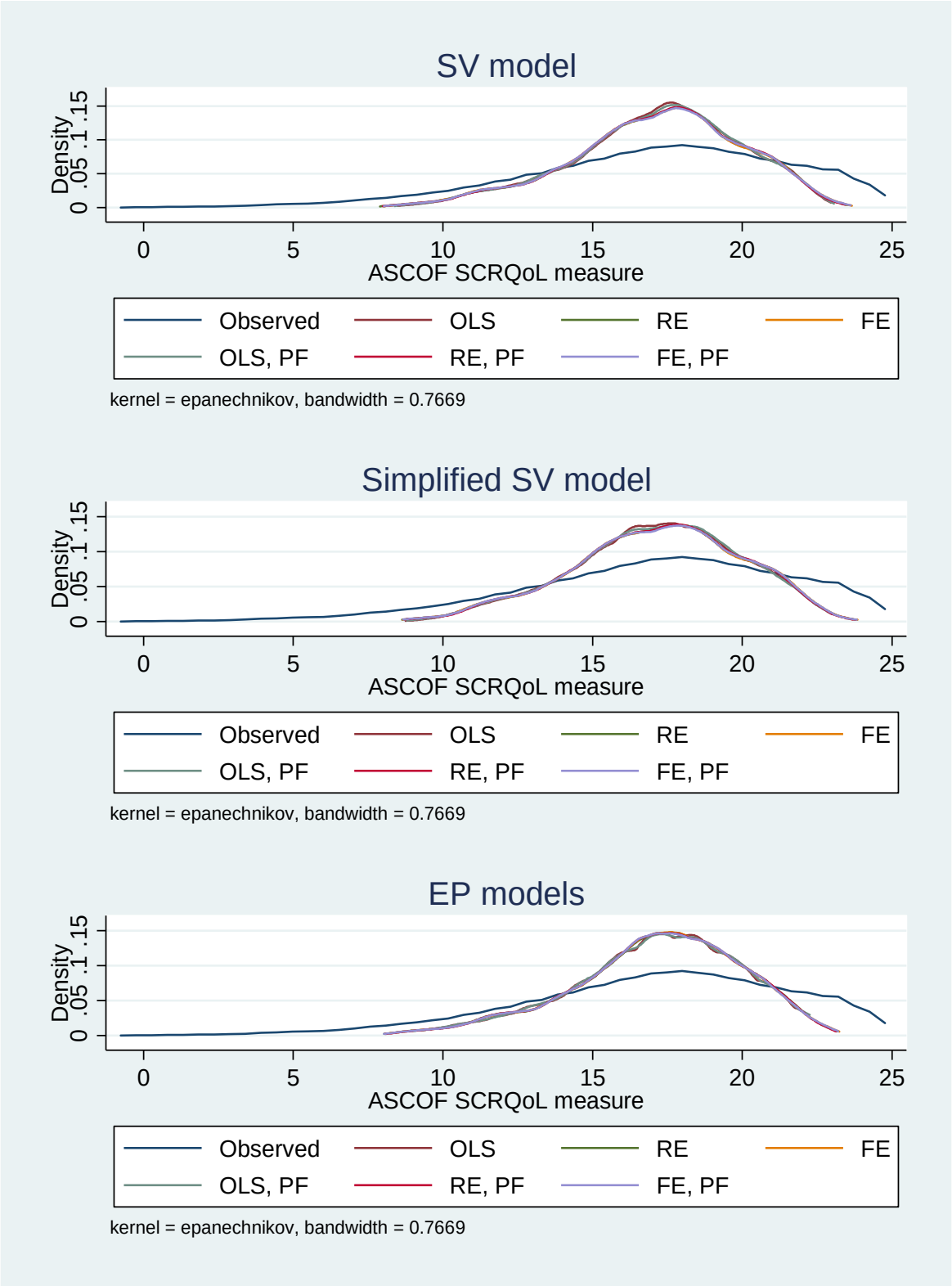

Figure 29: Distribution of predicted and actual scores, for the risk-adjustment and production function (PF) models, SCRQoL indicator

Discussion

In this chapter I have considered how production function models compare to risk-adjustment models for explaining heterogeneity in ASC outcomes. I compared the results by examining the model coefficients, distribution of predictions, predictive ability and model fit. I have used data from the 2010-11 ASCS to explore the question, but in comparison to Chapter 6, here I focused only on the 18 to 64 sub-group due to problems with the indicator of service intensity for the other client groups. As in the previous analysis I estimated the base production function and risk-adjustment models by OLS and used two specifications for the risk adjustors. To address potential bias in OLS estimates caused by clustering of the data by CASSR, I also considered RE and FE regression.

In this analysis I found high convergent validity between all the regression models, including between the production function and risk-adjustment models. There were only very marginal gains in predictive validity from the production function model over the risk-adjustment model and from the RE or FE models over the OLS. The model specification had a similarly fairly small effect on predictive validity, although the theoretically-driven models were marginally preferable. The specification, however, was important for the construct validity of the production function models, with only the theoretically-driven models providing valid models. I discuss these findings in some more detail below, and conclude by considering the implications of these findings for adjusting the SCRQoL PI. The policy and wider implications of this analysis are discussed in Chapter 9.

The effect of using a production function model to estimate SCRQoL

The production function model has marginally better predictive ability than the risk-adjustment model, but there was in general very little difference between the models on the model fit statistics. Although the coefficients for the risk adjustors did differ between the models, in most cases, the differences were small. From a purely statistical perspective there is not much to gain from using a production function model compared to the risk adjustment model to model the SCRQoL outcome.

A benefit of the production function model, however, is that it provides a direct estimate of the effectiveness of social care through the coefficient on the service quantity term. The analysis here showed that model specification was very important for generating a plausible estimate of the effect of social care on SCRQoL. The statistically-driven specification of risk adjustors produced an estimate of the effect of service quantity (the

budget variable) that was not significant. This should be compared to the significant positive estimate for the effect of service quantity on SCRQoL. Comparison of the specification suggests that the indicators of functional ability (ADL and IADL indicators) are important, which would fit with findings from previous studies (Davies et al. 2000a, Fernández 2005, Forder et al. 2014). The production function models require a more careful theoretical approach to selecting risk adjustors.

I expected that the service quantity indicator would be correlated with the risk adjustors that capture needs-related characteristics of service users and that this would be reflected by differences in the coefficients for the risk adjustors between the production function and risk-adjustment models. This was indeed the case and the coefficients of most risk adjustors changed in the expected fashion. Contrary to expectations, however, the variables concerning perceptions of health and home design all had more positive coefficients in the production function models than the risk adjustment models. This suggests that the correlation between the service intensity indicator and these variables is negative, such that poorer health or worse home design predicts on average less intense services other things being equal. Further investigation of this relationship in this dataset and other available datasets shows that for this sub-group the budget variable is negatively correlated with perceptions of health and home design, even after controlling for potential confounding factors.

Although this finding is counterintuitive where these variables are understood to be measures of need, if these variables are instead interpreted as outcomes then the relationships are as expected, i.e. more intense services lead to better self-perceived home design, better self-perceived health, and less anxiety/depression. While services can have clear health outcomes, the relationship with home design is less obvious. Equipment and minor adaptations may provide some part of this explanation; another may be that as a result of services people *perceive* their home design as less problematic. Services compensate for poor home design, which is as one would expect. Despite the possibility that these variables capture outcomes, I have included them in the model because they have a strong negative relationship with SCRQoL. Whether these variables should then be included as risk-adjustors is a matter for debate (Mukamel et al. 2008, Li et al. 2009, Iezzoni 2013), as I discuss in Chapter 9.

Another possibility is that the relationships observed between these variables and service intensity are simply an artefact of the data – a consequence of the fact that service

intensity is measured fairly inaccurately by the budget variable. This may provide a partial explanation for the finding, as I discuss below.

Limitations of the modelling and directions for future research

The most significant limitation of the modelling is the use of the budget variable as an indicator of service intensity. I described the budget variable as a cost-weighted utilisation measure, since it sums together all the components of the care package using the relative unit costs per component as weights. While this is true, since the cost-weighting is calculated on the basis of *local* unit costs the budget variable will reflect a range of factors aside from the intensity of provision. Notably, it will capture differences in the prices local areas choose to pay for social care services.

The price paid by local areas will to some extent the price paid will be a response to local economic conditions, e.g. in terms of the availability of providers, the availability of labour, local land prices and so on, but it will also depend on local commissioning policies and practices. I have included the Area Cost Adjustment, which is used in the resource allocation formulae that allocate central resources to local areas, as a mechanism for adjusting for price differences. The Area Cost Adjustment, however, is an imperfect and fairly insensitive measure for capturing differences in economic conditions (Blanchflower and Oswald 2005) and anyway will not adjust for differences due to the policies of local areas. The prices paid for social care services vary considerably across local areas and do not necessarily reflect local economic conditions. This is illustrated well be a recent report from the UK Homecare Association (2015), which shows wide variations in the price paid per hour of homecare by adjacent London councils where the economic conditions would be expected to be fairly similar. As well as introducing noise into the estimation of the relationship between service quantity and SCRQoL, this limitation of the budget variable is the reason why I have not been able to use any of the instrumental variables approach outlined by Forder et al. (2014, 2016) and discussed in Chapter 3. Both the spatially-lagged service use instrument and dummy variables for CASSRs will also capture pricing policies of the area that as well as being a reason why they will induce variation in the budget variable, may also have a direct effect on SCRQoL. For example, because areas that pay less may have a less happy workforce and less functional organisations. Not using an instrumental variables approach to estimate the production function is clearly a limitation of this analysis and would be expected to produce some bias in the model estimates, as I discussed in Chapter 3.

It is also the case that the budget variable will reflect differences in the allocation of components of care packages between authorities – cheaper packages all other things being equal may reflect substitution of more expensive components with cheaper components. This type of indicator therefore masks important substitution relationships between service forms, as the effect of inputs (service forms) on outcomes are not separately estimated. This may be one reason why I do not find evidence for user group marginal productivity effects as has been found in previous studies (Davies et al. 2000b, Fernández 2005). Another reason may be the inaccuracies in the reporting of the budget variable, which will make it harder to identify the signal from the noise.

Although it is clear that the inaccuracy of the budget variable is an impediment to accurate estimation of the production function models, the extent to which the results obtained here are biased can be gauged by comparing them to those from a recent study related to this thesis. The Identifying the Impact of Adult Social Care (IIASC) study (Forder et al. 2016), was designed to capture better information on service receipt and user characteristics to enable estimation of the production function using the instrumental variables technique discussed in Chapter 3. As well as collecting data from social care records, the IIASC study interviewed service users to capture information on their service use alongside a range of other data on their characteristics and SCRQoL. Since service use was collected directly from users, a cost-weighted utilisation measure[1] could be generated using average national unit costs. This is preferable to the budget variable from social care records that I used in this analysis, as average differences in the cost-weighted utilisation measure between areas are not influenced by area-level differences in the prices paid for social care. The production function models estimated for the IIASC study use the spatially-lagged service use variable described in Chapter 3 as an instrument for service quantity. The results from these instrumental variable models compare favourably to those presented here, with similar patterns to the relationships, and similar effects on performance assessment, as I discuss in Chapter 8. The most important difference is that the effect of service quantity on SCRQoL is greatly underestimated here, with the IIASC study estimating an effect about six times larger (Forder et al. 2016).

A further difference between the IIASC models and the models estimated here is that the difference between the coefficients for the self-perceived health variable from the production function and risk-adjustment models for the IIASC data are in the expected

[1] Sample size did not permit separate investigation of different components of the care package

direction. I examined the relationship between the self-perceived health, depression/anxiety, self-perceived home design variables and service intensity further using the IIASC dataset to see whether the relationships observed in the ASCS data were also present in the IIASC data. In this dataset the same negative raw correlations are found for these risk-adjustors with service intensity, irrespective of whether service intensity is measured using the self-reported cost-weighted utilisation measure or the council-reported budget variable. The raw correlations are, however, not quite as strong for the self-reported utilisation measure compared to the council-reported budget variable. When other user characteristics are taken into account the negative relationship with the self-perceived health and anxiety/depression variables disappears. In other words, when service intensity is measured using the self-reported utilisation variable, all other things being equal self-perceived health and anxiety/depression have no relationship with intensity. When the council-reported variable is used as an indicator of service intensity and other user characteristics are taken into account, only the negative relationship with anxiety/depression disappears. I interpret these findings as showing that the negative relationship between intensity and self-perceived health seems to be a consequence of using the council-reported budget variable; whereas the negative relationship with self-perceived home design reflects a real relationship.

The correlation between these two – council and self-reported – indicators of service intensity is around 0.56, which is not particularly high. This seems to be driven by a sub-group of users for whom this correlation is much lower. Users who received a review in between the date that the auxiliary data (which includes the budget variable) was extracted by the council and the date the survey was conducted have a lower correlation (0.42 for the reviewed group compared to 0.62 for the non-reviewed group) as do those who received an assessment in between the same two dates (0.15 for assessed group and 0.74 for unassessed group). When those who had either an assessment or review between the date of extraction of the auxiliary data and the date of the survey are excluded, the correlation between the two indicators of service intensity increases to 0.87. This provides a possible explanation for the positive relationship between self-perceived health and intensity found using the council-reported indicator and can help to explain the low estimate for the effect of service quantity on SCRQoL compared to the IIASC study. It may be that self-perceived health is an indicator of people whose conditions have deteriorated substantially over the period between data extraction and survey fieldwork and whose services have changed as a consequence of this deterioration. This would upwardly

bias the relationship between service intensity and self-perceived health in the manner observed in the ASCS analysis. It would also introduce further noise into the estimation of the relationship between service intensity and SCRQoL, which is already complicated by the other inaccuracies in the reporting of the budget variable and the differences in policies around prices paid for provision. Small numbers prevent me from examining this hypothesis any further. Nevertheless, they do suggest that there is some bias in the estimates obtained from the production function models here.

Concluding remarks

The production function model provides a direct estimate of the effectiveness of social care and has greater face validity than the risk-adjustment model; however, this analysis did not find strong statistical evidence to prefer the production function model over the risk-adjustment model. Part of the explanation for these results seems likely to lie in the difficulties encountered finding a good indicator of the quantity of social care provided. The budget variable was a poor indicator of the intensity of provision and is likely to have produced a lot of noise in the data. Despite the problems with the budget variable, I was able to estimate a production function model that largely had the expected relationships. The results from these models showed the importance of using a theoretically-driven approach to select risk-adjustors for a production function model to estimate a positive and significant effect of service intensity on SCRQoL. Indicators of functional ability seem to be critical risk adjustors for production function models. Indicators of self-perceived health and perceptions of home design seem in part to capture outcomes from care services. I discuss whether they should be included as risk adjustors in Chapter 9.

Although I have been able to estimate a production function model with the ASCS dataset, the model estimates do appear to be biased when compared to those obtained using a less noisy indicator of the intensity of service provision. A major limitation of the production function method is that it requires detailed data that is difficult to collect. Recent studies (the IIASC study), however, show that it is possible to collect appropriate data and estimate robust production function models using an instrumental variables approach to address endogeneity issues due to the effect of unobservable confounders. Until the ASCS collects data that provides a good indicator of service intensity, this clearly limits the applicability of the production function method for adjusting ASCOF PIs to account for heterogeneity in non-performance-related factors and for exploring social care outcomes more generally. Indeed, the greater simplicity and ease of estimation of the risk-

adjustment model, along with the less demanding data requirements make it in many ways a more attractive option than the production function model. The simplicity and transparency of the risk-adjustment model, however, needs to be balanced against the greater validity of the production function model. If the choice between the risk-adjustment and production function model does not have significant consequences for inferences about performance then the risk-adjustment model may be preferable. I explore this question in Chapter 8, albeit with some caution given the effect of data limitations on the confidence I have in the results presented here.

Chapter 8

What is the Effect of Adjusting for Case-Mix and Nonresponse on Inferences about Performance and Does the Method for Implementing Case-Mix Adjustment Matter?

Abstract

Objective: To explore the effect of adjusting for case-mix on performance assessment and consider whether the method for adjustment makes a difference. Additionally to consider the impact of nonresponse where indicators are case-mix adjusted.

Method: Data from the physically and sensorily disabled (PSD) subsample of the English 2010-11 ASCS are used to explore the effect of adjusting for case-mix on the SCRQoL indicator. Indirect standardisation is used to adjust the outcome indicator for case-mix. Various equations are used to estimate the expected outcome for each CASSR, using the risk-adjustment and production function models estimated in previous chapters. These models use different estimation methods and specifications for the set of risk adjustors included to enable investigation of the effect of method of adjustment on performance assessment. The impact of differences between unadjusted and adjusted SCRQoL scores is assessed through examining changes in rank position and outlier status. Analyses are repeated adjusting for nonresponse.

Results: Case-mix adjustment greatly reduced the number of outliers identified and had an important impact on the rank position of CASSRs. The effects were dependent on sub-group and covariate specification. Statistical estimation methods, including the choice of production function or risk-adjustment model, had little impact on rank position and outliers. Adjusting case-mix adjusted indicators for nonresponse, had in general little effect on the number of outliers identified and the rank position. The effect of adjusting for nonresponse on SCRQoL, however, was driven by bias correction due to patterns of missingness, and not increasing precision. For one CASSR this produced more extreme changes.

Conclusions: Without risk adjustment the SCRQoL indicator is likely to falsely identify a large number of CASSRs as outliers. The set of risk-adjustors included in risk-adjustment models has an important impact on conclusions but other methodological choices were unimportant for these data. The effect of nonresponse should be investigated to explore the effect of patterns of missingness on performance assessment.

Introduction

In ASC, CASSRs may serve different populations as a consequence of social and economic factors that affect where people live or as a consequence of policies affecting eligibility for publicly-funded care. This heterogeneity in the population organisations serve is often referred to as case-mix. Where case-mix characteristics directly impact on outcomes, for example because they make it more likely that someone will have a worse outcome from care, they can confound inferences about performance. It is possible to use statistical methods to adjust for these differences in case-mix. In Chapters 6 and 7, I have explored different methods for modelling social care outcomes measured by the SCRQoL PI. In this chapter I build upon those analyses and explore the effect of adjusting for case-mix on performance assessment.

To adjust for case-mix, I use indirect standardisation. As described in Chapter 3, this method applies the average effects of risk factors on outcomes estimated from a reference population to each organisation. In this thesis, the reference population is the PSD subsample from the 2010-11 ASCS. I carried out the analysis of the average effects of risk factors on the SCRQoL outcome in Chapters 6 and 7. I used a variety of different specifications and regression methods, to reflect as far as possible the important exogenous factors as predicted by my theoretical framework, the clustered structure of the data and the skewed distribution of the indicator. Here I calculate the expected outcome for each CASSR given their case-mix using the results from these statistical models, and use this to standardise the observed outcome for each CASSR. I refer to the 'indirectly' standardised outcome as the adjusted PI. The main aim of this analysis is to ascertain whether adjusting for differences in the case-mix of CASSRs affects performance assessment. A secondary aim is to explore whether the different estimation methods and model specifications have important implications for performance assessment. This analysis can be used to understand the analytical steps required to improve the validity of survey-based PIs.

In this chapter I also explore the effect of nonresponse on performance assessment in the context of case-mix adjustment. Adjusting for case-mix increases the percentage of data lost due to nonresponse, as cases missing data for risk-adjustors will also be excluded from the estimation of PIs. For this reason it is possible that adjusting for nonresponse will have a different impact in the context of case-mix adjustment to its impact on the raw PI scores, as explored in Chapter 5. To explore the effect of nonresponse on performance assessment, I assume that the missingness mechanism is missing-at-random (MAR). This is the same assumption used in Chapter 5, but I have had to adapt the methods because the adjusted PIs

are only available for the sub-sample of respondents with PSD. The aim of this aspect of the analysis is to understand whether adjusting for nonresponse in the context of case-mix adjustment affects performance assessment.

The chapter is organised as follows. I start by discussing the methods I use to calculate the adjusted PIs. I then present the results in two parts. First, I focus on a comparison of the effects of adjustment on performance assessment across the three sub-groups and the PSD sub-sample using the risk-adjustment models from Chapter 6. In the second part, I compare the effects of adjustment using the risk-adjustment and production function models from Chapter 7 on performance assessment. In each section I also look at the effect of case-mix adjustment in the context of adjustment for nonresponse. In the discussion I focus on the key findings and limitations of the study. The implications for policy are discussed in more detail in Chapter 9.

Empirical strategy

I use the 2010-11 ASCS data to explore the effect of adjusting for case-mix and nonresponse on inferences about performance and focus on the ASCOF SCRQoL indicator. The analysis has three steps, of which the first is to generate adjusted PIs under the assumption that the missingness mechanism is missing completely at random (MCAR), i.e. by complete case analysis (CCA). The adjusted PIs are generated from the risk adjustment and production function regression models developed in chapters 6 and 7, respectively. Because of this, they are sub-group specific and limited to the PSD population. They are not available for the full ASCS sample. I therefore report four adjusted PIs for each CASSR: one for each sub-group and an indicator for the PSD subsample.

The second step is to generate PIs that are adjusted for both case-mix and nonresponse. Here, I draw on the analysis from Chapter 5, but because the adjusted PIs are only calculated for the PSD subsample, I adjust only for the effect of item nonresponse on inferences about performance. It would not be appropriate to apply the weights to the sub-groups or PSD subsample as the sum of the weights would depend on the ratio of the size of the sub-group or subsample to the total respondent sample for each CASSR. This may distort the results. I therefore use the CCA/MI method discussed in Chapter 5, which assumes that the missingness mechanism for item nonresponse is MAR and the missingness mechanism for unit nonresponse is MCAR. The analysis for the effects of nonresponse is therefore more limited than that carried out in Chapter 5. The third step is to compare the unadjusted and adjusted PIs. In this step I explore whether performance assessment is affected by adjustment

for case-mix or for nonresponse, and whether the method and specification used has an effect on the results.

Since the methods used for the risk-adjustment and production function models are described in detail in Chapters 6 and 7, respectively, and the method for MI is explained in Chapter 5, I do not repeat those discussions here. In this chapter, I describe only the three steps set out above. All analyses are conducted in Stata 14.

Step 1: generating adjusted PIs under MCAR assumption

To calculate an adjusted PI score, I use indirect standardisation. Specifically I calculate adjustment factors using the 'error' approach and the two 'ratio' approaches, discussed in Chapter 3. These adjustment factors are then scaled by the national average SCRQoL score to give an adjusted PI on the same scale as the unadjusted PI. Since a detailed explanation of these methods is given in Chapter 3 I cover them here only briefly. The 'error' adjustment factor is calculated as the difference between the observed and expected outcome averaged across individuals within a CASSR (see equation (9), Chapter 3). The 'individual ratio' adjustment factor is calculated as the ratio of the observed to the expected outcome averaged across individuals within a CASSR (see equation (8), Chapter 3). The 'average ratio' adjustment factor produces as PI that is more conservative than the individual ratio adjustment factor. In this method, the adjustment factor is calculated as the ratio of the observed outcome averaged across individuals within a CASSR, to the expected outcome, calculated assuming CASSR average characteristics apply (see equation (10), Chapter 3). An indicator for the PSD subsample can also be estimated simply by averaging across the sub-group specific adjustment factors for each individual in the subsample.

The expected outcome is obtained using the results from regression models specified in Chapters 6 and 7. Where the risk-adjustment model is estimated by OLS or by fractional response (FR) regression, the expected outcome is simply the linear prediction[1]. Where the risk-adjustment model is estimated by FE and RE regression methods, the predicted outcome can be calculated assuming a null organisational effect (i.e. that u_j=0), so that differences in organisational effectiveness are not adjusted out of the PI (Li et al. 2009). To the extent that there are differences between CASSRs in the relationship between user characteristics and SCRQoL, the RE- and FE-based PIs will differ from OLS-based PIs.

[1] Converted back to the original scale in the case of FR regression.

Where the production function model is estimated, the intensity effect of services is included as a term in the regression, and therefore excluded from the error term. The linear prediction from the OLS regression (or the linear prediction assuming a null organisational effect for the RE and FE versions), would adjust out the effect of differences in service *quantity* on QoL. This would leave the adjusted PI reflecting only differences in quality (to the extent that quality is not a factor of quantity or, given the use of the budget variable in this analysis, cost). To generate PIs that are more comparable with the risk-adjusted PIs I instead assume a null service intensity effect (i.e. that $\ln(x_i) = 0$). This also means that the predicted outcome can be estimated for cases missing the budget variable, which makes it possible to apply the equation to the whole sample, even those CASSRs that have not provided any budget data. Here, however, I restrict the analysis to the sub-sample of 83 CASSRs with budget data, as it is not clear that the model estimates can be applied to the other CASSRs due to the data limitations discussed in Chapter 7. The production function model will produce different inferences about performance to the extent that there are differences among CASSRs in how they allocate services on the basis of need, i.e. in the relationship between $\boldsymbol{z_{ij}}$ and $\ln(x_i)$ in equation (17) Chapter 7.

Standard errors are produced for all PI estimates by bootstrapping the calculation of the average adjustment factor for each authority. In this instance I use 1000 replications.

Step 2: generating adjusted PIs under the MAR assumption

All of the PIs in step one are estimated by CCA, i.e. assuming that the missingness mechanism is MCAR. In this step I calculate the adjusted PIs assuming that the missingness mechanism for the respondent data is MAR. I use the MI dataset as described in Chapters 6 and 7 for estimation of the adjusted PIs. After imputation, I estimate the analysis models and predictions, as specified in Chapters 6 and 7, on each imputed dataset as set out above. The adjustment factors are calculated on each imputed dataset and combined using Rubin's rules to generate an adjusted PI for each CASSR, for each sub-group level and for the PSD sub-sample. As for the CCA, the PI calculation is bootstrapped to generate standard errors for each authority. Since coefficient estimates vary slightly between the adjustment models estimated by CCA and MI, I also generated predicted outcomes and PIs using the coefficients from the adjustment models estimated by CCA. Since the results are not sensitive to differences in the coefficients between the models estimated by CCA and MI these results are not shown here.

Step 3: comparing performance indicators

To assess the effect of adjusting PIs on inferences about the performance of organisations, I look at whether there are differences between the unadjusted PIs and adjusted PIs, in terms of (i) the ranking of CASSRs and (ii) the identification of outlying organisations. I apply the methods described in Chapter 5 to this analysis. I compare the ranking of CASSRs attained after adjustment to the ranking of CASSRs prior to adjustment, using caterpillar plots and correlation statistics. Specifically I use Pearson's correlation coefficient, Spearman's rank correlation coefficient and Kendall's tau statistic. Using the last of these statistics I can quantify the proportion of CASSR pairs that change order after the adjustment with the formula $100*(1-\tau)/2$ (Johnson et al. 2010).

To examine the effect of case-mix and nonresponse on the identification of outliers, I use a funnel plot, as described in Chapter 3. As in Chapter 5, to understand what is driving changes in the outlier status under different missingness assumptions I decompose the effect of nonresponse on performance assessment into the mean outcome, volume and variability components. I calculate the number of outliers under the different methods for case-mix adjustment and addressing missingness in the data, specifying whether cases are 'in control', high or low 'alerts', or high or low 'alarms' depending on where they lie in relation to the funnel plot control limits (see Chapter 3 for discussion). I also look at the number of 'movements' between the different outlier states. To assess the importance of the changes in the number of movements as a consequence of case-mix and nonresponse adjustment, I use the concept of 'false negatives' and 'false positives', as discussed in Chapter 3. The assumption of this analysis is that the adjusted data reflect 'true' effectiveness, the likelihood of which I reflect on in the discussion.

Results

In this section I present the results of the analysis to examine the effect of adjustment on performance assessment. First I present the results from the risk-adjustment models developed in Chapter 6 on the three sub-groups of the full dataset covering all CASSRs. I then compare the results from the risk-adjustment and production function-adjustment models developed in Chapter 7. Within each sub-section, I begin by looking at the convergent validity of the PIs generated using the different statistical models. I then examine the effect of adjustment on performance assessment.

It is first important to note that there are in fact substantial differences in the proportions of people across the CASSRs in each of the sub-groups explored in this analysis. As Table 37 shows, in one CASSR there are no cases in the care home sub-group. This means that for the care home sub-group the PI is missing for one CASSR. It also means that the adjusted PI for the PSD sub-sample has a different balance of sub-groups. I consider the implications of this imbalance between the sub-groups in the discussion.

Table 37: Proportion of CASSR sample in each population sub-group

	National	Proportion per CASSR		
Sub-group	**average**	**Minimum**	**Maximum**	**Median**
18 to 64	0.12	0.05	0.22	0.12
65 and over	0.49	0.24	0.69	0.49
Care home	0.12	0.00	0.40	0.11
Not PSD	0.27	0.08	0.47	0.26

The effect of risk adjustment using different statistical models on performance assessment

The average risk-adjusted SCRQoL PI scores for all variant regression models and specifications are shown in Table 38 for the 18 to 64 subgroup, Table 39 for the 65 and over sub-group and Table 40 for the care home sub-group. Comparing PI scores across the tables shows differences in the average score across the sub-groups, both before and after risk-adjustment. The care home sub-group report better outcomes than the 65 and over sub-group, who in turn report better outcomes than the 18 to 64 sub-group. Reflecting the dominance of the 65 and over sub-group in the PSD population, the average PI scores for the PSD sub-sample are similar to the 65 and over sub-group (Table 41).

There is high convergent validity between different regression methods at the population level. Inspection of pairwise correlations confirms this for the CASSR level, with Pearson and Spearman rank correlation coefficients in excess of 0.99 for all sub-groups (see Appendix 18, Table 99 to Table 101). There is also high convergent validity between PIs calculated using the error and two ratio methods, with Pearson and Spearman rank correlation coefficients all 0.98 or over (see Appendix 18, Table 102 to Table 104). Since the regression method and method for calculating the adjustment factor make little difference to the PI scores and ranking of CASSRs, for simplicity's sake, I show only the effect of risk-adjustment on ranking using adjusted-PIs derived from the OLS regression and calculated using the error method. Since the individual ratio method of PI calculation makes some

difference to the identification of outliers, however, I do present these results for comparison with the error method.

Table 38: Summary statistics for the SCRQoL indicator, estimated by different methods†, 18 to 64 sub-group

Type of risk adjustment method	Spec	No. of CASSRs	Complete case analysis				Multiply-imputed sample			
			Mean	Median	IQR	Range (min–max)	Mean	Median	IQR	Range (min–max)
0. HSCIC unadjusted	---	149	17.2	17.2	16.7-17.7	14.3, 19.5	17.2	17.2	16.8-17.4	14.3, 19.5
1. OLS regression	SV	149	17.1	17.1	16.8-17.5	15.6, 18.7	17.1	17.1	16.8-17.4	15.4, 18.4
	EP	149	17.1	17.1	16.7-17.5	15.3, 18.6	17.0	17.0	16.7-17.4	15.2, 18.5
2. FR regression	SV	149	17.1	17.1	16.7-17.5	15.7, 18.7	17.1	17.0	16.7-17.4	15.4, 18.4
	EP	149	17.1	17.1	16.7-17.5	15.3, 18.5	17.0	17.0	16.7-17.4	15.2, 18.4
3. FE regression	SV	149	17.1	17.1	16.8-17.5	15.6, 18.7	17.1	17.1	16.8-17.4	15.4, 18.4
	EP	149	17.1	17.1	16.7-17.5	15.2, 18.6	17.0	17.0	16.7-17.4	15.1, 18.5
4. RE regression	SV	149	17.1	17.1	16.7-17.5	15.6, 18.7	17.0	17.0	16.8-17.4	15.4, 18.4
	EP	149	17.1	17.1	16.7-17.5	15.2, 18.6	17.0	17.0	16.7-17.4	15.1, 18.5

Legend: † Indicators estimated using OLS risk-adjustment model and the 'error' method i.e. observed – expected; SV, theoretically-driven; EP, statistically-driven; IQR, interquartile range; --- not applicable

Table 39: Summary statistics for the SCRQoL indicator, estimated by different methods†, 65 and over sub-group

Regression method	Spec	No. of CASSRs	Complete case analysis				Multiply-imputed sample			
			Mean	Median	IQR	Range (min–max)	Mean	Median	IQR	Range (min–max)
0. HSCIC unadjusted		149	18.3	18.4	17.8-18.7	16.8, 20.1	18.3	18.4	17.8-18.7	16.8, 20.1
1. OLS regression	SV	149	18.2	18.2	17.9-18.4	17, 19.7	18.3	18.3	18.1-18.5	17.3, 19.2
	EP	149	18.2	18.2	18-18.5	17.2, 19.3	18.3	18.3	18.1-18.5	17.3, 19.3
2. FR regression	SV	149	18.2	18.2	17.9-18.4	17, 19.6	18.3	18.3	18.1-18.5	17.3, 19.3
	EP	149	18.2	18.2	18-18.5	17.1, 19.3	18.3	18.3	18.1-18.5	17.3, 19.2
3. FE regression	SV	149	18.2	18.2	17.9-18.4	17, 19.7	18.3	18.3	18.1-18.5	17.3, 19.2
	EP	149	18.2	18.2	17.9-18.5	17.2, 19.3	18.3	18.3	18.1-18.5	17.3, 19.3
4. RE regression	SV	149	18.2	18.2	17.9-18.4	17, 19.7	18.3	18.3	18.1-18.5	17.3, 19.2
	EP	149	18.2	18.2	18-18.5	17.2, 19.3	18.3	18.3	18.1-18.5	17.3, 19.3

Legend: † Indicators estimated using OLS risk-adjustment model and the 'error' method i.e. observed – expected; SV, theoretically-driven; EP, statistically-driven; IQR, interquartile range; --- not applicable

Table 40: Summary statistics for the SCRQoL indicator, estimated by different methods†, care home sub-group

Regression method	Spec	No. of CASSRs	Complete case analysis				Multiply-imputed sample			
			Mean	Median	IQR	Range (min, max)	Mean	Median	IQR	Range (min, max)
0. HSCIC unadjusted	---	148	19.6	19.6	19-20	17, 21.3	19.6	19.6	19-20	17, 21.3
1. OLS regression	SV	148	19.6	19.6	19.2-20	18.2, 22	19.5	19.5	19.1-19.8	17.9, 21.2
	EP	148	19.6	19.7	19.1-20.1	17.6, 21.2	19.4	19.5	19-19.8	17.2, 21.1
2. FR regression	SV	148	19.6	19.6	19.2-20	18.1, 21.9	19.5	19.4	19.1-19.8	18, 21.2
	EP	148	19.6	19.7	19.1-20.1	17.6, 21.2	19.4	19.5	19-19.8	17.2, 21.1
3. FE regression	SV	148	19.6	19.6	19.2-20	18.2, 22	19.5	19.5	19.1-19.8	17.9, 21.2
	EP	148	19.6	19.7	19.1-20.1	17.6, 21.2	19.4	19.5	19-19.8	17.3, 21.1
4. RE regression	SV	148	19.6	19.6	19.2-20	18.2, 22	19.5	19.5	19.1-19.8	17.9, 21.2
	EP	148	19.6	19.7	19.1-20.1	17.6, 21.2	19.4	19.5	19-19.8	17.3, 21.1

Legend: † Indicators estimated using OLS risk-adjustment model and the 'error' method i.e. observed – expected; SV, theoretically-driven; EP, statistically-driven; IQR, interquartile range; --- not applicable

Table 41: Summary statistics for the SCRQoL indicator, estimated by OLS regression†, all sub-groups

Regression method	Spec	No. of CASSRs	Complete case analysis				Multiply-imputed sample			
			Mean	Median	IQR	Range (min, max)	Mean	Median	IQR	Range (min, max)
0. HSCIC unadjusted	---	149	18.3	18.3	18-18.7	16.6, 19.6	18.3	18.3	18.1-18.6	16.6, 19.6
1. OLS regression	SV	149	18.3	18.3	18.1-18.5	17.3, 19.5	18.3	18.4	18.1-18.6	17.4, 19.1
	EP	149	18.3	18.3	18.1-18.5	17.2, 19.2	18.3	18.3	18-18.5	17.2, 19.2

Legend: † Indicators estimated using OLS risk-adjustment model and the 'error' method i.e. observed – expected; SV, theoretically-driven; EP, statistically-driven; IQR, interquartile range; --- not applicable

Risk-adjustment has an effect on PI scores, as is evidenced by the correlation statistics in Table 42 between the unadjusted and risk-adjusted PIs. Its effect varies across the sub-groups, with the greatest effect being for the 18 to 64 sub-group. Taking for example, the results for the risk-adjustment models using the theoretically-driven (SV) specification and estimated on the complete cases, the correlation between the unadjusted and adjusted PIs for the 18 to 64 sub-group is less than 0.7 and close to 28 per cent of CASSR pairs change order. By comparison, both the 65 and over and care home sub-groups have correlations over 0.75 with just over 20 per cent of CASSR pairs changing order. The picture is slightly different when the statistically-driven (EP) specification is used. Risk-adjustment has much less of an effect on PI scores for the care home sub-group compared to the other sub-groups. The correlation between the two specifications is also much lower for the care home sub-group at around 0.85 compared to over 0.9 for the other sub-groups. When the PI is calculated for the whole PSD sample, the results are fairly similar to the 65 and over sub-group.

The effect that risk adjustment has on PI scores, is in general to draw them towards the overall sample mean. This effect can clearly be seen for all the sub-groups in the caterpillar plots shown in Figure 30 (18 to 64 sub-group, SV specification), Figure 31 (65 and over, SV specification), Figure 32 and Figure 33 (care home sub-group, SV and EP specification respectively) and Figure 34 (all PSD sub-groups, SV specification) which appear more squashed after risk adjustment with more similar PI scores for CASSRs. It is also illustrated well by the funnel plots shown in Figure 35 (18 to 64 sub-group, SV specification), Figure 36 (65 and over, SV specification), Figure 37 and Figure 38 (care home sub-group, SV and EP specification respectively) and Figure 39 (all PSD sub-groups, SV specification), where the PI scores are more closely bunched and there are fewer outliers.

The colour coding of data points on the funnel plots shows the extent to which risk adjustment has an effect on ranking. The top ten and bottom ten CASSRs (as ascertained by ranking CASSRs on their unadjusted PI scores), are shown in each funnel plot with green and blue dots, respectively. It is clear that risk adjustment also has an effect on the ranking of CASSRs, with a number of CASSRs in each sub-group moving out of the top and bottom ten after risk-adjustment.

Table 42: Correlation statistics between unadjusted and adjusted indicators†

Population sub-group	Comparison	Pearson's R^2 (p-value)	Rho (p-value)	Tau (p-value)	% pairs change order
Complete case analysis					
18 to 64	SV-Unadjusted	0.664	0.615	0.446	27.7%
18 to 64	EP-Unadjusted	0.707	0.687	0.503	24.8%
18 to 64	SV-EP	0.931	0.915	0.765	11.8%
65 and over	SV-Unadjusted	0.771	0.777	0.584	20.8%
65 and over	EP-Unadjusted	0.798	0.803	0.611	19.5%
65 and over	SV-EP	0.951	0.943	0.814	9.3%
care home	SV-Unadjusted	0.762	0.702	0.529	23.6%
care home	EP-Unadjusted	0.929	0.907	0.756	12.2%
care home	SV-EP	0.854	0.814	0.636	18.2%
All	SV-Unadjusted	0.787	0.801	0.604	19.8%
All	EP-Unadjusted	0.812	0.827	0.638	18.1%
All	SV-EP	0.951	0.947	0.806	9.7%
Multiple imputation					
18 to 64	SV-Unadjusted	0.691	0.668	0.487	25.6%
18 to 64	EP-Unadjusted	0.692	0.678	0.490	25.5%
18 to 64	SV-EP	0.984	0.978	0.882	5.9%
65 and over	SV-Unadjusted	0.773	0.774	0.586	20.7%
65 and over	EP-Unadjusted	0.777	0.783	0.591	20.5%
65 and over	SV-EP	0.983	0.976	0.880	6.0%
care home	SV-Unadjusted	0.808	0.764	0.584	20.8%
care home	EP-Unadjusted	0.917	0.904	0.742	12.9%
care home	SV-EP	0.902	0.870	0.699	15.0%
All ‡	SV-Unadjusted	0.794	0.804	0.617	19.1%
All ‡	EP-Unadjusted	0.810	0.829	0.645	17.7%
All ‡	SV-EP	0.967	0.965	0.843	7.8%

Legend: † Indicators estimated using the error approach i.e. observed – expected; ‡ PIs estimated from OLS by CCA; SV, theoretically-driven; EP, statistically-driven; n=148 for care home sub-group and 149 for 18 to 64 and 65 and over sub-groups

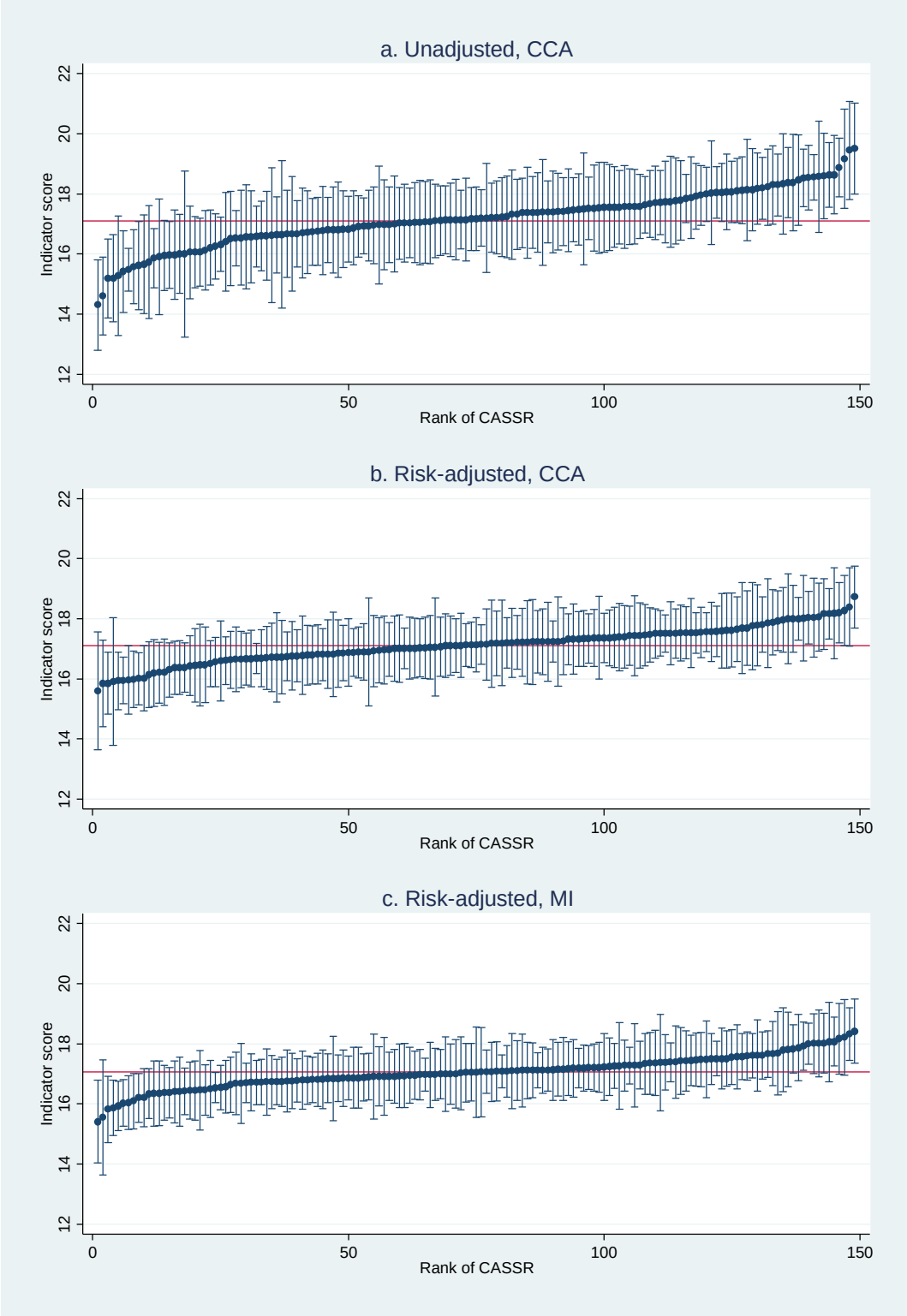

Figure 30: Caterpillar plots of SCRQoL PI scores, before and after adjustment, for the 18 to 64 sub-group, theoretically-driven specification

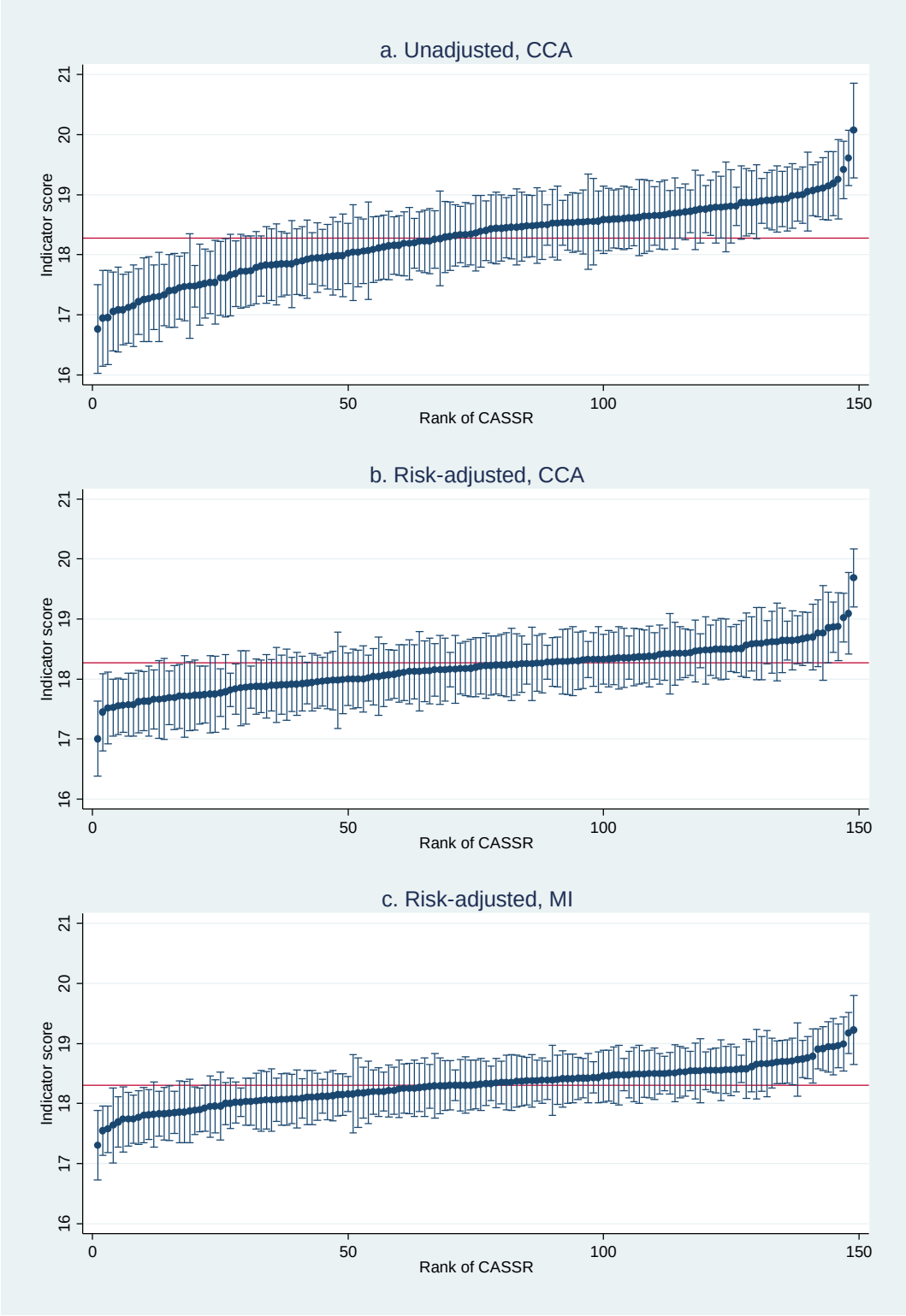

Figure 31: Caterpillar plots of SCRQoL PI scores, before and after adjustment, for the 65 and over sub-group, theoretically-driven specification

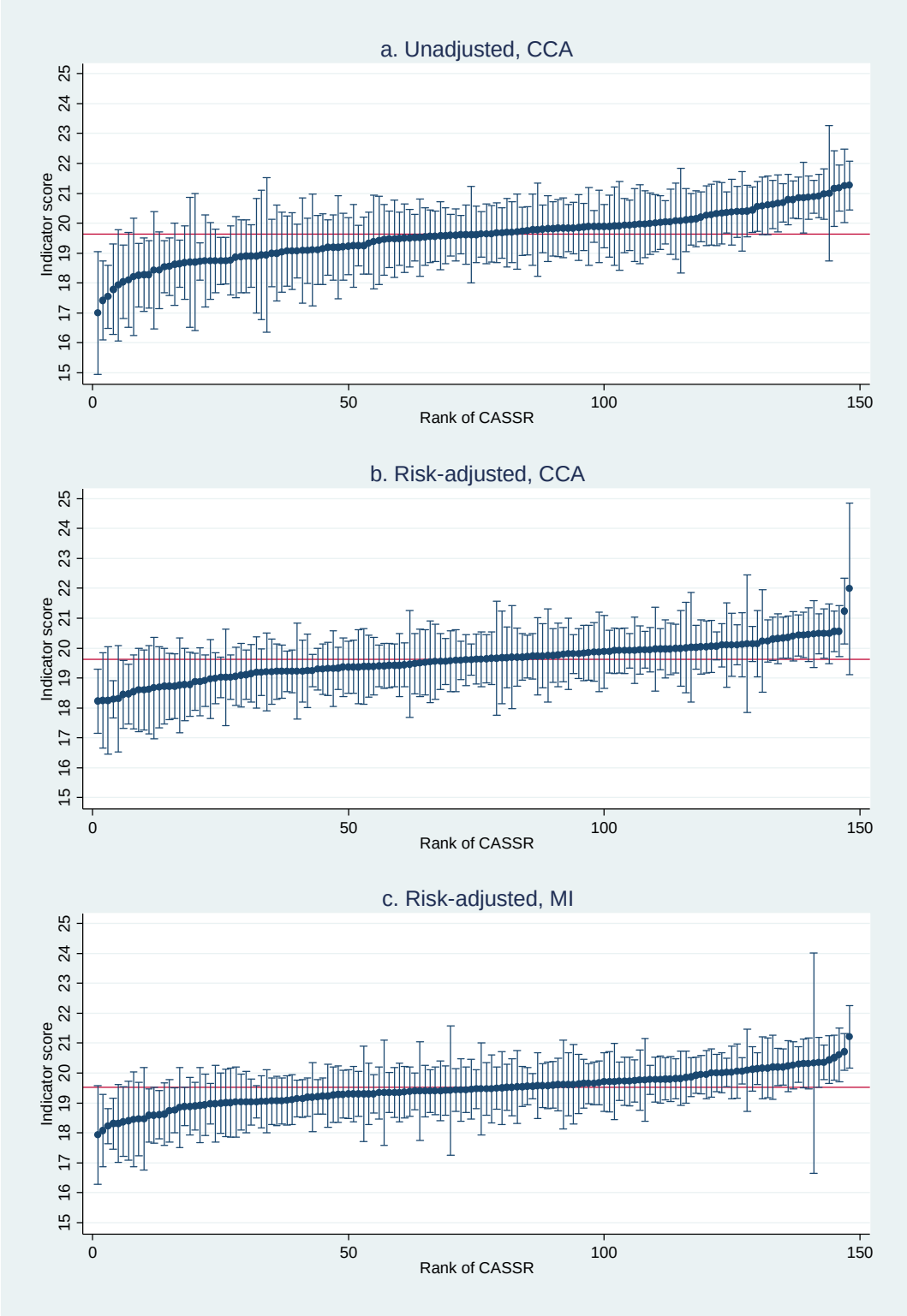

Figure 32: Caterpillar plots of SCRQoL PI scores, before and after adjustment, for the care home sub-group, theoretically-driven specification

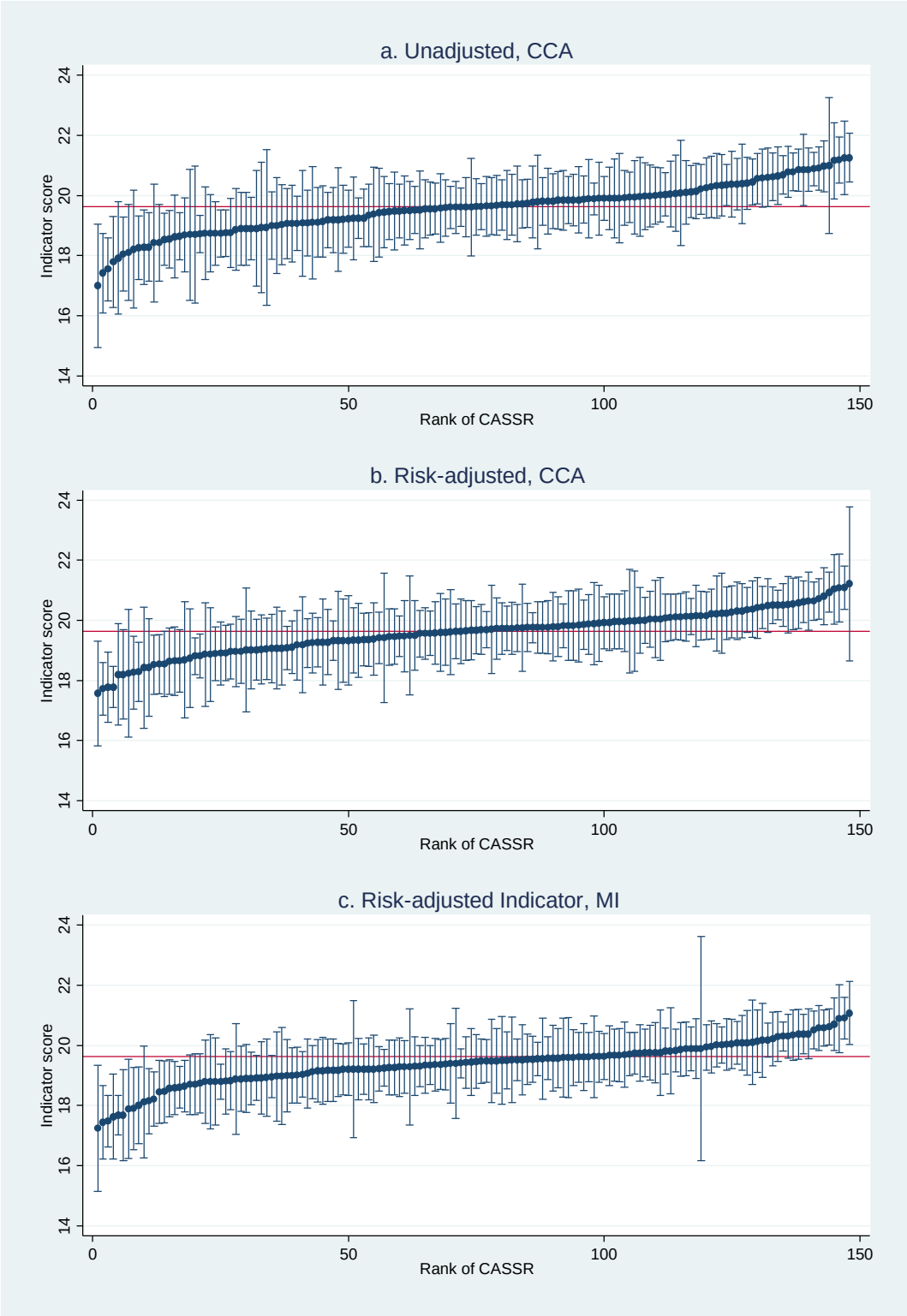

Figure 33: Caterpillar plots of SCRQoL PI scores, before and after adjustment, for the care home sub-group, statistically-driven specification

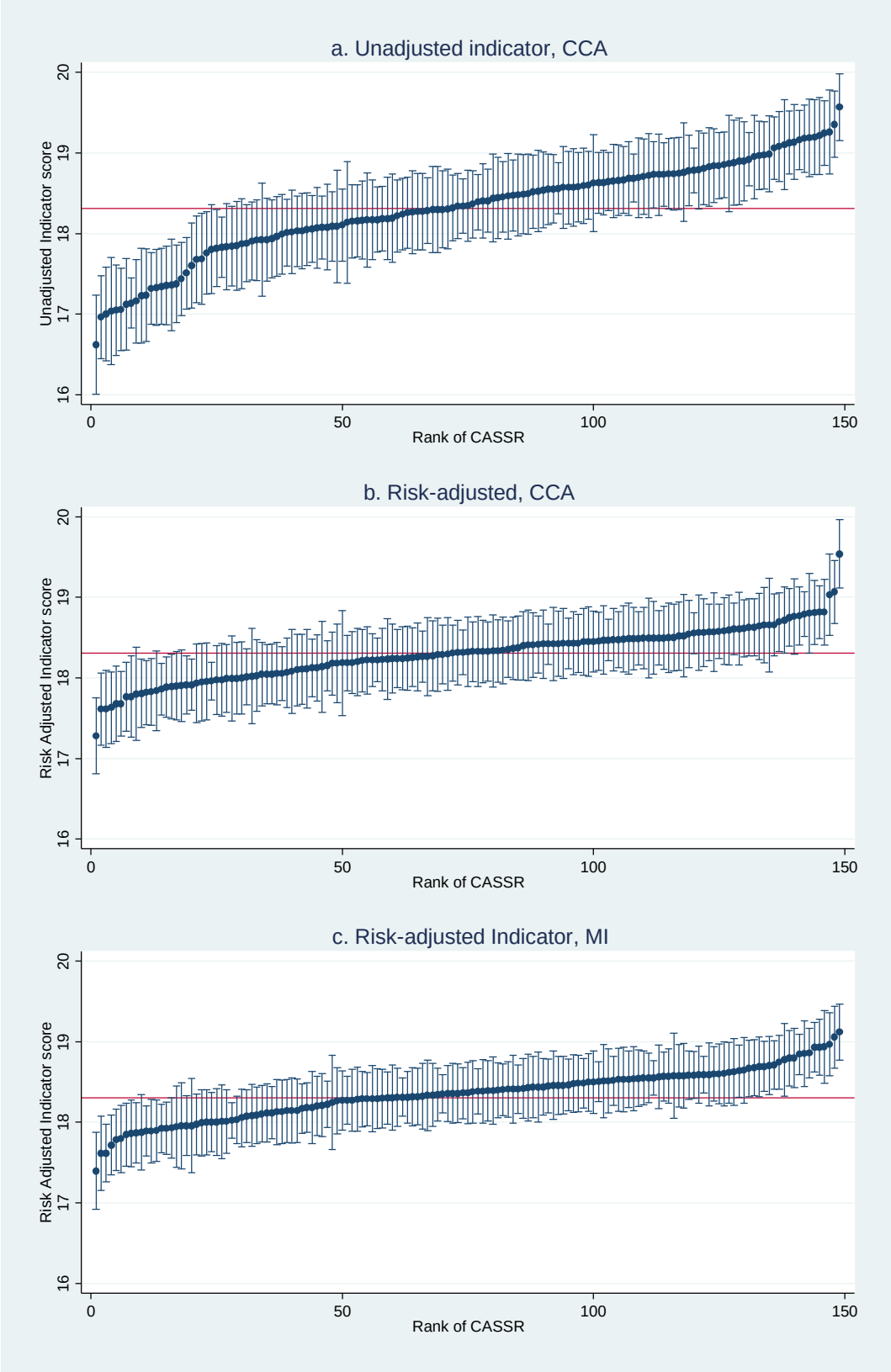

Figure 34: Caterpillar plots of SCRQoL PI scores, before and after adjustment, for all sub-groups, theoretically-driven specification

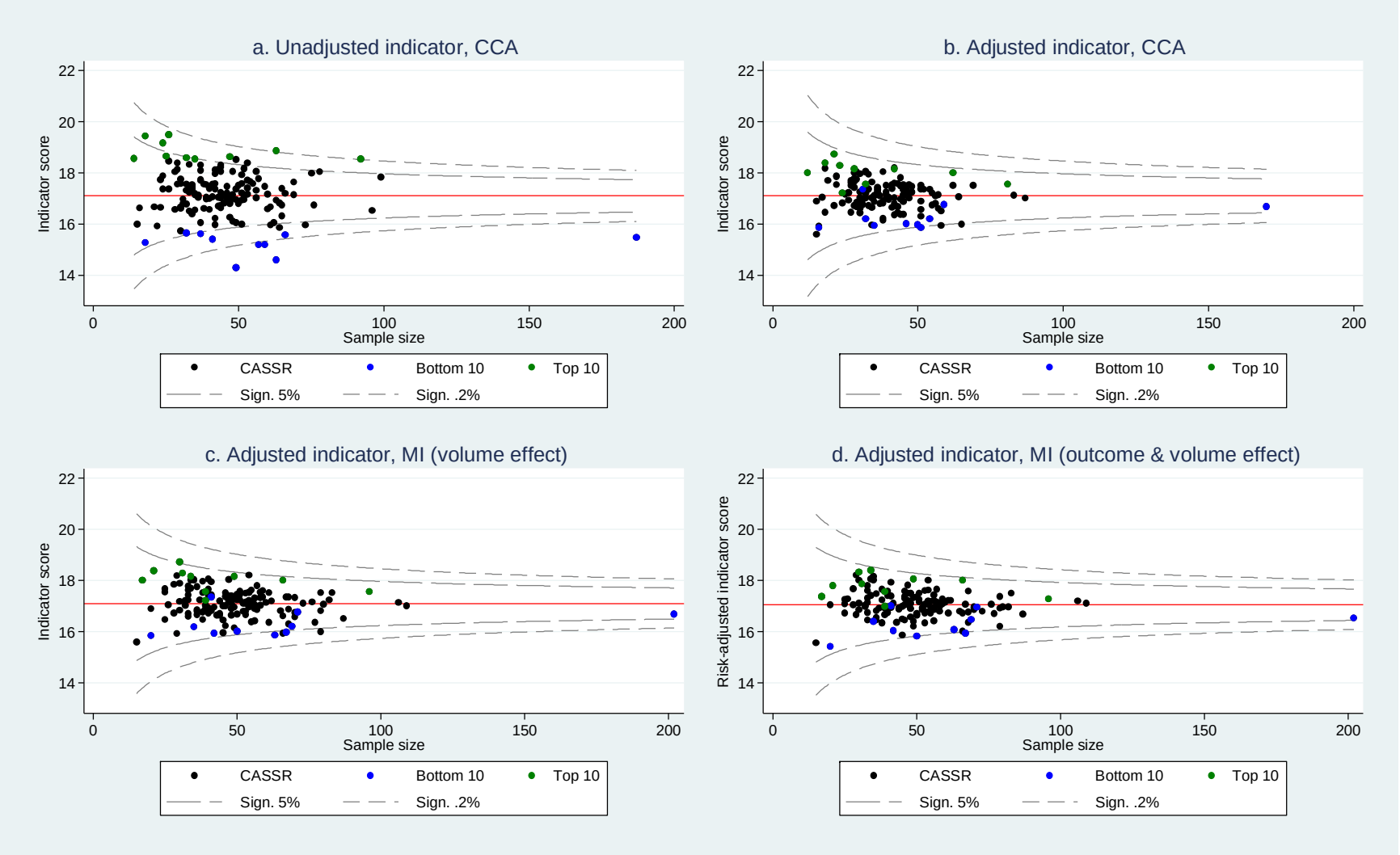

Figure 35: Funnel plots for SCRQoL PI scores, before and after adjustment, for the 18 to 64 sub-group, theoretically-driven specification

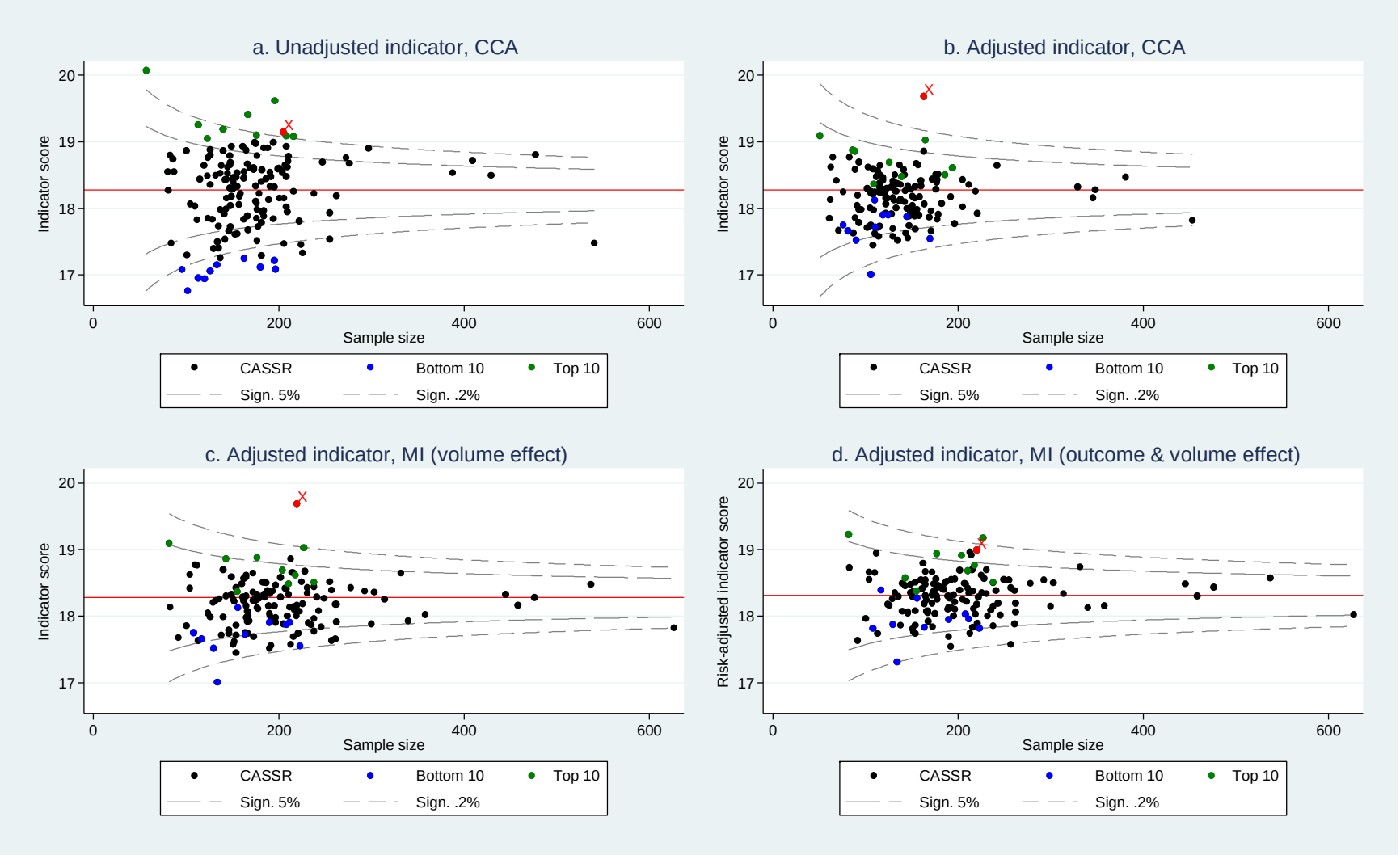

Figure 36: Funnel plots for SCRQoL PI scores, before and after adjustment, for the 65 and over sub-group, theoretically-driven specification

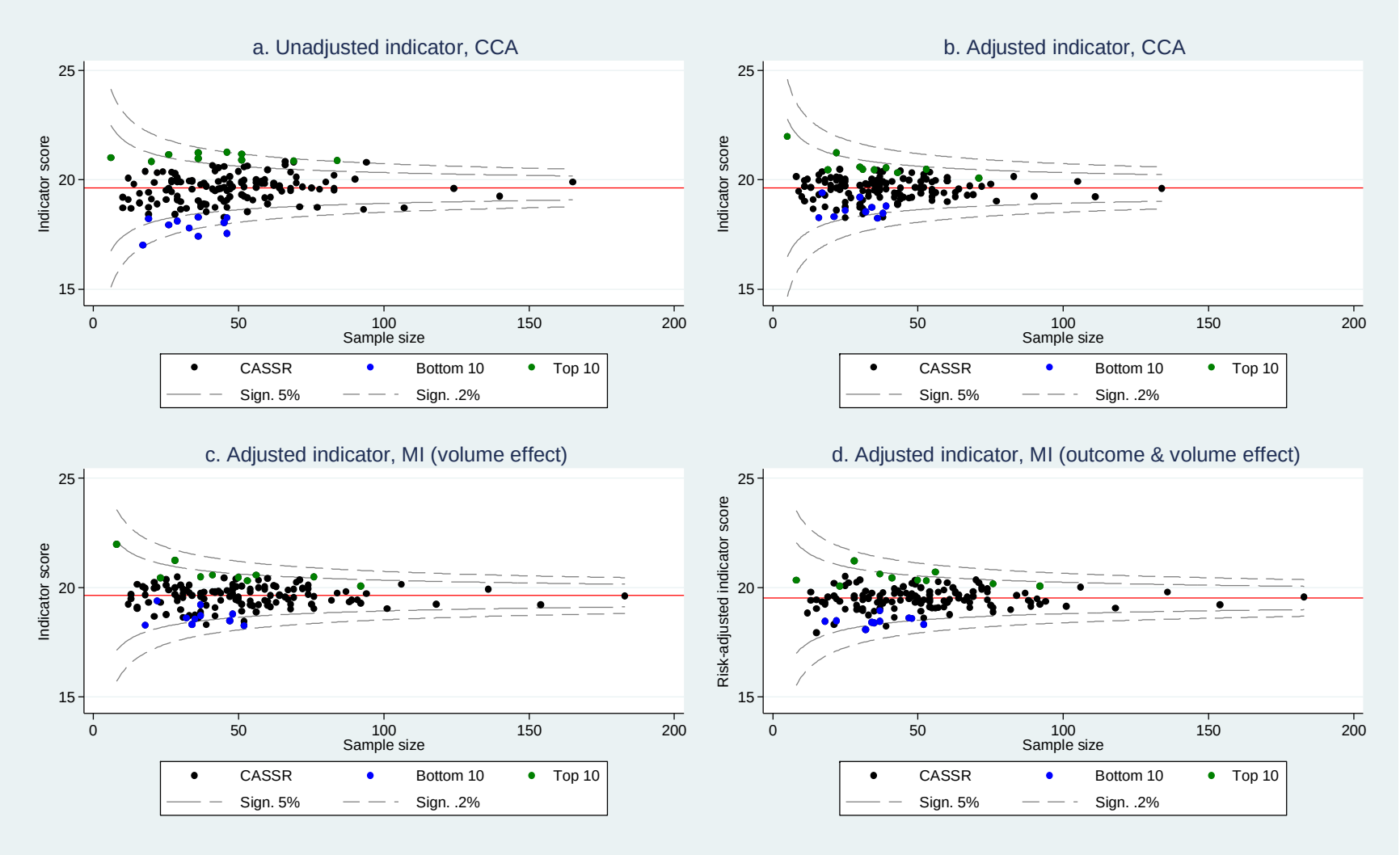

Figure 37: Funnel plots for SCRQoL PI scores, before and after adjustment, for the care home sub-group, theoretically-driven specification

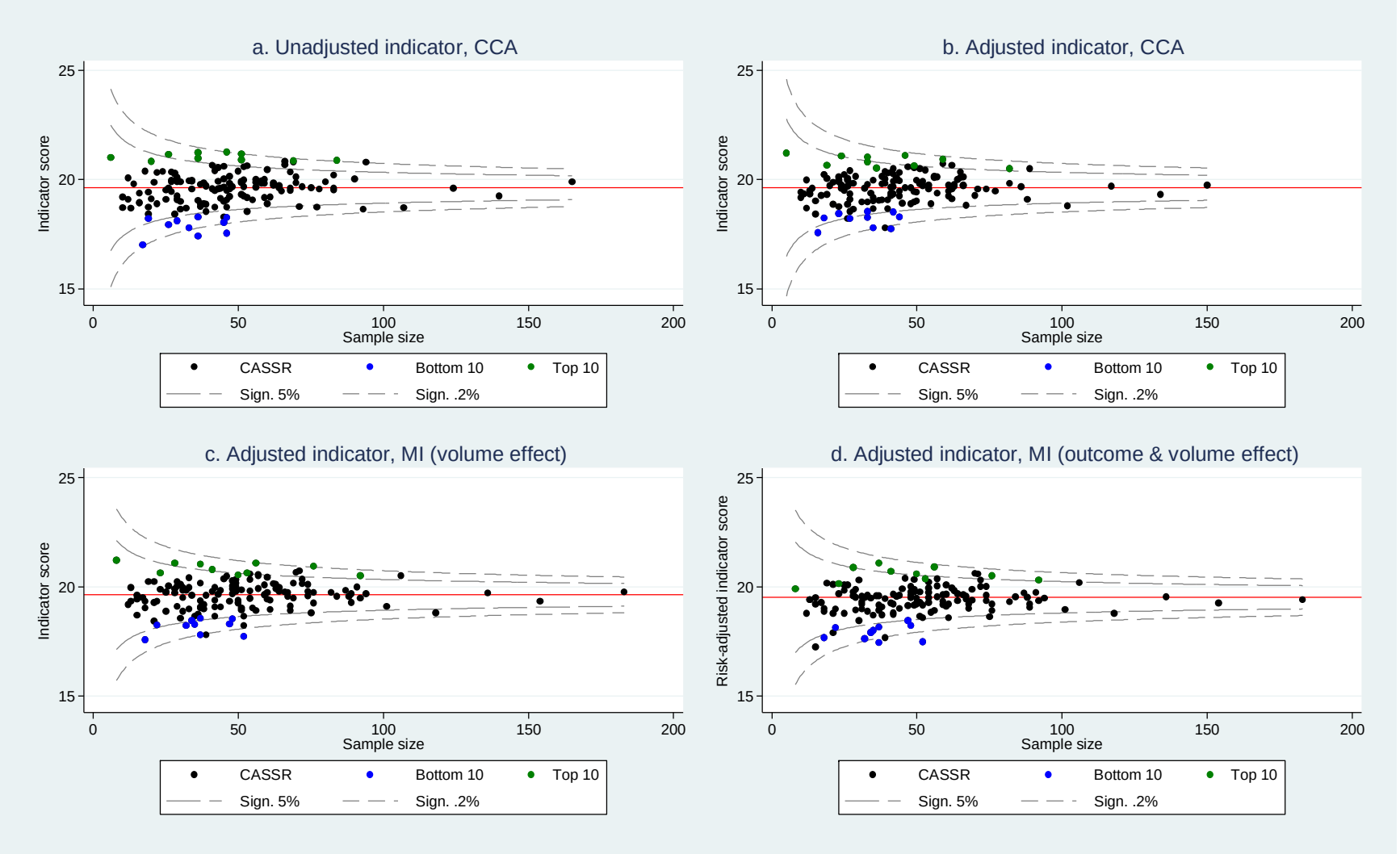

Figure 38: Funnel plots for SCRQoL PI scores, before and after adjustment, for the care home sub-group, statistically-driven specification

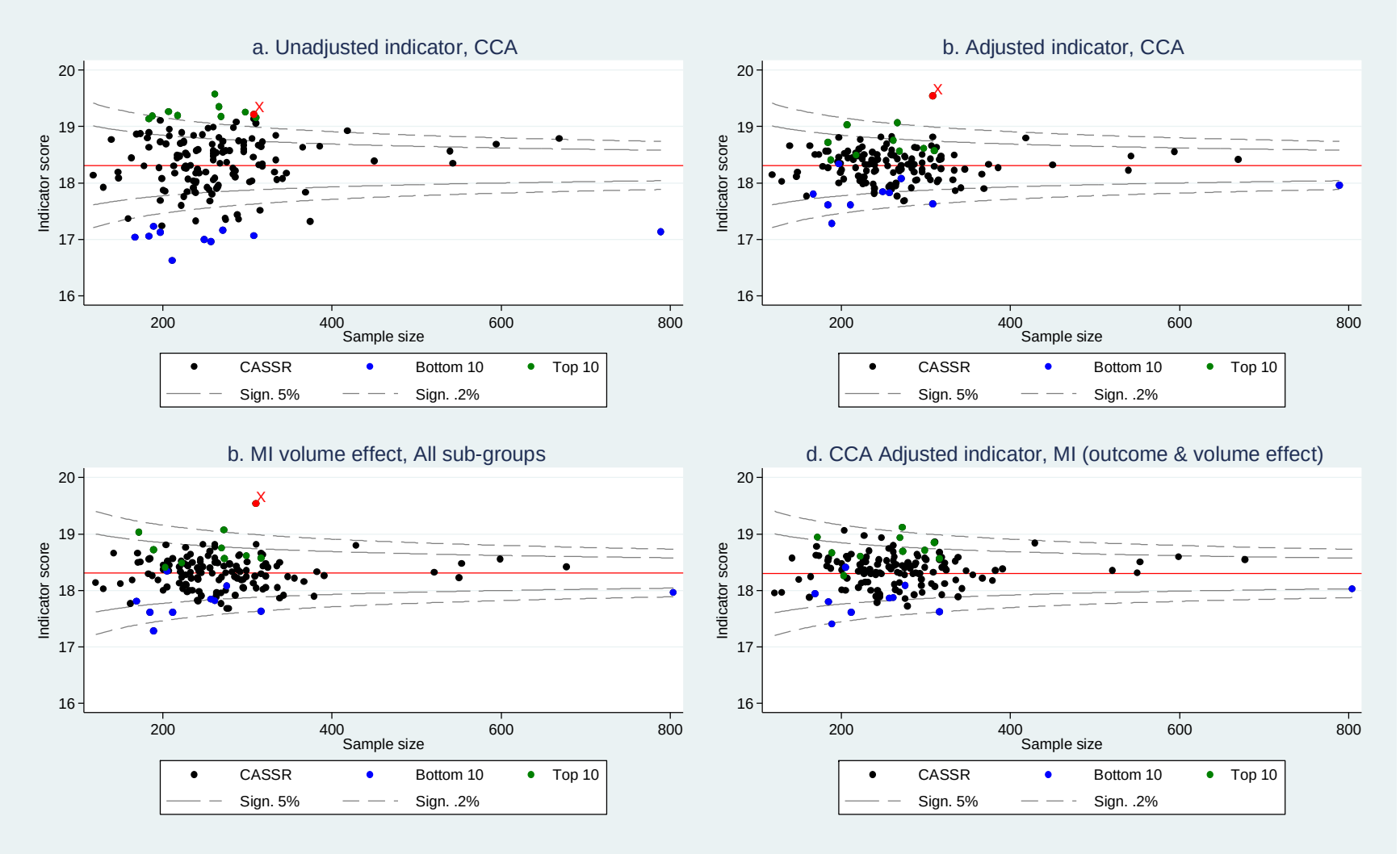

Figure 39: Funnel plots for SCRQoL PI scores, before and after adjustment, for all sub-groups, theoretically-driven specification

The effect of risk-adjustment on the identification of outliers is shown in more detail in Table 43 for the error method and Table 44 for the individual ratio method. Concentrating first on Table 43 , there is some variation across the sub-groups in the percentage of CASSRs identified as outliers, both pre and post risk adjustment. For the 18 to 64 sub-group only two per cent of CASSRs are identified as outliers following risk-adjustment, down from 17 per cent based on the unadjusted PI. A similar picture is presented for the care home sub-group, where PIs use the theoretically-driven (SV) risk-adjustment model. For this sub-group the percentage of CASSRs identified as outliers reduces from 22 to three. Where PIs for this sub-group are based on the statistically-driven (EP) risk-adjustment model, as expected, the percentage of outliers after risk-adjustment is higher at 14 per cent. For the 65 and over sub-group many more CASSRs are identified as outliers using the unadjusted PIs, at around 52 per cent. This reduces to 11 per cent after risk-adjustment. This suggests that both the specification of the model and the sub-group are important for determining which CASSRs are outliers. In the latter case this is probably due mostly to differences in the sample size between the sub-groups.

Comparing Table 43 and Table 44, it is clear that in general the number of outliers is not affected by the choice of method for standardising the observed outcome. Only for the PSD subsample is there a difference. Here slightly more CASSRs are identified as outliers after adjustment using the individual ratio method compared to the error method. Overwhelmingly the 'additional' outliers are in the lower end of the distribution.

Table 45 (18 to 64 sub-group), Table 46 (65 and over sub-group), Table 47 (care home sub-group) and Table 48 (PSD subsample), provide a different perspective on the effect of risk-adjustment on inferences about performance. These tables describe how CASSRs move between the various in control (normal) and out of control (alarm and alert) statuses. Across all sub-groups, following risk-adjustment the movements are overwhelmingly into an in control state from an out of control state rather than vice-versa. Of the movements into an in control state, a number are from an alarm state. These movements are particularly striking for the 65 and over sub-group, where many changes are from the low alarm state to the in control state. Taking as an example the SV specification, 42 of the 77 LAs identified as outliers using the crude PI are found to be in control when using the risk-adjusted PI. Of these 42 'false positives', 16 are identified as alarm status using the crude PI. By contrast only three of the 72 LAs identified as in control using the crude PI are identified as out of control using the risk-adjusted PI. All of these move into an alert state. For the PSD sub-sample the number of 'false positives' observed is slightly lower than the 65 and over sub-

group. Nevertheless 36 out of the 85 LAs identified as outliers are found to be 'false positives' after risk-adjustment, with half of the 'false positives' being in the alarm status when using the unadjusted PI score. The 'false positive rate' is also very high across all the sub-groups, being in all cases over ten per cent but reaching over 30 per cent for the 65 and over sub-group. Assuming the risk-adjusted PI scores are a more valid indicator of performance, this suggests the unadjusted PIs are highly misleading.

The effect of risk-adjustment on the ranking of CASSRs is illustrated to some extent in the funnel plots, which show the position of the top and bottom ten CASSRs. It is shown more clearly in Figure 40 (18 to 64 sub-group), Figure 41 (65 and over sub-group), Figure 42 (care home sub-group) and Figure 43 (PSD sub-sample). These histograms present the change in rank position following risk-adjustment for each CASSR. While it is possible to calculate from the tau-statistic in Table 42 that only 20 to 25 per cent of CASSR pairs change order for the individual sub-groups and slightly fewer than 20 per cent change order for the PSD sub-sample, it is clear that there are some quite large position changes of 50 or more places. Risk-adjustment therefore has an important effect on the ranking of CASSRs.

So far I have limited the discussion of results to a comparison between the unadjusted and risk-adjusted PIs calculated by CCA. I now consider the effect of adjustment for nonresponse in the context of risk adjustment. Taking nonresponse into account has a slightly moderating effect on risk adjustment. It reduces to some extent the differences between the unadjusted and risk-adjusted PI scores. This effect is shown in Table 42 where correlations are slightly higher between the unadjusted and risk-adjusted PIs estimated on the multiply-imputed dataset compared to those PIs estimated by CCA. Additionally, a smaller percentage of CASSR pairs change order and there is also more consistency between PIs calculated using the statistically- and theoretically-driven specifications.

There is only a small overall effect of adjusting for nonresponse on performance assessment. The number of outliers is very similar for the theoretically-driven (SV) model, although less so for the statistically-driven (EP) model for the care home sub-group (Table 43). The number of 'false positives' and 'false negatives' is also very similar, again except for the statistically-driven (EP) model results in the care home sub-group (Table 45 to Table 48). The clearest difference is in the changes in the ranking of CASSRs after risk-adjustment which are slightly less extreme when estimated on the MI sample compared to by CCA

(Figure 40 to Figure 43)[1]. From a national perspective taking nonresponse into account has a small effect on inferences regarding performance. This is despite the fraction of missing information due to item nonresponse being greater than 20 per cent (30 per cent for 65 and over, 22 per cent for 18 to 64, and 24 per cent for the care home sub-group).

For individual LAs, however, the effect of taking nonresponse into account can be quite significant. An example is given by the CASSR marked with an 'X' in Figure 36. This CASSR is remarkable in that it is also an exception to the general pattern found after risk-adjustment, since it moves further away from the national average in Figure 36b. This suggests that despite a good average SCRQoL score relative to other CASSRs it has an unfavourable population risk profile. However, when the PIs are estimated on the MI sample, the effect of risk-adjustment for this CASSR is reversed. When calculated on the MI sample, the PI score for this CASSR is slightly lower and it falls into the high alert rather than high alarm category (Figure 36d). This implies that the missing cases are likely to be people with a more favourable risk profile and this is borne out by an examination of the data.

Decomposition of the nonresponse effect into the mean outcome and volume effects suggests that the mean outcome effect is more important for this sample. A clear volume effect is observed by comparing graphs 'b' (risk-adjusted PI under CCA) and 'c' (risk-adjusted PI, MI volume effect) in Figure 35 to Figure 39. This volume effect is also observed in Table 43, which shows that there are slightly more outliers when the missing outcomes and case-mix data are taken into account. The volume effect is entirely cancelled out, however, by the mean outcome effect. This is illustrated by graphs 'c' (risk-adjusted PI, MI volume effect) and 'd' (risk-adjusted PI, MI) in Figure 35 to Figure 39, which show differences in the position of CASSRs relative to one another. This analysis suggests that some of the changes in PI scores following risk-adjustment and some of the effect of risk-adjustment on outliers and ranking is driven by a loss of data. How important this effect is for individual CASSRs depends on the pattern of missingness in the data. I reflect on the implications of this in the discussion.

[1] The distributional statistics confirm this. For the 18 to 64 sub-group, the standard deviation is 34 for the imputed dataset and 38 for CCA. For the 65 and over sub-group the corresponding figures are 28 and 29; and for the care home sub-group the figures are 27 and 33.

Table 43: Number (percentage) of outliers using unadjusted and risk-adjusted PIs, using error method

Missing data adjustment	Outlier status for SV specification					Outlier status for EP specification				
	High alarm y>-3 SD	High alert y>2 SD	'Normal' range	Low alert y>-2 SD	Low alarm y>-3 SD	High alarm y>-3 SD	High alert y>2 SD	'Normal' range	Low alert y>-2 SD	Low alarm y>-3 SD
18 to 64 sub-group (n=149)										
Unadjusted	2 (1%)	8 (5%)	123 (83%)	11 (7%)	5 (3%)	2 (1%)	8 (5%)	123 (83%)	11 (7%)	5 (3%)
OLS1, CCA	0 (0%)	0 (0%)	146 (98%)	3 (2%)	0 (0%)	0 (0%)	1 (1%)	148 (99%)	0 (0%)	0 (0%)
OLS1, MI VOL	0 (0%)	1 (1%)	144 (97%)	4 (3%)	0 (0%)	0 (0%)	3 (2%)	146 (98%)	0 (0%)	0 (0%)
OLS2, MI	0 (0%)	0 (0%)	147 (99%)	2 (1%)	0 (0%)	0 (0%)	1 (1%)	145 (97%)	3 (2%)	0 (0%)
65 and over sub-group (n=149)										
Unadjusted	7 (5%)	24 (16%)	72 (48%)	29 (19%)	17 (11%)	7 (5%)	24 (16%)	72 (48%)	29 (19%)	17 (11%)
OLS1, CCA	1 (1%)	3 (2%)	133 (89%)	11 (7%)	1 (1%)	2 (1%)	2 (1%)	135 (91%)	9 (6%)	1 (1%)
OLS1, MI VOL	1 (1%)	5 (3%)	125 (84%)	16 (11%)	2 (1%)	2 (1%)	6 (4%)	123 (83%)	15 (10%)	3 (2%)
OLS2, MI	1 (1%)	8 (5%)	132 (89%)	6 (4%)	2 (1%)	2 (1%)	7 (5%)	130 (87%)	9 (6%)	1 (1%)
Care home sub-group (n=148)										
Unadjusted	2 (1%)	13 (9%)	116 (78%)	15 (10%)	2 (1%)	2 (1%)	13 (9%)	116 (78%)	15 (10%)	2 (1%)
OLS1, CCA	0 (0%)	1 (1%)	144 (97%)	3 (2%)	0 (0%)	0 (0%)	8 (5%)	128 (86%)	10 (7%)	2 (1%)
OLS1, MI VOL	0 (0%)	2 (1%)	140 (95%)	6 (4%)	0 (0%)	1 (1%)	10 (7%)	123 (83%)	11 (7%)	3 (2%)
OLS2, MI	0 (0%)	2 (1%)	143 (97%)	3 (2%)	0 (0%)	0 (0%)	5 (3%)	121 (82%)	18 (12%)	4 (3%)
All sub-groups (n=149)										
Unadjusted	14 (9%)	28 (19%)	64 (43%)	25 (17%)	18 (12%)	14 (9%)	28 (19%)	64 (43%)	25 (17%)	18 (12%)
OLS1, CCA	2 (1%)	7 (5%)	127 (85%)	12 (8%)	1 (1%)	2 (1%)	8 (5%)	125 (84%)	12 (8%)	2 (1%)
OLS1, MI VOL	2 (1%)	7 (5%)	126 (85%)	12 (8%)	2 (1%)	2 (1%)	8 (5%)	125 (84%)	12 (8%)	2 (1%)
OLS1, MI	1 (1%)	11 (7%)	129 (87%)	6 (4%)	2 (1%)	2 (1%)	8 (5%)	123 (83%)	13 (9%)	3 (2%)

Key: OLS1, indicators are based on OLS regression by CCA; OLS2, indicators are based on OLS regression by MI; MI, VOL volume effect from multiple imputation; SV, theoretically-driven; EP, statistically-driven.

Table 44: Number (percentage) of outliers using unadjusted and risk-adjusted PIs, using individual ratio method

Missing data adjustment	Outlier status for SV specification					Outlier status for EP specification				
	High alarm y>-3 SD	High alert y>2 SD	'Normal' range	Low alert y>-2 SD	Low alarm y>-3 SD	High alarm y>-3 SD	High alert y>2 SD	'Normal' range	Low alert y>-2 SD	Low alarm y>-3 SD
18 to 64 sub-group (n=149)										
Unadjusted	2 (1%)	8 (5%)	123 (83%)	11 (7%)	5 (3%)	2 (1%)	8 (5%)	123 (83%)	11 (7%)	5 (3%)
OLS1, CCA	0 (0%)	1 (1%)	142 (95%)	6 (4%)	0 (0%)	0 (0%)	2 (1%)	145 (97%)	2 (1%)	0 (0%)
OLS2, MI	0 (0%)	0 (0%)	145 (97%)	4 (3%)	0 (0%)	0 (0%)	2 (1%)	142 (95%)	5 (3%)	0 (0%)
65 and over sub-group (n=149)										
Unadjusted	7 (5%)	24 (16%)	72 (48%)	29 (19%)	17 (11%)	7 (5%)	24 (16%)	72 (48%)	29 (19%)	17 (11%)
OLS1, CCA	1 (1%)	3 (2%)	130 (87%)	14 (9%)	1 (1%)	2 (1%)	3 (2%)	130 (87%)	13 (9%)	1 (1%)
OLS2, MI	1 (1%)	8 (5%)	128 (86%)	10 (7%)	2 (1%)	2 (1%)	7 (5%)	127 (85%)	11 (7%)	2 (1%)
Care home sub-group (n=148)										
Unadjusted	2 (1%)	13 (9%)	116 (78%)	15 (10%)	2 (1%)	2 (1%)	13 (9%)	116 (78%)	15 (10%)	2 (1%)
OLS1, CCA	0 (0%)	1 (1%)	142 (96%)	5 (3%)	0 (0%)	0 (0%)	8 (5%)	127 (86%)	10 (7%)	3 (2%)
OLS2, MI	0 (0%)	2 (1%)	140 (95%)	6 (4%)	0 (0%)	0 (0%)	6 (4%)	120 (81%)	17 (11%)	5 (3%)
All sub-groups (n=149)										
Unadjusted	14 (9%)	28 (19%)	64 (43%)	25 (17%)	18 (12%)	14 (9%)	28 (19%)	64 (43%)	25 (17%)	18 (12%)
OLS1, CCA	2 (1%)	8 (5%)	123 (83%)	12 (8%)	4 (3%)	2 (1%)	7 (5%)	123 (83%)	14 (9%)	3 (2%)
OLS1, MI	1 (1%)	11 (7%)	126 (85%)	9 (6%)	2 (1%)	2 (1%)	8 (5%)	118 (79%)	16 (11%)	5 (3%)

Key: OLS1, indicators are based on OLS regression by CCA; OLS2, indicators are based on OLS regression by MI; SV, theoretically-driven; EP, statistically-driven.

Table 45: Movements into and out of 'control' status following risk-adjustment†, 18 to 64 sub-group

Missing data adjustment	SV specification			EP specification		
Indicator	**CCA**	**MI, VOL**	**OLS2, MI**	**CCA**	**MI, VOL**	**OLS2, MI**
Movements into high 'alert' status from						
in control	0	1	0	0	1	0
high alarm	0	0	0	0	0	1
Movements into low 'alert' status from						
in control	1	1	0	0	0	1
low alarm	1	2	2	0	0	2
Movements into 'control' from						
low alert	5	5	6	6	6	6
high alert	6	6	6	5	4	6
low alarm	4	3	3	5	5	3
high alarm	2	2	2	2	2	1
'False positives'	17	16	17	18	17	16
'False negatives'	1	2	0	0	1	1
'Type I error rate'	11.6	11.1	11.6	12.2	11.6	11.0
'Type II error rate'	50.0	66.7	--	0.0	33.3	100.0

Legend: † Indicators are error method indicators; MI, VOL volume effect from multiple imputation; OLS2, indicators are based on OLS regression by MI; SV, theoretically-driven; EP, statistically-driven; zero movements into alarm status so these rows are not shown; -- not estimable because denominator is zero (no 'true positives' or 'false negatives')

Table 46: Movements into and out of 'control' status following risk adjustment†, 65 and over sub-group

Missing data adjustment	SV specification			EP specification		
Indicator	**CCA**	**MI, VOL**	**OLS2, MI**	**CCA**	**MI, VOL**	**OLS2, MI**
Movements into high 'alert' status from						
in control	1	1	1	0	1	1
high alarm	1	2	3	0	2	2
Movements into low 'alert' status from						
in control	2	4	1	3	5	3
low alarm	5	5	2	4	4	3
Movements into 'control' from						
low alert	9	7	11	11	9	10
high alert	17	16	14	17	16	15
low alarm	11	10	13	12	10	13
high alarm	5	4	3	5	3	3
'False positives'	42	37	41	45	38	41
'False negatives	3	5	2	3	6	4
'Type I error rate'	31.1	28.9	30.4	32.6	29.7	30.8
'Type II error rate'	37.5	35.7	22.2	42.9	40.0	36.4

Legend: † Indicators are error method indicators; MI, VOL volume effect from multiple imputation; OLS2, indicators are based on OLS regression by MI; SV, theoretically-driven; EP, statistically-driven; zero movements into alarm status so these rows are not shown.

Table 47: Movements into and out of 'control' status following risk adjustment†, care home sub-group

Indicator	SV specification			EP specification		
Missing data adjustment	**CCA**	**MI, VOL**	**OLS2, MI**	**CCA**	**MI, VOL**	**OLS2, MI**
Movements into high 'alarm' status from						
high alert	0	0	0	0	1	0
in control	0	0	0	0	0	0
Movements into low 'alarm' status from						
low alert	0	0	0	0	0	1
in control	0	0	0	1	1	1
Movements into high 'alert' status from						
in control	0	0	0	0	0	0
high alarm	0	0	0	2	2	0
Movements into low 'alert' status from						
in control	1	2	1	1	1	9
low alarm	1	1	1	1	0	0
Movements into 'control' from						
low alert	12	10	12	7	6	7
high alert	10	9	9	5	3	6
low alarm	1	1	1	0	0	0
high alarm	2	2	2	0	0	2
'False positives'	25	22	24	12	9	15
'False negatives'	1	2	1	2	2	10
'Type I error rate'	17.4	15.7	16.8	9.2	7.1	12.0
'Type II error rate'	33.3	28.6	25.0	13.3	11.1	45.5

Key: † Indicators are error method indicators; MI, VOL volume effect from multiple imputation; OLS2, indicators are based on OLS regression by MI; SV, theoretically-driven; EP, statistically-driven.

Table 48: Movements into and out of 'control' status following risk adjustment†, PSD sub-sample

Indicator	SV specification			EP specification		
Missing data adjustment	**CCA**	**MI, VOL**	**OLS1, MI**	**CCA**	**MI, VOL**	**OLS1, MI**
Movements into high 'alert' status from						
in control	0	0	1	0	0	0
high alarm	3	3	5	3	3	3
Movements into low 'alert' status from						
in control	2	2	0	2	2	3
low alarm	8	7	4	7	7	6
Movements into 'control' from						
low alert	6	6	7	6	6	6
high alert	12	12	10	11	11	11
low alarm	9	9	12	9	9	9
high alarm	9	9	8	9	9	9
'False positives'	36	36	37	35	35	35
'False negatives'	2	2	1	2	2	3
Type I error rate'	27.7	27.7	30.4	27.1	27.1	30.8
'Type II error rate'	25.0	22.2	25.0	20.0	20.0	50.0

Key: † Indicators are error method indicators; MI, VOL volume effect from multiple imputation; OLS1, indicators are based on OLS regression by CCA; SV, theoretically-driven; EP, statistically-driven, zero movements into high alarm and low alarm status so these rows are not shown.

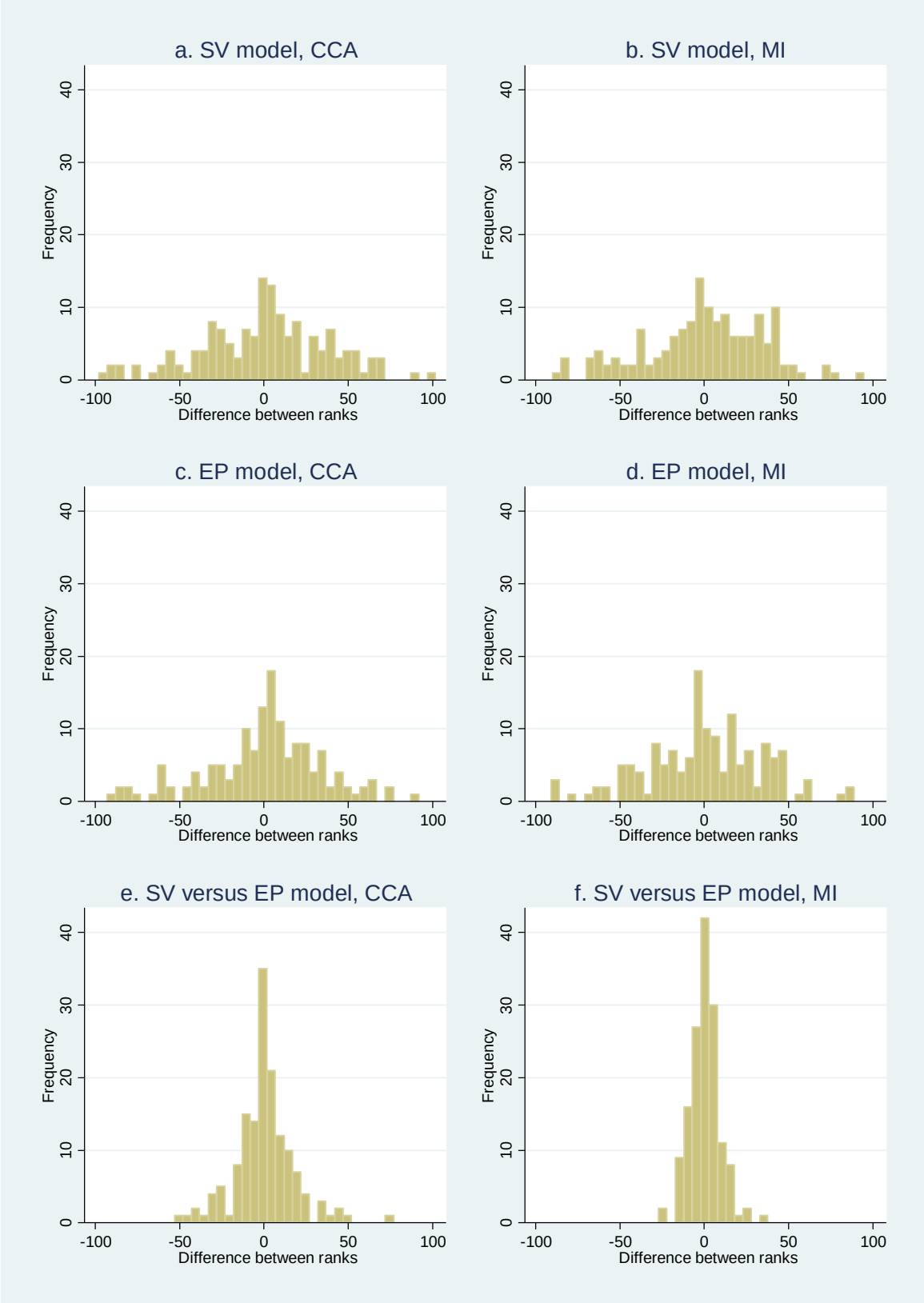

Figure 40: Distribution of CASSR changes in rank for the 18 to 64 sub-group

Key: a to d, unadjusted - risk-adjusted; e and f, SV-EP adjusted; SV, theoretically-driven; EP, statistically-driven

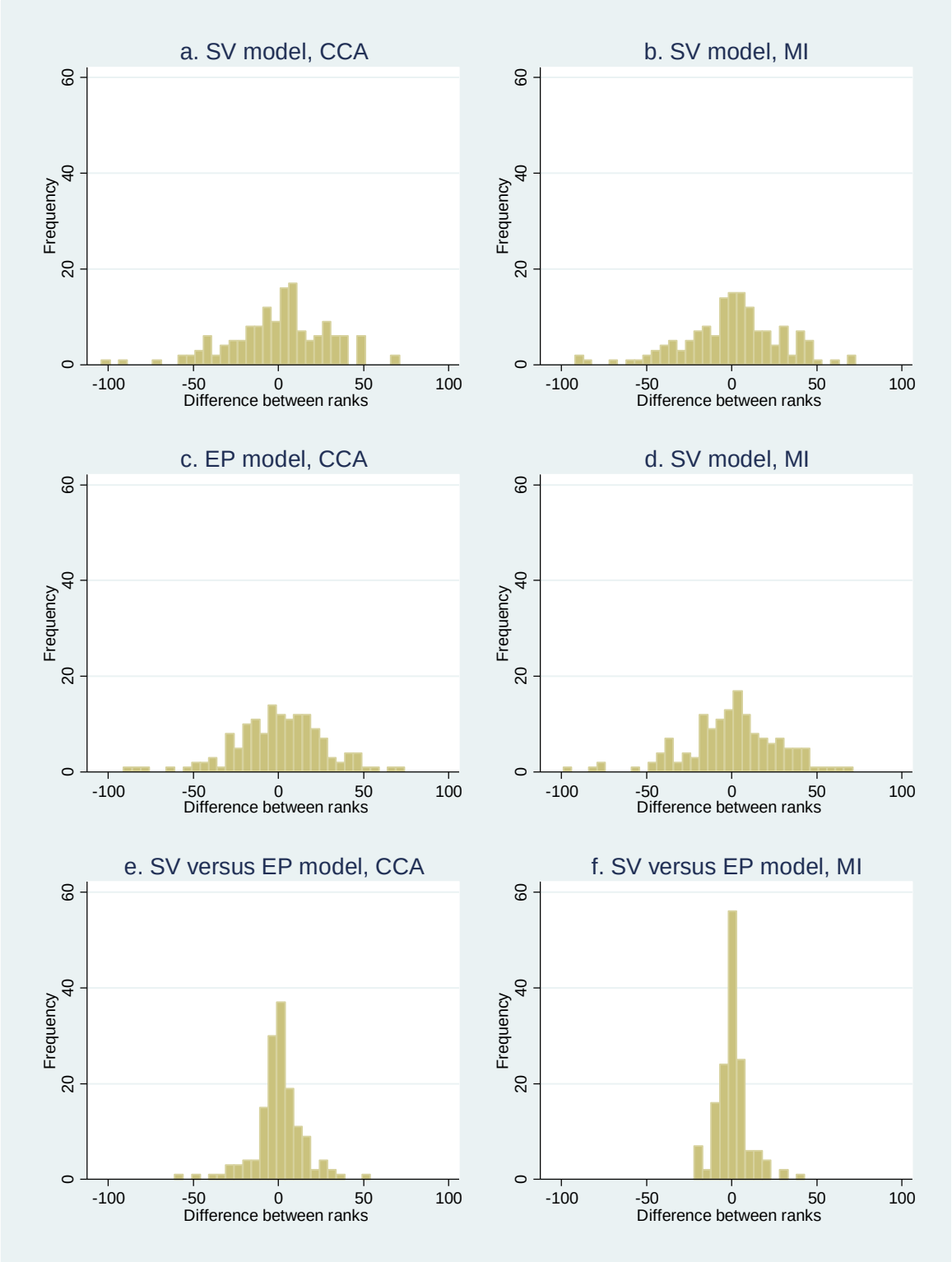

Figure 41: Distribution of CASSR changes in rank for the 65 and over sub-group

Key: a to d, unadjusted - risk-adjusted; e and f, SV-EP adjusted; SV, theoretically-driven; EP, statistically-driven

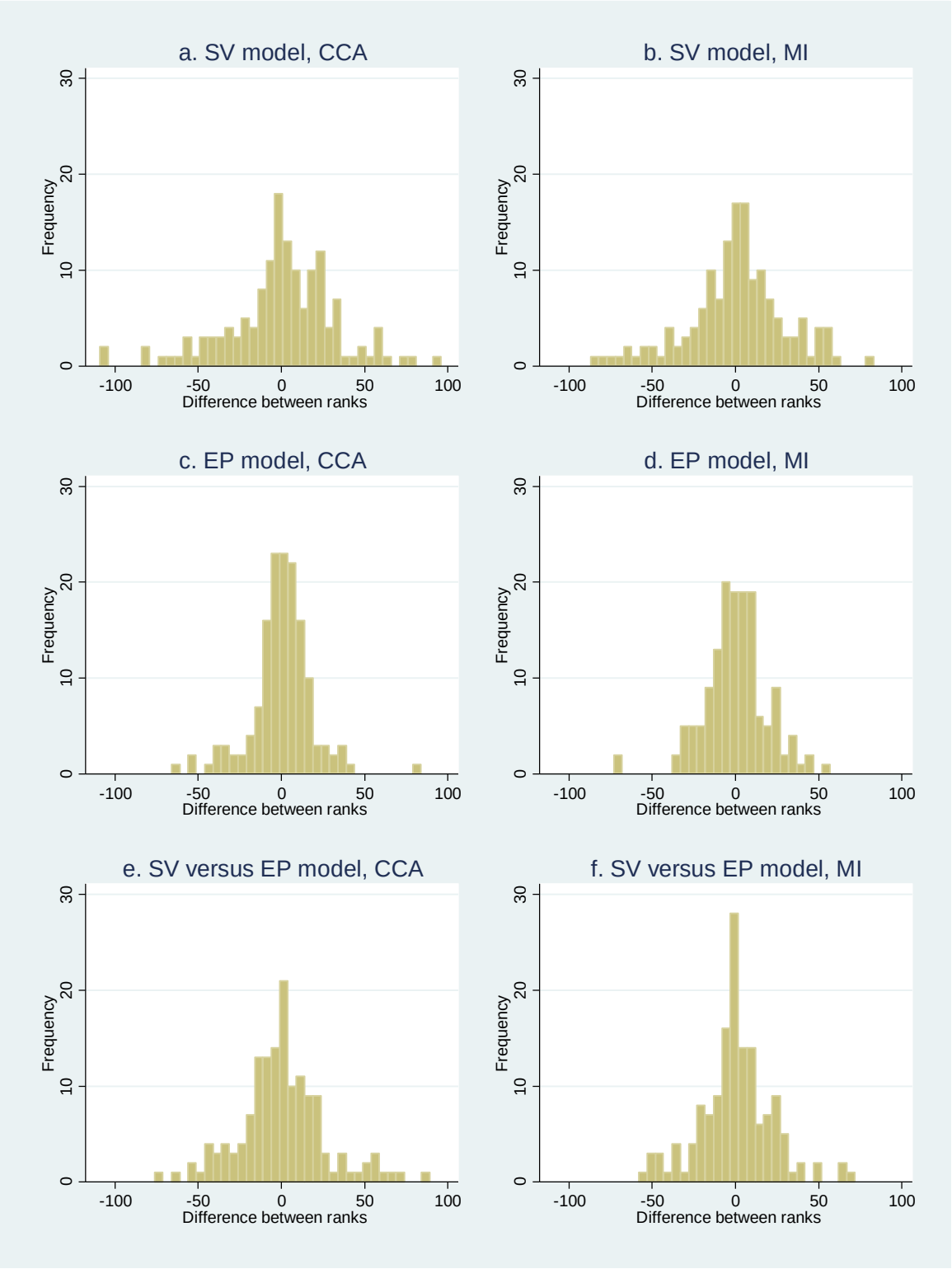

Figure 42: Distribution of CASSR changes in rank for the care home sub-group

Key: a to d, unadjusted - risk-adjusted; e and f, SV-EP adjusted; SV, theoretically-driven; EP, statistically-driven

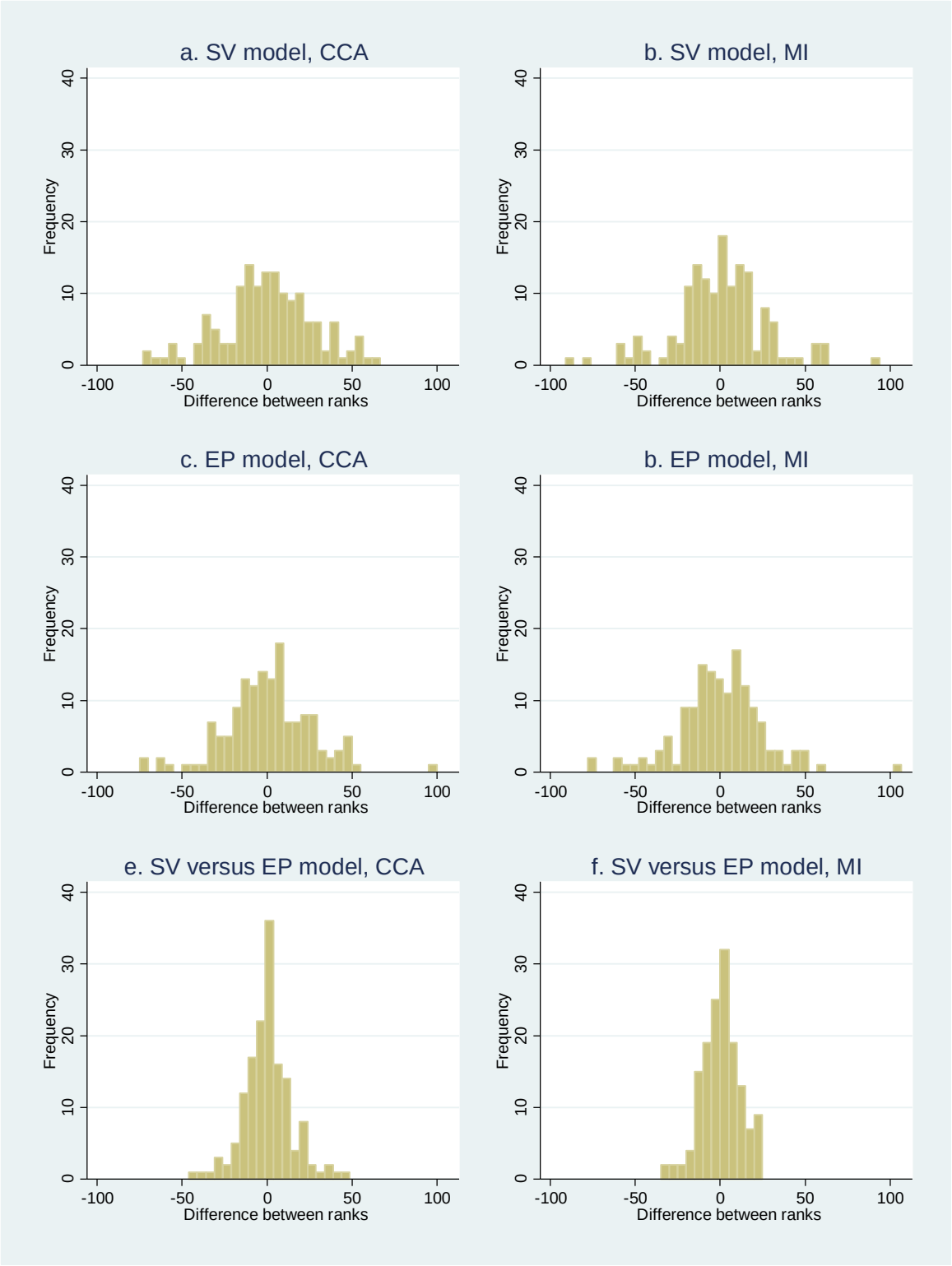

Figure 43: Distribution of CASSR changes in rank for all sub-groups

Key: a to d, unadjusted - risk-adjusted; e and f, SV-EP adjusted; SV, theoretically-driven; EP, statistically-driven

The effect of risk-adjustment and adjustment using a production function approach on performance assessment

The average adjusted SCRQoL PI score is shown in Table 49 for PIs estimated using the risk-adjustment and production function adjustment models and calculated using the error method. Although not shown in these tables, there is little difference in PI scores generated using the ratio method as compared to the error method. Pearson and Spearman rank correlation coefficients are all 0.96 or over (see Appendix 18, Table 105). Hence, all results presented here are generated using the error method.

Consistent with the results from the sample containing all CASSRs, there are few differences between the average risk-adjusted PI scores estimated by the different regression methods (Table 49). Inspection of pairwise correlations of the PI scores estimated using OLS, FE and RE regression, shows perfect correlation for all variant risk-adjustment regression models (see Appendix 18, Table 107). Pairwise correlations for the production function adjusted PIs calculated by different regression methods are similarly perfectly correlated for the OLS and RE models, but the pairwise comparisons with the FE regressions have lower correlations at around 0.94. Interestingly, the FE-estimated production function adjusted PIs are near perfectly correlated (0.99 or over) with the risk-adjusted PIs. Because of the pattern of correlations between the risk-adjusted and production function adjusted PIs estimated using the different regression methods, the remainder of the results are presented only for the PIs estimated by OLS.

There are clear differences in the average PI scores between the risk-adjusted and production function adjusted PIs (Table 49). This reflects the differences between the production function and risk-adjustment models in the coefficient estimates for the risk-adjustors, which affects estimates of effectiveness. The correlation between the two approaches, however, is high at over 0.9, with slightly under 10 per cent of CASSR pairs changing order (Table 50). The high correlation suggests that the choice between the risk adjustment and production function approaches makes only a small difference to inferences about performance.

Table 49: Summary statistics for the SCRQoL indicator, estimated by different methods†, partial dataset for 18 to 64 sub-group

Type of risk adjustment method	Spec	No. of CASSRs	Complete case analysis				Multiply-imputed sample			
			Mean	Median	IQR	Range (min–max)	Mean	Median	IQR	Range (min–max)
0. HSCIC unadjusted	---	83	17.1	17.1	16.6-17.9	14.3, 19.4	17.1	17.1	16.6-17.3	14.3, 19.4
1a. OLS regression	SSV	83	17.0	17.0	16.6-17.5	15.5, 18.2	17.0	17.0	16.6-17.3	15.2, 18.2
	EP	83	17.0	17.0	16.6-17.4	15.1, 18.5	17.0	17.0	16.7-17.3	15, 18.3
1b. FE regression	SSV	83	17.0	17.1	16.6-17.5	15.5, 18.2	17.0	17.0	16.6-17.3	15.2, 18.3
	EP	83	17.0	17.0	16.6-17.4	15, 18.5	17.0	17.0	16.7-17.3	15, 18.3
1c. RE regression	SSV	83	17.0	17.0	16.6-17.4	15.5, 18.2	16.9	17.0	16.6-17.3	15.2, 18.2
	EP	83	17.0	17.0	16.6-17.4	15.1, 18.5	16.9	17.0	16.6-17.3	15, 18.3
2a. OLS production	SSV	83	17.8	17.9	17.4-18.2	16.2, 18.9	18.1	18.1	17.7-18.3	16.7, 19.1
function	EP	83	17.4	17.4	17-17.8	15.9, 18.7	17.2	17.1	16.9-17.6	15.8, 18.4
2b. FE production	SSV	83	17.8	17.8	17.4-18.2	16.3, 19	18.1	18.1	17.7-18.5	16.3, 19.3
function	EP	83	17.4	17.4	17-17.8	15.5, 18.9	17.2	17.3	16.9-17.6	15.4, 18.6
2c. RE production	SSV	83	17.8	17.9	17.4-18.2	16.2, 18.9	18.1	18.1	17.7-18.3	16.7, 19.1
function	EP	83	17.4	17.4	17-17.8	15.9, 18.7	17.2	17.1	16.9-17.6	15.8, 18.4

Legend: † Indicators estimated using the 'error' method i.e. observed – expected; SV, significant variables; SSV, theoretically-driven; EP, statistically-driven; IQR, interquartile range; --- not applicable

Table 50: Correlation statistics between risk-adjusted and production function-adjusted indicators†, 18 to 64 sub-group (n=83)

Specification	Pearson's R^2 (p-value)	Rho (p-value)	Tau (p-value)	% pairs change order
Complete case analysis				
SSV	0.951	0.943	0.823	8.9%
EP	0.951	0.935	0.820	9.0%
Multiple imputation				
SSV	0.941	0.918	0.788	10.6%
EP	0.946	0.916	0.774	11.3%

Legend: † Indicators estimated using OLS regression and calculated using the error approach i.e. observed – expected; SSV, theoretically-driven; EP, statistically-driven

Table 51: Correlation statistics between unadjusted indicators and adjusted indicators†, 18 to 64 sub-group (n=83)

Adjustment method	Specification	Pearson's R^2 (p-value)	Rho (p-value)	Tau (p-value)	% pairs change order
Complete case analysis					
OLS, RA	SSV-Unadjusted	0.702	0.658	0.474	26.3%
	EP-Unadjusted	0.722	0.688	0.505	24.8%
	SSV-EP	0.940	0.933	0.785	10.8%
OLS, PF	SSV-Unadjusted	0.610	0.599	0.427	28.6%
	EP-Unadjusted	0.633	0.621	0.450	27.5%
	SSV-EP	0.929	0.918	0.758	12.1%
Multiple imputation					
OLS, RA	SSV-Unadjusted	0.729	0.697	0.505	24.7%
	EP-Unadjusted	0.712	0.675	0.490	25.5%
	SSV-EP	0.982	0.975	0.868	6.6%
OLS, PF	SSV-Unadjusted	0.637	0.617	0.441	27.9%
	EP-Unadjusted	0.640	0.628	0.451	27.5%
	SSV-EP	0.970	0.957	0.838	8.1%

Legend: † Indicators estimated using the error approach i.e. observed – expected; SSV, theoretically-driven; EP, statistically-driven; PF, production function.

The correlation between the unadjusted PIs and adjusted PIs varies slightly according to whether a risk-adjustment or production function approach is used. The correlation is slightly lower for the production function approach, and a slightly higher per cent of CASSRs change order (see Table 51). Interestingly, despite the insignificant service effect in the statistically-driven (EP) production function models, these PIs still have a slightly lower correlation with the unadjusted indicator compared to the risk-adjustment model with the

same covariate specification. It seems that despite the inadequate risk-adjustment in the statistically-driven models, the inclusion of the service effect still has a small effect on the results. There is also a surprisingly high correlation, at over 0.9, between the PIs estimated from the production function models, under the two (EP and SSV) specifications. The results obtained for the risk-adjusted PIs are very similar to those found for the full dataset (c.f. Table 51 and Table 42), suggesting that these results are not subject to bias arising from the reduced sample of CASSRs.

The effect of production function-adjustment, as it was for risk-adjustment, is to draw PI scores towards the overall sample mean (Figure 46 and Figure 47). The benchmark used in the funnel plots is different for the risk-adjusted and production function-adjusted PIs, as the distribution of PI scores is shifted quite significantly upwards for the production function models. This means that it is less meaningful to compare the numbers of outliers between the two models. What is clear, however, is that after adjustment by either approach there are many fewer outliers. Looking at the theoretically-driven (SSV) specification the risk-adjusted PIs have two outliers and there are no outliers for the production function adjusted PIs (Table 52). Table 53 shows the same information, but this time illustrating how CASSRs move into and out of control status following adjustment. Around a quarter of the movements are from low alarm status to in control, which demonstrates that some of the changes in PI score post-adjustment are fairly large. This analysis suggests that the unadjusted PIs identify many 'false positives', with 'false positive rates' over ten per cent.

The extent of the movements after adjustment can be seen more clearly in the funnel plots (Figure 46 and Figure 47). As before, to trace the movements following adjustment, the top ten CASSRs, as identified using the unadjusted PI, are marked green and the bottom ten are marked in blue. There is a lot of movement out of the bottom and top ten. After production function adjustment only four of the bottom ten and five of the top ten remain the same and after risk adjustment only five of the bottom ten and seven of the top ten remain the same. Most of the movements in placing are fairly small, but some are much greater with maximum movements of over 50 places (see Figure 50 and Figure 51). Importantly, placings are very similar irrespective of whether PIs are production function adjusted or risk-adjusted, with only a handful of CASSRs having movements greater than ten positions (Figure 52). The top and bottom ten CASSRs are also fairly consistent, with eight of the ten CASSRs remaining the same for the different adjustment models. This suggests that the production function adjustment and risk adjustment approaches in general identify the same CASSRs as outliers and produce the same ordering of CASSRs.

Considering now the effect of adjusting for nonresponse, this has a small effect on inferences about performance in the context of case-mix adjustment using the production function method. The correlation between unadjusted and production function adjusted PIs is very similar, albeit marginally higher after adjustment with a slightly lower percentage of CASSR pairs changing order (Table 51). The correlation between production function adjusted and risk-adjusted PIs is also very similar whether estimated by CCA or using MI (Table 50). When the production function adjusted PIs are calculated on the multiply-imputed dataset from the theoretically-driven (SSV) specification, the number of outliers is very similar (Table 52), as are the number of 'false positives' and 'false negatives' (Table 53). There are slightly less extreme changes in the ranking of CASSRs when adjusted PIs are estimated by MI, compared to by CCA (Figure 51)[1]. There is very little effect for individual CASSRs for this particular sub-group. This suggests that taking nonresponse into account has only a small effect on performance assessment, for this particular sub-group.

[1] The distributional statistics confirm this: the standard deviation is 19.5 for the MI dataset and 21.5 for CCA.

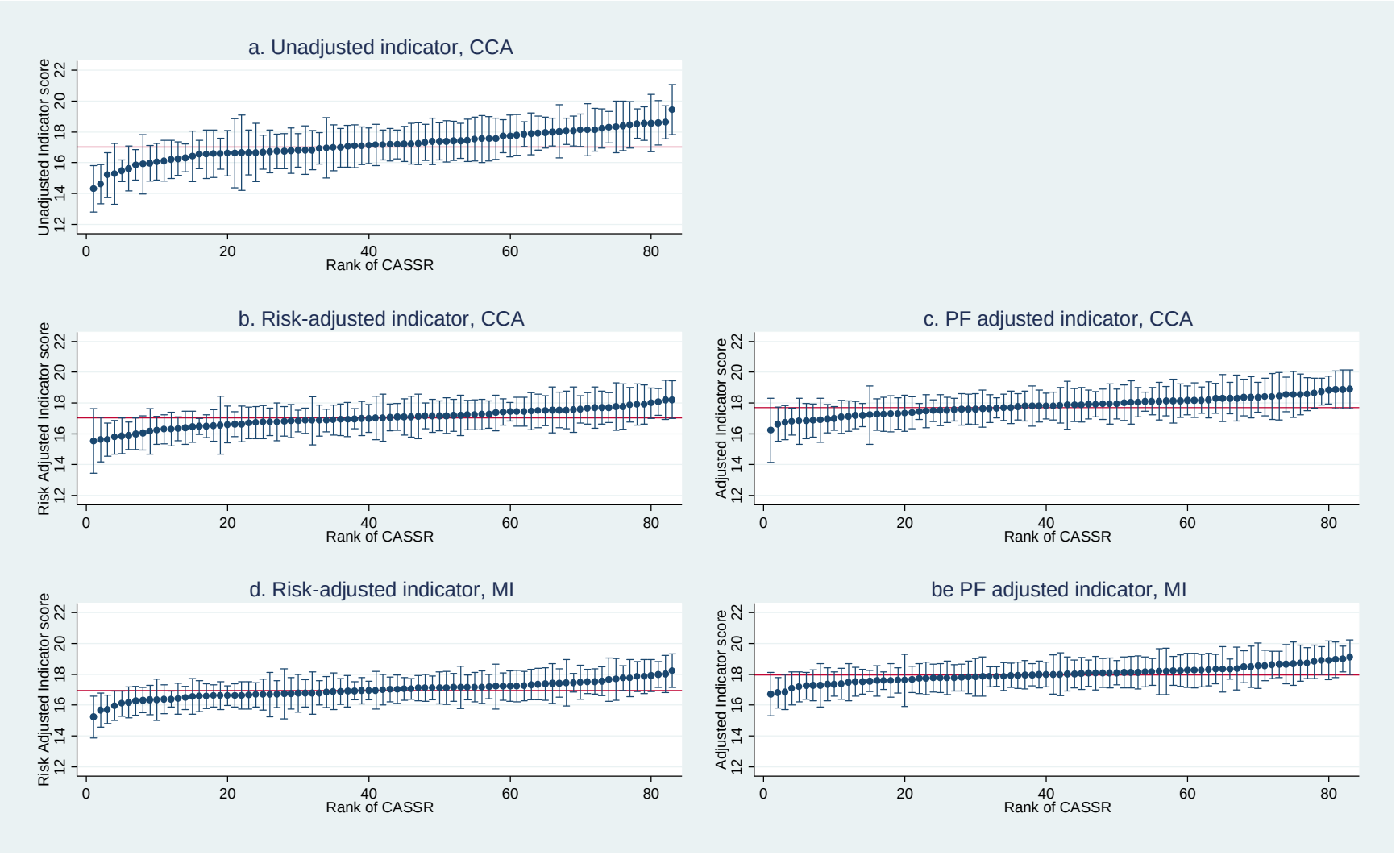

Figure 44: Caterpillar plots showing SCRQoL scores for CASSRs, theoretically-driven specification, 18 to 64 sub-group

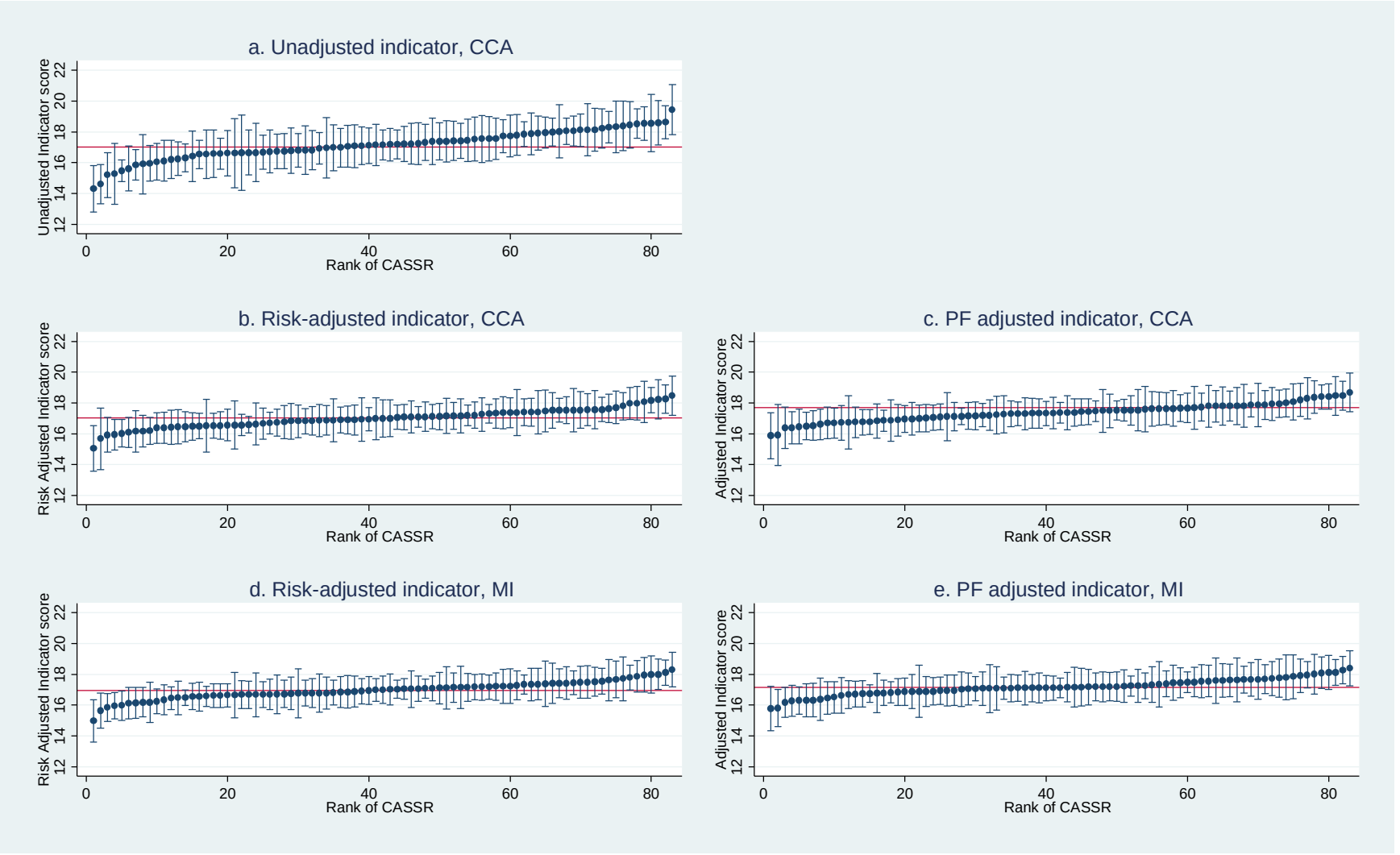

Figure 45: Caterpillar plots showing SCRQoL scores for CASSRs, statistically-driven specification, 18 to 64 sub-group

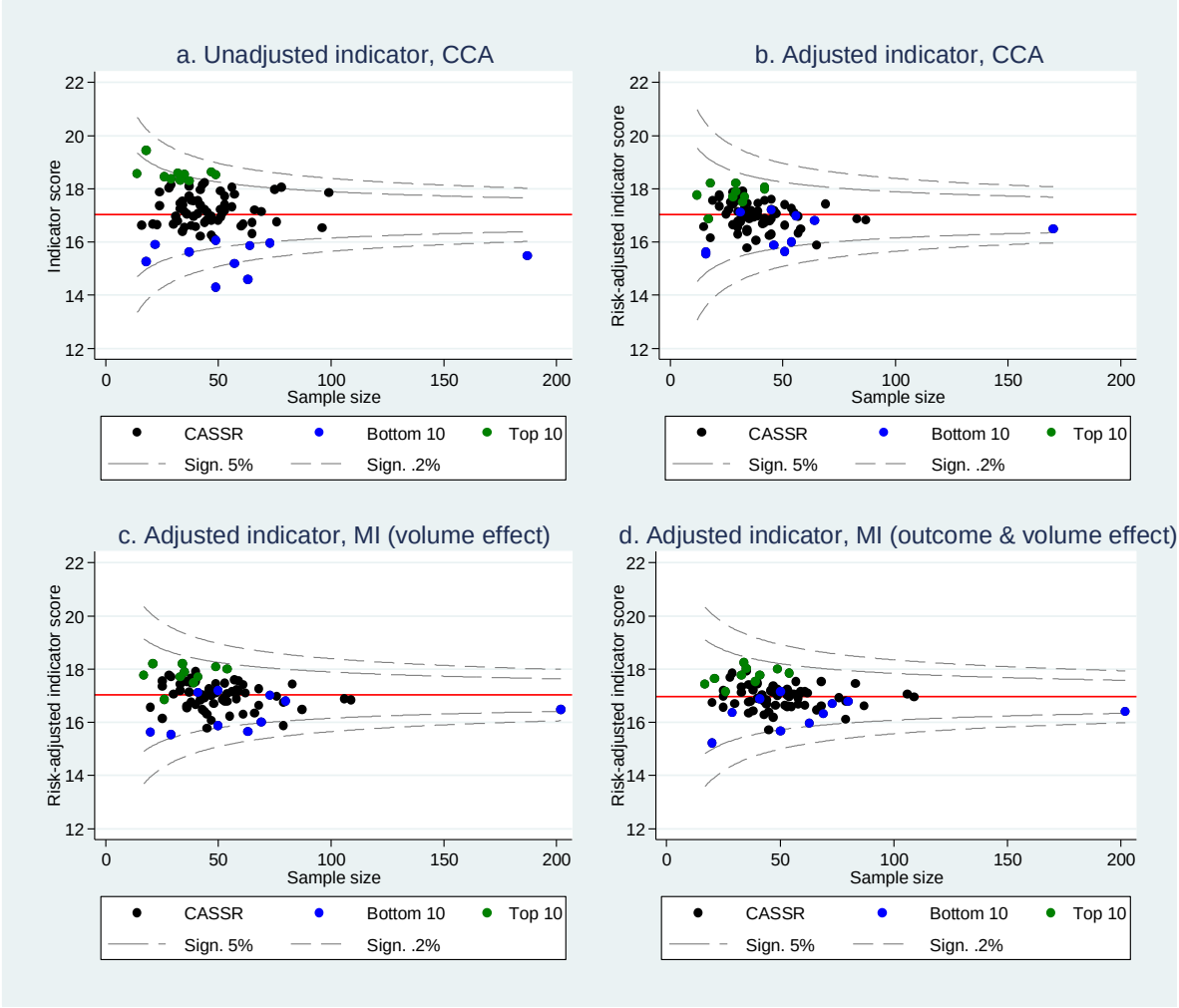

Figure 46: Funnel plots for SCRQoL PI, before and after risk-adjustment, theoretically-driven specification, 18 to 64 sub-group

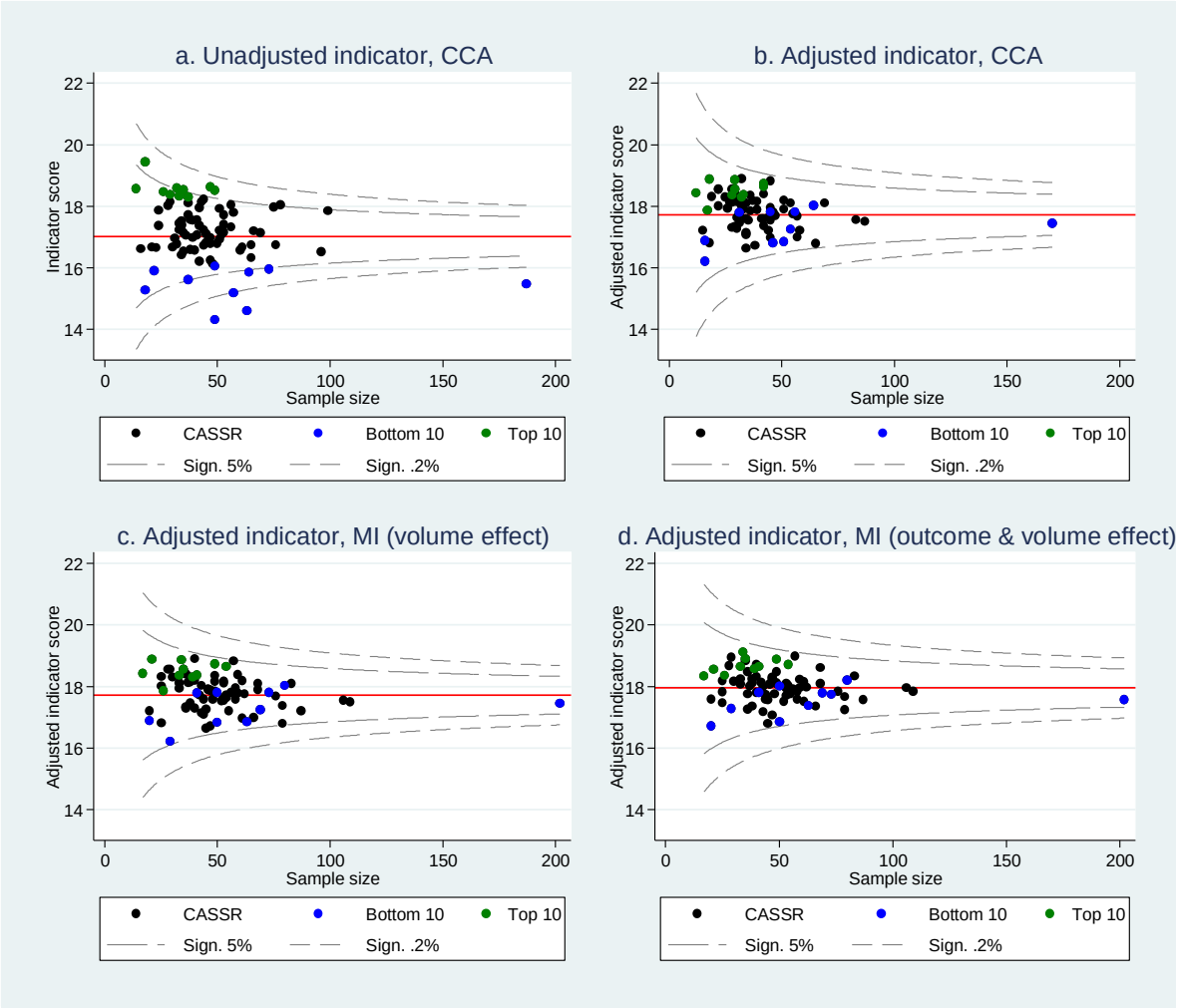

Figure 47: Funnel plots for SCRQoL PI, before and after production function adjustment, theoretically-driven specification, 18 to 64 sub-group

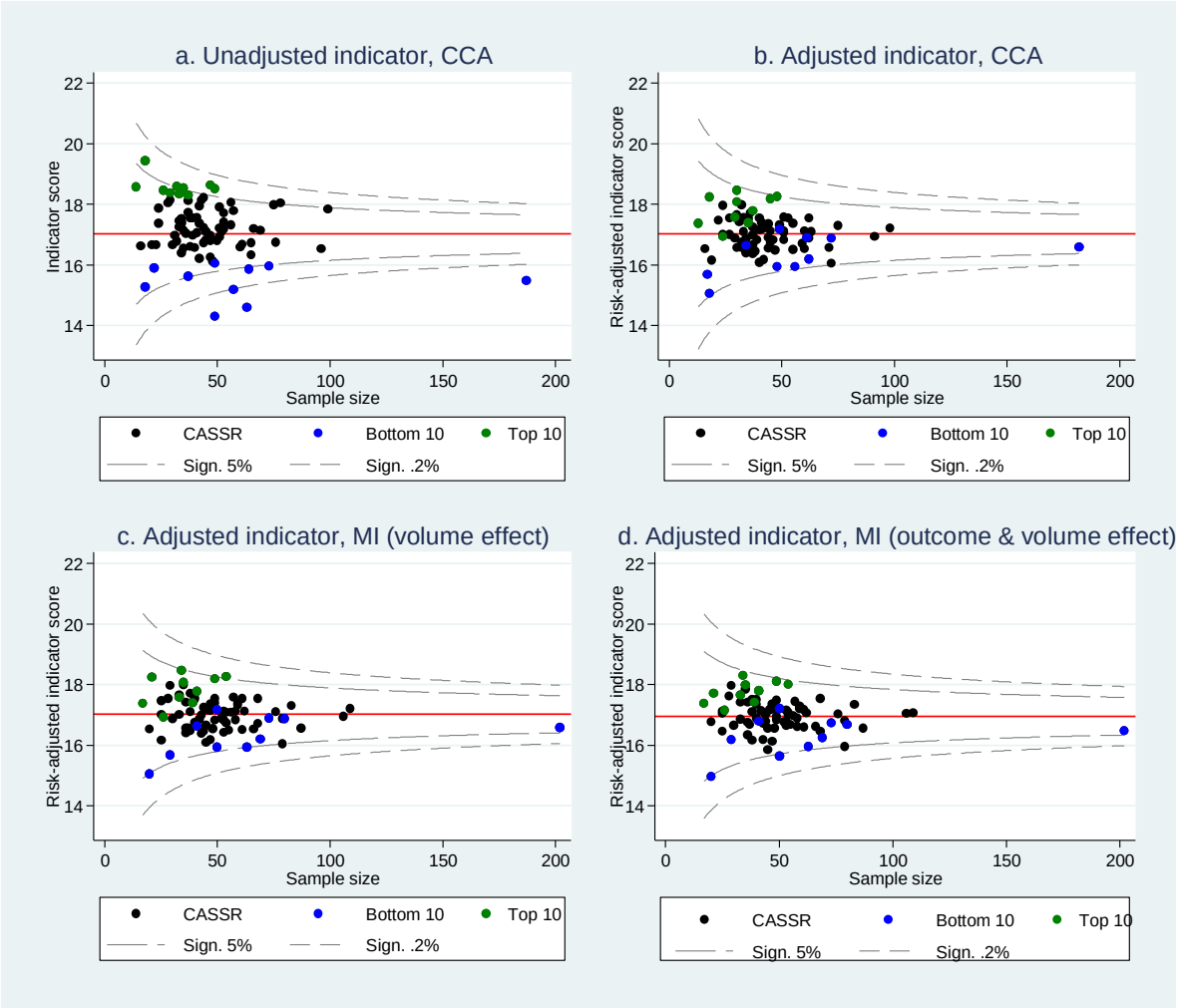

Figure 48: Funnel plots for SCRQoL PI, before and after risk-adjustment, statistically-driven specification, 18 to 64 sub-group

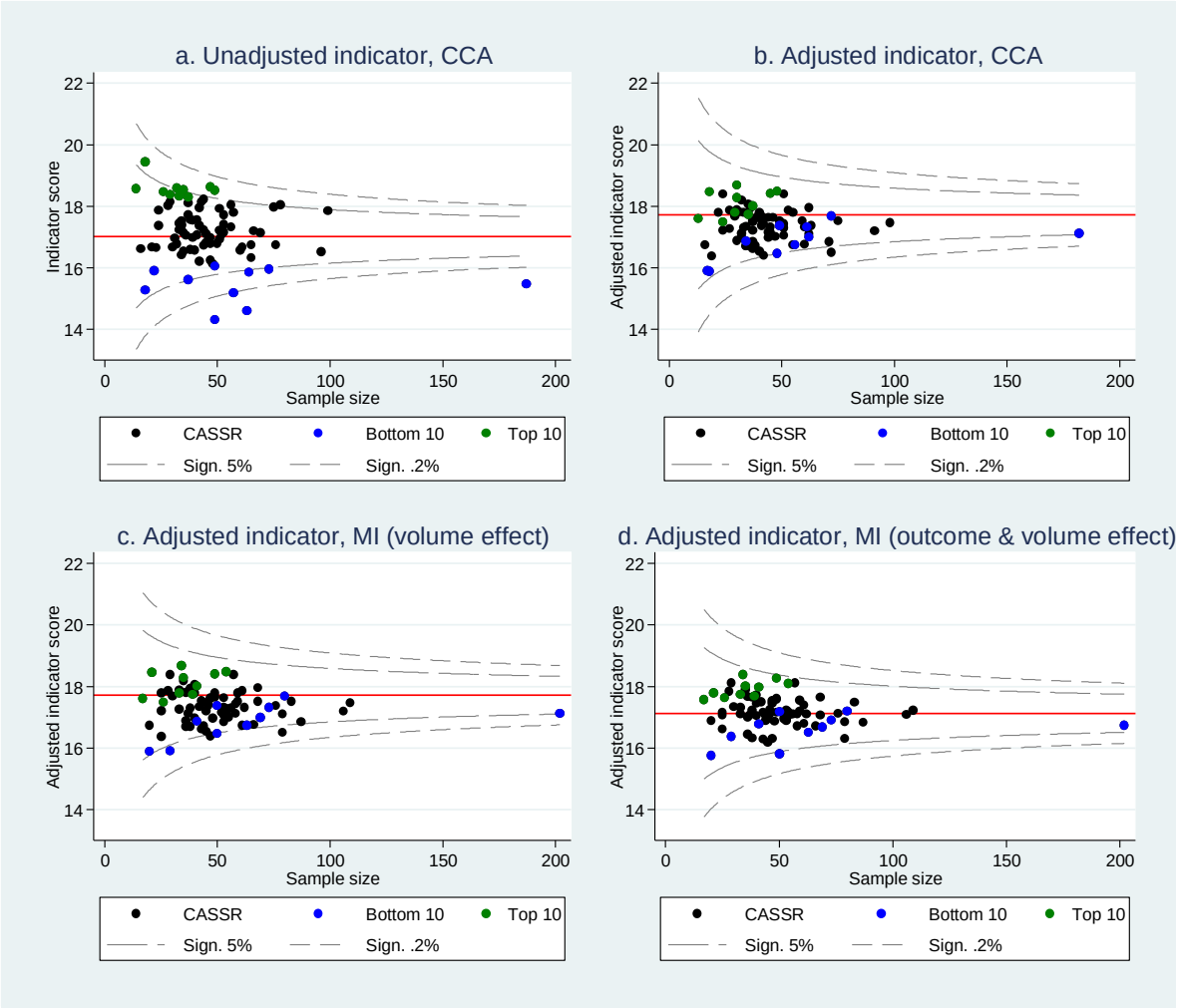

Figure 49: Funnel plots for SCRQoL PI, before and after production function adjustment, statistically-driven specification, 18 to 64 sub-group

Table 52: Number (percentage) of CASSRs identified as outliers for unadjusted and adjusted indicators (n=83)

Types of adjustments to indicators	Outlier status for Simplified, SV specification					Outlier status for EP specification				
	High alarm y>-3 SD	High alert y>2 SD	'Normal' range	Low alert y>-2 SD	Low alarm y>-3 SD	High alarm y>-3 SD	High alert y>2 SD	'Normal' range	Low alert y>-2 SD	Low alarm y>-3 SD
Unadjusted	0 (0%)	6 (7%)	67 (81%)	6 (7%)	4 (5%)	0 (0%)	6 (7%)	67 (81%)	6 (7%)	4 (5%)
Adjusted indicators										
OLS RA1, CCA	0 (0%)	0 (0%)	81 (98%)	2 (2%)	0 (0%)	0 (0%)	0 (0%)	83 (100%)	0 (0%)	0 (0%)
OLS RA1, MI VOL	0 (0%)	0 (0%)	81 (98%)	2 (2%)	0 (0%)	0 (0%)	1 (1%)	81 (98%)	1 (1%)	0 (0%)
OLS RA2, MI	0 (0%)	0 (0%)	82 (99%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	80 (96%)	3 (4%)	0 (0%)
OLS PF1, CCA	0 (0%)	0 (0%)	83 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	83 (100%)	0 (0%)	0 (0%)
OLS PF1, MI VOL	0 (0%)	0 (0%)	82 (99%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	83 (100%)	0 (0%)	0 (0%)
OLS PF2, MI	0 (0%)	0 (0%)	83 (100%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	82 (99%)	1 (1%)	0 (0%)

Legend: OLS RA1, RA OLS model estimated on CCA dataset; OLS RA2, RA OLS estimated on MI dataset; MI, VOL volume effect from multiple imputation; OLS PF1, OLS production function estimated on CCA dataset; OLS PF2, OLS production function estimated on MI dataset; SSV, theoretically-driven; EP, statistically-driven. Average SCRQoL score used as target for risk-adjustment models, calculated on CCA or MI dataset as appropriate. Average SCRQoL score plus difference between average SCRQoL and average production function-adjusted score used as target for production function models, calculated on CCA or MI dataset as appropriate.

Table 53: Movements into and out of 'control' status following risk adjustment and adjustment for nonresponse†

Indicator	Simplified SV specification						EP specification					
	OLS, RA			OLS, PF			OLS, RA			OLS, PF		
Missing data adjustment	CCA	MI, VOL	MI 2	CCA	MI, VOL	MI 2	CCA	MI, VOL	MI 2	CCA	MI, VOL	MI 2
Movements into low 'alert' status from												
in control	1	1	0	0	1	0	0	1	2	0	0	0
low alarm	1	1	1	0	0	0	0	0	1	0	0	1
Movements into 'control' from												
low alert	2	2	2	2	2	2	2	2	2	2	2	2
high alert	6	6	6	6	6	6	6	5	6	6	6	6
low alarm	3	3	3	4	4	4	4	4	3	4	4	3
high alarm	0	0	0	0	0	0	0	0	0	0	0	0
'False positives'	11	11	11	12	12	12	12	11	11	12	12	11
'False negatives'	1	1	0	0	1	0	0	1	2	0	0	0
'Type I error rate'	13.6	13.6	13.4	14.5	14.6	14.5	14.5	13.6	13.8	14.5	14.5	13.4
'Type II error rate'	100.0	100.0	--	--	100.0	--	--	50.0	100.0	--	--	--

Legend: † Indicators are based on OLS regression; MI, VOL volume effect from multiple imputation; MI 2, MI using OLS estimates from MI models; RA, risk-adjustment; PF, production function; SSV, theoretically-driven; EP, statistically-driven; zero movements into alarm status and high alert status, so these rows are not shown; -- not estimable because denominator is zero (no 'true positives' or 'false negatives').

Figure 50: Distribution of CASSR changes in rank (Unadjusted – Risk-adjusted) for the 18 to 64 sub-group

Legend: simplified SV, theoretically-driven; EP, statistically-driven

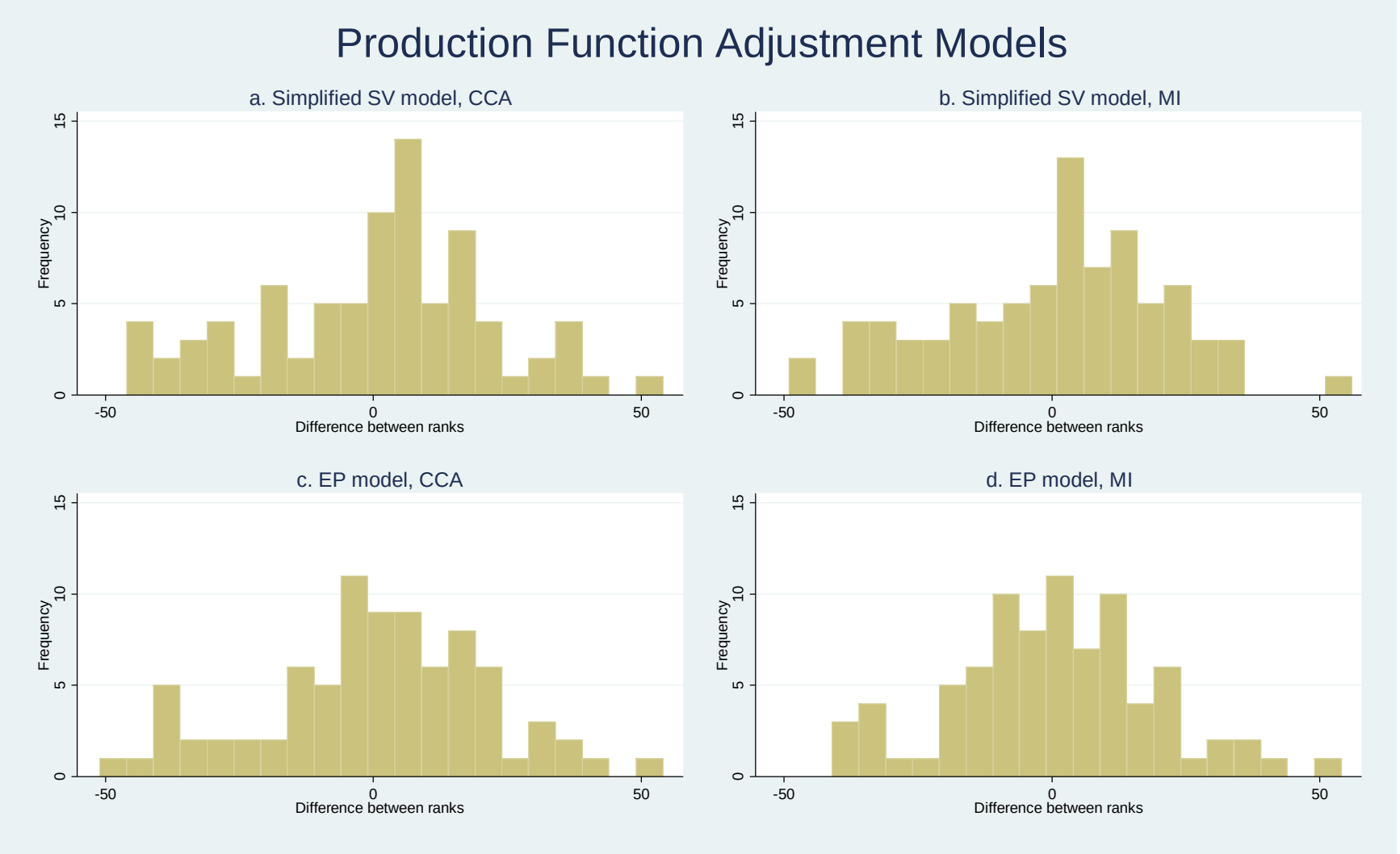

Figure 51: Distribution of CASSR changes in rank (Unadjusted – production function-adjusted) for the 18 to 64 sub-group

Legend: simplified SV, theoretically-driven; EP, statistically-driven

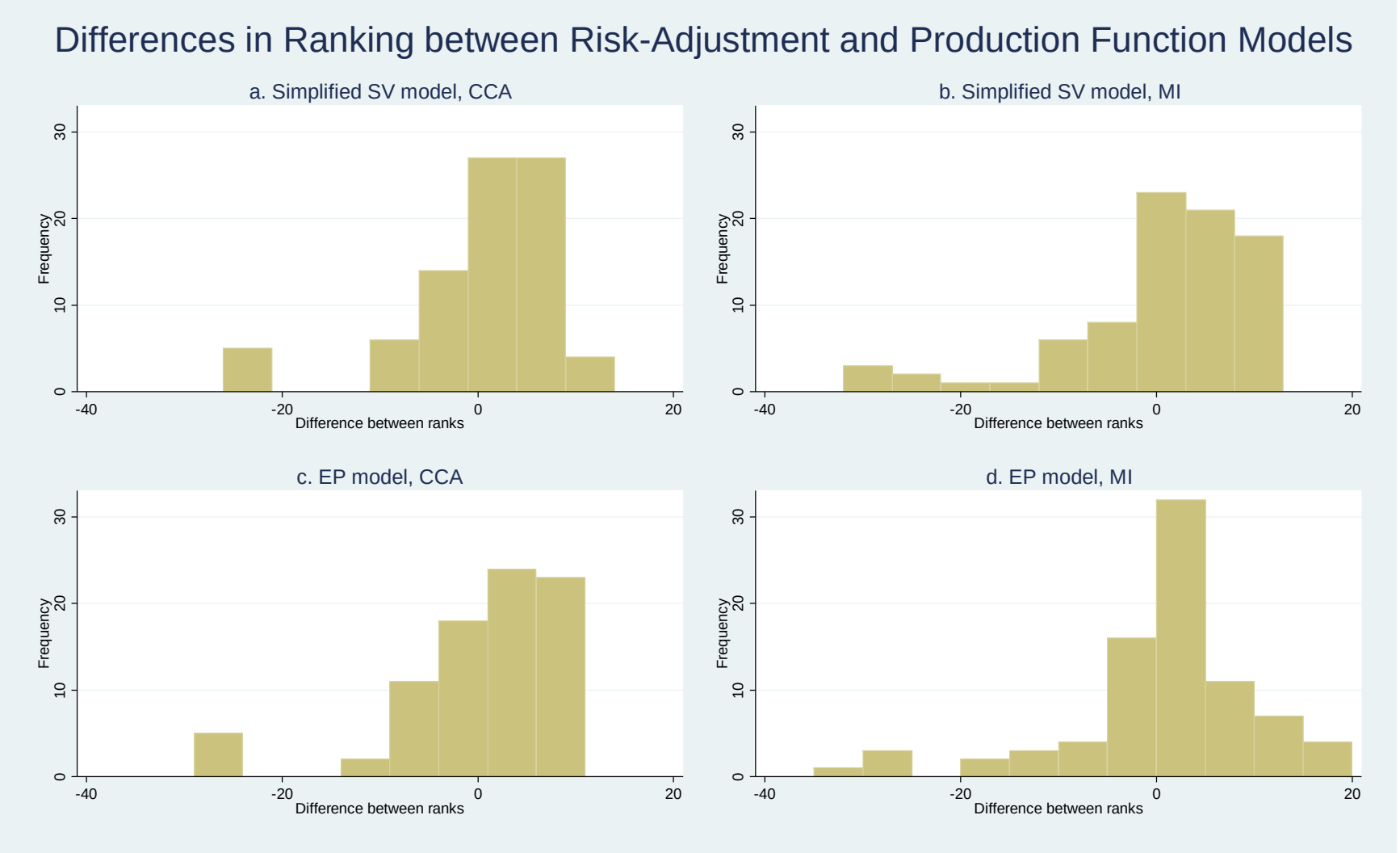

Figure 52: Distribution of CASSR changes in rank (Risk-adjusted – production function-adjusted) for the 18 to 64 sub-group

Legend: simplified SV, theoretically-driven; EP, statistically-driven

Discussion

The primary goal of this analysis was to answer the question of whether case-mix adjustment affects the interpretation of PIs from performance surveys. I have argued that this question is important because the differences in case-mix between CASSRs are alternative explanations for any observed differences in PIs scores. In this analysis I focused on the SCRQoL indicator derived from the ASCS data. Reflecting the decision I took to develop sub-group specific regression models for the PSD subsample of the ASCS, I examined whether adjustment has an effect on the performance of CASSRs for each of the PSD sub-groups and the overall PSD subsample. To adjust the SCRQoL indicator, I standardised the observed outcome for each CASSR by its expected outcome, as calculated from the regression models in Chapter 6 and 7. I used several different regression methods to estimate the expected outcome, to reflect as far as possible the clustered structure of the data and skewed distribution of the SCRQoL indicator. I also explored models with different variable specifications, including a version of the model akin to a production function that included a term for the effect of services. Additionally, the PIs were calculated under two different assumptions about the missingness mechanism – MCAR and MAR – to understand whether nonresponse has an impact on performance assessment where there is adjustment for case-mix. The intention is that this analysis will inform decisions about the analytical steps that should be taken to improve the validity of survey-based PIs.

This analysis has shown that adjusting the SCRQoL indicator for differences in the case-mix of CASSRs has an important effect on performance assessment, resulting in many changes in the rank position of CASSRS and their outlier status. The method used to adjust PIs has only a small effect on performance assessment. Model specification is slightly more important than the choice of statistical method, which is consistent with much of the literature in this area (Hannan et al. 1997, Huang et al. 2005, Mukamel et al. 2008, Li et al. 2009, Iezzoni 2013). In what follows, I discuss these results in more detail and consider the main limitations of the analysis presented here. I consider the implications of this analysis for policy and research in Chapter 9.

The effect of adjustment for case-mix on performance assessment

This analysis has shown that adjusting for differences in the case-mix of CASSRs has clear effects on performance assessment. There were some variations in the effects of case-mix adjustment on performance assessment across sub-groups and according to the method

used for adjustment, but in general the effects were highly consistent. Notably, case-mix adjustment results in changes to the ranking of CASSRs, with some CASSRs seeing quite large changes in rank position after adjustment. There was also a lot of movement in the top and bottom ten positions. Case-mix adjustment also substantially reduces the number of CASSRs that are outliers. It seems likely that sample size makes a difference to the importance of the effect of case-mix on outlier analysis, as the 'false positive rate' for the larger 65 and over sub-group and the PSD subsample was much higher than the 'false positive rate' for the two smaller (18 to 64 and care home) sub-groups. For all analyses, the 'false positive rate' was substantially over the five per cent level that is generally considered acceptable. Assuming that the case-mix adjusted SCRQoL PIs are more valid indicators of performance than the unadjusted PIs, this analysis provides strong evidence for using case-mix adjusted PIs for performance assessment.

I reflect on the question of whether the case-mix adjusted PIs are more valid indicators of performance than the unadjusted PIs in Chapter 9. For now, I focus on the implications of the analysis for the choice of adjustment method. In general, the method used for standardising the observed outcome made only a marginal difference to performance assessment. Some method choices, however, did matter more than others and I discuss each of these decisions, i.e. concerning the estimation method, the specification of the model, the choice between the production function or the risk-adjustment model, and the method for standardising the observed outcome, in turn.

I take first the estimation method used for modelling the SCRQoL outcomes. The choice between the OLS, RE, FE and FR regression had no consequences for performance assessment. Although the RE, FE and FR models may have greater face validity than the OLS estimation method, at least for these data, this was not important for performance assessment. For the SCRQoL PI, as far as I have been able to test here, the OLS provides a reasonable basis for adjustment.

The model specification for these data, by contrast, did have some consequences for performance assessment. I used two specifications for the statistical models: one based on a more theoretically-driven approach for retaining variables in the model (the SV model) and another based on a more statistically-driven approach for retaining variables in the model (the EP model). There was, in general, a fairly high correlation between PIs based on the two model specifications and fairly good agreement between the two specifications over which CASSRs were identified as outliers, this is despite some differences in the list of risk-adjustors. For the 18-64 age group the statistically-driven

model included self-perceived health, self-perceived home design and anxiety/depression, while the theoretically-driven model included these variables and ADLs, the help with finances/paperwork IADL, practical help, gender and the source and type of assistance the person had to complete the questionnaire. For the over 65 sub-group, the statistically-driven model included self-perceived health, self-perceived home design, anxiety/depression and ADLs. In addition the theoretically-driven specification included: the help with finances/paperwork IADL, practical help, gender, whether the questionnaire was completed by a proxy respondent, the source and type of assistance the person had to complete the questionnaire, and whether the person made private contributions to their care needs. It is notable that the two specifications had a similar R^2 for both of these sub-groups.

There was less good agreement on the rank position and outlier status of CASSRs using the two specifications for the care home sub-group. It is notable that for the care home sub-group, the statistically-driven specification included many fewer variables as risk-adjustors compared to the theoretically-driven specification and had a substantially lower R^2. The statistically-driven specification included only anxiety/depression and practical help, whereas the theoretically-driven specification also included self-perceived health, gender, age group, whether the questionnaire was answered by a proxy respondent, whether the person had help to complete the questionnaire and the source and type of help given, and whether the person made additional contributions to the cost of their care. The degree of difference between PIs based on these two model specifications for the care home group, suggests that the statistically-driven method, may sub-optimally adjust for case-mix differences between CASSRs for this sub-group.

The statistical method used to generate the statistically-driven specification, was based on the method proposed by Zaslavksy et al. (2001), in which variables are selected on the basis of their explanatory power. Here explanatory power refers to the predictive power of the variable and the extent to which it varies across CASSRs. The criterion used to select variables on the basis of explanatory power seems fairly arbitrary and it may be too stringent for the care home dataset. Sensitivity analysis may help to understand the appropriate criterion to use for different datasets; however, the sub-optimal adjustment of the SCRQoL PI for the care home sub-group suggests that a more theoretically-driven approach to retaining variables in the model is preferable (Iezzoni 2013). These findings also confirm the results of previous research that stresses the importance of the

specification of the case-mix adjustment models for performance assessments (Arling et al. 2007, Mukamel et al. 2008).

A further choice with respect to the statistical modelling is between the risk-adjustment and production function adjustment models. Although there were quite large differences in the estimates of the absolute scores for the adjusted SCRQoL indicator with the production function and risk-adjustment methods the choice of model made little difference to the ranking of CASSRs or the identification of outliers. This suggests that the relative values of PIs do not change substantially as a result of the choice of model. One explanation for this is that there is little difference in the relationship between the intensity of services that people receive (as measured by the budget variable) and outcomes across CASSRs given the observed risk-adjustors; another is the data limitations discussed in Chapter 7. The implication of this analysis is that the risk-adjusted PI provides a fairly good indicator of the relative performance of CASSRs. Although, as I noted in Chapter 7, the production function model provides a better estimate of effectiveness, so the choice between the two approaches depends on how the data are to be used, as I discuss in Chapter 9.

The choice between whether to standardise the observed outcome using the error method (observed – expected) or one of the two ratio methods (observed/expected) had little effect on performance assessment. The ranking of CASSRs was almost identical and the CASSRs identified as outliers were also in most cases identical. For the PSD subsample, however, the ratio method (where ratios were estimated for individuals and then averaged across cases within a CASSR) identified more CASSRs as low outliers than the error method (and the more conservative version of the ratio method where expected outcomes are estimated for CASSRs given their 'average' risk profile). This reflects the weighting implicit in the construction of the indicator, which gives more weight to large differences between observed and expected outcomes where observed SCRQoL is low than it does to large differences between observed and expected outcomes where observed SCRQoL is high. It is not clear why only the PSD subsample was affected, although the larger sample size used for estimating the CASSRs may be a factor. The validity of the individual ratio approach depends on how certain we can be that individuals with poor SCRQoL could have better outcomes. Given the difficulties capturing the tail of the distribution, discussed in Chapter 6, there is reason to doubt that this is the case. Consequently, as I discuss in Chapter 9, the ratio method may not be the fairest way of constructing PIs for these data.

The effect of adjusting for nonresponse, where PIs are adjusted for case-mix

Taking nonresponse into account has a small effect on inferences about performance, with the effect depending to some extent on the specification of the adjustment model and on the sub-group. In general, for these data, adjusting for nonresponse seems to moderate the effects of adjustment for case-mix. Adjusted PIs estimated on the MI dataset tend to have higher correlations with the unadjusted PIs, identifying more CASSRs as outliers and having fewer changes in rank position compared to the unadjusted PIs. It should be emphasised that the differences between the adjusted PIs estimated on the MI dataset and by CCA are in all cases fairly small. Nevertheless, this analysis suggests that some of the differences between the unadjusted and case-mix adjusted PIs arise due to loss of data.

Importantly decomposition of the effect of adjusting for nonresponse suggests that the main driver of the differences observed between the case-mix adjusted PIs estimated by CCA and on the MI dataset is a mean outcome effect and *not* a volume effect. In other words the correction for bias has a more important effect on performance assessment than the increase in the precision of the estimated PIs. The implication is that for the SCRQoL PI, for these sub-groups, the loss of data associated with adjusting PIs for case-mix has a biasing effect on the resultant PIs. It was also notable that, in this dataset, one CASSR was significantly affected by bias arising from nonresponse. The effect of nonresponse in the presence of case-mix adjustment seems to be highly dependent on the individual patterns of missingness within CASSRs. I consider the implications of this for policy in Chapter 9.

Limitations of the methods and data and some directions for future research

This analysis exploring the effects of adjustment on performance assessment is dependent on the regression models developed in Chapters 6 and 7, since these form the basis for estimating the expected outcomes used to standardise the observed outcome. The findings are therefore subject to the limitations associated with the models, which were discussed in some detail in Chapters 6 and 7. Notably, the concerns I raised over the poor fit of the models to the data, the possibility of omitted variable bias in the estimates due to fact that some theoretically important variables are unobserved, the small numbers in some of the sub-groups which prevented a validation study and the accuracy of the budget variable.

All of these problems with the statistical models mean there may be some bias in the case-mix adjusted PIs. The strong correlation between the PIs adjusted using the various different methods, however, is a strength of the analysis and suggests that the

results are fairly robust to violations of the model assumptions. The results presented here also agree with those reported by Forder et al. (2016) for the IIASC study[1] (discussed in Chapter 7). Despite the very different results for the estimate of effectiveness found between the IIASC study and the production function models presented in Chapter 7, the IIASC study found a similar level of agreement between production function adjusted and risk-adjusted PIs and a similar correlation between the production function adjusted PI and the unadjusted PIs. This suggests that despite the problems with the statistical models, the findings from this study provide a good indication of the relationship between PIs based on the production function and risk-adjustment models.

The small sample size of the care home and 18 to 64 sub-groups needs further consideration since it has implications beyond the robustness of the models. The identification of outliers depends on the precision of the estimates and it seems likely that the small sample size of most of the CASSRs in these sub-groups hindered the outlier analysis. The funnel plots for the care home and 18 to 64 sub-groups still show a spread of points after adjustment, but very few (and usually zero) CASSRs are outliers after adjustment. This is in contrast to the larger 65 and over sub-group where a number of CASSRs are still identified as outliers after adjustment. The small numbers limited the power of the outlier analysis for these groups, and therefore my ability to understand the difference between the methods for adjustment on the identification of outliers within these sub-groups. An important implication of this is that the sub-groups are too small to warrant separate reporting and analysis. I reflect on the implications of this for case-mix adjustment of PIs in Chapter 9.

The proportions of people within each of the PSD sub-groups varies quite substantially across CASSRs, which raises questions about how comparable the adjusted PIs are for the PSD subsample where the three sub-groups are combined into one PI[2].

---

[1] The IIASC study (Forder et al. 2016), was designed to capture better information on service receipt and user characteristics to enable a more thorough investigation of the relationship between service use, needs-related characteristics and outcomes in the context of outcomes adjustment. As well as collecting data from social care records, the IIASC study interviewed service users to capture information on their service use alongside a range of other data on their characteristics and SCRQoL. Since service use was collected directly from users, a cost-weighted utilisation measure[1] could be generated using average national unit costs. This is more accurate than the budget variable from social care records that I used in this analysis, as it is not (directly) influenced by regional price differences.

[2] The method for combining the sub-group results together is consistent with indirect standardisation. It involved calculating adjustment factors for each individual using the sub-group specific regression equation and averaging across the adjustment factors to generate an expected outcome and hence adjusted-PI. This method effectively adjusts for sub-group specific differences in outcomes, through the sub-group specific

Indirect standardisation provides fairly good adjustment for case-mix where the samples have similar case-mix profiles, but where there are large differences in the case-mix of organisations indirectly standardised PIs are not comparable (Glance et al. 2000, Julious et al. 2001, Rixom 2002, Julious and George 2007). Indirect standardisation tends to be used for case-mix adjustment as the alternative more correct method, direct standardisation, is usually not feasible. In direct standardisation, the effects of case-mix variables on observed outcomes within each CASSR would be applied to a reference population to estimate the expected outcome for the reference population had the effects observed for the CASSRs applied. This method requires large sample sizes for each CASSR and cannot be used with large numbers of variables or with continuous variables (Nicholl et al. 2013). In Chapter 9 I discuss directions for future research that could improve this facet of the analysis.

The adjustment for the effects of nonresponse on the indicators was restricted in this chapter to an analysis of the effects of item nonresponse on performance assessment using MI. In Chapter 5, I explored the effect of adjusting for both item and unit nonresponse on PI scores, using the combined IPW/MI method, which also investigated the effects of accounting for unit nonresponse on performance assessment. I did not apply the IPW/MI approach to these data because the strategy of conducting the analysis by sub-group is not compatible with the sample-level IPW analysis. Applying the weights to the sub-groups could distort adjusted-PIs as the total of the effective weights would depend on the ratio of the size of the sub-group to the total respondent sample, which would vary for each CASSR. The lack of weighting for unit nonresponse means, however, that differences between CASSRs in PI scores may be a consequence of unobserved heterogeneity arising from differences in the characteristics of respondents and nonrespondents between CASSRs. If IPW/MI adjustment has the same effects in this context as it did in Chapter 5, additional adjustment for unit nonresponse may have produced marginally greater differences in the results.

Concluding remarks

A clear implication of this analysis is that it is important to adjust the SCRQoL PI for case-mix, to avoid spurious conclusions regarding the relative performance of CASSRs. I make

---

regression equations. It is equivalent to a regression for the entire sample including interaction terms between the sub-group and all covariates.

this assertion on the assumption that the adjusted-PI has greater validity as an indicator of performance than the unadjusted SCRQoL PI. There is no benchmark against which to assess the validity of the unadjusted and adjusted indicators. In Chapter 9, I discuss why the analyses from across Chapters 6, 7 and 8 suggest that the adjusted-PIs are more valid indicators of performance than the unadjusted PI.

I have also discussed the implications of this analysis for the method that should be used to adjust the SCRQoL PI for case-mix. While the estimation procedures appear to make little difference to performance assessment, the choice of risk adjustors is important. The specification that selected risk-adjustors on the basis of theoretical and empirical relationships appeared to optimally adjust for case-mix. Although the others aspects of the method were not important for conclusions regarding performance assessment, they do have consequences for the face validity of the results. I consider the trade-offs between the different choices and their implications for the ASCOF and other studies in Chapter 9.

The effect of adjusting for nonresponse on inferences about the performance of CASSRs warrants further investigation for the SCRQoL indicator. The pattern of missingness drives the results reported here, which means for some CASSRs with particular patterns of missingness adjustment for nonresponse can have significant implications for performance assessment. In adjusting for nonresponse I make an assumption about the missingness mechanism in order to 'observe' the effects of nonresponse on PIs. In this analysis I made the assumption that the missingness mechanism is MAR for item missingness among the respondents, i.e. dependent on the value of the observed variables, and MCAR for unit missingness. As I have suggested this assumption is a limitation of the methods used here and these assumptions may both be false: item missingness may in fact be NMAR and unit missingness may be either MAR or NMAR. I consider the implications of this limitation for the study conclusions and future research in Chapter 9.

Chapter 9
Discussion, Policy Implications and Conclusions

Introduction

In this thesis I set out to answer the following research question: *if survey-based indicators are to provide valid evidence concerning performance for routine use, what strategies should be used to address the confounding influence of observed and unobserved heterogeneity?* The validity of survey-based indicators is drawn into question due to the possibility that variations in scores may be explained by factors unrelated to the performance of services. I have used data from the Adult Social Care Survey (ASCS) to examine two possible sources of heterogeneity in indicators drawn from this survey for the Adult Social Care Outcomes Framework (ASCOF): nonresponse and the characteristics of service users, often referred to as the 'case-mix'. I found that both nonresponse and differences in the case-mix of CASSRs influenced absolute ASCOF PI scores, but case-mix had a more important impact on performance assessment. Although generally less important, the effect of adjusting for nonresponse varied depending on the PI, which I argue is a consequence of differences between PIs in the number of cases with missing information. The method used to adjust for case-mix had very little effect on performance assessment, but the choice of risk-adjustors was important. The choice of method to adjust for nonresponse also had some implications for conclusions. In what follows, I argue that for valid inferences about performance it is necessary to (i) statistically adjust for differences in case-mix between CASSRs; (ii) to explore the effects of nonresponse on PIs and consider implementing post-hoc adjustment; and (iii) to consider implementing strategies to increase response rates from groups that are currently underrepresented in the ASCS.

The rest of this chapter is organised as follows. First, I discuss the main findings from the empirical analyses in this thesis from a policy perspective. Secondly, I consider the strengths and limitations of the analyses and set out some directions for future research. Finally, I conclude by reflecting on the contribution of this thesis to ASC policy and performance assessment research more broadly. I consider some of the questions raised by this thesis for these two fields of research and suggest an agenda for future studies in these two fields.

Key findings and their implications for policy

The analysis in the empirical sections of this thesis addressed the core research question through four sub-questions:

i. What is the most appropriate method for modelling ASC outcomes?
ii. What is the effect of adjusting for case-mix on inferences about performance and does the method of implementing adjustment matter?
iii. What is the effect of nonresponse on inferences about performance and does the method used to address differences in samples due to nonresponse matter?
iv. Who is missing from performance surveys and how can response rates be improved?

The first two questions relate to the problem of differences in the case-mix of CASSRs and the second two, to the problem of nonresponse. In both the case-mix and nonresponse analyses, where the concern was to establish the effect of adjustment on inferences about performance, I examined the effect on the ranking of CASSRs and on which CASSRs were identified as outliers.

The strategies explored in the empirical chapters were designed to improve the validity of the indicators. Should policymakers be solely concerned with validity then, the answer to the research question would depend only on the question of whether the strategies were effective in this regard. As I have discussed in the opening chapter, however, policymakers need to balance the validity of indicators against other qualities like feasibility and acceptability; they also need to consider the views of different stakeholders and the different uses for the data. The methods proposed here, particularly those to address nonresponse, improve validity but sometimes at the cost of feasibility (e.g. affecting the cost of data collection or the timeliness of the analysed data) and acceptability (e.g. affecting usability or transparency). When competing policy priorities are taken into account, a more nuanced answer to the research question is needed. In this section I discuss the key findings from the study, organised around the sub-questions above. I reflect on how the potential uses of the PIs affect whether, and how, policymakers may wish to address the effects of nonresponse and differences in the case-mix of CASSRs on survey-based PIs.

The role of observed differences in user characteristics on performance assessment

The case-mix analysis focused on the SCRQoL indicator, which measures the effect of ASC services on users' QoL. As is indicated by the research sub-questions, in the first

stage I explored different methods for modelling ASC outcomes. Due to the complexity of estimating a single model on a sample as diverse as the ASCS, all analyses were conducted on a subsample of the ASCS (people with a physical and sensory disability, PSD) and separate models were estimated for three sub-groups (care home residents, adults aged 18 to 64 in private households, and adults aged 65 and over in private households) within this subsample.

The statistical modelling identified a number of factors beyond the direct control of CASSRs that were important predictors of variations in SCRQoL. As suggested by the conceptual framework (Chapter 2) and consistent with previous research studies (Davies et al. 2000b, Fernández 2005, Glendinning et al. 2008a, Forder and Caiels 2011a, Forder et al. 2014, van Leeuwen et al. 2014), the factors that were found to be important included: underlying health and disabling conditions variables, such as ADLs and self-perceived health, perceptions of the accessibility of the home environment, variables capturing receipt of practical help and reporting-related factors, such as whether the person received assistance to complete the questionnaire. These factors explained a sizeable proportion of variation in SCRQoL scores. This underlines the importance of attending to variations in these characteristics for risk adjustment and when assessing the effectiveness of social care interventions more generally.

The analysis in Chapter 8 found that statistically adjusting PIs for differences between CASSRs in the characteristics of users had a substantial impact on performance assessment. There was a marked effect on PI scores for individual CASSRs, which had a substantial impact upon the ranking of CASSRs and meant that many fewer CASSRs were identified as outliers after adjustment. The effect of user characteristics on the absolute and relative performance of CASSRs as measured by SCRQoL, suggests problems for all uses of the crude SCRQoL PI. For example, if the unadjusted SCRQoL PI were to be used for benchmarking, investigation would frequently reveal that poor performance was explained by an unlucky case-mix and conversely that good performance was explained by a fortunate case-mix. Likewise were the unadjusted SCRQoL PI to be used for accountability purposes it would present a false impression of the relative performance of a large proportion of CASSRs. The dominance of SCRQoL outcomes by risk factors would also make it difficult for CASSRs to use the unadjusted PI for internal research and evaluation as inferences regarding effectiveness would be confounded by the characteristics of service users. Without adjustment for case-mix, there is a fairly high probability that resources will be wasted, either because end-users will make sub-optimal

decisions or because, having collected the performance data, they may not use it due to questions over its validity (Moynihan and Pandey 2010, Kroll 2015). There is therefore a strong argument for adjusting the SCRQoL PI to take account of the observed variations between CASSRs in case-mix.

The importance of methods choices for case-mix adjustment

I turn now to the question of how policymakers should adjust the SCRQoL PI for differences in the case-mix of CASSRs. The adjusted PIs were all generated by indirectly standardising the observed outcome for each CASSR by the expected outcome for the CASSR. In all cases the expected outcome was calculated using the results from a regression model estimated on a reference population. I estimated a number of regression models to produce a range of alternative expected outcomes and therefore adjusted PIs. I explored several different estimation methods, including OLS, fixed effects, random effects and fractional response regression, in order to address the negatively-skewed distribution of the SCRQoL indicator and the clustering of responses within CASSRs. Additionally I tested two specifications with slightly different selections of explanatory variables. This produced a number of alternative adjusted PIs that were very strongly correlated. Indeed, for all sub-groups, the differences between the crude and adjusted PIs were much greater than the differences between the alternative adjusted PIs. The strong correlation is good from the perspective of validity since it strengthens conclusions regarding the effect of case-mix adjustment on performance assessment. It also implies that the method used for indirect standardisation matters very little for the validity of performance assessment using the SCRQoL PI.

In most situations, the best method for adjustment will therefore be the one that provides the simplest explanation of the data. On this basis, OLS is the preferred estimation procedure for these data as it provides the simplest explanation of the data and produces results that are not significantly different to those from the more complex estimation procedures. Faced with similar results, some researchers have argued for fixed or random-effects models on the basis that these models have greater face validity (Hannan et al. 2005, Glance et al. 2006, Li et al. 2009). However, for ASC, simplicity of explanation is of paramount importance as social services departments have very few employees with research skills (Rainey et al. 2015). I therefore favour the OLS estimation procedure.

In Chapter 8 I argued that a model specification with more extensive risk-adjustment is desirable. I used two different methods to specify which risk adjustors to retain in the models: a more statistically-driven approach (the 'EP' approach) and a more theoretically-driven approach (the 'SV' approach).[1] The results discussed in Chapter 8 showed that for the two private household sub-groups the two model specifications were very similar. For the care home sub-group the specifications had different effects on performance assessment. As the statistically-driven specification excluded many theoretically-important variables and had a much lower R^2 it is my view that it under-adjusts for the effects of case-mix and has poor validity. Consistent with other studies (Arling et al. 2007, Mukamel et al. 2008, Iezzoni 2013), the analysis presented here supports a more extensive and theoretically-grounded approach to the retention of risk adjustors.

I estimated the SCRQoL outcome using a production function model to address the possibility that the risk-adjustment model produces downwardly-biased estimates of effectiveness. As discussed in Chapters 2 and 3, bias arises because the needs-related characteristics of service users (included in the model as risk factors for SCRQoL-outcomes) strongly predict the care package provided. A production function model is useful because it addresses confounding by directly including an indicator of service quantity in the model. Despite it being important from a theoretical perspective to include services in the statistical model, I showed in Chapter 8 that it did not make a large difference to inferences about the *relative* performance of CASSRs. The similarity in inferences regarding relative performance from these two models implies that there is little difference in the relationship between the care package that people receive and SCRQoL-outcomes across CASSRs given the observed service user characteristics. By corollary, the risk-adjustment model would appear to be adequate for assessing the relative performance of CASSRs using the SCRQoL PI.

It is important to note that while *relative* performance was more or less unaffected, the indicator estimates did vary between the production function adjusted and risk-adjusted PIs. This reflects the results from the production function models in Chapter 7, which showed that the indicator of service quantity was a significant predictor of SCRQoL. Theoretically, as outlined in Chapter 3, the production function model should provide a better estimate of effectiveness and for some uses of the SCRQoL PI the choice between

---

[1] The methods used for data reduction are explained in detail in Chapter 6 and 7.

these two versions of the adjusted PI could produce different results. This may occur where the interest is in exploring differences between groups of users within CASSRs, or across CASSRs, for internal research and evaluation. Although there is little evidence that CASSRs use the ASCS data in this way at present (Heath et al. 2015), such uses have been documented for NHS PROMs (Devlin and Appleby 2010, Basser 2015) and are cited as one of the main benefits of PROMs over, for example, satisfaction and the more delivery-oriented measures (PREMs) (Smith and Street 2013). Given how important a supportive environment is for ensuring the use of measures (Moynihan 2009, Moynihan and Pandey 2010, Kroll 2015), the production function adjusted PI, which gets closer to a value-added from social care indicator, may be a better PI for supporting local uses of the data.

To standardise the observed outcome, I argued in Chapter 8 that the difference between the observed and expected outcomes (referred to as the error method) should be used rather than the ratio of the observed outcome to the expected outcomes (referred to as the ratio method). This is because in some circumstances the ratio method (where the ratios were estimated for individuals and averaged to generate a PI for the CASSR) identified more CASSRs as outliers, particularly as low outliers. This reflects the weighting implicit in the ratio PI, which effectively punishes CASSRs more for poor performance serving people with low observed SCRQoL than for poor performance serving people with high observed SCRQoL. This is only fair where the statistical model has good explanatory power over the full distribution. Given the difficulties capturing the tail of the distribution, discussed in Chapter 6, the ratio method seemed likely to result in unfair treatment where CASSRs have a number of cases with extreme low values for SCRQoL. Similar concerns have been raised about the NHS PROMs PIs where the statistical model likewise does not explain the low-scoring cases well (Ara et al. 2013, NHS England Analytical Team 2013).

The role of adjustment for nonresponse on performance assessment

There was clear evidence that the missing-at-random (MAR) assumption is more plausible than the missing completely at random (MCAR) assumption for the ASCS data. I found that both unit nonresponse (i.e. participation in the survey) and item nonresponse (i.e. response to individual questions) were dependent on the observed characteristics of the sample members. In theory, approaches that address nonresponse to the ASCS should enhance the validity of the ASCOF PIs.

One strategy to address nonresponse is post-hoc adjustment. In this study I used an adjustment method that combined inverse propensity weighting (IPW) for unit nonresponse with multiple imputation (MI) for item nonresponse (referred to as IPW/MI). To explore the effects of nonresponse on performance assessment, I compared the results obtained using the indicators adjusted for nonresponse to those obtained by complete case analysis (referred to as CCA/CCA). I also compared the IPW/MI method, to two partial applications of the method that addressed only unit nonresponse (IPW/CCA) and only item nonresponse (CCA/MI).

The importance of nonresponse for performance assessment depended on the method used for post-hoc adjustment. The largest estimated bias due to nonresponse was generally found where IPW/MI was used. The greatest effects on performance assessment, i.e. on changes in the rank position of CASSRs and in outlier status, were also observed for this method compared to IPW/CCA and CCA/MI for most of the survey-based ASCOF PIs. In terms of precision, the CCA/MI approach was generally precision-enhancing while the IPW/CCA approach was always precision-reducing. For the SCRQoL PI, IPW/MI was on average precision-enhancing. The precision-enhancing effects of CCA/MI and IPW/MI are important since loss of precision from nonresponse adjustment has been given as a reason for not adjusting indicators in other performance surveys (Roland et al. 2009). It suggests that these two methods are more promising options for adjusting PIs than IPW/CCA.

In general, IPW/CCA had a similar effect on performance assessment across the PIs. By contrast, the effect of IPW/MI and CCA/MI varied across the ASCOF PIs. For these two methods the effects were greatest for the SCRQoL indicator, which had the highest rate of item nonresponse. For this indicator the IPW/MI method adjusted for bias from nonresponse without any loss of precision for the majority of CASSRs. Indeed the effect of adjusting the SCRQoL PI for nonresponse using the IPW/MI method is on average precision-enhancing. However, IPW/MI was not on average precision-enhancing for the other ASCOF PIs. I argued in Chapter 5 that the precision-enhancing effects for the SCRQoL PI may be a consequence of higher rates of item nonresponse for this indicator. This would be consistent with simulation studies which have shown that MI tends to have more effect with increasing rates of missingness (McKee et al. 1999). Interestingly White and Carlin (1999) suggest that MI starts to have an impact on results when the fraction of incomplete cases is around ten per cent, which is the level of item nonresponse for the

SCRQoL PI. The implication of this analysis is that it may be appropriate to adjust the SCRQoL PI for nonresponse using IPW/MI, but not the other ASCOF PIs.

While nonresponse appeared to explain some of the heterogeneity in PIs, even for the SCRQoL indicator the number of changes in rank position of CASSRs and the outlier status of CASSRs makes it difficult to determine whether post-hoc adjustment is warranted. This decision depends on whether the difference between the crude PIs and the nonresponse-adjusted PIs are meaningful and substantial. I have discussed at some length in Chapter 5 both the difficulty of establishing what is a meaningful change for as diverse a population as the ASCS[2], and the problem, due to sample size, with using statistical criteria to judge what is a substantial score change for CASSRs. Neither of these avenues provided much insight into the question of whether to adjust the PIs for nonresponse, so I suggested that a more useful approach is to assume that the adjusted score is the 'true' score and consider the 'false positive (and negative) rate' that can be tolerated for the sample of CASSRs.

The 'false positive rate' that can be tolerated will depend on how the PIs are used in the policy process. Under New Labour (1997-2010), PIs were used to control the behaviours of LAs, with serious negative consequences for poor performance. Such conditions put a large weight on the accuracy of scores over all other criteria and, therefore, on ensuring 'false positives' are minimised. A 'type one error' rate of one per cent may be acceptable under such a system, which would suggest that all the PIs, except the safety indicator, should be adjusted for nonresponse. In the current performance measurement system, PIs are used more as 'intelligence'[3], and, although PIs are published in caterpillar plots, negative consequences do not follow directly from poor performance on the ASCOF PIs (Department of Health 2010c, Department of Health 2014). Under such a system the acceptability of the PIs, as judged by transparency and simplicity, may have more weight than measurement accuracy. Even for the SCRQoL PI, where the 'type one error rate' is seven per cent, and therefore over the standard acceptable five per cent limit, the loss of transparency may outweigh the benefits from improved accuracy. For most uses of the nationally-reported PIs, i.e. benchmarking and accountability, the 'false positive rate' is probably within acceptable bounds for all PIs, except the SCRQoL

[2] This point has been noted by other researchers for other performance surveys (Valderas et al. 2012, Fernandez et al. 2013a).
[3] This is a term used by Hood (2007) to describe different ways of using performance information to improve the delivery of public services.

indicator. This suggests that policymakers may want to consider adjusting the SCRQoL indicator for nonresponse using the IPW/MI method.

There are, however, various problems with using IPW/MI for performance assessment studies. While IPW is fairly easy to use in analysis and has a straightforward logic, this is not the case for MI. To analyse multiply-imputed datasets requires a fairly high level of statistical knowledge and specialist statistical software. Most end-users of the data would not be able to analyse such data nor could they easily replicate the method. Consequently, it could not be incorporated by CASSRs into any local analysis, e.g. evaluating local programmes or other local research. It would be possible to use the IPW/MI method to adjust nationally-reported figures, but given the increased resources required for analysis, the possible effect on the publication timetable, and difficulties explaining the method to most end-users, there is a need to consider whether the increase in validity outweighs these disadvantages. Policymakers would need to consider what level of accuracy is appropriate for accountability and benchmarking.

A remaining question is what steps, if any, should be taken to address nonresponse where there is adjustment for case-mix. I used a CCA/MI procedure to account for the effects of nonresponse in the analysis of the case-mix adjusted PIs, as the IPW/MI method was not compatible with the sub-group analysis. The analysis suggested there was some bias and some loss of precision in PI estimates due to loss of data. Neither of the effects, however, had much of an impact on inferences about performance for the overwhelming majority of CASSRs. This is despite a much higher fraction of incomplete data in the CASSR samples used to estimate the case-mix adjusted PIs than in the rest of the nonresponse analysis, due to missingness in the variables included as quasi-inputs in the models. One CASSR in the 65 and over sub-group was an exception to this. Its performance substantially worsened after accounting for nonresponse. This finding suggests that nonresponse may be an important source of heterogeneity in survey-based indicators and that it is important to investigate the effects of nonresponse on performance assessment where PIs are adjusted for case-mix. Prior to the publication of the adjusted PIs it would seem sensible to conduct detailed research to rule out the possibility that differences in performance are due to an unfortunate pattern of missingness in the data.

The representativeness of the ASCS and implications for survey design

The complexities of post-hoc adjustment mean that policymakers may prefer to explore whether it is possible to alter the design of the survey to increase response rates. This is an

attractive option where only a fairly small fraction of the respondent data is incomplete, as is the case for many of the single item ASCOF PIs that are not adjusted for case-mix. The analysis pointed to a number of groups of users that were under-represented in the current ASCS. These included people with mental health problems, working age groups (particularly 18 to 34 year olds) and nursing home residents. To a lesser extent people receiving low-level forms of service provision, people from black and minority ethnic groups and people from more deprived areas are also targets. There was also the suggestion that people with severe disabilities were more likely to be missing. All of these groups could be targeted with "tailored" strategies (Dillman et al. 2009) to improve their participation and therefore the representativeness of the ASCS data.

Although I was able to identify a number of groups of users that were under-represented in the ASCS, the analysis provided less evidence for strategies to increase response rates. Only one strategy that varied between CASSRs – chasing nonrespondents – was found to be significantly positively associated with CASSR response rates. The results of this study therefore suggest that the requirement to chase nonrespondents should be more stringently enforced. All other strategies that varied between CASSRs were not significant in the regression analysis. Interestingly, however, I found a fairly large CASSR effect on response propensity. One interpretation of this effect is that CASSRs play an important role in determining response rates due to the control they have over the survey design and management. Further studies are required to test this hypothesis, but if this is the case then it suggests that more could be done to improve response rates.

Limitations of the analysis and directions for future research

The arguments presented above around the requirements for valid inferences about performance all depend on the assumption that the adjusted PIs are more valid measures of performance than the crude PIs. If the adjusted PIs are biased such that 'true' performance lies closer to the crude PI than the adjusted PI, then there is no benefit from post-hoc adjustment. Since there is no benchmark against which to assess the validity of the unadjusted and adjusted indicators, arguments must appeal to theory to establish the validity of the statistical models that underpin the adjusted indicators and, where the research was designed to establish the effectiveness of interventions, the internal validity of the PIs (Mukamel 1997, Cartwright 2007, Li et al. 2009, Clarke et al. 2015). I was constrained in the statistical analysis by features of the design of the ASCS and the quality of some of the variables. Some of the decisions I took to address these limitations have

implications for the study conclusions and the generalisability of the findings, which I discuss here.

Limitations of the statistical modelling of ASC outcomes

The purpose of the risk-adjustment and production function modelling was to estimate the effectiveness of social care interventions provided by CASSRs in terms of service users' SCRQoL. As I discussed in Chapter 6, the internal validity of this aspect of the study was compromised to some extent by data limitations, which meant it was only possible to include four out of the five sets of risk factors for SCRQoL outcomes that were hypothesised as important. Aside from basic socio-demographic characteristics, there were no variables available for the group of influences that I labelled as personal characteristics and motivations, which included factors such as self-care ability. Data limitations also meant there were no indicators for factors such as health conditions, which have been shown to be important in previous studies (Glendinning et al. 2008a, Forder and Caiels 2011b, Forder et al. 2014, Forder et al. 2016).

These unobserved variables are likely to account for some of the unexplained variation in SCRQoL and therefore lead to under-adjustment for case-mix. The extent of the problem depends on the degree of correlation between the unobserved and observed variables. Where there is a high level of correlation between observed and unobserved variables, model estimates will be biased. The implication of the bias in model coefficients depends on the type of model. Where risk-adjustment models are used to estimate effectiveness, the interest is in the error term. In such circumstances, a high correlation between the observed and unobserved variables would suggest that the estimates of effectiveness are less biased. Where production function models are used, if the unobserved characteristics are correlated with the observed risk factors and the service quantity (as is likely to be the case where the observed characteristics imperfectly capture the selection mechanism through which users are allocated care packages) then estimates of effectiveness are likely to be biased. Forder and Caiels (2011a) argue that such 'endogeneity bias' is a problem for estimating effectiveness using production function models. In a series of papers, Forder et al. (2014, 2016) use an instrumental variables approach to address this problem. As discussed in Chapters 3 and 7 there were no suitable instruments in the ASCS dataset. This would, however, be one area for further research if a suitable dataset could be identified.

There was some evidence that a number of the risk factors were also capturing outcomes from care. I selected variables for inclusion in the model on the basis that they influenced social care outcomes, but were not directly (at least in the short-term) influenced by the CASSR. While some variables are clearly exogenous, for example age, most of the variables could be considered both risk factors and social care outcomes. For example, the person's home design makes delivery of care more difficult but could be a target of intervention, as improvements through adaptations to the home will help the person to be more independent. Similarly, functional disability affects QoL directly, but is the key outcome of rehabilitative programmes. Health too, including mental health, affects QoL directly but can be considered as an outcome since one aim of care provision is to prevent the negative health consequences of functional decline, e.g. bed sores and social isolation. There was some evidence that these variables, particularly home design and some of the health variables (anxiety/depression), were also capturing some of the impact of social care. This raises questions about whether they should be included as risk-adjustors, since to the extent that these variables are also targets of social care services, PIs based on these regression models will 'over-adjust' for the effect of these factors on performance (Berlowitz and Intrator 2013, Iezzoni 2013).

The models of SCRQoL were also compromised by the poor fit of the models to the data. Even the fractional response regression did not predict the negatively-skewed shape of the distribution well. Despite various attempts to improve model fit (discussed in Chapter 6), predictions were particularly poor for cases in the tail of the SCRQoL distribution. It is likely that the models provide an upwardly-biased prediction of expected SCRQoL for these cases. This would downwardly bias PIs, with the extent of the problem depending on the number of extreme cases. To guide policymakers, it is important to gain a better understanding of the impact of this modelling limitation on inferences regarding performance. Activities that would support such a goal include sensitivity analysis and following up 'high risk' cases to establish more information about their characteristics. Since 'high risk' users are often most instructive for quality improvement purposes, such activities are likely to support the use of the PIs for performance improvement (Shahian et al. 2001).

Additionally, it would be useful to explore whether alternative modelling approaches could provide a better fit to the data. One option that could be tried is a censored regression model (e.g. the Tobit model). I did not apply this model here as the censoring assumption of these models, means that the results are difficult to interpret for

multi-attribute utility measures where the maximum score is understood as perfect quality of life (Pullenayegum et al. 2010). Since the ASCOF focuses currently on the SCRQoL measure rather than ASCOT, its utility-weighted counterpart, this problem around interpretation is likely to be less limiting.

A second option that could be explored is to separately predict each of the attributes that comprise the SCRQoL measure and then combine the risk-adjusted predictions from each of these models to generate an expected outcome for each case. Ara et al. (2013, 2014) apply this method to the NHS GP patient survey data to predict each of the five questions that comprise the EQ-5D. They find that this method provides a better fit to the bimodal EQ-5D GP patient survey data than an OLS regression. A benefit of this approach is that it also makes it possible to explore variations in the attributes of SCRQoL. Given many of the SCRQoL domains are also ASCOF PIs, this approach would also expand the number of ASCOF PIs that are case-mix adjusted. A disadvantage of this approach is that the method is more complicated, which would affect the transparency of the adjusted PI.

The additional data requirements of the production function model led to further limitations in the modelling of SCRQoL-outcomes. The production function model requires a measure of service quantity, which was hard to find. I used the budget variable in the auxiliary data provided by CASSRs from their records, which captures the amount of money allocated by the CASSR to each individual for their care package (often referred to as the 'personal budget'). It was not a very accurate measure of service quantity, for the reasons discussed in Chapter 7. To minimise the effect of some of the problems of the budget variable on the results, I restricted the production function modelling to the 18 to 64 PSD sub-group. This limits the generalisability of the analysis to sub-groups other than the estimation sample. Despite limitations with the budget variable as an indicator of service quantity, the conclusions from this analysis appear sound. As I discussed in Chapters 7 and 8, the results compare favourably with those from the Identifying the Impact of Adult Social Care (IIASC) study[4] (Forder et al. 2016) that was closely related to the research in this thesis. Although the IIASC study also had some limitations (e.g. its focus on people in private households and the fairly small sample size for estimation), the similarity between the results from the IIASC study and this thesis suggests that although the adjusted PIs

[4] The IIASC study was designed to provide a PI for ASCOF that captured the effectiveness of social care. This study was designed in light of the preliminary modelling for this thesis (Malley and Fernandez 2014), to capture data that would make it possible to overcome some of the limitations of the analysis discussed here. Specifically, detailed data on resource inputs was collected and data was collected to allow for instrumental variables estimation within the production function models.

may be subject to some bias this is not significant enough to undermine the validity of the conclusions from this thesis.

Limitations of the method for case-mix adjustment

Indirect standardisation is a very flexible method that can be applied easily to adjust PIs for differences in the case-mix of CASSRs. A notable benefit of this approach is that the model for generating expected outcomes is estimated on a reference population and then applied to others, so there is no need to rerun the statistical analysis for each new wave of the survey. The applicability of the approach, however, depends on two factors. First, the statistical model estimated on the reference population must be generalisable to other samples. This was not something I explored in this analysis due to the small number of cases in two of the three sub-groups. Generalisability of the models could therefore be improved by validating the models on a 'holdout' sample. This strategy guards against over-fitting of the model (Ash et al. 2013), and would be important to carry out before using the models for adjustment of ASCOF PIs in practice.

Secondly, if indirect standardisation is to provide a basis for making fairer comparisons between CASSRs, the CASSRs must have a fairly similar case-mix profile (Glance et al. 2000, Julious et al. 2001, Rixom 2002, Julious and George 2007). As I discussed in Chapter 8, there are quite large differences between CASSRs in the proportions of the sample in the sub-groups used for estimating the models. This raises questions about whether the indirectly-standardised PIs for the PSD subsample are comparable and whether the conclusions are more valid than those drawn from the unadjusted PIs (Glance et al. 2000, Nicholl et al. 2013). This question is important because it is clear that the PIs cannot be reported at the sub-group level. As I have argued in Chapter 8, the sample size was too small for many of the sub-groups to have adequate power for outlier analysis. Indeed, this was a limitation of the outlier analysis for the care home and 18 to 64 sub-groups. To have adequate power for benchmarking analysis, a single adjusted PI needs to be reported for each CASSR.

An alternative to indirect standardisation is 'direct standardisation', in which the observed outcome is standardised to the case-mix of the reference population. This method is more appropriate where the case-mix of CASSRs is very different, but it is generally infeasible for performance assessment studies because the sample size of organisations is too small and continuous variables cannot be used to standardise. In recent years, alternative methods have been developed that combine direct and indirect

standardisation. As in indirect standardisation, a regression is used to estimate the observed outcome. Rather than being used to estimate the expected outcome, however, the predictions from this regression are used to estimate a risk profile for each organisation, which can be standardised to the reference population by using adjustment cells or risk categories (Canadian Institute for Health Information 2013, Nicholl et al. 2013). This method enables users of the data to make fairer comparisons of the performance of organisations. It should be possible to adapt this method so that it can be applied to the ASCOF PIs. It would be useful to investigate whether this approach to standardisation has any effect on inferences concerning performance.

Limitations of the nonresponse adjustment analysis

Theory suggests that where data are missing and are not MCAR, estimates produced under the MAR assumption will lead to more valid inferences than estimates produced under the MCAR assumption (Rubin 1987). The validity of the estimates, however, depends on the correct specification of the underlying response propensity and imputation models (McKee et al. 1999, Seaman and White 2013, Seaman and White 2014, Mukaka et al. 2016). As I discussed in Chapter 5, the response propensity and imputation models were subject to some data limitations that are likely to affect the validity of the adjusted PIs. I review these here and discuss how they may affect the conclusions from this thesis.

For both the response propensity and imputation model, it is possible that estimates are biased because missingness is NMAR (i.e. dependent on the value of unobserved factors). For the response propensity models, the lack of good indicators for key variables theorised to affect survey participation, such as the availability of informal care, health, disability type and severity, is an important limitation. Viewed in the context of the poor predictive validity of the model, it seems fairly plausible that missingness is NMAR, unless survey participation is a quasi-random process. From a theoretical perspective, the imputation model for the respondent population has good validity. The imputation model was congenial with the analysis where the crude PIs were used. There were some differences between the imputation and analysis models where there was adjustment for case-mix, which could have introduced some bias into the estimation. The plausibility of the MAR assumption was strengthened, however, by the large number of covariates in the imputation model. The robustness of the results to NMAR could be tested by conducting sensitivity analysis around the MI estimates (Carpenter et al. 2007, Kirkham 2008, Carpenter and Kenward 2013, Gomes et al. 2016). Given the fairly low levels of item

nonresponse for most of the analyses such efforts did not seem warranted in this case. It may, however, be important for future applications of the methods and for the analysis of the effects of nonresponse in the presence of adjustment for case-mix.

These limitations in the statistical modelling probably introduced a degree of bias into the adjusted PIs, but, as I argued in Chapter 5, a more important limitation was the method used to address missingness in the auxiliary data. The auxiliary data is used to estimate the response propensity models that underpin IPW. It is necessary for the auxiliary data to be fully observed for weights to be available for the whole dataset. I used MI to recover the missing data, but the strategy I used had some weaknesses, which mean that the estimates may be subject to some bias and that the IPW/MI method was probably more precision-reducing than I found. Given the small effects of IPW/MI for most of the ASCOF PIs, these problems raise questions about whether the adjusted PIs are more valid than the crude PIs, which it is not possible to address without further research. In a recent paper, Seaman and White (2014) suggest a way of addressing missingness in the auxiliary data in the context of IPW/MI that is not subject to the problems identified with the method I used. The method could be applied as an extension to this research to check the robustness of the results.

A further limitation of the nonresponse analysis was the limited way in which I was able to address nonresponse in the context of the case-mix analysis. I explored only the effect of adjusting for item nonresponse through MI, which implicitly assumes that unit nonresponse is MCAR. I did not apply the IPW/MI approach to the case-mix adjusted PIs because the strategy of conducting the analysis by sub-group is not compatible with the sample-level IPW analysis. Applying the weights to the sub-groups would distort adjusted PIs as the sum of the weights would depend on the ratio of the size of the sub-group to the total respondent sample, which would vary for each CASSR. The lack of weighting for unit nonresponse means it is a source of unobserved heterogeneity in the case-mix adjusted PIs for the sub-groups and PSD subsample. If IPW/MI adjustment has the same effects in this context as it does for the raw PIs, it is likely that nonresponse would have produced a greater degree of bias in the case-mix adjusted PIs although at the expense of less precise estimates.

Limitations of the analysis into factors influencing survey participation

As I have already mentioned, there were some limitations in the availability of variables to capture key characteristics of users theorised to affect survey participation. In addition to

the lack of variables available to reflect receipt of practical help, the dataset had poor indictors for investigating the effect of disability and health conditions on response propensity. The effect of proxy indicators for these variables, when examined alongside research from other studies (e.g. Kauppi et al. 2005, Hutchings et al. 2012, Peters et al. 2014b), did suggest reasons to be concerned about the representativeness of the ASCS for more disabled groups. However, further research is needed to fully understand the effect of health and disability on participation in the ASCS. Since performance surveys seem to be most frequently carried out in health and allied sectors, in populations with high prevalence rates of long-term conditions and disability, many surveys would benefit from research that seeks to better understand how disability and health affect response propensity and performance assessment.

A further weakness of this analysis was that very few of the survey attributes that varied between CASSRs had a significant relationship with response rates. As I discussed in Chapter 4, there is, however, a need to be cautious about interpreting the lack of significance of the effect of the various CASSR-level variables tested as evidence that they are not important in predicting response propensity. There were not many observations at the CASSR level and the binary variables that describe survey attributes have in most cases very skewed distributions, so the power to detect differences is limited. Indeed, the exploratory analysis found that the social environment variables and many of the survey attributes did have relationships with response propensity. Furthermore, many of the indicators of survey attributes were not particularly sensitive measures, as they did not distinguish the variety of ways in which each attribute could have been implemented. This will dampen the association between these variables and response propensity. It is possible that more detailed data and more CASSR data points would have improved the power of this analysis to detect effects where they are present. There may be the potential to improve the analysis presented here by including more waves of the ASCS in future studies.

The observational design of the study, however, is a factor that works against the potential for further analysis to identify effective strategies for increasing response rates. The internal validity of observational studies is weakened by the possibility that there are alternative explanations for the observed effect. The strong CASSR effect found in this study may be evidence that CASSRs have an important role to play in determining response rates. An alternative explanation is that response propensity is affected by unobserved area-level correlates. Equally, there may be alternative explanations for the

finding that chasing nonrespondents affected CASSR response rates. Considered alongside evidence from experimental research that has also found that chasing nonrespondents increases response rates (Edwards et al. 2002), the causal explanation for the relationship between chasing nonrespondents and response rates is strengthened. It is notable, however, that this strategy is not targeted at groups with poor response rates. It may therefore not do much to improve the representativeness of the respondent sample, as untargeted strategies may simply encourage participation from people 'like' those that are already participants (Groves and Cialdini 1991, Groves et al. 1992, Groves et al. 2000, Groves et al. 2004, Groves and Heeringa 2006, Dillman et al. 2009). Some studies, however, have suggested that those who respond to later mail-outs have different characteristics to those who respond to earlier mail-outs (Strayer et al. 1993, Sheldon et al. 2007). This suggests that chasing nonrespondents may be important for ensuring representativeness, although further studies would be needed to confirm this.

Conclusions and wider implications for adult social care policy and performance assessment research

This thesis makes original contributions in three areas. First, this is the first study to explore the need for case-mix adjustment of social care performance indicators and provides the only assessment of the value of case-mix adjustment for indicators of the care home population. As well as being the first application of risk-adjustment methods to social care, this thesis has also pioneered the use of a production function method to generate case-mix adjusted indicators. The methods developed here were extended by Forder et al. (2016) in a related study. Secondly, this thesis provides the first and only exploration of the impact of nonresponse on performance assessment using social care performance indicators. In this thesis I have pioneered the use of a combined inverse propensity weighting/multiple imputation method to explore the effect of nonresponse on indicators. Thirdly, this thesis has provided the first detailed analysis of the characteristics that are associated with participation in the ASCS. I have also explored the role CASSRs have in determining participation to the ASCS and have questioned whether such analysis has insights for improving response rates. The main recommendations from this thesis are shown in Box 3.

In this section, I conclude by revisiting the policy implications, and considering the contribution of this thesis to ASC policy and performance assessment research, more generally. Because of the nature of the empirical analysis, the results have the most

implications for ASC policy. As I noted in the introductory chapter, however, related areas of policy research, e.g. health care, face many of the problems for performance assessment that I have sought to overcome in this thesis. This analysis therefore holds lessons for health policy research around performance assessment, which I draw out here. I also discuss some of the questions that are raised by this research for ASC policy and practice around performance assessment and for future research around performance assessment more broadly. This discussion is organised around the themes of case-mix and nonresponse.

Conclusions concerning the importance of differences in case-mix between organisations for performance assessment and implications for policy and future research studies

This thesis and the related IIASC study (Forder et al. 2016), together, provide the first evidence about the effect of adjusting the ASCOF survey-based PIs for differences between CASSRs in the characteristics of service users. Even accounting for the limitations of the model, this thesis provides strong evidence for the need to adjust the SCRQoL PI for differences in the case-mix of CASSRs.

Since many of the ASCOF survey-based PIs are drawn from the items that comprise the multi-item SCRQoL indicator, it seems likely that adjustment will also be necessary for the other survey-based ASCOF indicators. Exactly how adjustment should be performed and its effect on inferences regarding the performance of CASSRs, however, is a question for future analysis. Although studies suggest that the types of variables available in the ASCS dataset will be important predictors of these PIs (Bauld et al. 2000, Chesterman et al. 2001, Jones et al. 2007, Malley et al. 2012), it is not clear that the same covariates will be relevant. Indeed where the PIs are the more delivery-oriented, PREMs, or satisfaction measures, the conceptual framework will need to be adapted to reflect the characteristics of these measures and their relationship to ASC outcomes.

In this thesis, I explored various approaches to case-mix adjustment. This analysis had an exploratory flavour and the limitations of the study are such that the methods require more investigation before they can be applied in practice. A key reason for this is that the analysis was restricted to the PSD subpopulation of the ASCS. Further research is therefore required to develop statistical models for the other sub-groups or to identify alternative methods for adjusting for differences between CASSRs in the case-mix of these sub-groups. Although the IIASC study (Forder et al. 2016) provides some answers for several sub-groups, this study is also limited in its coverage of the ASCS population.

Additionally, this research raised a number of deeper questions about the purpose of the statistical modelling and the method used for case-mix adjustment that need consideration.

In this respect an important finding was the suggestion that some of the variables included as risk-adjustors in the models also captured social care outcomes. A negative consequence of adjusting the SCRQoL PI for factors that are also outcomes from social care is that over time the adjusted PI would discount the effect of these variables. Achievements in these areas would therefore be understated (Kane 2001) and there is the possibility that such a regime would, perversely, discourage investment in these areas (Mukamel et al. 2008). In ASC, this could mean CASSRs avoiding investment in improvements to the housing stock, or in improvements to people's underlying health and level of functional ability. On the other hand, if policymakers exclude all factors that are affected by ASC, the list of risk-adjustors would probably be very short and explain very little of the variation in SCRQoL (Arling et al. 2005). Most importantly a model that excluded all the factors affected by social care would fail to adjust for the effect of needs-related characteristics on outcomes and would therefore have poor validity.

Mukamel et al. (2008) suggest a solution to this problem that could be applied to ASCOF and indeed to other indicator sets that also suffer from this problem. They suggest taking a liberal approach to the inclusion of variables as risk-adjustors in the model, but countering the negative consequences of over-adjustment, by stipulating that where risk-adjustors are subject to these problems they should be included as additional PIs. There are clear benefits to such an approach: accomplishments across multiple areas could be factored into performance assessment and it would improve the identification of poor care. It would also guard against perverse incentives and encourage investment in improving a broader set of outcomes. A first step, however, would be to re-estimate the case-mix adjustment models excluding the problem variables to understand the effect this has on inferences about performance. Should this significantly affect performance assessment, policymakers may wish to consider expanding the set of ASCOF PIs to ensure that the system incentivises the right practices.

This thesis, with the IIASC study (Forder et al. 2016), have also provided the first evidence concerning the value of using production function models to adjust for differences in the case-mix of organisations. Although the production function and risk-adjustment models produce very similar conclusions regarding the relative performance of organisations, the production function model provides a better estimate of effectiveness. Such a value-added indicator is likely to be more appropriate for internal research and

evaluation studies by CASSRs, but its benefits are not clear for other uses of the measure. More research is, however, needed to overcome some of the limitations of both this thesis and the IIASC study and explore further the potential applications of this method in the context of performance assessment.

The DH has suggested that the production function models estimated in the IIASC study will be used to form the basis of an adjusted SCRQoL indicator for future ASCOF publications (NHS Digital 2016a). As I have already noted, some thought needs to be given to the question of how to incorporate into the adjusted PIs the observed outcomes of individuals within the sub-groups not investigated in this thesis or the IIASC study. Further analysis is required to develop statistical models to underpin any case-mix adjustment, but this will be difficult for the smaller client groups (e.g. substance misuse clients) due to the number of cases. An alternative would be to exclude the smaller client groups from the adjusted PIs, but this option is unlikely to be viewed favourably given the original intention of the survey to capture all ASC users (Department of Health 2009).

Consideration also needs to be given to the integration of the care home sub-group within the adjusted PI. I did not estimate a production function model for this group, because the budget variable captures price paid by the CASSR for placements not intensity. Indeed it is difficult to capture differences in the intensity of care provision for care home residents, and studies that have examined cost and production relations in care homes have all measured resource inputs at the facility level (Knapp 1978a, Knapp 1978b, Knapp 1979, Darton and Knapp 1984, Degenholtz et al. 2006, Shippee et al. 2015). Such an approach is infeasible with the current design of the ASCS, since it is not possible to link the questionnaire data to data about the characteristics of care homes. Arguably, a risk-adjustment model is more appropriate for care homes as the service is more homogeneous.

In theory, it would be possible to extend the production function method of adjustment to the sub-groups of the ASCS that were not analysed in either this thesis or the IIASC study. The lack of routinely collected data that is suitable for this purpose, however, means this is not feasible in the short-term.[5] The lack of a suitable dataset also

[5] Although there have been attempts to improve the quality of the budget variable in successive waves of the ASCS, it is an optional auxiliary data item and CASSRs tend not to record it. Indeed the proportion of CASSRs opting to report it has declined each year (83 in 2010-11, compared to 51 in 2011-12, 47 in 2012-13 and 37 in 2013-14). In any case improving the reporting of the budget variable would not address the many others problems associated with using this variable as an indicator of resource inputs, such as the delay between the extraction of the auxiliary data and the survey and the effect of CASSR policies and local prices

means it will not be possible to replicate or improve upon the models, which hampers the progress that can be made understanding the effectiveness of ASC. It seems important, therefore, to invest in ways of improving the collection of information about service utilisation in ASC. Currently, data collections seem to be moving in reverse gear with respect to this goal, as traditional service packages are redefined as personal budgets and the data held by CASSRs becomes progressively less detailed. A potential way forward would be for the data collection regime to draw upon information held by providers about the detail of individual care packages. This would also support the estimation of production functions to compare care home facilities. As authorities seek to develop plans to manage their local markets for quality, as part of the requirements of the *Care Act 2014*, a case can be made for capturing these data at a local level to support commissioning and broader 'market-shaping' activities. Indeed it is hard to see how CASSRs could use the adjusted PIs to explore the effectiveness and cost-effectiveness of their services without more detailed information about individual care packages. Such data would also have clear value for the Care Quality Commission and the Competition and Markets Authority.

Production function models may be useful for case-mix adjustment in other settings. Such models are likely to be useful where PIs assess the performance of 'complex' interventions, are based on cross-sectional data and are drawn from surveys of diverse populations. In all of these situations, the characteristics of the service users are likely to both determine features of the intervention received and directly affect the success of the intervention. Examples from outside ASC are services for people with long-term conditions, which seem likely to be subject to this problem. The NHS GP practice survey may also benefit from this technique as it has a diverse population, which complicates the estimation of service effectiveness using PROMs. It is likely that the difficulties I faced here finding a good indicator of services for the production function models will limit the application of the method in other settings. Nevertheless, smaller scale studies may be possible and desirable to understand whether estimates of effectiveness are biased by heterogeneity in the provision of care services. This seems important given the interest the government has shown in using the GP practice survey as a source of data to demonstrate

---

on the measure. The IIASC study included a phase of primary data collection to provide a suitable dataset for estimating the production functions (Forder et al. 2016). The difficulties experienced by the researchers in obtaining the IIASC sample, however, suggest that large scale primary data collection is unlikely to be feasible for future studies unless significant resources are available to fund the fieldwork.

the effectiveness of health care services (Ara et al. 2013, Fernandez et al. 2013a) and in establishing a PROM for long-term conditions (Peters et al. 2014b, Hunter et al. 2015).

Conclusions about the effects of nonresponse on performance assessment and implications for policy and future research studies

This thesis is the first to examine whether nonresponse to the ASCS has an effect on performance assessment using the ASCOF PIs. It is also the first to examine in any detail the factors influencing participation in the ASCS. I found that nonresponse explains a small amount of heterogeneity in PIs. The importance of nonresponse seems likely to be greater (i) the higher the level of item nonresponse for the PI, and (ii) where there is adjustment for case-mix and the organisation has an unfortunate pattern of missingness within the set of variables used for adjustment. While adjustments for nonresponse are not clearly indicated, this study does suggest the need to investigate the effects of nonresponse on PIs thoroughly before publication of datasets to rule out the possibility that nonresponse is a cause of differences between CASSRs on PIs.

The findings from this study are somewhat different to the small number of studies that have looked at the effects of nonresponse on performance assessment (Elliott et al. 2005, Roland et al. 2009, Gomes et al. 2016). Elliott et al. (2005) employed IPW/IPW[6] and found that nonresponse is not important particularly where PIs are adjusted for case-mix. Gomes et al. (2016) used MI/MI[7] and found conversely that nonresponse is important even where PIs are adjusted for case-mix. I used an IPW/MI method and found that adjusting for nonresponse may be important for some indicators, but not all. Although the conclusions from these studies seem contradictory, one interpretation is that the more important MI is to the method for addressing nonresponse, the more necessary post-hoc adjustment is found to be. There could be other explanations for the different conclusions from these studies, but this interpretation is consistent with the comparison between CCA/MI, IPW/CCA and IPW/MI conducted in this study. This interpretation is also consistent with theoretical expectations based on the greater efficiency of MI compared to IPW, and Seaman et al.'s (2012) simulation study into the relative effects of IPW/MI, MI/MI and IPW/IPW on survey estimates. This implies that the method used to adjust for

[6] This is where IPW is used to adjust for both unit and item nonresponse.
[7] This is where MI is used to adjust for both unit and item nonresponse.

nonresponse and explore its effects matters for conclusions regarding the influence of nonresponse over survey estimates.

The effect that the choice of method for investigating nonresponse may have on conclusions about the effect of nonresponse on PIs and the need for post-hoc adjustment has important consequences for future studies. An important lesson of this analysis is that researchers should explore the impact of both unit and item nonresponse on each PI using appropriate methods. In the context of performance surveys, the IPW/MI approach is likely to have much broader applicability than the MI/MI approach used by Gomes et al. and it is also likely to be more efficient than an IPW/IPW approach (Seaman et al. 2012). MI has more modelling assumptions (Carpenter et al. 2006) and will be infeasible for most performance surveys due to high rates of unit nonresponse and limited auxiliary data. Indeed MI is generally not recommended for such patterns of missingness (McKee et al. 1999). The PROMs data analysed by Gomes et al. (2016) is highly unusual for performance surveys in that it contains a large amount of auxiliary data, and has measurements at two time points.

I argued that the complexity of the IPW/MI method means that it will probably be used only in exceptional circumstances to adjust for nonresponse to performance surveys. It would be useful for future research to explore the conditions under which adjustment using IPW/MI might be necessary. There are also situations in which IPW/MI, or IPW approaches more generally will not be appropriate. For example, if a form of direct standardisation were to be used for case-mix adjustment, IPW would be redundant. This is because the adjusted PIs would be standardised to a common average case-mix, so weighting would not adjust for nonresponse. Strategies to address nonresponse therefore need to be considered alongside those to address case-mix to ensure the optimal mix of methods for performance assessment.

I have argued that future research should explore strategies for increasing response rates to the ASCS. Although there was little evidence from this study that unit nonresponse had a large effect on performance assessment, increasing survey participation rates is important for reasons of face validity, precision and cost. I was able to offer very few recommendations for CASSRs about ways of increasing response rates from this analysis, aside from the suggestion to chase nonrespondents. Further research in this area would benefit from a randomised controlled trial design rather than the observational approach used here, since the randomisation between control and trial groups minimises the potential for alternative explanations for effects. This experimental design has been

used successfully to identify a number of strategies that are effective at increasing response rates (Edwards et al. 2002, Dillman et al. 2009). Such methods could be implemented within the ASCS to identify methods for improving response rates

Although very few of the survey attributes that varied between CASSRs were significant in the regression models, the exploratory analysis suggested some fruitful avenues for further research to explore strategies to improve responses rates. These included (i) the potential for chasing nonrespondents by email or face-to-face interview to improve response rates, (ii) the role of incentives, and (iii) the potential for targeting specific groups (e.g. people with sensory impairments, cognitive impairments and low English literacy) with appropriately modified versions of the questionnaire. Better data about client characteristics would be required to target accessible or translated versions of the surveys at the right people, but such an investment would have many other benefits for understanding nonresponse. Given the persistent declines in response rates for the ASCS, such research and associated developments to data systems would be a worthwhile investment.

It is important to recognise that for some sub-groups of the ASCS population, it may be difficult to increase response rates. For example, high rates of dementia in care homes (Darton et al. 2006, Matthews et al. 2016) mean it is likely that residents do not respond because they lack the mental capacity to do so. Indeed, in the ASCS, CASSRs were instructed to remove people from the sampling frame who are unable to respond due to a lack of mental capacity. Since these people are far more likely to be in care homes, this means that the sample is unrepresentative of the care home population even before nonresponse is taken into account, as people with dementia are systematically excluded. Considered alongside the need for different case-mix adjustment methods and the very different characteristics of the care home population compared to the private household population, a fairly strong argument can be made for focusing the ASCS on the private household population.

For the care home population alternative methods for monitoring quality in care homes are likely to provide a more representative picture (Qureshi and Rowlands 2004). One such technique is the ASCOT care homes instrument, which uses observational techniques alongside interviews to score resident outcomes (Netten et al. 2012c, Towers et al. 2015). The cost of using observational methods for routine performance assessment is likely to be prohibitive, but such methods could be used as an adjunct to reviews and to inspections to gather data on the QoL and experiences of residents. Indeed a number of

CASSRs are beginning to use ASCOT in this way and there is also interest from care home providers (Johnstone and Page 2014, Towers et al. 2015, Towers et al. 2016).

Another alternative would be to take a more clinically- and facility-oriented approach to quality measurement, with similarity to the approaches used in the USA (Mor 2005) and Germany (Schulz 2012, Garms-Homolova and Busse 2014). In these countries the staff within the residential facilities collect data concerning clinical outcomes, such as pressure sores and weights loss, for each resident. In the USA, quality indicators are collected on a quarterly basis, providing a longitudinal dataset with which to track the trajectory of a person's health over the course of the intervention and beyond. This can help to resolve some estimation problems for quality of life risk adjustment as discussed in Chapter 3, since data on needs-related characteristics from earlier time points can be used to adjust for the effect of these characteristics on change in quality of life outcomes (Black and Jenkinson 2009, Smith and Street 2013). However, in practice stability in people's conditions and therefore needs-related characteristics may mean that this design is not as successful for explaining change in quality of life outcomes. Research would need to establish the appropriate time period between data collection points and develop an appropriate risk-adjustment model.

The data used for measuring the quality and outcomes of nursing home residents in the USA suffer from other kinds of reporting problems with the data, and also require risk-adjustment (Mor 2005, Stevenson and Bramson 2014). Indeed there is a large literature around risk-adjustment of these indicators for the US government-sponsored consumer comparison website 'Nursing Home Compare' (Zimmerman et al. 1995, Mukamel 1997, Mor et al. 2003, Arling et al. 2007, Mukamel et al. 2008, Li et al. 2009, Li et al. 2010). The US quality indicators have also been criticised for a lack of focus on QoL outcomes and steps have been taken to address this imbalance (Kane et al. 2004, Arling et al. 2005). Given these criticisms of the US system, it would be of benefit to policymakers in many different countries for research to systematically explore the advantages and disadvantages of the various approaches through careful comparative analysis.

This study has also raised important questions about the role of CASSRs in determining response rates. Much of the between-CASSR variation in response rates was unexplained by the analysis, even after controlling for differences in individual characteristics. It will be important for the ASCS to understand what is driving differences between CASSRs in response rates. In particular, research should try to unpick whether differences are due to variations in the management of the survey or to area-level

correlates. Area-level correlates have been identified as important in previous studies (Goyder et al. 1992, Johnson et al. 2006), and this study identified both types of factors as important. The large CASSR effect suggests that, as a precautionary measure, it would be sensible to work harder to ensure consistency in processes for managing and running the ASCS across CASSRs. These findings also have implications for other surveys with similar designs, such as the US CAHPS surveys and some of the NHS surveys. This thesis suggests that more attention should be paid to the effect of organisations on response rates in performance surveys. The role of the survey guidance is clearly important for maintaining consistency and research may have a role to play in determining the best design for such guidance and the best ways of ensuring organisations comply with it.

Box 3: List of recommendations

For analysis for policymakers

- The SCRQoL PI should be adjusted for case-mix. Research should also investigate whether other ASCOF indicators require case-mix adjustment.
- Standard risk-adjustment approaches provide an adequate approach to case-mix adjust indicators where the aim is to compare the relative performance of CASSRs. The production function approach, particularly where instrumental variables estimation is feasible, is a better method for case-mix adjustment where the aim is to generate a value-added indicator.
- Where risk-adjustment is used OLS models provide appropriate inferences regarding performance and is easy to understand.
- Risk adjustors should be selected for inclusion on the basis of theoretical plausibility and statistical significance. The aim should be to maximise the amount of variation explained.
- Indicators based on the differences between observed and expected outcomes are preferable to indicators based on the ratio of observed to expected outcomes.
- The effect of nonresponse on the rank position and outlier status of CASSRs should be explored in detail for each survey using methods that adjust for both item and unit nonresponse (i.e. IPW/MI methods). The importance or lack thereof of nonresponse for performance assessment cannot be assumed on the basis of the analysis of previous years' data.

- NHS Digital should enforce CASSRs to chase nonrespondents and send out at least one reminder. They may also want to consider randomised control trials to explore strategies to improve participation.
- Alternative methods for gathering the quality of care for residents of care homes and, in particular, nursing homes should be considered given the exclusion of these groups from the sample, the poorer response rates for nursing home residents.

For CASSRs to improve participation

- CASSRs should tailor their strategies for improving response rates to improve participation among people with mental health problems, working age groups (particularly 18 to 34 year olds) and nursing home residents. To a lesser extent CASSRs could also target improving participation from people from black and minority ethnic groups and people from more deprived areas. Such approaches would improve the representativeness of the ASCS.

Bibliography

Aitkin, M. and Longford, N. (1986). "Statistical modelling issues in school effectiveness studies." Journal of the Royal Statistical Society. Series A (General) **149**(1): 1-43.

Akaike, H. (1974). "A new look at the statistical model identification." IEEE Transactions on Automatic Control **19**(6): 716-723.

Andresen, E. M., Vahle, V. J. and Lollar, D. (2001). "Proxy reliability: Health-related quality of life (HRQoL) measures for people with disability." Quality of Life Research **10**(7): 609-619.

Andrews, R. (2004). "Analysing deprivation and local authority performance: The implications for CPA." Public Money and Management **24**(1): 19-26.

Andrews, R., Boyne, G., Law, J. and Walker, R. M. (2005). "External constraints on local service standards: The case of comprehensive performance assessment in English local government." Public Administration **83**(3): 639-656.

Ara, R., Kearns, B., Vanhout, B. A. and Brazier, J. E. (2014). "Predicting preference-based utility values using partial proportional odds models." BMC Research Notes **7**(1): 1-9.

Ara, R., Van Hout, B., B, K. and Brazier, J. (2013). Case-mix methodology for the NHS outcomes framework GP patient survey questionnaire data. Sheffield, Policy Research Unit in Economic Evaluation in Health and Care Interventions,.

Arling, G., Kane, R. L., Lewis, T. and Mueller, C. (2005). "Future development of nursing home quality indicators." The Gerontologist **45**(2): 147-156.

Arling, G., Lewis, T., Kane, R. L., Mueller, C. and Flood, S. (2007). "Improving quality assessment through multilevel modeling: The case of Nursing Home Compare." Health Services Research **42**(3 Pt 1): 1177-1199.

Ash, A. S., Schwartz, M., Peköz, E. A. and Hanchate, A. D. (2013). Comparing outcomes across providers. Risk adjustment for measuring health care outcomes. L. I. Iezzoni. Chicago, Illinois, Health Administration Press**:** 335-378.

Austin, P. C. (2002a). "Bayesian extensions of the Tobit model for analyzing measures of health status." Medical Decision Making **22**(2): 152-162.

Austin, P. C. (2002b). "A comparison of methods for analyzing health-related quality-of-life measures." Value in Health **5**(4): 329-337.

Austin, P. C., Alter, D. A. and Tu, J. V. (2003). "The use of fixed-and random-effects models for classifying hospitals as mortality outliers: A Monte Carlo assessment." Medical Decision Making **23**(6): 526-539.

Austin, P. C., Escobar, M. and Kopec, J. A. (2000). "The use of the Tobit model for analyzing measures of health status." Quality of Life Research **9**(8): 901-910.

Axinn, W. G., Link, C. F. and Groves, R. M. (2011). "Responsive survey design, demographic data collection, and models of demographic behavior." Demography **48**(3): 1127-1149.

Baldock, J. (1997). "Social care in old age: More than a funding problem." Social Policy & Administration **31**(1): 73-89.

Bamford, C. H., Qureshi, H., Nicholas, E. and Vernon, A. (1999). Outcomes of social care for disabled people and carers. Outcomes in Community Care Practice. Social Policy Research Unit. York, Social Policy Research Unit.

Barberger-Gateau, P., Dartigues, J.-F. and Letenneur, L. (1993). "Four instrumental activities of daily living score as a predictor of one-year incident dementia." Age and ageing **22**(6): 457-463.

Basser, M. (2015). Benefits case study: 'Patient reported outcome measures (PROMs)' outputs. Improving health outcomes for patients undergoing knee replacement, hip replacement, varicose vein and groin hernia treatments. Leeds, Health and Social Care Information Centre,.

Basu, A. and Manca, A. (2012). "Regression estimators for generic health-related quality of life and quality-adjusted life years." Medical Decision Making **32**(1): 56-69.

Bauld, L., Chesterman, J. and Judge, K. (2000). "Measuring satisfaction with social care amongst older service users: Issues from the literature." Health and Social Care in the Community **8**(5): 316-324.

Baum, C. (2000). Xttest3: Stata module to compute modified Wald statistic for groupwise heteroskedasticity, Statistical Software Components, Boston College Department of Economics.

Baum, C. F. (2008). "Stata tip 63: Modeling proportions." Stata Journal **8**(2): 299-303.

Beadle-Brown, J., Ryan, S., Windle, K., Holder, J., Turnpenny, A., Smith, N., Richardson, L. and Whelton, B. (2012). Engagement of people with long-term conditions in health and social care research: Barriers and facilitators to capturing the views of seldom-heard populations. Discussion paper 2850. Canterbury, Kent, QORU.

Berlowitz, D. and Intrator, O. (2013). Risk adjustment for long-term care Risk adjustment for measuring health care outcomes. L. I. Iezzoni. Chicago, Illinois, Health Administration Press**:** 423-442.

Berthoud, R. (2006). The employment rates of disabled people. Department for Work and Pensions. Research report no 298. London, Department for Work and Pensions.

Bevan, G. and Hood, C. (2006a). "Have targets improved performance in the English NHS?" BMJ **332**(7538): 419.

Bevan, G. and Hood, C. (2006b). "What's measured is what matters: Targets and gaming in the english public health care system." Public Administration **84**(3): 517-538.

Bird, S. M., Sir David, C., Farewell, V. T., Harvey, G., Tim, H. and Peter C, S. (2005). "Performance indicators: Good, bad, and ugly." Journal of the Royal Statistical Society: Series A (Statistics in Society) **168**(1): 1-27.

Black, N. (2013). "Patient reported outcome measures could help transform healthcare." BMJ : British Medical Journal **346**.

Black, N. and Jenkinson, C. (2009). "Measuring patients' experiences and outcomes." BMJ **339**.

Blake, M., Bowes, A., Gill, V., Husain, F. and Mir, G. (2016). "A collaborative exploration of the reasons for lower satisfaction with services among Bangladeshi and Pakistani social care users." Health & Social Care in the Community: n/a-n/a.

Blanchflower, D. G. and Oswald, A. J. (2005). "Regional wages and the need for a better area cost adjustment." Public Money & Management **25**(2): 86-88.

Blanchflower, D. G. and Oswald, A. J. (2008). "Is well-being U-shaped over the life cycle?" Social Science & Medicine **66**(8): 1733-1749.

Bonsang, E. (2009). "Does informal care from children to their elderly parents substitute for formal care in europe?" Journal of Health Economics **28**(1): 143-154.

Bourdieu, P. (1986). The forms of capital. Handbook of theory and research for the sociology fo education. J. Richardson. New York, Greenwood**:** 46-58.

Bovaird, T. and Löffler, E. (2003). "Evaluating the quality of public governance: Indicators, models and methodologies." International Review of Administrative Sciences **69**(3): 313-328.

Bowling, A. (1995). "What things are important in people's lives? A survey of the public's judgements to inform scales of health related quality of life." Social Science & Medicine **41**(10): 1447-1462.

Bowling, A., Banister, D., Sutton, S., Evans, O. and Windsor, J. (2002). "A multidimensional model of the quality of life in older age." Aging & Mental Health **6**(4): 355-371.

Bowling, A. and Windsor, J. (2001). "Towards the good life: A population survey of dimensions of quality of life." Journal of Happiness Studies **2**(1): 55-82.

Bowling, A. N. N. and Gabriel, Z. (2007). "Lay theories of quality of life in older age." Ageing and Society **27**(6): 827-848.

Boyne, G. A. (2002). "Concepts and indicators of local authority performance: An evaluation of the statutory frameworks in England and Wales." Public Money and Management **22**(2): 17-24.

Brown, K. and Coulter, P. B. (1983). "Subjective and objective measures of police service delivery." Public Administration Review **43**(1): 50-58.

Browne, W. J. (2015). MCMC estimation in mlwin, version 2.32. Bristol, Centre for Multilevel Modelling, University of Bristol.

Browne, W. J., Subramanian, S. V., Jones, K. and Goldstein, H. (2005). "Variance partitioning in multilevel logistic models that exhibit overdispersion." Journal of the Royal Statistical Society: Series A (Statistics in Society) **168**(3): 599-613.

Burchardt, T., Obolenskaya, P. and Vizard, P. (2015). The coalition's record on adult social care: Policy, spending and outcomes 2010-2015. Working paper 17. London, STICERD, London School of Economics.

Burnham, K. P. and Anderson, D. R. (2004). "Multimodel inference: Understanding AIC and BIC in model selection." Sociological Methods & Research **33**(2): 261-304.

Buttle, F. (1996). "SERVQUAL: Review, critique, research agenda." European Journal of Marketing **30**(1): 8-32.

Byford, S. and Sefton, T. (2003). "Economic evaluation of complex health and social care interventions." National Institute Economic Review **186**(1): 98-108.

Cameron, A. C. and Trivedi, P. K. (2005). Microeconometrics: Methods and applications. Cambridge, Cambridge University Press.

Campbell, S. M., Braspenning, J., Hutchinson, A. and Marshall, M. (2002). "Research methods used in developing and applying quality indicators in primary care." Quality and Safety in Health Care **11**(4): 358-364.

Canadian Institute for Health Information (2013). CCRS quality indicators risk adjustment methodology. Ottawa, Ontario, Canadian Institute for Health Information,.

Cañette, I. and Marchenko, Y. (2013, June 2013). "How can I combine results other than coefficients in e(b) with multiply imputed data?" Retrieved 10/03/2016, 2016, from http://www.stata.com/support/faqs/statistics/combine-results-with-multiply-imputed-data/.

Carpenter, J. and Kenward, M. (2013). Multiple imputation and its application. Chichester, John Wiley & Sons.

Carpenter, J. R., Kenward, M. G. and Vansteelandt, S. (2006). "A comparison of multiple imputation and doubly robust estimation for analyses with missing data." Journal of the Royal Statistical Society: Series A (Statistics in Society) **169**(3): 571-584.

Carpenter, J. R., Kenward, M. G. and White, I. R. (2007). "Sensitivity analysis after multiple imputation under missing at random: A weighting approach." Stat Methods Med Res **16**(3): 259-275.

Carr-Hill, R. A. (1992). "The measurement of patient satisfaction." J Public Health **14**(3): 236-249.

Cartwright, N. (2007). "Are RCTs the gold standard?" BioSocieties **2**(1): 11-20.

Challis, D., Clarkson, P. and Warburton, R. (2006). Performance indicators in social care for older people. Aldershot, Ashgate.

Chang, L. C., Lin, S. W. and Northcott, D. N. (2002). "The NHS performance assessment framework: A "balanced scorecard" approach?" Journal of Management in Medicine **16**(5): 345-358.

Chesterman, J., Bauld, L. and Judge, K. (2001). "Satisfaction with the care-managed support of older people: An empirical analysis." Health & Social Care in the Community **9**(1): 31-42.

Clark, D. E., Hannan, E. L. and Raudenbush, S. W. (2010). "Using a hierarchical model to estimate risk-adjusted mortality for hospitals not included in the reference sample." Health Services Research **45**(2): 577-587.

Clarke, P., Crawford, C., Steele, F. and Vignoles, A. (2015). "Revisiting fixed- and random-effects models: Some considerations for policy-relevant education research." Education Economics **23**(3): 259-277.

Clarkson, P. (2010). "Performance measurement in adult social care: Looking backwards and forwards." British Journal of Social Work **40**(1): 170-187.

Cleary, P. (1998). "Satisfaction may not suffice! A commentary on "a patient's perspective"." International Journal of Technology Assessment in Health Care **14**(1): 35-37.

Cleary, P. D. (1999). "The increasing importance of patient surveys: Now that sound methods exist, patient surveys can facilitate improvement." BMJ: British Medical Journal **319**(7212): 720.

Clemes, M. D., Gan, C. E. C. and Kao, T.-H. (2008). "University student satisfaction: An empirical analysis." Journal of Marketing for Higher Education **17**(2): 292-325.

Cohen, G. and Duffy, J. C. (2002). "Are nonrespondents to health surveys less healthy than respondents?" Journal of Official Statistics **18**(1): 13.

Comas-Herrara, A., Pickard, L., Wittenberg, R., Malley, J. and King, D. (2010). The long-term care system for the elderly in England. ENEPRI research report no. 74, CEPS.

Commission for Social Care Inspection (2004). Social services performance assessment framework indicators, 2003-2004. London, Commission for Social Care Inspection.

Connell, B. R., Sanford, J. A., Long, R. G., Archea, C. K. and Turner, C. S. (1993). "Home modifications and performance of routine household activities by individuals with varying levels of mobility impairments." Technology and Disability **2**(4): 9-18.

Converse, P. (1970). Attitudes and non-attitudes: Continuation of a dialogue. The quantitative analysis of social problems. E. Tuffe. Reading, MA, Addison-Wesley**:** 168-189.

Coulter, A. (2006). "Can patients assess the quality of health care?" BMJ **333**(7557): 1-2.

Coulter, A., Locock, L., Ziebland, S. and Calabrese, J. (2014). "Collecting data on patient experience is not enough: They must be used to improve care." British Medical Journal **348**.

Couper, M. (1998). Measuring survey quality in a CASIC environment. Invited paper presented at the Joint Statistical Meetings of the American Statistical Association. Dallas.

Couper, M. and Groves, R. (1996). "Social environmental impacts on survey cooperation." Quality and Quantity **30**(2): 173-188.

Cronin Jr, J. J. and Taylor, S. A. (1992). "Measuring service quality: A reexamination and extension." Journal of Marketing **56**(3): 55-68.

D'Agostino, R. B., Belanger, A. and D'Agostino, R. B. (1990). "A suggestion for using powerful and informative tests of normality." The American Statistician **44**(4): 316-321.

Darby, C., Hays, R. D. and Kletke, P. (2005). "Development and evaluation of the CAHPS® hospital survey." Health Services Research **40**(6p2): 1973-1976.

Darton, R., Forder, J., Bebbington, A., Netten, A., Towers, A.-M. and Williams, J. (2006). Analysis to support the development of the relative needs formula for older people: Final report. PSSRU discussion paper no. 2265/3. Canterbury, Kent, PSSRU.

Darton, R. and Knapp, M. (1984). "The cost of residential care for the elderly: The effects of dependency, design and social environment." Ageing and Society **4**(2): 157-183.

Davies, B. (1985). Production of welfare approach. Discussion paper 400. Canterbury, Personal Social Services Research Unit.

Davies, B., Fernandez, J.-L. and Saunders, R. (1998). Community care in England and france: Reforms and the improvement of equity and efficiency. Aldershot, UK, Ashgate.

Davies, B., Fernandez, J. and Nomer, B. (2000a). Equity and efficiency policy in community care: Needs, service productivities, efficiencies, and their implications. Canterbury, Kent, Ashgate.

Davies, B., Fernández, J. L. and Nomer, B. (2000b). Equity and efficiency policy in community care. Aldershot, Ashgate.

Davies, B. and Knapp, M. (1981). Old people's homes and the production of welfare. London, Routledge and Keegan Paul.

Davies, B. P. and Challis, D. J. (1986). Matching resources to needs in community care. Aldershot, Ashgate.

Dawson, J., Doll, H., Fitzpatrick, R., Jenkinson, C. and Carr, A. J. (2010). "The routine use of patient reported outcome measures in healthcare settings." BMJ **340**.

De Leeuw, E. and De Heer, W. (2002). Trends in household survey nonresponse: A longitudinal and international comparison. Survey nonresponse. R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little. New York, Wiley**:** 41-54.

De Lepeleire, J., Aertgeerts, B., Umbach, I., Pattyn, P., Tamsin, F., Nestor, L. and Krekelbergh, F. (2004). "The diagnostic value of IADL evaluation in the detection of dementia in general practice." Aging & Mental Health **8**(1): 52-57.

De Vellis, R. (2003). Scale development : Theory and applications. London, Sage.

Deeks, J., Dinnes, J., D'amico, R., Sowden, A., Sakarovitch, C., Song, F., Petticrew, M. and Altman, D. G. (2003). "Evaluating non-randomised intervention studies." Health Technology Assessment **7**(27): 186.

Deeks, J. J., Macaskill, P. and Irwig, L. (2005). "The performance of tests of publication bias and other sample size effects in systematic reviews of diagnostic test accuracy was assessed." Journal of Clinical Epidemiology **58**(9): 882-893.

Degenholtz, H. B., Kane, R. A., Kane, R. L., Bershadsky, B. and Kling, K. C. (2006). "Predicting nursing facility residents' quality of life using external indicators." Health Services Research **41**(2): 335-356.

Delong, E. R., Peterson, E. D., Delong, D. M., Muhlbaier, L. H., Hackett, S. and Mark, D. B. (1997). "Comparing risk-adjustment methods for provider profiling." Statistics in Medicine **16**(23): 2645-2664.

Deming, W. (1994). The new economics for industry, government, education. Cambridge, MA, MIT Press.

Deming, W. E. (1986). Out of the crisis. Cambridge, MA, Massachusetts Institute of Technology Press.

Department for Communities and Local Government (2010). Annex h: Area cost adjustment factors. Local government finance (England). The local government finance report (England) 2010/2011. Hc280 London, TSO.

Department of Health (1998). Modernising social services: Promoting independence, improving protection, raising standards. Cmnd 4169. London, The Stationery Office.

Department of Health (2003). Social services performance assessment framework indicators 2002–2003. London, Department of Health.

Department of Health (2009). Making policy count: Developing performance indicators for health and social care partnerships. Position paper social care. Department for Health. London.

Department of Health (2010a). Prioritising need in the context of 'putting people first': A whole system approach to eligibility for social care – guidance on eligibility criteria for adult social care, England 2010. London, Department of Health.

Department of Health (2010b). Transparency in outcomes: A framework for adult social care. A consultation on proposals. London, Department of Health,.

Department of Health (2010c). Transparency in outcomes: A framework for quality in adult social care. The 2012/13 adult social care outcomes framework, Department of Health.

Department of Health (2010d). A vision for adult social care: Capable communities and active citizens. London, Department of Health.

Department of Health (2011a). The adult social care outcomes framework. Handbook of definitions. Version 2 (November 2011). Department of Health. London.

Department of Health (2011b). Patient reported outcome measures (PROMs) in England. A methodology for identifying potential outliers Department of Health. London.

Department of Health (2012). The adult social care outcomes framework 2013/14. Department of Health. London.

Department of Health (2014). The adult social care outcomes framework 2015/16. London, Department of Health.

Devlin, N. and Appleby, J. (2010). Getting the most out of proms: Putting health outcomes at the heart of NHS decision-making. London, King's Fund.

Diewert, W. E. (1971). "An application of the shephard duality theorem: A generalized leontief production function." Journal of Political Economy **79**(3): 481-507.

Dillman, D. A., Smyth, J. D. and Christian, L. M. (2009). Internet, mail and mixed-mode surveys: The tailored design method. New York, John Wiley and sons.

Dixit, A. (2002). "Incentives and organizations in the public sector: An interpretative review." The Journal of Human Resources **37**(4): 696-727.

Donnelly, M., Wisniewski, M., Dalrymple, J. F. and Curry, A. C. (1995). "Measuring service quality in local government: The SERVQUAL approach." International Journal of Public Sector Management **8**(7): 15-20.

Durrant, G. B. and Steele, F. (2009). "Multilevel modelling of refusal and non-contact in household surveys: Evidence from six UK government surveys." Journal of the Royal Statistical Society: Series A (Statistics in Society) **172**(2): 361-381.

Easterlin, R. A. (2006). "Life cycle happiness and its sources: Intersections of psychology, economics, and demography." Journal of Economic Psychology **27**(4): 463-482.

Eddings, W. and Marchenko, Y. (2006). "Diagnostics for multiple imputation in Stata." Stata Journal **12**(3): 353-367.

Edwards, P., Roberts, I., Clarke, M., Diguiseppi, C., Pratap, S., Wentz, R. and Kwan, I. (2002). "Increasing response rates to postal questionnaires: Systematic review." BMJ **324**.

Eijkenaar, F. and Van Vliet, R. C. J. A. (2014). "Performance profiling in primary care: Does the choice of statistical model matter?" Medical Decision Making **34**(2): 192-205.

Elliott, M. N., Beckett, M. K., Chong, K., Hambarsoomians, K. and Hays, R. D. (2008). "How do proxy responses and proxy-assisted responses differ from what Medicare beneficiaries might have reported about their health care?" Health Services Research **43**(3): 833-848.

Elliott, M. N., Edwards, C., Angeles, J., Hambarsoomians, K. and Hays, R. D. (2005). "Patterns of unit and item nonresponse in the CAHPS® hospital survey." Health Services Research **40**(6p2): 2096-2119.

Elliott, M. N., Swartz, R., Adams, J., Spritzer, K. L. and Hays, R. D. (2001). "Case-mix adjustment of the national CAHPS benchmarking data 1.0: A violation of model assumptions?" Health Services Research **36**(3): 555-573.

Elliott, M. N., Zaslavsky, A. M., Goldstein, E., Lehrman, W., Hambarsoomians, K., Beckett, M. K. and Giordano, L. (2009). "Effects of survey mode, patient mix, and nonresponse on CAHPS® hospital survey scores." Health Services Research **44**(2p1): 501-518.

Elwyn, G., Buetow, S., Hibbard, J. and Wensing, M. (2007). "Respecting the subjective: Quality measurement from the patient's perspective." BMJ **335**(7628): 1021-1022.

Epstein, A. M., Hall, J. A., Tognetti, J., Son, L. H. and Conant, L., Jr. (1989). "Using proxies to evaluate quality of life: Can they provide valid information about patients' health status and satisfaction with medical care?" Medical Care **27**(3): S91-S98.

Erevelles, S. and Leavitt, C. (1992). "A comparison of curent models of consumer satisfaction/dissatisfaction." Journal of Consumer Satisfaction, dissatisfaction and complaining behaviour **5**: 104-114.

Eselius, L. L., Cleary, P. D., Zaslavsky, A. M., Huskamp, H. A. and Busch, S. H. (2008). "Case□mix adjustment of consumer reports about managed behavioral health care and health plans." Health Services Research **43**(6): 2014-2032.

Ferguson, I. (2012). "Personalisation, social justice and social work: A reply to simon duffy." Journal of Social Work Practice **26**(1): 55-73.

Fernández, J.-L. (2005). Utilisation and service productivities in community social care for older people: Patterns and policy implications. PhD, London School of Economics and Political Science.

Fernandez, J.-L. and Forder, J. (2015). "Local variability in long-term care services: Local autonomy, exogenous influences and policy spillovers." Health Economics **24**: 146-157.

Fernandez, J.-L., Forder, J., Trukeschitz, B., Rokosova, M. and Mcdaid, D. (2009). How can european states design efficient, equitable and sustainable funding systems for long-term care for older people? Policy brief no. 11. Copenhagen, World Health Organisation Europe.

Fernandez, J.-L., Hughes, A., Watson, K., Forder, J. and Fitzpatrick, R. (2013a). Report on using the GPPS to assess trends in EQ-5D scores for people with long-term conditions. Canterbury, Kent, Quality and Outcomes Research Unit and South East Public Health Observatory.

Fernández, J.-L. and Knapp, M. (2004). Production relations in social care. Long-term care: Matching resources and needs. M. Knapp, D. Challis, J.-L. Fernandez and A. Netten. Aldershot, Ashgate**:** 171-182.

Fernandez, J.-L., Snell, T. and Wistow, G. (2013b). Changes in the patterns of social care provision in England: 2005/6 to 2012/13. PSSRU discussion paper 2867. London, PSSRU, London School of Economics.

Fitzpatrick, R. (1991). "Surveys of patients satisfaction: I--important general considerations." BMJ : British Medical Journal **302**(6781): 887-889.

Forder, J. and Caiels, J. (2011a). "Measuring the outcomes of long-term care." Soc. Sci. Med. **73**.

Forder, J., Malley, J., Rand, S., Vadean, F., Jones, K. and Netten, A., Eds. (2016). IIASC report: Interpreting outcomes data for use in the adult social care outcomes framework (ASCOF). Discussion paper 2892. University of Kent, Canterbury, PSSRU.

Forder, J., Malley, J., Towers, A.-M. and Netten, A. (2014). "Using cost-effectiveness estimates from survey data to guide commissioning: An application to home care." Health Economics **23**(8): 979-992.

Forder, J. E. and Caiels, J. (2011b). "Measuring the outcomes of long-term care." Social Science & Medicine **73**(12): 1766-1774.

Francis, J. and Netten, A. (2004). "Raising the quality of home care: A study of service users' views." Social policy and administration **38**(3 June): 290-305.

Frey, B. S., Benz, M. and Stutzer, A. (2004). "Introducing procedural utility: Not only what, but also how matters." Journal of Institutional and Theoretical Economics (JITE) **160**(3): 377-401.

Frey, B. S. and Stutzer, A. (2005). "Beyond outcomes: Measuring procedural utility." Oxford Economic Papers **57**(1): 90-111.

Gale, C. P., Cattle, B. A., Moore, J., Dawe, H., Greenwood, D. C. and West, R. M. (2011). "Impact of missing data on standardised mortality ratios for acute myocardial infarction: Evidence from the myocardial ischaemia national audit project (MINAP) 2004–7." Heart **97**(23): 1926-1931.

Garms-Homolova, V. and Busse, R. (2014). Monitoring the quality of long-term care in germany. Regulating long-term care quality. An international comparison. V. Mor, T. Leone and A. Maresso. Cambridge, Cambridge University Press**:** 67-101.

Gelman, A. and Pardoe, I. (2006). "Bayesian measures of explained variance and pooling in multilevel (hierarchical) models." Technometrics **48**(2): 241-251.

Gill, V., Husain, F., Vowden, K., Aznar, C. and Blake, M. (2014). Satisfaction with social care services among black and minority ethnic populations. Exploring satisfaction with adult social care services amongst Pakistani, Bangladeshi and white British people. London, Natcen Social Research.

Gitlin, L. N., Corcoran, M., Winter, L., Boyce, A. and Hauck, W. W. (2001). "A randomized, controlled trial of a home environmental intervention: Effect on efficacy and upset in caregivers and on daily function of persons with dementia." The Gerontologist **41**(1): 4-14.

Gitlin, L. N., Miller, K. S. and Boyce, A. (1999). "Bathroom modifications for frail elderly renters: Outcomes of a community-based program." Technology & Disability **10**(3): 141-149.

Glance, L. G., Dick, A., Osler, T. M., Li, Y. and Mukamel, D. B. (2006). "Impact of changing the statistical methodology on hospital and surgeon ranking: The case of the New York state cardiac surgery report card." Medical Care **44**(4): 311-319.

Glance, L. G., Osler, T. and Shinozaki, T. (2000). "Effect of varying the case mix on the standardized mortality ratio and W statistic: A simulation study." Chest **117**(4): 1112-1117.

Glendinning, C., Challis, D., Fernandez, J.-L., Jacobs, S., Jones, K., Knapp, M., Manthorpe, J., Moran, N., Netten, A., Stevens, M. and Wilberforce, M. (2008a). Evaluation of the Individual Budgets pilot programme. Final report. York, Social Policy Research Unit, University of York.

Glendinning, C., Clarke, S., Hare, P., Maddison, J. and Newbronner, L. (2008b). "Progress and problems in developing outcomes-focused social care services for older people in England." Health & Social Care in the Community **16**(1): 54-63.

Goldstein, H. (1997). "Methods in school effectiveness research." School Effectiveness and School Improvement **8**(4): 369-395.

Goldstein, H. (2003). Multilevel statistical models. London, Arnold.

Goldstein, H., Browne, W. and Rasbash, J. (2002). Partitioning variation in multilevel models. Bristol, Bristol University.

Goldstein, H. and Healy, M. J. R. (1995). "The graphical presentation of a collection of means." Journal of the Royal Statistical Society. Series A (Statistics in Society) **158**(1): 175-177.

Goldstein, H. and Spiegelhalter, D. J. (1996). "League tables and their limitations: Statistical issues in comparisons of institutional performance." Journal of the Royal Statistical Society. Series A (Statistics in Society): 385-443.

Gomes, M., Gutacker, N., Bojke, C. and Street, A. (2014). Addressing missing data in patient-reported outcome measures (PROMs): Implications for comparing provider performance. CHE research paper 101. York, CHE, University of York.

Gomes, M., Gutacker, N., Bojke, C. and Street, A. (2016). "Addressing missing data in patient-reported outcome measures (PROMS): Implications for the use of PROMS for comparing provider performance." Health Economics **25**(5): 515-528.

Gormley, W. T. J. and Weimer, D. L. (1999). Organizational report cards. Cambridge, MA, Harvard University Press.

Goyder, J., Boyer, L. and Martinelli, G. (2006). "Integrating exchange and heuristic theories of survey nonresponse." Bulletin de méthodologie sociologique(92): 28-44.

Goyder, J., Lock, J. and Mcnair, T. (1992). "Urbanization effects on survey nonresponse: A test within and across cities." Quality and Quantity **26**(1): 39-48.

Greene, J. and Hibbard, J. H. (2012). "Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes." Journal of General Internal Medicine **27**(5): 520-526.

Greene, W. H. (2012). Econometric analysis. Harlow, Essex, Pearson.

Greenfield, S., Kaplan, S. H., Kahn, R., Ninomiya, J. and Griffith, J. L. (2002). "Profiling care provided by different groups of physicians: Effects of patient case-mix (bias) and physician-level clustering on quality assessment results." Annals of Internal Medicine **136**(2): 111-121.

Griffiths, R. (1988). "Does the public service serve? The consumer dimension." Public Administration **66**(2): 195-204.

Groves, R. (2006). "Nonresponse rates and nonresponse bias in household surveys." Public Opinion Quarterly **70**(5): 646-675.

Groves, R. and Couper, M. (1998). Nonresponse in household interview surveys. New York, John Wiley.

Groves, R. and Heeringa, S. (2006). "Responsive design for household surveys: Tools for actively controlling survey errors and costs." Journal of the Royal Statistical Society A **169**(3): 439-457.

Groves, R., Singer, E. and Corning, A. (2000). "Leverage-saliency theory of survey participation: Description and an illustration." The Public Opinion Quarterly **64**(3): 299-308.

Groves, R. M. and Cialdini, R. B. (1991). Toward a useful theory of survey participation. Proceedings of the Section on Survey Methods Research, American Statistical Association.

Groves, R. M., Cialdini, R. B. and Couper, M. P. (1992). "Understanding the decision to participate in a survey." The Public Opinion Quarterly **56**(4): 475-495.

Groves, R. M. and Peytcheva, E. (2008). "The impact of nonresponse rates on nonresponse bias: A meta-analysis." Public Opinion Quarterly **72**(2): 167-189.

Groves, R. M., Presser, S. and Dipko, S. (2004). "The role of topic interest in survey participation decisions." The Public Opinion Quarterly **68**(1): 2-31.

Gutiérrez-Romero, R., Haubrich, D. and Mclean, I. (2008). "The limits of performance assessments of public bodies: External constraints in English local government." Environment and Planning - Part C **26**(4): 767-787.

Gutiérrez Romero, R., Haubrich, D. and Mclean, I. (2010). "To what extent does deprivation affect the performance of English local authorities?" International Review of Administrative Sciences **76**(1): 137-170.

Hanley, R. J., Wiener, J. M. and Harris, K. M. (1991). "Will paid home care erode informal support?" J Health Polit Policy Law **16**(3): 507-521.

Hannan, E. L., Racz, M. J., Jollis, J. G. and Peterson, E. D. (1997). "Using Medicare claims data to assess provider quality for CABG surgery: Does it work well enough?" Health Services Research **31**(6): 659-678.

Hannan, E. L., Wu, C., Delong, E. R. and Raudenbush, S. W. (2005). "Predicting risk-adjusted mortality for CABG surgery: Logistic versus hierarchical logistic models." Medical Care **43**(7): 726-735.

Haubrich, D. and Mclean, I. (2006). "Assessing public service performance in local authorities through CPA - a research note on deprivation." National Institute Economic Review **197**(1): 93-a-105.

Hauser, R. M. (2005). "Survey response in the long run: The Wisconsin longitudinal study." Field Methods **17**(1): 3-29.

Health and Social Care Information Centre (2013). Personal social services adult social care survey, England 2012-13, final release, HSCIC.

Health and Social Care Information Centre (2014). Personal social services adult social care survey, England, 2013-14, final release, HSCIC.

Heath, C., Malley, J., Razik, K., Jones, K., Forder, J., Fox, D., Caiels, J. and Beecham, J. (2015). How can MAX help local authorities to use social care data to inform local policy? Maximising the value of survey data in adult social care [MAX] project. Canterbury PSSRU, University of Kent.

Hébert, R., Bravo, G., Korner-Bitensky, N. and Voyer, L. (1996). "Refusal and information bias associated with postal questionnaires and face-to-face interviews in very elderly subjects." Journal of Clinical Epidemiology **49**(3): 373-381.

Heckman, J. J. and Smith, J. A. (1995). "Assessing the case for social experiments." The Journal of Economic Perspectives **9**(2): 85-110.

Heinrich, C. J. (2003). Measuring public sector performance and effectiveness. Handbook of public administration. B. G. Peters and J. Pierre. London, Sage**:** 25-37.

Heller, A., Elliott, M. N., Haviland, A. M., Klein, D. J. and Kanouse, D. E. (2009). "Patient activation status as a predictor of patient experience among Medicare beneficiaries." Medical Care **47**(8): 850-857.

Hellström, Y. and Hallberg, I. R. (2001). "Perspectives of elderly people receiving home help on health, care and quality of life." Health Soc. Care Community **9**.

Hellström, Y., Persson, G. and Hallberg, I. R. (2004). "Quality of life and symptoms among older people living at home." Journal of Advanced Nursing **48**(6): 584-593.

Hendershot, G. E., Colpe, L. J. and Hunt, P. C. (2003). "Persons with activity limitations: Non response and proxy response in the US national health interview survey on disability." Research in Social Science and Disability **3**: 41-54.

Henn, M., Weinstein, M. and Wring, D. (2002). "A generation apart? Youth and political participation in Britain." The British Journal of Politics and International Relations **4**(2): 167-192.

Herzog, A. R. and Rodgers, W. L. (1988). "Age and response rates to interview sample surveys." Journal of Gerontology **43**(6): S200-S205.

Heywood, F. (2004). "The health outcomes of housing adaptations." Disability & Society **19**(2): 129-143.

Heywood, F. and Turner, L. (2007). Better outcomes, lower costs. Implications for health and social care budgets of investment in housing adaptations, improvements and equipment: A review of the evidence. London, Department for Work and Pensions.

Höfler, M., Pfister, H., Lieb, R. and Wittchen, H.-U. (2005). "The use of weights to account for non-response and drop-out." Social Psychiatry and Psychiatric Epidemiology **40**(4): 291-299.

Holland, P. W. (1986). "Statistics and causal inference." Journal of the American Statistical Association **81**(396): 945-960.

Hood, C. (2006). "Gaming in targetworld: The targets approach to managing British public services." Public Administration Review **66**(4): 515-521.

Hood, C. (2007). "Public service management by numbers: Why does it vary? Where has it come from? What are the gaps and the puzzles?" Public Money and Management **27**(2): 95-102.

Hood, C. and Jackson, M. (1991). Administrative argument. Dartmouth, Aldershot.

Horton, N. J. and Kleinman, K. P. (2007). "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models." The American Statistician **61**(1): 79-90.

Hosmer, D. and Lemeshow, S. (2000). Applied logistic regression. New York, John Wiley and Sons.

House, J. S. and Wolf, S. (1978). "Effects of urban residence on interpersonal trust and helping behavior." Journal of Personality and Social Psychology **36**(9): 1029-1043.

Huang, I.-C., Dominici, F., Frangakis, C., Diette, G. B., Damberg, C. L. and Wu, A. W. (2005). "Is risk-adjustor selection more important than statistical approach for provider profiling? Asthma as an example." Medical Decision Making **25**(1): 20-34.

Huang, I., Frangakis, C., Atkinson, M. J., Willke, R. J., Leite, W. L., Vogel, W. B. and Wu, A. W. (2008). "Addressing ceiling effects in health status measures: A comparison of techniques applied to measures for people with HIV disease." Health Serv Res **43**(1p1): 327-339.

Huber, P. J. (1967a). The behavior of maximum likelihood estimates under non-standard conditions. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability.

Huber, P. J. (1967b). The behavior of maximum likelihood estimates under nonstandard conditions. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Berkeley, CA, University of California Press. **1:** 221-233.

Humphries, R., Ruth, T., Holder, H., Hall, P. and Charles, A. (2016). Social care for older people. Home truths. London, The King's Fund and Nuffield Trust.

Hunter, C., Fitzpatrick, R., Jenkinson, C., Darlington, A.-S. E., Coulter, A., Forder, J. E. and Peters, M. (2015). "Perspectives from health, social care and policy stakeholders on the value of a single self-report outcome measure across long-term conditions: A qualitative study." BMJ Open **5**(5).

Hutchings, A., Neuburger, J., Grosse Frie, K., Black, N. and Van Der Meulen, J. (2012). "Factors associated with non-response in routine use of patient reported outcome measures after elective surgery in England." Health Qual Life Outcomes **10**.

Hutchings, B. L., Olsen, R. V. and Moulton, H. J. (2008). "Environmental evaluations and modifications to support aging at home with a developmental disability." Journal of Housing For the Elderly **22**(4): 286-310.

Hwang, E., Cummings, L., Sixsmith, A. and Sixsmith, J. (2011). "Impacts of home modifications on aging-in-place." Journal of Housing For the Elderly **25**(3): 246-257.

Iezzoni, L. I., Ed. (2013). Risk adjustment for measuring health care outcomes. Chicago, Illinois, Health Administration Press.

Iglesias, C. P., Birks, Y. F. and Torgerson, D. J. (2001). "Improving the measurement of quality of life in older people: The York SF-12." QJM **94**(12): 695-698.

Intriligator, M. D., Bodkin, R. and Hsiao, C. (1996). Econometric models, techniques, and applications. New Jersey, Prentice Hall.

Iwarsson, S., Horstmann, V. and Slaug, B. (2007). "Housing matters in very old age -- yet differently due to ADL dependence level differences." Scandinavian Journal of Occupational Therapy **14**(1): 3-15.

James, O. (2004). "The UK core executive's use of public service agreements as a tool of governance." Public Administration **82**(2): 397-419.

Janssen, C. G., Schuengel, C. and Stolk, J. (2005). "Perspectives on quality of life of people with intellectual disabilities: The interpretation of discrepancies between clients and caregivers." Quality of Life Research **14**(1): 57-69.

Jenkinson, C., Coulter, A. and Bruster, S. (2002a). "The Picker Patient Experience Questionnaire: Development and validation using data from in-patient surveys in five countries." International Journal for Quality in Health Care **14**(5): 353-358.

Jenkinson, C., Coulter, A., Bruster, S., Richards, N. and Chandola, T. (2002b). "Patients' experiences and satisfaction with health care: Results of a questionnaire study of specific aspects of care." Quality and Safety in Health Care **11**(4): 335-339.

Jenkinson, C. and Fitzpatrick, R. (2013). Patient reported outcomes. Understanding and using health experiences. Improving patient care. S. Ziebland, A. Coulter, J. Calabrese and L. Locock. Oxford, Oxford University Press**:** 72-80.

Jenkinson, C., Fitzpatrick, R., Peto, V., Greenhall, R. and Hyman, N. (1997). "The Parkinson's disease questionnaire (PDQ-39): Development and validation of a Parkinson's disease summary index score." Age Ageing **26**(5): 353-357.

Johnson, M. L., Rodriguez, H. P. and Solorio, M. R. (2010). "Case-mix adjustment and the comparison of community health center performance on patient experience measures." Health Services Research **45**(3): 670-690.

Johnson, T. P., Cho, Y. I., Campbell, R. T. and Holbrook, A. L. (2006). "Using community-level correlates to evaluate nonresponse effects in a telephone survey." Public Opinion Quarterly **70**(5): 704-719.

Johnstone, L. and Page, C. (2014). "Using Adult Social Care Outcomes Toolkit (ASCOT) in the assessment and review process." Research, Policy and Planning **30**(3): 179-192.

Jones, K., Netten, A., Francis, J. and Bebbington, A. (2007). "Using older home care user experiences in performance monitoring." Health & Social Care in the Community **15**(4): 322-332.

Jowell, R. and Park, A. (1998). Young people, politics and citizenship. Working paper 67. Oxford, CREST, University of Oxford.

Julious, S., Nicholl, J. and George, S. (2001). "Why do we continue to use standardized mortality ratios for small area comparisons?" Journal of Public Health **23**(1): 40-46.

Julious, S. A. and George, S. (2007). "Are hospital league tables calculated correctly?" Public Health **121**(12): 902-904.

Kaldenberg, D. O., Koenig, H. F. and Becker, B. W. (1994). "Mail survey response rate patterns in a population of the elderly: Does response deteriorate with age?" The Public Opinion Quarterly **58**(1): 68-76.

Kalton, G. and Flores-Cervantes, I. (2003). "Weighting methods." Journal of Official Statistics **19**(2): 81-97.

Kane, Rosalie, A., Kling, Kristen, C., Bershadsky, Boris, Robert, L., Giles, Katherine, Degenholtz, Howard, B., Jiexin, L. I. U., Cutlers and Lois, J. (2003). "Quality of life measures for nursing home residents." Journals of Gerontology. Series A, Biological sciences and medical sciences **58**(3): 9.

Kane, R. A. (2001). "Long-term care and a good quality of life: Bringing them closer together." Gerontologist **41**(3): 293-304.

Kane, R. L., Bershadsky, B., Kane, R. A., Degenholtz, H. H., Liu, J., Giles, K. and Kling, K. C. (2004). "Using resident reports of quality of life to distinguish among nursing homes." The Gerontologist **44**(5): 624-632.

Kaplan, R. S. and Norton, D. P. (1992). "The balanced scorecard: Measures that drive performance." Harvard business review(Jan/Feb): 71-79.

Katz, S., Downs, T. D., Cash, H. R. and Grotz, R. C. (1970). "Progress in development of the index of ADL." The Gerontologist **10**(1 Part 1): 20-30.

Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A. and Jaffe, M. W. (1963). "Studies of illness in the aged: The index of ADL: A standardized measure of biological and psychosocial function." JAMA **185**(12): 914-919.

Kauppi, M., Sokka, T. and Hannonen, P. (2005). "Survey nonresponse is associated with increased mortality in patients with rheumatoid arthritis and in a community population." The Journal of Rheumatology **32**(5): 807-810.

Kelly, J. M. (2005). "The dilemma of the unsatisfied customer in a market model of public administration." Public Administration Review **65**(1): 76-84.

Kirkham, J. J. (2008). "A comparison of hospital performance with non-ignorable missing covariates: An application to trauma care data." Statistics in Medicine **27**(27): 5725-5744.

Kirkpatrick, I. and Martinez Lucio, M. (1995a). Introduction: The politics of quality in the public sector. The politics of quality in the public sector. The management of change. I. Kirkpatrick and M. Martinez Lucio. London, Routledge**:** 1-15.

Kirkpatrick, I. and Martinez Lucio, M. (1995b). The uses of 'quality' in the British government's reform of the public sector. The politics of quality in the public sector. The management of change. I. Kirkpatrick and M. Martinez Lucio. London, Routledge**:** 16-43.

Klein, D. J., Elliott, M. N., Haviland, A. M., Saliba, D., Burkhart, Q., Edwards, C. and Zaslavsky, A. M. (2011). "Understanding nonresponse to the 2007 Medicare CAHPS survey." Gerontologist **51**(6): 843-855.

Knapp, M. (1978a). "Cost functions for care services for the elderly." Gerontologist **18**: 30-36.

Knapp, M. (1978b). "Economies of scale in residential care." International Journal of Social Economics **5**: 81-92.

Knapp, M. (1984). The economics of social care. Basingstoke, Macmillan.

Knapp, M., Hardy, B. and Forder, J. (2001). "Commissioning for quality: Ten years of social care markets in England." Journal of social policy **30**(2): 283-306.

Knapp, M. and Smith, J. (1985). "The costs of residential child care: Explaining variations in the public sector." Policy & Politics **13**(2): 127-154.

Knapp, M. R. J. (1979). "On the determination of the manpower requirements of old people's homes." Social Policy & Administration **13**(3): 219-236.

Kodde, D. A. and Palm, F. C. (1986). "Wald criteria for jointly testing equality and inequality restrictions." Econometrica **54**(5): 1243-1248.

Kreft, I. and De Leeuw, J. (1998). Introducing multilevel modeling. London, Sage.

Kroll, A. (2015). "Drivers of performance information use: Systematic literature review and directions for future research." Public Performance & Management Review **38**(3): 459-486.

Kroll, T., Wyke, S., Jahagirdar, D. and Ritchie, K. (2014). "If patient-reported outcome measures are considered key health-care quality indicators, who is excluded from participation?" Health Expectations **17**(5): 605-607.

Lasek, R. J., Barkley, W., Harper, D. L. and Rosenthal, G. E. (1997). "An evaluation of the impact of nonresponse bias on patient satisfaction surveys." Med Care **35**(6): 646-652.

Lau, L. J. (1974). Application of duality theory: A comment. Frontiers in quantitative economics. M. D. Intriligator and D. A. Kendrick. Amsterdam, North-Holland. **Vol II:** 176-199.

Lawton, M. P. (1983). "Environment and other determinants of well-being in older people." The Gerontologist **23**(4): 349-357.

Lawton, M. P. and Brody, E. M. (1969). "Assessment of older people: Self-maintaining and instrumental activities of daily living." The Gerontologist **9**(3 Part 1): 179-186.

Leadbeater, C. (2004). Personalisation through participation: A new script for public services. London, Demos.

Leckie, G. and Charlton, C. (2013). "Runmlwin - a program to run the mlwin multilevel modelling software from within Stata." Journal of Statistical Software **52**(11): 1-40.

Li, C. (2013). "Little's test of missing completely at random." Stata Journal **13**(4): 795-809.

Li, Y., Cai, X., Glance, L. G., Spector, W. D. and Mukamel, D. B. (2009). "National release of the nursing home quality report cards: Implications of statistical methodology for risk adjustment." Health Serv Res **44**(1): 79-102.

Li, Y., Schnelle, J., Spector, W. D., Glance, L. G. and Mukamel, D. B. (2010). "The "Nursing Home Compare" measure of urinary/fecal incontinence: Cross-sectional variation, stability over time, and the impact of case mix." Health Services Research **45**(1): 79-97.

Lilford, R., Mohammed, M. A., Spiegelhalter, D. and Thomson, R. (2004). "Use and misuse of process and outcome data in managing performance of acute medical care: Avoiding institutional stigma." The Lancet **363**(9415): 1147-1154.

Linder-Pelz, S. (1982). "Toward a theory of patient satisfaction." Social Science and Medicine **16**: 577-582.

Little, R. J. and Rubin, D. B. (2002). Statistical analysis with missing data. Hoboken, NJ, Wiley.

Little, R. J. A. (1986). "Survey nonresponse adjustments for estimates of means." International Statistical Review **54**: 139-157.

Little, R. J. A. (1988). "A test of missing completely at random for multivariate data with missing values." Journal of the American Statistical Association **83**: 1198-1202.

Little, R. J. A. and Vartivarian, S. (2003). "On weighting the rates in non-response weights." Statistics in Medicine **22**: 1589-1599.

Lo Sasso, A. T. and Johnson, R. W. (2002). "Does informal care from adult children reduce nursing home admissions for the elderly?" Inquiry **39**(3): 279-297.

Local Government Association (2013) "Towards excellence in adult social care. Statement of purpose ".

Lovelock, C. and Gummesson, E. (2004). "Whither services marketing?: In search of a new paradigm and fresh perspectives." Journal of Service Research **7**(1): 20-41.

Lymbery, M. (2010). "A new vision for adult social care? Continuities and change in the care of older people." Critical Social Policy **30**(1): 5-26.

Lynn, P. (2012). The propensity of older respondents to participate in a general purpose survey Understanding Society Working Paper Series. No. 2012 – 03 Institute for Social and Economic Research, University of Essex

Malley, J., Caiels, J., Fox, D., Mccarthy, M., Smith, N., Beadle-Brown, J., Netten, A. and Towers, A.-M. (2010). A report on the development studies for the national adult social care user experience survey, PSSRU discussion paper 2721. Canterbury, Personal Social Services Research Unit, University of Kent.

Malley, J. and Fernández, J.-L. (2010). "Measuring quality in social care services: Theory and practice." Annals of Public and Co-operative Economics **81**(4): 559-582.

Malley, J. and Fernandez, J. (2014). Generating adjusted indicators from social care survey data. PSSRU Discussion Paper 2873. London, Personal Social Services Research Unit, London School of Economics and Political Sciences.

Malley, J. and Netten, A. (2008). Measuring user experience of social care services: A discussion of three approaches. A report to the department of health. Discusssion paper 2529. Canterbury, Personal Social Services Research Unit.

Malley, J. and Netten, A. (2009). Putting People First: Development of the Putting People First user experience survey. Discussion paper 2637. Canterbury, Personal Social Services Research Unit.

Malley, J., Towers, A.-M., Netten, A., Brazier, J., Forder, J. and Flynn, T. (2012) "An assessment of the construct validity of the ASCOT measure of social care-related quality of life with older people." Health and Quality of Life Outcomes **10** DOI: 10.1186/1477-7525-10-21.

Manning, W. G., Basu, A. and Mullahy, J. (2005). "Generalized modeling approaches to risk adjustment of skewed outcomes data." Journal of Health Economics **24**(3): 465-488.

Mattei, A. (2009). "Estimating and using propensity score in presence of missing background data: An application to assess the impact of childbearing on wellbeing." Statistical Methods and Applications **18**(2): 257-273.

Matthews, F. E., Bennett, H., Wittenberg, R., Jagger, C., Dening, T., Brayne, C. and Cognitive Function, A. S. C. (2016). "Who lives where and does it matter? Changes in the health profiles of older people living in long term care and the community over two decades in a high income country." PLoS ONE **11**(9): e0161705.

Mcadam, R., Hazlett, S. A. and Casey, C. (2005). "Performance management in the UK public sector: Addressing multiple stakeholder complexity." International Journal of Public Sector Management **18**(3): 256-273.

Mcadam, R. and Walker, T. (2003). "An inquiry into balanced scorecards within Best Value implementation in UK local government." Public Administration **81**(4): 873-892.

Mckee, M., Britton, A., Black, N., Mcpherson, K., Sanderson, C. and Bain, C. (1999). "Interpreting the evidence: Choosing between randomised and non-randomised studies." BMJ **319**(7205): 312-315.

Mclean, I., Haubrich, D. and Gutierrez-Romero, R. (2007). "The perils and pitfalls of performance measurement: The CPA regime for local authorities in England." Public Money and Management **27**(2): 111-118.

Mclennan, D., Barnes, H., Noble, M., Davies, J., Garratt, E. and Dibben, C. (2011). The English indices of deprivation 2010. London, Communities and Local Government.

Mead, N., Bower, P. and Roland, M. (2008). "The General Practice Assessment Questionnaire (GPAQ) – development and psychometric characteristics." BMC Family Practice **9**(1): 1-11.

Micheli, P. and Neely, A. (2010). "Performance measurement in the public sector in England: Searching for the golden thread." Public Administration Review **70**(4): 591-600.

Miller, E., Cooper, S.-A., Cook, A. and Petch, A. (2008). "Outcomes important to people with intellectual disabilities." Journal of Policy and Practice in Intellectual Disabilities **5**(3): 150-158.

Mitra, R. and Reiter, J. P. (2011). "Estimating propensity scores with missing covariate data using general location mixture models." Statistics in Medicine **30**(6): 627-641.

Mohammed, M. A., Cheng, K. K., Rouse, A. and Marshall, T. (2001). "Bristol, Shipman, and clinical governance: Shewhart's forgotten lessons." Lancet **357**(9254): 463-467.

Mohammed, M. A. and Deeks, J. J. (2008). "In the context of performance monitoring, the caterpillar plot should be mothballed in favor of the funnel plot." The Annals of Thoracic Surgery **86**(1): 348.

Mold, A. (2010). "Patient groups and the construction of the patient-consumer in Britain: An historical overview." Journal of social policy **39**(04): 505-521.

Mold, A. (2011). "Making the patient-consumer in margaret thatcher's britain." The Historical Journal **54**(02): 509-528.

Molenberghs, G. (2007). "Editorial: What to do with missing data?" Journal of the Royal Statistical Society: Series A (Statistics in Society) **170**(4): 861-863.

Mor, V. (2005). "Improving the quality of long-term care with better information." Milbank Quarterly **83**(3): 333-364.

Mor, V., Berg, K., Angelelli, J., Gifford, D., Morris, J. and Moore, T. (2003). "The quality of quality measurement in U.S. Nursing homes." The Gerontologist **43**(suppl 2): 37-46.

Moscone, F., Knapp, M. and Tosetti, E. (2007). "Mental health expenditure in England: A spatial panel approach." Journal of Health Economics **26**(4): 842-864.

Moynihan, D. P. (2009). "Through a glass, darkly: Understanding the effects of performance regimes." Public Performance & Management Review **32**(4): 592-603.

Moynihan, D. P. and Pandey, S. K. (2010). "The big question for performance management: Why do managers use performance information?" Journal of Public Administration Research and Theory **20**(4): 849-866.

Mukaka, M., White, S. A., Terlouw, D. J., Mwapasa, V., Kalilani-Phiri, L. and Faragher, E. B. (2016). "Is using multiple imputation better than complete case analysis for estimating a prevalence (risk) difference in randomized controlled trials when binary outcome observations are missing?" Trials **17**: 341.

Mukamel, D. B. (1997). "Risk-adjusted outcome measures and quality of care in nursing homes." Medical Care **35**(4): 367-385.

Mukamel, D. B., Glance, L. G., Li, Y., Weimer, D. L., Spector, W. D., Zinn, J. S. and Mosqueda, L. (2008). "Does risk adjustment of the CMS quality measures for nursing homes matter?" Medical Care **46**(5): 532-541.

Murtaugh, C. M., Peng, T., Aykan, H. and Maduro, G. (2007). "Risk adjustment and public reporting on home health care." HealtH Care FinanCing review **28**(3): 77.

National Consumer Council (1983). Measuring the performance of local authorities in England and Wales -- some consumer principles. London, National Consumer Council,.

Neely, A., Gregory, M. and Platts, K. (1995). "Performance measurement system design: A literature review and research agenda." International Journal of Operations & Production Management **15**(4): 80-116.

Nelson, P. (1970). "Information and consumer behavior." Journal of Political Economy **78**(2): 311-329.

Netten, A. (2011). Overview of outcome measurement for adults using social care services and support. Methods review 6. London, School for Social Care Research.

Netten, A., Beadle-Brown, J., Caiels, J., Forder, J., Malley, J., Smith, N., Towers, A.-M., Trukeschitz, B., Welch, E. and Windle, K. (2011). Adult Social Care Outcomes Toolkit (ASCOT) main guidance v 2.1. PSSRU discussion paper 2716/3. Canterbury, Personal Social Services Research Unit, University of Kent.

Netten, A., Burge, P., Malley, J., Potoglou, D., Towers, A.-M., Brazier, J., Flynn, T. and Forder, J. (2012a). "Outcomes of social care for adults: Developing a preference-weighted measure." Health Technology Assessment **16**(16).

Netten, A. and Davies, B. (1990). "The social production of welfare and consumption of social services." Journal of Public Policy **10**(3): 331-347.

Netten, A., Jones, K., Knapp, M., Fernandez, J. L., Challis, D., Glendinning, C., Jacobs, S., Manthorpe, J., Moran, N., Stevens, M. and Wilberforce, M. (2012b). "Personalisation through Individual Budgets: Does it Work and for whom?" British Journal of Social Work **42**(8): 1556-1573.

Netten, A., Trukeschitz, B., Beadle-Brown, J., Forder, J., Towers, A.-M. and Welch, E. (2012c). "Quality of life outcomes for residents and quality ratings of care homes: Is there a relationship?" Age and ageing **41**(4): 512-517.

NHS Digital. (2016a). "Measures from the adult social care outcomes framework, England - 2014-15, final release." Retrieved 10th January 2017, from http://content.digital.nhs.uk/catalogue/PUB18657.

NHS Digital (2016b). Personal social services adult social care survey England 2015-16. Leeds, NHS Digital.

NHS England Analytical Team (2013). Patient reported outcome measures (PROMs). An alternative aggregation method for case-mix adjustment. NHS England. London, NHS England.

NHS Information Centre (2010). Personal social services adult social care survey guidance document. Leeds, The Health and Social Care Information Centre.

Nicholl, J. (2007). "Case-mix adjustment in non-randomised observational evaluations: The constant risk fallacy." J Epidemiol Community Health **61**(11): 1010-1013.

Nicholl, J., Jacques, R. M. and Campbell, M. J. (2013). "Direct risk standardisation: A new method for comparing casemix adjusted event rates using complex models." BMC Medical Research Methodology **13**(1): 133.

Nocon, A. and Qureshi, H. (1996a). Outcomes of community care. Measuring outcome in the public sector. P. C. Smith. London, Taylor and Francis**:** 74-93.

Nocon, A. and Qureshi, H. (1996b). Outcomes of community care for users and carers: A social services perspective. Buckingham, Open University Press.

Normand, S.-L. T., Glickman, M. E. and Gatsonis, C. A. (1997). "Statistical methods for profiling providers of medical care: Issues and applications." Journal of the American Statistical Association **92**(439): 803-814.

Nunnally, J. C. (1967). Psychometric theory. New York, New York : McGraw-Hill.

Nutley, S., Walter, I. and Davies, H. T. O. (2003). "From knowing to doing: A framework for understanding the evidence-into-practice agenda." Evaluation **9**(2): 125-148.

Nuttall, D., Parkin, D. and Devlin, N. (2015). "Inter-provider comparison of patient-reported outcomes: Developing an adjustment to account for differences in patient case mix." Health Economics **24**(1): 41-54.

Nygren, C., Oswald, F., Iwarsson, S., Fänge, A., Sixsmith, J., Schilling, O., Sixsmith, A., Széman, Z., Tomsone, S. and Wahl, H.-W. (2007). "Relationships between objective and perceived housing in very old age." The Gerontologist **47**(1): 85-95.

O'malley, A. J., Zaslavsky, A. M., Elliott, M. N., Zaborski, L. and Cleary, P. D. (2005). "Case-mix adjustment of the CAHPS® hospital survey." Health Services Research **40**(6p2): 2162-2181.

O'donnell, O., Van Doorslaer, E., Wagstaff, A. and Lindelow, M. (2008). Measuring and explaining inequity in health service delivery. Analyzing health equity using household survey data: A guide to techniques and their implementation. Washington, D.C., The World Bank.

Office for National Statistics (2001). Population density (UV02), from 2001 census. Office for National Statistics.

Oliver, R. L. (1997). Satisfaction: A behavioral perspective on the consumer. New York, Irwin/McGraw-Hill.

Paganini-Hill, A., Hsu, G., Chao, A. and Ross, R. K. (1993). "Comparison of early and late respondents to a postal health survey questionnaire." Epidemiology **4**(4): 375-379.

Papke, L. E. and Wooldridge, J. M. (1996). "Econometric methods for fractional response variables with an application to 401 (K) plan participation rates." Journal of Applied Econometrics: 619-632.

Parasuraman, A., Berry, L. L. and Zeithaml, V. A. (1991). "Refinement and reassessment of the SERVQUAL scale." Journal of Retailing **67**(4): 420-451.

Parasuraman, A., Zeithmal, V. and Berry, L. (1985). "A conceptual model of service quality and its implications for future research." Journal of Marketing **49**: 41-50.

Parkin, D. and Devlin, N. J. (2012). "Using health status to measure NHS performance: Casting light in dark places." BMJ Quality & Safety **21**(4): 355-356.

Parks, R. (1984). "Linking objective and subjective measures of performance." Public Administration Review **44**(March/April): 118-127.

Patmore, C. (2004). "Quality in home care for older people: Factors to pay heed to." Quality in Ageing and Older Adults **5**(1): 32-40.

Perneger, T. V., Chamot, E. and Bovier, P. A. (2005). "Nonresponse bias in a survey of patient perceptions of hospital care." Medical Care **43**(4): 374-380.

Perry, J. and Felce, D. (2002). "Subjective and objective quality of life assessment: Responsiveness, response bias, and resident:Proxy concordance." Mental Retardation **40**(6): 445-456.

Peters, M., Crocker, H., Dummett, S., Jenkinson, C., Doll, H. and Fitzpatrick, R. (2014a). "Change in health status in long-term conditions over a one year period: A cohort survey using patient-reported outcome measures." Health and Quality of Life Outcomes **12**(1): 1-10.

Peters, M., Crocker, H., Jenkinson, C., Doll, H. and Fitzpatrick, R. (2014b). "The routine collection of patient-reported outcome measures (PROMs) for long-term conditions in primary care: A cohort survey." BMJ Open **4**(2).

Petticrew, M. and Roberts, H. (2003). "Evidence, hierarchies, and typologies: Horses for courses." Journal of Epidemiology and Community Health **57**(7): 527-529.

Pfeffer, N. and Coote, A. (1991). Is quality good for you? A critical review of quality assurance in welfare services. London, IPPR.

Pickard, L. (2001). "Carer break or carer-blind? Policies for informal carers in the UK." Social Policy & Administration **35**(4): 441-458.

Pickard, L. (2012). "Substitution between formal and informal care: A 'natural experiment' in social policy in Britain between 1985 and 2000." Ageing and Society **32**(7): 1147-1175.

Pickard, L., Wittenberg, R., Comas-Herrera, A., King, D. and Malley, J. (2012). "Mapping the future of family care: Receipt of informal care by older people with disabilities in England to 2032." Social Policy and Society **11**(4): 533-545.

Pierre, U., Wood-Dauphinee, S., Korner-Bitensky, N., Gayton, D. and Hanley, J. (1998). "Proxy use of the Canadian SF-36 in rating health status of the disabled elderly." Journal of Clinical Epidemiology **51**(11): 983-990.

Pollitt, C. (1988). "Bringing consumers into performance measurement: Concepts, consequences and constraints." Policy & Politics **16**(2): 77-87.

Potoglou, D., Burge, P., Flynn, T., Netten, A., Malley, J., Forder, J. and Brazier, J. E. (2011). "Best–worst scaling vs. Discrete choice experiments: An empirical comparison using social care data." Social Science & Medicine **72**(10): 1717-1727.

Pregibon, D. (1980). "Goodness of link tests for generalized linear models." Applied Statistics: 15-14.

Pregibon, D. (1981). "Logistic regression diagnostics." The Annals of Statistics: 705-724.

Pullenayegum, E. M., Tarride, J.-E., Xie, F., Goeree, R., Gerstein, H. C. and O'reilly, D. (2010). "Analysis of health utility data when some subjects attain the upper bound of 1: Are Tobit and clad models appropriate?" Value in Health **13**(4): 487-494.

Qu, Y. and Lipkovich, I. (2009). "Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach." Stat Med **28**(9): 1402-1414.

Qureshi, H. and Henwood, M. (2000). Older people's definitions of quality services. York, Joseph Rowntree Foundation

Qureshi, H. and Nicholas, E. (2001). "A new conception of social care outcomes and its practical use in assessment with older people." Research, Policy and Planning **19**(2): 11-26.

Qureshi, H., Patmore, C., Nicholas, E. and Bamford, C. H. (1998). Overview: Outcomes of social care for older people. Outcomes in Community Care Practice Series. York, Social Policy Research Unit.

Qureshi, H. and Rowlands, O. (2004). "User satisfaction surveys and cognitive questions testing in the public sector: The case of personal social services in England." International Journal of Social Research Methodology: Theory and Practice **7**(4): 273-287.

Rabe-Hesketh, S. and Skrondal, A. (2008). Multilevel and longitudinal modelling using Stata. College Station, Texas, Stata press.

Raftery, A. E. (1995). "Bayesian model selection in social research." Sociological Methodology **25**: 111-163.

Rahi, J. S., Manaras, I., Tuomainen, H. and Lewando Hundt, G. (2004). "Engaging families in health services research on childhood visual impairment: Barriers to, and degree and nature of bias in, participation." British Journal of Ophthalmology **88**(6): 782-787.

Rainey, C., Woolham, J. and Stevens, M. (2015). Research capacity, knowledge, skills and use in councils with adult social care responsibilities. Findings from an online survey of research active local government staff. London, London School of Economics.

Ramsey, J. B. (1969). "Tests for specification errors in classical linear least-squares regression analysis." Journal of the Royal Statistical Society. Series B (Methodological): 350-371.

Rao, M., Clarke, A., Sanderson, C. and Hammersley, R. (2006). "Patients' own assessments of quality of primary care compared with objective records based measures of technical quality of care: Cross sectional study." BMJ **333**(7557): 19.

Raynes, N., Temple, B., Glenister, C. and Coulthard, L. (2001). Quality at home for older people. Involving service users in defining home care specifications. York, The Policy press and Joseph rowntree Foundation.

Raynes, N. V. (1998). "Involving residents in quality specification." Ageing and Society **18**(1): 65-78.

Renfroe, E. G., Heywood, G., Foreman, L., Schron, E., Powell, J., Baessler, C., Warwick, D., Morris, M. and Hallstrom, A. (2002). "The end-of-study patient survey: Methods influencing response rate in the avid trial." Controlled Clinical Trials **23**(5): 521-533.

Revelli, F. (2006). "Performance rating and yardstick competition in social service provision." Journal of Public Economics **90**(3): 459-475.

Rixom, A. (2002). "Performance league tables : Use of indirect standardisation is inappropriate." BMJ : British Medical Journal **325**(7357): 177-178.

Roland, M., Elliott, M., Lyratzopoulos, G., Barbiere, J., Parker, R. A., Smith, P., Bower, P. and Campbell, J. (2009). "Reliability of patient responses in pay for performance schemes: Analysis of national general practitioner patient survey data in England." British Medical Journal **339**.

Royston, P. (1991). "Sg3.5: Comment on sg3.4 and an improved d'agostino test." Stata Technical Bulletin **3**: 23-22.

Rti International (2017). MDS 3.0 quality measures user's manual version 11.0. Baltimore, MD, Centers for Medicare and Medicaid Services.

Rubin, D. B. (1974). "Estimating causal effects of treatments in randomized and nonrandomized studies." Journal of Educational Psychology **66**(5): 688-701.

Rubin, D. B. (1976). "Inference and missing data." Biometrika **63**(3): 581-592.

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. New York, J. Wiley & Sons.

Rubin, D. B. (1996). "Multiple imputation after 18+ years." Journal of the American Statistical Association **91(434)**: 473-489.

Rubin, D. B. (2003). "Discussion on multiple imputation." International Statistical Review **71**: 619-625.

Ryan, M. (1999). "Using conjoint analysis to take account of patient preferences and go beyond health outcomes: An application to in vitro fertilisation." Social Science & Medicine **48**(4): 535-546.

Sangl, J., Buchanan, J., Cosenza, C., Bernard, S., Keller, S., Mitchell, N., Brown, J., Castle, N., Sekscenski, E. and Larwood, D. (2007). "The development of a CAHPS® instrument for nursing home residents (NHCAHPS)." Journal of Aging & Social Policy **19**(2): 63-82.

Scheidt, R. J. and Norris-Baker, C. (2003). "Many meanings of community: Contributions of M. Powell Lawton." Journal of Housing for the Elderly **17**(1/2): 55-66.

Schmidt, S., Power, M., Green, A., Lucas-Carrasco, R., Eser, E., Dragomirecka, E. and Fleck, M. (2010). "Self and proxy rating of quality of life in adults with intellectual disabilities: Results from the DISQOL study." Research in developmental disabilities **31**(5): 1015-1026.

Schulz, E. (2012). Quality assurance policies and indicators for long-term care in the european union. Country report: Germany. ENEPRI research report no. 104, ENEPRI.

Schwarz, G. (1978). "Estimating the dimension of a model." The Annals of Statistics **6**(2): 461-464.

Seaman, S. and White, I. (2014). "Inverse probability weighting with missing predictors of treatment assignment or missingness." Communications in Statistics: Theory & Methods **43**(16): 3499-3515.

Seaman, S. R. and White, I. R. (2013). "Review of inverse probability weighting for dealing with missing data." Statistical Methods in Medical Research **22**(3): 278-295.

Seaman, S. R., White, I. R., Copas, A. J. and Li, L. (2012). "Combining multiple imputation and inverse-probability weighting." Biometrics **68**(1): 129-137.

Shahian, D. M., Normand, S.-L., Torchiana, D. F., Lewis, S. M., Pastore, J. O., Kuntz, R. E. and Dreyer, P. I. (2001). "Cardiac surgery report cards: Comprehensive review and statistical critique1." The Annals of Thoracic Surgery **72**(6): 2155-2168.

Shahian, D. M. and Normand, S. L. (2008). "Comparison of "risk-adjusted" hospital outcomes." Circulation **117**(15): 1955-1963.

Sheldon, H., Graham, C., Pothecary, N. and Rasul, F. (2007). Increasing response rates amongst black and minority ethnic and seldom heard groups. A review of literature relevant to the national acute patients' survey, Picker Institute Europe.

Shippee, T. P., Henning-Smith, C., Kane, R. L. and Lewis, T. (2015). "Resident- and facility-level predictors of quality of life in long-term care." The Gerontologist **55**(4): 643-655.

Shiu, E., Vaughan, L. and Donnelly, M. (1997). "Service quality: New horizons beyond SERVQUAL: An investigation of the portability of SERVQUAL into the voluntary and local government sectors." International Journal of Nonprofit and Voluntary Sector Marketing **2**(4): 324-331.

Sikkes, S. A., Visser, P. J., Knol, D. L., De Lange-De Klerk, E. S., Tsolaki, M., Frisoni, G. B., Nobili, F., Spiru, L., Rigaud, A. S., Frolich, L., Rikkert, M. O., Soininen, H., Touchon, J., Wilcock, G., Boada, M., Hampel, H., Bullock, R., Vellas, B., Pijnenburg, Y. A., Scheltens, P., Verhey, F. R. and Uitdehaag, B. M. (2011). "Do instrumental activities of daily living predict dementia at 1- and 2-year follow-up? Findings from the development of screening guidelines and diagnostic criteria for predementia Alzheimer's disease study." J Am Geriatr Soc **59**(12): 2273-2281.

Sitzia, J. and Wood, N. (1997). "Patient satisfaction: A review of issues and concepts." Social Science & Medicine **45**(12): 1829.

Sixsmith, A. and Sixsmith, J. (2008). "Ageing in place in the United Kingdom." Ageing International **32**(3): 219-235.

Skinner, C. J. and D'arrigo, J. (2011). "Inverse probability weighting for clustered nonresponse." Biometrika **98**(4): 953-966.

Smith, A. E., Sim, J., Scharf, T. and Phillipson, C. (2004). "Determinants of quality of life amongst older people in deprived neighbourhoods." Ageing and Society **24**(5): 793-814.

Smith, P. (1990). "The use of performance indicators in the public sector." Journal of the Royal Statistical Society. Series A (Statistics in Society): 53-72.

Smith, P. C. and Street, A. D. (2013). "On the uses of routine patient-reported health outcome data." Health Economics **22**(2): 119-131.

Smith, T. W. (2011). "The report of the international workshop on using multi-level data from sample frames, auxiliary databases, paradata and related sources to detect and adjust for nonresponse bias in surveys." International Journal of Public Opinion Research **23**(3): 389-402.

Song, J., Belin, T. R., Lee, M. B., Gao, X. and Rotheram-Borus, M. J. (2001). "Handling baseline differences and missing items in a longitudinal study of HIV risk among runaway youths." Health Services and Outcomes Research Methodology **2**(3): 317-329.

Spiegelhalter, D. J. (2005a). "Funnel plots for comparing institutional performance." Stat Med **24**(8): 1185-1202.

Spiegelhalter, D. J. (2005b). "Handling over-dispersion of performance indicators." Quality and Safety in Health Care **14**(5): 347-351.

Stevenson, D. and Bramson, J. (2014). Regulation of long-term care in the United States. The challenges in regulating long-term care quality: An international comparison. V. Mor, T. Leone and A. Maresso. Cambridge, Cambridge University Press.

Stewart, J. and Walsh, K. (1994). "Performance measurement: When performance can never be finally defined." Public Money and Management **14**(2): 45-49.

Stipak, B. (1979). "Citizen satisfaction with urban services: Potential misuse as a performance indicator." Public Administration Review **39**(January/February): 46-52.

Strayer, M., Kuthy, R. and Sutton, S. (1993). "Elderly nonrespondents to a mail survey: A telephone follow-up." Spec Care Dentist **13**(6): 245-248.

Streiner, D. L., Norman, G. R. and Cairney, J. (2014). Health measurement scales: A practical guide to their development and use, Oxford University Press, USA.

Talbot, C. (2007). Performance management. The Oxford handbook of public management. E. Ferlie, L. Lynn and C. Pollitt. Oxford, Oxford University Press.

Talbot, C. (2010). Theories of performance: Organizational and service improvement in the public domain. Oxford, Oxford University Press.

Tanner, B., Tilse, C. and De Jonge, D. (2008). "Restoring and sustaining home: The impact of home modifications on the meaning of home for older people." Journal of Housing For the Elderly **22**(3): 195 - 215.

The Health and Social Care Information Centre (2012). Personal social services: Expenditure and unit costs - England 2010-11- final release Leeds, The Health and Social Care Information Centre,.

The Information Centre for Health and Social Care (2012a). Community care statistics 2010-11: Social services activity report, England. Leeds, Health and Social Care Information Centre,.

The Information Centre for Health and Social Care (2012b). Personal social services adult social care survey, England 2010-11 (final release). Leeds, The Health and Social Care Information Centre,.

The Information Centre for Health and Social Care (2012c). Personal social services: Expenditure and unit costs -England 2010-11- final release. Leeds, The Health and Social Care Information Centre,.

The NHS Information Centre (2010). Personal social services adult social care survey guidance document. Leeds, Health and Social Care Information Centre.

Towers, A.-M., Holder, J., Smith, N., Crowther, T., Netten, A., Welch, E. and Collins, G. (2015). "Adapting the adult social care outcomes toolkit (ASCOT) for use in care home quality monitoring: Conceptual development and testing." BMC Health Services Research **15**(1): 1-12.

Towers, A., Smith, N., Palmer, S. and Welch, E. (2016). "Giving care home staff feedback on residents' outcomes: Can it be used to inform practice and improve residents' quality of life?" BMC Health Services Research.

Trickey, F., Maltais, D., Gosselin, C. and Robitaille, Y. (1994). "Adapting older persons' homes to promote independence." Physical & Occupational Therapy In Geriatrics **12**(1): 1-14.

Triemstra, M., Winters, S., Kool, R. and Wiegers, T. (2010). "Measuring client experiences in long-term care in the netherlands: A pilot study with the consumer quality index long-term care." BMC Health Services Research **10**(1): 95.

Trukeschitz, B. (2011). "Worauf es letztlich ankommt: Ergebnisqualität in der langzeitpflege und betreuung [quality of outcomes in long-term care]." Kurswechsel **26**(4): 22-35.

United Kingdom Homecare Association (UKHCA) (2015). The homecare deficit. A report on the funding of older people's homecare across the United Kingdom. Wallington, UKHCA.

Vaarama, M. (2009). "Care-related quality of life in old age." European Journal of Ageing **6**(2): 113-125.

Valderas, J. M., Fitzpatrick, R. and Roland, M. (2012). "Using health status to measure NHS performance: Another step into the dark for the health reform in England." BMJ Quality & Safety **21**(4): 352-353.

Van Buuren, S. (2007). "Multiple imputation of discrete and continuous data by fully conditional specification." Statistical Methods in Medical Research **16**(3): 219-242.

Van Houtven, C. H. and Norton, E. C. (2004). "Informal care and health care use of older adults." Journal of Health Economics **23**(6): 1159-1180.

Van Houtven, C. H. and Norton, E. C. (2008). "Informal care and Medicare expenditures: Testing for heterogeneous treatment effects." Journal of Health Economics **27**(1): 134-156.

Van Leeuwen, K. M., Malley, J., Bosmansc, J. E., Jansena, A. P. D., Ostelo, R. W., Van Der Horsta, H. E. and Netten, A. (2014). "What can local authorities do to improve the social care-related quality of life of older adults living at home? Evidence from the adult social care survey." Health & Place **29**: 104-113.

White, H. (1980). "A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity." Econometrica **48**: 817–830.

White, I. R., Royston, P. and Wood, A. M. (2011). "Multiple imputation using chained equations: Issues and guidance for practice." Statistics in Medicine **30**(4): 377-399.

Williams, B. (1994). "Patient satisfaction: A valid concept?" Social Science & Medicine **38**(4): 509.

Willis, R., Khambhaita, P., Pathak, P. and Evandrou, M. (2015). "Satisfaction with social care services among south asian and White British older people: The need to understand the system." Ageing & Society **FirstView**: 1-24.

Wimbush, E. (2011). "Implementing an outcomes approach to public management and accountability in the UK—are we learning the lessons?" Public Money & Management **31**(3): 211-218.

Wisniewski, M. (1996). "Measuring service quality in the public sector: The potential for SERVQUAL." Total Quality Management **7**(4): 357-366.

Yeung, E. Y. W., Partridge, M. and Irvine, F. (2016). "Satisfaction with social care: The experiences of people from chinese backgrounds with physical disabilities." Health & Social Care in the Community **24**(6): e144-e154.

Yuan, Y. and Little, R. J. A. (2007). "Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response." Journal of the Royal Statistical Society: Series C (Applied Statistics) **56**(1): 79-97.

Zaslavsky, A. M., Zaborski, L. B. and Cleary, P. D. (2002). "Factors affecting response rates to the consumer assessment of health plans study survey." Med Care **40**(6): 485-499.

Zaslavsky, A. M., Zaborski, L. B., Ding, L., Shaul, J. A. and Et Al. (2001). "Adjusting performance measures to ensure equitable plan comparisons." Health Care Financing review **22**(3): 109-126.

Zimmerman, D. R., Karon, S. L., Arling, G., Clark, B. R., Collins, T., Ross, R. and Sainfort, F. (1995). "Development and testing of nursing home quality indicators." Health Care Financing review **16**(4): 107-127.

Appendix 1: Specification of the Imputation Equations for Multiple Imputation of the ASCS Respondent Sample

I imputed the missing data in the respondent sample, due to missingness in key variables, including the outcome indicators, covariates included in the adjustment analysis and the variables defining the sub-groups investigated in the adjustment models, i.e. age group, primary client group and location of care. I imputed binary variables using logistic regression (logit); ordinal variables using proportional odds models (ologit); and imputed variables with more than two nominal categories using multinomial logit models (mlogit).

The imputation models included all theoretically-relevant variables for the case-mix adjustment (CMA) models. To ensure the plausibility of the MAR assumption, other variables that were not tested for inclusion in the adjustment model but were available in the dataset were also included in the imputation models. (Variables with more than 50 per cent of the cases missing (i.e. secondary client group, religion and budget) were not considered for the imputation models (McKee et al. 1999, Rubin 2003).) However, due to size of the dataset and number of variables with missing data (and therefore requiring imputation) a pragmatic approach had to be taken to the specification of the imputation equations to make the imputation computationally feasible. Therefore, particularly in the case of the variables imputed using mlogits, covariates were omitted for the imputation equation where the predictive significance of the variable was found to be less than 0.1. I also omitted variables that were highly collinear or perfectly predicted the outcome (e.g. in the case of helpstat).

A set of dummy variables for the CASSR were included in the imputation models (except those using mlogit due to difficulties with estimation) to capture CASSR effects. The derived variables (i.e. the ADL scales and the ASCOT measure) were not included in the imputation equations as their derivation could not be specified in the way required by the software. However, imputation of each individual item within the scales should preserve the complexity of the relationships. Additionally not all relevant interactions could be included in the imputation models where these were specified between multi-item measures.

Due to missingness in many of the additional variables included in the imputation models to ensure the plausibility of the MAR assumption, imputation models had to be specified for these additional variables. All covariates for the adjustment models and all additional variables were included in these models, although in the case of the mlogits a

reduced set of variables were used for computational feasibility. Only variables with a p-value of 0.1 or less were included in the models.

To provide valid imputations of the service receipt variables, information on CASSR-level service provision and gross ASC expenditure (The Information Centre for Health and Social Care 2012c, The Information Centre for Health and Social Care 2012a) were included in the imputation equations for these variables. These authority-level data were introduced to allow for shifts in the relationship between service receipt and the other variables by CASSR, particularly important given some service receipt variables (meals, short-term residential care, personal budgets, professional support, equipment and other services) were completely missing within a number of CASSRs. Of the 20 instances where service receipt data was completely missing for all cases within a CASSR, in ten service receipt and expenditure information was available; in a further seven the national data indicated that the service was not provided by that CASSR (i.e. recipients were zero), so all missing values for the relevant service for the respective CASSRs were replaced by a zero (not receiving the service); unfortunately in a further three cases no service receipt information was available at the national level for that service. Although total gross expenditure was available for these CASSRs the imputed values may be more unreliable for the missing services for these CASSRs.

Due to the missingness in the service receipt and gross ASC expenditure and deprivation variables, these variables also had to be imputed. The continuous variables were imputed using predictive mean matching (pmm), due to the skewed nature of the distribution for these variables. Given these additional variables are not being used in the adjustment models, the outcome variables were not included in the imputation equations and only CASSR-level variables were included to mitigate excessive CASSR-level variability. All available (i.e. fully observed or in the imputation model) CASSR-level variables were used.

Appendix 2: Additional Information on the Multiple Imputation of the ASCS Respondent Sample

Checking convergence of the multiple chains

Prior to imputation proper, I checked for convergence of the multiple chains, by examining plots summarising the distribution (means and standard deviations) of imputed values against iteration numbers. For this purpose I used the chainonly and savetrace options in Stata alongside the mi impute chained command. The variables did not appear to show any trends, and convergence appears to be achieved around 20 to 40 iterations.

In addition I examined the behaviour of three chains, each obtained using a different set of initial values, to check the convergence and stability of the algorithm, using the add(3) option instead of chainonly. The variables did not appear to show any trends, and the three chains seem to oscillate around the same point, providing some evidence of convergence of the algorithm. Convergence appears to be achieved after around 20 to 40 iterations.

Checking the fit of the imputation models

I compared the imputed values with the observed ones for each of the continuous variables using the user-written command midiagplot (Eddings and Marchenko 2006), to check the fit of the imputation model. Inspection of these plots revealed that predictive mean matching does a relatively good job of imputing the skewed distribution of the CASSR-level aggregate service receipt and gross expenditure variables.

For all other variables, to ensure the imputations produce sensible results I compared the distributional statistics calculated on the casewise-deleted sample to those calculated on the multiply-imputed sample for each variable. Table 54 shows the distributional statistics for the outcome indicators. The mean and standard errors are very similar regardless of the sample for all the outcome indicators, across all analysis subgroups (not shown). Similarly for the variables considered in the CMA models, the distributional statistics calculated on the casewise-deleted sample are very similar to those calculated on the multiply-imputed sample, for all sample subgroups and variables (see Table 55). The ADL score measures show the largest differences but in all cases the estimates for the mean are still identical to two significant figures. Although less important as the variables are only used to improve the MAR assumption in the imputation procedure, the distributional statistics for all the other variables included in the imputation models are shown in Table 56.

Table 54: Comparison of distributional statistics for outcome indicators for the respondent sample on casewise-deleted and multiply-imputed samples

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
ASCOF SCRQoL	54,350	18.62	0.02	18.58	0.02
Satisfaction: Ex. Sat	57,929	18,489	31.9	19,321	31.7
Satisfaction: V. sat	57,929	19,139	33.0	20,163	33.0
Satisfaction: Q. sat	57,929	14,965	25.8	15,860	26.0
Satisfaction: Neither	57,929	3,410	5.9	3,629	5.9
Satisfaction: Q. dissat	57,929	1,046	1.8	1,115	1.8
Satisfaction: V. dissat	57,929	391	0.7	418	0.7
Satisfaction: Ex. dissat	57,929	489	0.8	520	0.9
Control: preferred	59,478	17,742	29.8	18,225	29.9
Control: needs met	59,478	26,738	45.0	27,424	44.9
Control: some needs	59,478	12,096	20.3	12,404	20.3
Control: high needs	59,478	2,902	4.9	2,973	4.9
Pers care: preferred	59,688	32,823	55.0	33,549	55.0
Pers care: needs met	59,688	23,707	39.7	24,237	39.7
Pers care: some needs	59,688	2,710	4.5	2,780	4.6
Pers care: high needs	59,688	448	0.8	460	0.8
Food: preferred	59,373	37,512	63.2	38,531	63.1
Food: needs met	59,373	18,650	31.4	19,178	31.4
Food: some needs	59,373	2,627	4.4	2,714	4.4
Food: high needs	59,373	584	1.0	603	1.0
Accom: preferred	59,355	37,842	63.8	38,844	63.7
Accom: needs met	59,355	18,638	31.4	19,207	31.5
Accom: some needs	59,355	2,503	4.2	2,590	4.2
Accom: high needs	59,355	372	0.6	386	0.6
Safety: preferred	59,499	36,811	61.9	37,752	61.9
Safety: needs met	59,499	18,234	30.6	18,705	30.7
Safety: some needs	59,499	3,157	5.3	3,236	5.3
Safety: high needs	59,499	1,297	2.2	1,333	2.2
Soc part: preferred	59,469	24,835	41.8	25,471	41.7
Soc part: needs met	59,469	21,043	35.4	21,593	35.4
Soc part: some needs	59,469	10,493	17.6	10,777	17.7
Soc part: high needs	59,469	3,098	5.2	3,186	5.2
Occ: preferred	58,949	17,359	29.4	17,872	29.3
Occ: needs met	58,949	19,679	33.4	20,380	33.4
Occ: some needs	58,949	17,520	29.7	18,209	29.8
Occ: high needs	58,949	4,391	7.4	4,565	7.5
Dignity: preferred	57,569	32,765	56.9	34,557	56.6
Dignity: needs met	57,569	18,681	32.4	19,842	32.5
Dignity: some needs	57,569	5,458	9.5	5,894	9.7
Dignity: high needs	57,569	665	1.2	733	1.2
Information: v. easy	42,884	11,174	26.1	11,860	26.2
Information: fairly easy	42,884	20,719	48.3	21,872	48.3
Information: fairly diff	42,884	7,478	17.4	7,879	17.4
Information: v. diff	42,884	3,513	8.2	3,702	8.2

Table 55: Comparison of distributional statistics for adjustment model covariates for the respondent sample on casewise-deleted and multiply-imputed samples

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
Male	60,985	22,143	36.3	22,158	36.3
Age: 18-24	60,953	1,553	2.5	1,555	2.5
Age: 25-34	60,953	2,542	4.2	2,545	4.2
Age: 35-44	60,953	3,892	6.4	3,898	6.4
Age: 45-54	60,953	5,516	9.0	5,524	9.1
Age: 55-64	60,953	6,390	10.5	6,400	10.5
Age: 65-74	60,953	8,189	13.4	8,201	13.4
Age: 75-84	60,953	14,923	24.5	14,942	24.5
Age: 85 and over	60,953	17,948	29.4	17,962	29.4
Physical/sensory	60,778	43,501	71.6	43,685	71.6
Mental health problem	60,778	6,809	11.2	6,842	11.2
Learning disability	60,778	8,876	14.6	8,900	14.6
Substance misuse	60,778	136	0.2	137	0.2
Vulnerable person	60,778	1,456	2.4	1,462	2.4
ADLs diff score	56,092	2.84	0.01	2.82	0.01
ADLs can't score	56,092	1.28	0.01	1.27	0.01
ADL, indoors: can	58,480	32,097	54.9	33,510	54.9
ADL, indoors: diff	58,480	17,413	29.8	18,138	29.7
ADL, indoors: can't	58,480	8,970	15.3	9,377	15.4
ADL, bed/chair: can	58,645	33,117	56.5	34,527	56.6
ADL, bed/chair: diff	58,645	15,601	26.6	16,177	26.5
ADL, bed/chair: can't	58,645	9,927	16.9	10,322	16.9
ADL, feed self: can	58,707	47,036	80.1	48,897	80.1
ADL, feed self: diff	58,707	8,118	13.8	8,415	13.8
ADL, feed self: can't	58,707	3,553	6.1	3,715	6.1
ADL, wash: can	58,773	21,068	35.8	21,968	36.0
ADL, wash: diff	58,773	15,837	26.9	16,473	27.0
ADL, wash: can't	58,773	21,868	37.2	22,585	37.0
ADL, un/dress: can	58,679	28,247	48.1	29,530	48.4
ADL, un/dress: diff	58,679	15,985	27.2	16,555	27.1
ADL, un/dress: can't	58,679	14,447	24.6	14,941	24.5
ADL, WC/toilet: can	58,710	38,967	66.4	40,587	66.5
ADL, WC/toilet: diff	58,710	10,113	17.2	10,444	17.1
ADL, WC/toilet: can't	58,710	9,630	16.4	9,995	16.4
ADL, face/hands: can	58,865	44,204	75.1	45,888	75.2
ADL, face/hands: diff	58,865	8,232	14.0	8,487	13.9
ADL, face/hands: can't	58,865	6,429	10.9	6,651	10.9
IADL, finance: can	58,344	16,404	28.1	17,446	28.6
IADL, finance: diff	58,344	10,646	18.2	11,171	18.3
IADL, finance: can't	58,344	31,294	53.6	32,410	53.1
SPH: very good	57,599	6,026	10.5	6,336	10.4
SPH: good	57,599	14,428	25.0	15,246	25.0
SPH: fair	57,599	25,697	44.6	27,297	44.7
SPH: bad	57,599	8,673	15.1	9,211	15.1
SPH: very bad	57,599	2,775	4.8	2,936	4.8
pain: none	58,509	17,182	29.4	17,901	29.3
pain: moderate	58,509	31,676	54.1	33,095	54.2

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
pain: extreme	58,509	9,651	16.5	10,030	16.4
anx/dep: none	57,843	27,325	47.2	28,925	47.4
anx/dep: moderate	57,843	26,240	45.4	27,624	45.3
anx/dep: extreme	57,843	4,278	7.4	4,477	7.3
SPHD: meets needs	58,481	30,572	52.3	31,883	52.2
SPHD: most needs	58,481	19,717	33.7	20,601	33.8
SPHD: some needs	58,481	6,699	11.5	6,985	11.4
SPHD: inappropriate	58,481	1,493	2.6	1,557	2.6
PH, in h'hold	57,291	23,051	40.2	24,527	40.2
PH, outside h'hold	57,291	27,872	48.6	29,836	48.9
PH, none	57,291	11,486	20.0	12,481	20.5
No assistance	57,004	18,229	32.0	20,036	32.8
Proxy	57,004	4,161	7.3	4,353	7.1
Assistance	57,004	34,614	60.7	36,637	60.0
Assist, in h'hold	56,690	8,801	15.5	9,374	15.4
Assist, out h'hold	56,690	15,545	27.4	16,581	27.2
Assist, care worker	56,690	8,780	15.5	9,361	15.3
Assist, read	57,004	22,515	39.5	23,761	38.9
Assist, translate	57,004	8,586	15.1	8,952	14.7
Assist, write	57,004	20,248	35.5	21,078	34.5
Assist, talk through	57,004	14,332	25.1	15,187	24.9
Add, own money	56,135	16,717	29.8	18,413	30.2
Add, family pays	56,135	4,052	7.2	4,427	7.3
Add, none	56,135	36,012	64.2	39,383	64.5

Table 56: Comparison of distributional statistics for variables included to improve the MAR assumption, on casewise-deleted and multiply-imputed samples for the respondent sample

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
Nursing home	60,493	2,619	4.3	2,623	4.3
Residential care home	61,000	9,769	16.0	9,770	16.0
Home care	60,469	21,848	36.1	21,956	36.0
Day care	60,096	8,599	14.3	8,741	14.3
Meals	57,974	3,079	5.3	3,167	5.2
Direct Payment	59,681	5,392	9.0	5,474	9.0
Professional support	57,464	8,214	14.3	8,442	13.8
Short-term residential	59,543	805	1.4	825	1.4
Equipment	58,983	13,412	22.7	13,622	22.3
Other services	58,375	4,887	8.4	5,036	8.3
Ethnicity: white	60,166	55,002	91.4	55,776	91.4
Ethnicity: mixed	60,166	328	0.5	333	0.5
Ethnicity: black	60,166	2,319	3.9	2,356	3.9
Ethnicity: Asian	60,166	1,961	3.3	1,995	3.3
Ethnicity: other	60,166	556	0.9	565	0.9
QoL: so good	59,669	5,398	9.0	5,526	9.1
QoL: very good	59,669	13,954	23.4	14,278	23.4
QoL: good	59,669	16,264	27.3	16,623	27.2
QoL: alright	59,669	18,165	30.4	18,568	30.4
QoL: bad	59,669	3,997	6.7	4,097	6.7
QoL: very bad	59,669	1,228	2.1	1,257	2.1
QoL: so bad	59,669	663	1.1	678	1.1
SP help effect: better	57,723	32,585	56.5	34,299	56.2
SP help effect: none	57,723	17,430	30.2	18,449	30.2
SP help effect: sl worse	57,723	6,745	11.7	7,235	11.9
SP help effect: worse	57,723	963	1.7	1,043	1.7
Services help, control	55,500	30,522	55.0	32,947	54.0
Services help, pers. care	55,500	36,585	65.9	38,883	63.7
Services help, meals	55,500	27,745	50.0	29,434	48.2
Services help, social	55,500	25,175	45.4	26,845	44.0
Services help, do things	55,500	22,775	41.0	24,291	39.8
Services help, safe	55,500	31,630	57.0	33,728	55.3
Services help, home	55,500	29,602	53.3	31,757	52.0
Services help, other	55,500	8,949	16.1	10,034	16.4
Unsafe talk, family	58,179	43,264	74.4	45,331	74.3
Unsafe talk, friend	58,179	12,225	21.0	12,795	21.0
Unsafe talk, care worker	58,179	21,444	36.9	22,290	36.5
Unsafe talk, manager	58,179	12,434	21.4	12,920	21.2
Unsafe talk, soc. worker	58,179	9,997	17.2	10,401	17.0
Unsafe talk, other person	58,179	5,113	8.8	5,364	8.8

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
Unsafe talk, no-one	58,179	1,360	2.3	1,481	2.4
Unsafe talk, DK	58,179	1,914	3.3	2,084	3.4
Outdoor access: all places	57,859	17,669	30.5	18,540	30.4
Outdoor access: difficult	57,859	15,714	27.2	16,628	27.2
Outdoor access: unable	57,859	12,209	21.1	12,917	21.2
Outdoor access: not leave	57,859	12,267	21.2	12,941	21.2

Checking sufficiency of the number of imputations

Following imputation, I carried out checks to ensure the number of imputations was sufficient for reporting the outcome indicators and key variables used in the adjustment models. As well as examining the efficiency of the estimates, using the fraction of missing information (FMI), I also examined the reproducibility of the estimates, by looking at the Monte Carlo error of the estimates as proposed by White et al (2011). I estimate the FMI as follows:

$$FMI = \frac{B}{B+W},$$

where B and W, refer to between- and within-imputations variance respectively. The Monte Carlo error for each estimate is defined as the standard deviation of the estimate across repeated runs of the same imputation. For a 1-dimensional parameter estimate, the Monte Carlo error is calculated as √B/m, where m is the number of imputations. Monte Carlo errors for other statistics may be computed using a jackknife procedure. All of these statistics are automatically generated by using the mcerror, dftable and vartable options in Stata alongside the mi estimate command. I calculated Monte Carlo errors and the FMI for the mean of each imputed variable and for the estimates from each specification of the adjustment models.

Sufficiency of the number of imputations for reporting mean statistics

Table 57 shows the estimates of the Monte Carlo error (MC error) for the mean for each of the outcome variables for the respondent sample. Table 58 breaks down the variance into its between- and within-imputation components and provides estimates of the FMI and the relative efficiency of the estimates. The statistics in both tables imply that 20 imputations are sufficient for reporting sample means for the outcome indicators. Specifically, the MC error of the mean is less than ten per cent of the standard error of the mean, the relative efficiency of 20 imputations compared to an infinite number of imputations is over 99% and the FMI is <0.2. The smallest degrees of freedom and largest increase in the SE are found for the

satisfaction and information indicators, but statistics for these suggest that 20 imputations is more than sufficient. Results are similar for the three subsamples used for the analysis (i.e. older people receiving support in their own homes, younger people receiving support in their own homes, and care home residents), with statistics implying that 20 imputations are more than sufficient for reporting means associated with these variables (not shown).

Table 57: Monte Carlo estimates of the mean for all outcome indicators and effect of imputation on the standard error of the mean

Variable	MC error of mean	MC error of mean/SE *100	Degrees of freedom	% increase in SE
ASCOT score	0.000	2.299	43,267	0.56
ASCOF SCRQoL	0.000	2.180	45,887	0.50
Satisfaction indicator	0.000	4.863	6,792	2.58
Control indicator	0.000	3.340	21,873	1.19
Safety indicator	0.000	3.933	13,793	1.67
Information indicator	0.000	4.655	7,570	2.36

Table 58: Imputation variance and efficiency associated with the mean for each outcome indicator

	Imputation variance					Relative
Variable	Within	Between	Total	RVI	FMI	efficiency
ASCOT score	0.000	0.000	0.000	0.011	0.011	0.999
ASCOF SCRQoL	0.000	0.000	0.000	0.010	0.010	1.000
Satisfaction indicator	0.000	0.000	0.000	0.052	0.050	0.998
Control indicator	0.000	0.000	0.000	0.024	0.024	0.999
Safety indicator	0.000	0.000	0.000	0.034	0.033	0.998
Information indicator	0.000	0.000	0.000	0.048	0.046	0.998

Sufficiency of the number of imputations for reporting case-mix adjustment models

I also explored the sufficiency of the imputations for all the case-mix adjustment models, by examining the MC error for the mean for each of the model estimates. I also calculated estimates of the FMI and the relative efficiency of the estimates. In all cases the statistics imply that 20 imputations are sufficient for reporting estimates for the model parameters. The MC error of the mean for each model parameter was less than ten per cent of the standard error of the mean, the relative efficiency of 20 imputations compared to an infinite number of imputations was over 99% and the FMI was <0.2.

Appendix 3: Cluster-Specific Nonignorable Nonresponse Checks

Table 59 shows the correlation (r) between CASSR level response rates and PI scores for all the ASCOF survey-based PIs. The starred correlation coefficients indicate that for some of the PI cluster-specific nonignorable nonresponse may be a problem.

Table 59: Correlation between CASSR response rates and PI scores

	SCRQoL	Satisfaction	Control	Safety	Information
Obs.	149	149	149	149	149
r	0.0685	0.0685	0.1396*	0.1984**	0.2209***
p-value	0.4063	0.4063	0.0894	0.0153	0.0068

* p<.1; ** p<.05; *** p<.01

Appendix 4: Output of the Fixed Effects Multinomial Logistic Models for Inverse Propensity Weighting

The results from the model used for response propensity weighting are shown in Table 60 and are very similar across the CCA and multiply-imputed samples. The biggest difference is for the constant for the blank form equation, which is very much smaller and found to be not significant in the multiply-imputed sample. Pseudo (McFadden's) R-squared for the multiply-imputed sample is estimated following methods proposed by Canette and Marchenko (2013). The proportion correctly classified is estimated following the methods proposed by White et al (2011) and implemented in Stata to generate predicted probabilities, which are then compared to observed classifications to generate the proportion correctly classified. Fit statistics for models estimated on both CCA and multiply-imputed samples are very similar.

Table 60: Multinomial logistic regression models of response propensity, with fixed effects for CASSRs, under two assumptions regarding the missing data mechanism

	Complete case analysis (n=125,753)				Multiple imputation (n=150,672)			
	Blank form†		*Nonrespondent†*		*Blank form†*		*Nonrespondent†*	
	Beta	Robust SE	Beta	Robust SE	Beta	SE	Beta	SE
Fixed part								
Mental health‡	0.472***	0.032	0.478***	0.020	0.463***	0.030	0.468***	0.018
Learning disability‡	-0.616***	0.046	-0.275***	0.023	-0.659***	0.043	-0.279***	0.021
Age, 31-39§	-0.163**	0.061	-0.274***	0.033	-0.224***	0.056	-0.285***	0.030
Age, 40-49§	-0.240***	0.058	-0.460***	0.031	-0.303***	0.053	-0.458***	0.028
Age, 50-64§	-0.267***	0.058	-0.630***	0.031	-0.320***	0.053	-0.621***	0.028
Age, 65-74§	-0.042	0.056	-0.685***	0.031	-0.100	0.051	-0.684***	0.028
Age, 75-84§	0.110*	0.054	-0.683***	0.030	0.057	0.049	-0.683***	0.027
Age, 85 and over§	0.261***	0.054	-0.672***	0.030	0.213***	0.049	-0.673***	0.027
White	0.121**	0.041	-0.209***	0.023	0.128***	0.040	-0.211***	0.022
Count of service types	-0.237***	0.049	-0.100***	0.029	-0.249***	0.046	-0.100***	0.027
Count of service types – sq	0.047***	0.011	0.024***	0.007	0.053***	0.010	0.024***	0.006
Nursing Home	0.226***	0.053	0.297***	0.034	0.180***	0.049	0.265***	0.031
Residential Home	-0.239***	0.043	-0.076**	0.025	-0.268***	0.040	-0.059**	0.023
Home Care	-0.407***	0.039	-0.039	0.022	-0.428***	0.036	-0.049**	0.021
Low-level services	0.078	0.041	0.073**	0.025	0.060	0.039	0.081***	0.023
Direct Payment	-0.399***	0.052	-0.131***	0.028	-0.418***	0.048	-0.142***	0.026
Equipment	0.087*	0.037	-0.050*	0.022	0.051	0.035	-0.057**	0.021
Constant	-1.634***	0.153	1.300***	0.079	0.076	0.105	-2.094***	0.228
CASSR dummy variables	Not shown				Not shown			
Model fit statistics								
Log likelihood			-108,345				n/a	
AIC			217,278				n/a	
BIC			220,142				n/a	
Wald test (χ^2) / F-test			13,167 (d.f. 292)***				46.08 (d.f. 330)***	
McFadden's R^2			0.103				0.098	
Proportion correctly classified			0.554				0.555	

Legend: †Base category: Respondent; ‡ Base category: Physically disabled (or vulnerable person); § Base category: 18-30 years; * p<.05; ** p<.01; *** p<.001

Appendix 5: Distribution of Predicted Response Propensity and the Inverse Propensity Weights

The distribution of outcome probabilities from the response propensity models and the response propensity weights are shown in Table 61 and Table 62, respectively. For the weights, the coefficient of variation is very small and there do not appear to be very large outlier values. For this reason I have simply used the inverse of the propensity score and not trimmed the weights or used adjustment cells (Little 1986, Kalton and Flores-Cervantes 2003). The inverse response propensity score weights are calibrated by multiplying by a post-stratification adjustment to ensure the population totals for each CASSR remain the same after weighting and the average of the weights equals one.

Table 61: Distribution of outcome probabilities from MNL response propensity models

Sample	Outcome	Mean	Std. Dev.	Min	Max
CCA (n=125,753)	P(respondent)	0.408	0.107	0.059	0.936
	P(blank form)	0.110	0.130	0.000	0.870
	P(nonrespondent)	0.481	0.158	0.000	0.869
MI (n=150,672)	P(respondent)	0.405	0.105	0.082	0.939
	P(blank form)	0.108	0.124	0.000	0.866
	P(nonrespondent)	0.487	0.155	0.000	0.877

Key: MNL, multinomial logit; CCA, complete case analysis; MI, multiply-imputed sample

Table 62: Distribution of the response propensity weights from MNL response propensity models

Sample	Mean	Std. Dev.	Min	Max	CV	DEFF
CCA (n=125,750)†	1	0.230	0.440	2.944	0.230	1.053
MI (n=150,672)	1	0.229	0.434	2.695	0.229	1.052

Key: MNL, multinomial logit; CCA, complete case analysis; MI, multiply-imputed sample; †3 further cases are excluded from the weighted sample since they are for an authority with only three cases with full information

Appendix 6: Specification of the Imputation Equations for Multiple Imputation of the Auxiliary Data for Estimation of the Response Propensity Models

The response propensity models were first run on the CCA sample to identify likely variables for the final model and therefore the candidate variables for imputation. Binary variables were imputed using logistic regression (logit); ordinal variables using proportional odds models (ologit); variables with more than two nominal categories were imputed using multinomial logit models (mlogit); and continuous variables were imputed using predictive mean matching (pmm) because of the skewed nature of the distribution of the CASSR-level variable capturing the percentage of the sample removed due to the individual lacking mental capacity. Due to size of the dataset and number of variables with missing data, therefore requiring imputation, a pragmatic approach had to be taken to the specification of the imputation equations to make the imputation computationally feasible.

In general, the imputation equations for each variable include all of the variables used in the response propensity models. However, variables were omitted where the predictive significance of the variable was found to be less than 0.15. In addition, the CASSR-level variables related to the removal of people lacking capacity and population density include only CASSR-level variables and respondent status. This approach was taken to mitigate excessive variability in the imputed cases within CASSRs, which would occur if individual-level variables were used. Since we have to include respondent status, this solution is not ideal and some variability within CASSRs is observed in the imputed dataset.

Ideally a multilevel model would have been used to impute these CASSR-level variables, and indeed all the variables, given random-effects models are applied to the multiply-imputed data. However, the mi impute command in Stata 12 does not support imputation of multilevel models. Alternative recommended approaches such as using dummy variables for CASSRs or imputing each CASSR separately were not practical solutions. Neither option resolves the problem with imputing CASSR-level variables; the number of CASSRs makes the dummy variable solution computationally too slow and leads to collinearity problems where the equations contain other CASSR-level variables; separate imputation for each CASSR sample is infeasible, as some variables were missing for entire CASSRs we therefore needed to borrow information from the entire sample to impute sensible values for the missing variables for these CASSRs. This means that the multiply-

imputed dataset may slightly underestimate relationships between the CASSR-level covariate and the individual-level variables and may underestimate the degree of clustering within CASSRs. However, given each variable is missing only a small fraction of cases the underestimation should not be serious.

The derived variables (i.e. countservices, countservices2, and low-level services) were not included in the imputation equations as their derivation could not be specified straightforwardly in the way required by the software. However, to ensure the plausibility of the MAR assumption, where possible any variables not included in the response propensity model that nevertheless were strongly correlated with the variables to be imputed were also included. For example, sex was included in both the age group and primary client group equations; and additional CASSR-level variables including data on service receipt and gross adult social care expenditure (The Information Centre for Health and Social Care 2012c, The Information Centre for Health and Social Care 2012a), population density (Office for National Statistics 2001) and the size of the eligible population (provided by CASSRs in their data returns) were included in the equation for the percentage of people removed due to lack of capacity.

The CASSR-level variables on service receipt and gross adult social care expenditure were also included in the imputation equations for the service receipt variables. These authority-level data were introduced to allow for shifts in the relationship between service receipt and the other variables by CASSR, particularly important given some service receipt variables (meals, short-term residential care, personal budgets, professional support, equipment and other services) were completely missing within a number of CASSRs. Of the 20 instances where service receipt data was completely missing for all cases within a CASSR, in ten service receipt and expenditure information was available; in a further seven the national data indicated that the service was not provided by that CASSR (i.e. recipients were zero), so all missing values for the relevant service for the respective CASSRs were replaced by a zero (not receiving the service); unfortunately in three cases no service receipt information was available at the national level for that service. Although total gross expenditure was available for these CASSRs the imputed values may be more unreliable for the missing services for these CASSRs.

Due to the missingness in the service receipt and gross adult social care expenditure, these variables also had to be imputed. The continuous variables were imputed using predictive mean matching (pmm), due to skewed nature of the distribution for these variables. Where the additional variables are not used in the response propensity models, the respondent

status was not included in the imputation equations and only CASSR-level variables were included to mitigate excessive CASSR-level variability. All available CASSR-level variables were used.

Appendix 7: Information on Convergence and the Sufficiency of the Multiple Imputation Procedure for the Auxiliary Data for Estimation of the Response Propensity Models

Checking convergence of the multiple chains

Prior to imputation proper, checks for convergence of the multiple chains were carried out, by examining plots summarising the distribution (means and standard deviations) of imputed values against iteration numbers. This is achieved using the chainonly and savetrace options in Stata alongside the mi impute chained command. In addition I examined the behaviour of three chains, each obtained using a different set of initial values, to check the convergence and stability of the algorithm, using the add(3) option instead of chainonly. The variables did not appear to show any trends, and the three chains seem to oscillate around the same point, providing some evidence of convergence of the algorithm. Convergence appears to be achieved after around 20 iterations.

Checking the fit of the imputation models

I compare the imputed values with the observed ones for each of the continuous variables using the user-written command midiagplot (Eddings and Marchenko 2006), to check the fit of the imputation model. Inspection of the graphs produced for each imputed dataset revealed that the predictive mean matching performed well in imputing the skewed distribution of the CASSR-level service receipt variables.

For all other variables, to ensure the imputations produce sensible results I compared the distributional statistics calculated on the CCA sample to those calculated on the multiply-imputed sample for each variable. Table 63 shows the distributional statistics for all the model covariates. The percentages, mean and standard errors are very similar regardless of the sample for all the variables. Although less important as the variables are only used to improve the MAR assumption in the imputation procedure, the distributional statistics for all the other variables included in the imputation models are shown in Table 64. Again, the percentages, mean and standard errors are very similar regardless of the sample for all the variables.

Table 63: Comparison of distributional statistics for response propensity model covariates on casewise-deleted and multiply-imputed samples

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
Physical disability†	150,040	105,685	70.4	106,142	70.4
Mental health user‡	150,040	24,299	16.2	24,428	16.2
Learning disability	150,040	20,056	13.4	20,102	13.3
18-34 years	150,515	13,064	8.7	13,078	8.7
35-44 years	150,515	10,983	7.3	10,996	7.3
45-54 years	150,515	14,057	9.3	14,076	9.3
55-64 years	150,515	14,981	10.0	15,002	10.0
65-74 years	150,515	19,059	12.7	19,080	12.7
75-84 years	150,515	35,473	23.6	35,509	23.6
85 years and over	150,515	42,898	28.5	42,933	28.5
White	148,119	133,235	90.0	135,521	89.9
Count of services	128,056	1.4	0.00	1.4	0.00
Nursing Care home	149,382	7,732	5.2	7,736	5.1
Residential Care home	150,608	23,167	15.4	23,167	15.4
Low level services	128,056	25,888	20.2	30,653	20.3
Direct Payment	147,184	12,314	8.4	12,485	8.3
Short-Term Res. care	146,944	2,205	1.5	2,299	1.5
Equipment	145,129	30,936	21.3	31,596	21.0
Home care	149,128	49,819	33.4	50,101	33.3
Day care	148,148	20,296	13.7	20,673	13.7
Meals	142,605	7,834	5.5	8,158	5.4
Personal budget	138,966	19,593	14.1	21,403	14.2
Professional support	141,272	23,822	16.9	25,215	16.7
Other services	144,037	11,621	8.1	12,182	8.1
% sample removed	150,601	10.9	0.03	10.8	0.06
Pop density	149,584	24.5	0.07	24.3	0.07

Legend: †Also includes ‘vulnerable people’ client group ‡ Also include ‘substance misuse’ client group

Table 64: Comparison of distributional statistics for covariates used to improve MAR assumption, on casewise-deleted and multiply-imputed samples

	Number of complete cases	Imputed		Casewise	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
Male	148,217	55,993	37.2	55,082	37.2
CASSR home care	146,619	2,181	4.42	2,168	4.45
CASSR day care	146,619	919	1.85	909	1.82
CASSR meals	145,313	334	1.53	330	1.15
CASSR short-term res.	146,619	113	0.37	112	0.36
CASSR direct payments	144,617	639	1.39	639	1.28
CASSR prof. support	143,229	1,504	4.90	1,486	3.94
CASSR equipment	146,619	1,614	4.33	1,597	4.24
CASSR other	145,487	591	3.55	585	2.86
CASSR gross expend	149,725	134,964	239.94	134,798	237.40

Checking sufficiency of the number of imputations

Following imputation, checks were also carried out to ensure the number of imputations was sufficient for reporting the mean of each outcome variable and for the estimates from each of the response propensity models, using the methods set out in Appendix 2.

The Monte Carlo errors and the FMI statistics are reported for the MNL response propensity models, with CASSRS entered as fixed effects, in Table 65 to Table 68. In contrast to the estimates for the mean of the imputed variables the error is very small for all variables and is in all cases less than ten per cent of the standard error of beta. The Monte Carlo error for the test statistic, t, and the associated probability, p, (not shown here) are also within the required bounds to ensure reproducibility of the data. The Monte Carlo errors for beta, t and p are proportionally largest for the effect of the count of services variable when the outcome is blank form, and the degrees of freedom are smallest for this estimate (see Table 65). The estimates of the FMI and relative efficiency are also all within acceptable levels, with the largest FMI being 0.079, again for the estimate of the effect of the count of services variable when the outcome is blank form. This suggests that the number of imputations is more than sufficient to ensure reproducibility of the MNL model estimates.

Table 65: Monte Carlo estimates for the beta coefficients for each variable in the MNL response propensity model and the effect of imputation on the standard error of beta (Outcome=Blank form)

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Outcome=Blank form				
Physical disability†	0.001	2.231	1.74E+05	0.530
Mental health user‡	0.000	0.917	6.08E+06	0.090
Learning disability	0.000	0.780	1.20E+07	0.060
18-34 years	0.000	0.911	6.25E+06	0.090
35-44 years	0.001	1.174	2.27E+06	0.150
45-54 years	0.001	1.088	3.07E+06	0.120
55-64 years	0.000	0.981	4.65E+06	0.100
65-74 years	0.001	1.086	3.10E+06	0.120
White	0.002	3.938	1.79E+04	1.670
Count of services	0.003	6.121	3.07E+03	4.180
Count of services - sq	0.001	5.264	5.61E+03	3.040
Nursing care home	0.001	3.017	5.20E+04	0.970
Residential care home	0.002	3.959	1.75E+04	1.690
Home care	0.002	4.860	7.72E+03	2.580
Low level services	0.002	5.436	4.93E+03	3.260
Direct Payment	0.002	3.970	1.73E+04	1.700
Equipment	0.002	5.422	4.98E+03	3.240
Northumberland	0.000	0.213	2.10E+09	0.000
Gateshead	0.000	0.149	8.70E+09	0.000
Newcastle upon Tyne	0.000	0.337	3.30E+08	0.010
North Tyneside	0.000	0.251	1.10E+09	0.010
South Tyneside	0.001	0.363	2.50E+08	0.010
Sunderland	0.001	0.371	2.30E+08	0.010
Hartlepool	0.001	0.361	2.50E+08	0.010
Middlesbrough	0.001	0.369	2.30E+08	0.010
Redcar and Cleveland	0.001	0.423	1.30E+08	0.020
Stockton on Tees	0.001	0.390	1.90E+08	0.020
Durham	0.000	0.305	5.00E+08	0.010
Darlington	0.001	0.328	3.70E+08	0.010
Barnsley	0.000	0.304	5.10E+08	0.010
Doncaster	0.001	0.326	3.80E+08	0.010
Rotherham	0.000	0.266	8.60E+08	0.010
Sheffield	0.001	0.620	2.90E+07	0.040
Bradford	0.000	0.296	5.60E+08	0.010
Calderdale	0.000	0.252	1.10E+09	0.010
Kirklees	0.001	0.373	2.20E+08	0.010

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Leeds	0.000	0.293	5.80E+08	0.010
Wakefield	0.001	0.546	4.90E+07	0.030
East Riding	0.000	0.252	1.10E+09	0.010
Kingston-upon-Hull	0.001	0.096	5.20E+10	0.000
N E Lincolnshire	0.001	0.433	1.20E+08	0.020
North Lincolnshire	0.001	0.402	1.60E+08	0.020
North Yorkshire	0.000	0.291	6.00E+08	0.010
York	0.001	0.443	1.10E+08	0.020
Bolton	0.000	0.329	3.70E+08	0.010
Bury	0.001	0.000	2.50E+23	0.000
Manchester	0.000	0.176	4.50E+09	0.000
Oldham	0.001	0.773	1.20E+07	0.060
Rochdale	0.001	0.369	2.30E+08	0.010
Salford	0.000	0.461	9.50E+07	0.020
Stockport	0.000	0.276	7.40E+08	0.010
Tameside	0.001	0.474	8.50E+07	0.020
Trafford	0.001	0.432	1.20E+08	0.020
Wigan	0.000	0.244	1.20E+09	0.010
Knowsley	0.001	0.338	3.30E+08	0.010
Liverpool	0.001	0.539	5.10E+07	0.030
Sefton	0.001	0.395	1.80E+08	0.020
St Helens	0.000	0.416	1.40E+08	0.020
Wirral	0.001	0.835	8.86E+06	0.070
Halton	0.001	0.407	1.60E+08	0.020
Warrington	0.000	0.277	7.40E+08	0.010
Lancashire	0.000	0.249	1.10E+09	0.010
Blackburn with Darwen	0.000	0.160	6.60E+09	0.000
Blackpool	0.000	0.217	2.00E+09	0.000
Cheshire East	0.000	0.347	3.00E+08	0.010
Cheshire West and Chester	0.001	0.466	9.10E+07	0.020
Warwickshire	0.001	0.383	2.00E+08	0.020
Birmingham	0.000	0.263	9.00E+08	0.010
Coventry	0.000	0.344	3.10E+08	0.010
Dudley	0.000	0.394	1.80E+08	0.020
Sandwell	0.001	0.351	2.80E+08	0.010
Solihull	0.000	0.229	1.60E+09	0.010
Walsall	0.001	0.459	9.80E+07	0.020
Wolverhampton	0.001	0.461	9.50E+07	0.020
Staffordshire	0.001	0.423	1.30E+08	0.020
Stoke-on-Trent	0.001	0.688	1.90E+07	0.050
Herefordshire	0.002	1.201	2.07E+06	0.150

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Worcestershire	0.000	0.281	6.90E+08	0.010
Shropshire	0.001	0.372	2.20E+08	0.010
The Wrekin	0.000	0.339	3.30E+08	0.010
Lincolnshire	0.001	0.375	2.20E+08	0.010
Northamptonshire	0.000	0.326	3.80E+08	0.010
Derbyshire	0.001	0.383	2.00E+08	0.020
Derby	0.001	0.366	2.40E+08	0.010
Leicestershire	0.001	0.283	6.70E+08	0.010
Leicester	0.001	0.604	3.20E+07	0.040
Rutland	0.001	0.338	3.30E+08	0.010
Nottinghamshire	0.001	0.492	7.30E+07	0.030
Nottingham	0.001	0.314	4.40E+08	0.010
Hertfordshire	0.000	0.316	4.30E+08	0.010
Norfolk	0.000	0.319	4.10E+08	0.010
Oxfordshire	0.001	0.513	6.20E+07	0.030
Suffolk	0.001	0.431	1.20E+08	0.020
Luton	0.000	0.288	6.30E+08	0.010
Buckinghamshire	0.001	0.859	7.93E+06	0.080
Milton Keynes	0.001	0.412	1.50E+08	0.020
Bracknell Forest	0.000	0.334	3.50E+08	0.010
West Berkshire	0.001	0.342	3.10E+08	0.010
Reading	0.001	0.395	1.80E+08	0.020
Slough	0.001	0.294	5.80E+08	0.010
Windsor & Maidenhead	0.001	0.405	1.60E+08	0.020
Wokingham	0.000	0.238	1.40E+09	0.010
Essex	0.001	0.715	1.70E+07	0.050
Southend	0.001	0.332	3.60E+08	0.010
Thurrock	0.001	0.390	1.90E+08	0.020
Cambridgeshire	0.000	0.299	5.40E+08	0.010
Peterborough	0.000	0.217	1.90E+09	0.000
Bedford	0.001	0.384	2.00E+08	0.020
Central Bedfordshire	0.000	0.156	7.20E+09	0.000
Camden	0.001	0.093	5.90E+10	0.000
Greenwich	0.001	0.389	1.90E+08	0.020
Hackney	0.001	0.382	2.00E+08	0.020
Hammersmith & Fulham	0.001	0.531	5.40E+07	0.030
Islington	0.001	0.471	8.80E+07	0.020
Kensington & Chelsea	0.001	0.430	1.30E+08	0.020
Lambeth	0.001	0.681	2.00E+07	0.050
Lewisham	0.001	0.331	3.60E+08	0.010
Southwark	0.001	0.668	2.20E+07	0.050

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Tower Hamlets	0.001	0.611	3.10E+07	0.040
Wandsworth	0.001	0.551	4.70E+07	0.030
Westminster	0.001	0.446	1.10E+08	0.020
Barking & Dagenham	0.001	0.446	1.10E+08	0.020
Barnet	0.001	0.359	2.60E+08	0.010
Bexley	0.001	0.275	7.60E+08	0.010
Brent	0.002	1.060	3.42E+06	0.120
Bromley	0.000	0.169	5.30E+09	0.000
Croydon	0.001	0.595	3.40E+07	0.040
Ealing	0.001	0.634	2.70E+07	0.040
Enfield	0.000	0.242	1.30E+09	0.010
Haringey	0.001	0.402	1.60E+08	0.020
Harrow	0.001	0.321	4.10E+08	0.010
Havering	0.001	0.388	1.90E+08	0.020
Hillingdon	0.001	0.563	4.30E+07	0.030
Hounslow	0.001	0.931	5.73E+06	0.090
Kingston-upon-Thames	0.001	0.474	8.50E+07	0.020
Merton	0.001	0.293	5.90E+08	0.010
Newham	0.001	0.105	3.50E+10	0.000
Redbridge	0.001	0.419	1.40E+08	0.020
Sutton	0.000	0.193	3.10E+09	0.000
Waltham Forest	0.001	0.448	1.10E+08	0.020
Isle of Wight	0.000	0.194	3.00E+09	0.000
Surrey	0.001	0.459	9.70E+07	0.020
West Sussex	0.000	0.213	2.10E+09	0.000
Dorset	0.001	0.366	2.40E+08	0.010
Bournemouth	0.001	0.478	8.30E+07	0.020
Poole	0.001	0.465	9.20E+07	0.020
Hampshire	0.001	0.478	8.20E+07	0.020
Portsmouth	0.000	0.264	8.90E+08	0.010
Southampton	0.001	0.736	1.50E+07	0.060
East Sussex	0.000	0.232	1.50E+09	0.010
Brighton & Hove	0.000	0.218	1.90E+09	0.000
Wiltshire	0.001	0.441	1.10E+08	0.020
Swindon	0.001	0.289	6.20E+08	0.010
Kent	0.001	0.404	1.60E+08	0.020
Medway Towns	0.001	0.000	4.30E+23	0.000
Cornwall	0.001	0.508	6.50E+07	0.030
Gloucestershire	0.001	0.446	1.10E+08	0.020
Somerset	0.001	0.437	1.20E+08	0.020
Bath & N E Somerset	0.000	0.242	1.20E+09	0.010

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Bristol	0.001	0.388	1.90E+08	0.020
North Somerset	0.000	0.210	2.20E+09	0.000
South Gloucestershire	0.001	0.453	1.00E+08	0.020
Devon	0.000	0.393	1.80E+08	0.020
Plymouth	0.000	0.206	2.40E+09	0.000
Torbay	0.001	0.529	5.50E+07	0.030
Constant	0.003	2.946	5.72E+04	0.920

†Also includes ‘vulnerable people’ client group ‡ Also include ‘substance misuse’ client group

Table 66: Monte Carlo estimates for the beta coefficients for each variable in the MNL response propensity model and the effect of imputation on the standard error of beta (Outcome=Nonrespondent)

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Outcome=Nonrespondent				
Physical disability†	0.000	1.741	4.69E+05	0.320
Mental health user‡	0.000	1.110	2.84E+06	0.130
Learning disability	0.000	1.242	1.81E+06	0.160
18-34 years	0.000	1.052	3.51E+06	0.120
35-44 years	0.000	1.086	3.10E+06	0.120
45-54 years	0.000	1.041	3.66E+06	0.110
55-64 years	0.000	1.121	2.73E+06	0.130
65-74 years	0.000	1.263	1.69E+06	0.170
White	0.001	3.428	3.12E+04	1.260
Count of services	0.001	4.770	8.32E+03	2.480
Count of services - sq	0.000	4.357	1.20E+04	2.060
Nursing care home	0.001	2.186	1.89E+05	0.510
Residential care home	0.001	3.147	4.39E+04	1.060
Home care	0.001	3.666	2.39E+04	1.440
Low level services	0.001	4.678	8.99E+03	2.380
Direct Payment	0.001	3.068	4.86E+04	1.000
Equipment	0.001	4.671	9.05E+03	2.370
Northumberland	0.000	0.122	1.90E+10	0.000
Gateshead	0.000	0.143	1.00E+10	0.000
Newcastle upon Tyne	0.000	0.150	8.40E+09	0.000
North Tyneside	0.000	0.143	1.00E+10	0.000
South Tyneside	0.000	0.167	5.60E+09	0.000
Sunderland	0.000	0.158	6.90E+09	0.000
Hartlepool	0.000	0.161	6.40E+09	0.000
Middlesbrough	0.000	0.189	3.40E+09	0.000
Redcar and Cleveland	0.000	0.170	5.10E+09	0.000
Stockton on Tees	0.000	0.193	3.10E+09	0.000
Durham	0.000	0.000	2.90E+26	0.000
Darlington	0.000	0.170	5.20E+09	0.000
Barnsley	0.000	0.126	1.70E+10	0.000
Doncaster	0.000	0.195	3.00E+09	0.000
Rotherham	0.000	0.104	3.70E+10	0.000
Sheffield	0.000	0.203	2.50E+09	0.000
Bradford	0.000	0.136	1.30E+10	0.000
Calderdale	0.000	0.116	2.40E+10	0.000
Kirklees	0.000	0.153	7.90E+09	0.000

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Leeds	0.000	0.131	1.50E+10	0.000
Wakefield	0.000	0.209	2.20E+09	0.000
East Riding	0.000	0.125	1.80E+10	0.000
Kingston-upon-Hull	0.000	0.192	3.20E+09	0.000
N E Lincolnshire	0.000	0.168	5.40E+09	0.000
North Lincolnshire	0.000	0.158	7.00E+09	0.000
North Yorkshire	0.000	0.104	3.70E+10	0.000
York	0.000	0.158	6.90E+09	0.000
Bolton	0.000	0.138	1.20E+10	0.000
Bury	0.000	0.179	4.20E+09	0.000
Manchester	0.000	0.131	1.50E+10	0.000
Oldham	0.001	0.220	1.80E+09	0.010
Rochdale	0.000	0.160	6.60E+09	0.000
Salford	0.000	0.150	8.60E+09	0.000
Stockport	0.000	0.118	2.20E+10	0.000
Tameside	0.000	0.175	4.60E+09	0.000
Trafford	0.000	0.185	3.70E+09	0.000
Wigan	0.000	0.135	1.30E+10	0.000
Knowsley	0.000	0.153	7.80E+09	0.000
Liverpool	0.000	0.204	2.50E+09	0.000
Sefton	0.000	0.171	5.10E+09	0.000
St Helens	0.000	0.170	5.20E+09	0.000
Wirral	0.001	0.251	1.10E+09	0.010
Halton	0.000	0.155	7.40E+09	0.000
Warrington	0.000	0.144	1.00E+10	0.000
Lancashire	0.000	0.127	1.70E+10	0.000
Blackburn with Darwen	0.000	0.124	1.80E+10	0.000
Blackpool	0.000	0.114	2.50E+10	0.000
Cheshire East	0.000	0.000	3.20E+26	0.000
Cheshire West and Chester	0.000	0.171	5.10E+09	0.000
Warwickshire	0.000	0.155	7.50E+09	0.000
Birmingham	0.000	0.154	7.70E+09	0.000
Coventry	0.000	0.148	8.90E+09	0.000
Dudley	0.000	0.167	5.50E+09	0.000
Sandwell	0.000	0.158	7.00E+09	0.000
Solihull	0.000	0.108	3.10E+10	0.000
Walsall	0.000	0.188	3.40E+09	0.000
Wolverhampton	0.000	0.194	3.00E+09	0.000
Staffordshire	0.000	0.168	5.40E+09	0.000
Stoke-on-Trent	0.001	0.239	1.30E+09	0.010
Herefordshire	0.001	0.268	8.30E+08	0.010

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Worcestershire	0.000	0.112	2.70E+10	0.000
Shropshire	0.000	0.164	5.90E+09	0.000
The Wrekin	0.000	0.151	8.40E+09	0.000
Lincolnshire	0.000	0.170	5.20E+09	0.000
Northamptonshire	0.000	0.142	1.00E+10	0.000
Derbyshire	0.000	0.172	5.00E+09	0.000
Derby	0.000	0.182	3.90E+09	0.000
Leicestershire	0.000	0.174	4.70E+09	0.000
Leicester	0.001	0.220	1.80E+09	0.010
Rutland	0.000	0.128	1.60E+10	0.000
Nottinghamshire	0.000	0.162	6.30E+09	0.000
Nottingham	0.000	0.170	5.20E+09	0.000
Hertfordshire	0.000	0.131	1.50E+10	0.000
Norfolk	0.000	0.134	1.30E+10	0.000
Oxfordshire	0.000	0.183	3.90E+09	0.000
Suffolk	0.000	0.148	8.90E+09	0.000
Luton	0.000	0.148	9.00E+09	0.000
Buckinghamshire	0.000	0.197	2.90E+09	0.000
Milton Keynes	0.000	0.148	8.90E+09	0.000
Bracknell Forest	0.000	0.145	9.80E+09	0.000
West Berkshire	0.000	0.153	7.80E+09	0.000
Reading	0.000	0.162	6.20E+09	0.000
Slough	0.000	0.204	2.50E+09	0.000
Windsor & Maidenhead	0.000	0.174	4.70E+09	0.000
Wokingham	0.000	0.113	2.70E+10	0.000
Essex	0.000	0.195	3.00E+09	0.000
Southend	0.000	0.163	6.20E+09	0.000
Thurrock	0.000	0.197	2.80E+09	0.000
Cambridgeshire	0.000	0.126	1.70E+10	0.000
Peterborough	0.000	0.094	5.60E+10	0.000
Bedford	0.000	0.176	4.50E+09	0.000
Central Bedfordshire	0.000	0.104	3.70E+10	0.000
Camden	0.001	0.263	9.10E+08	0.010
Greenwich	0.000	0.197	2.80E+09	0.000
Hackney	0.000	0.208	2.30E+09	0.000
Hammersmith & Fulham	0.001	0.217	2.00E+09	0.000
Islington	0.000	0.194	3.00E+09	0.000
Kensington & Chelsea	0.000	0.172	5.00E+09	0.000
Lambeth	0.001	0.251	1.10E+09	0.010
Lewisham	0.001	0.214	2.10E+09	0.000
Southwark	0.001	0.231	1.50E+09	0.010

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Tower Hamlets	0.001	0.274	7.60E+08	0.010
Wandsworth	0.001	0.225	1.70E+09	0.010
Westminster	0.000	0.180	4.10E+09	0.000
Barking & Dagenham	0.000	0.198	2.80E+09	0.000
Barnet	0.000	0.167	5.60E+09	0.000
Bexley	0.001	0.238	1.30E+09	0.010
Brent	0.001	0.351	2.80E+08	0.010
Bromley	0.000	0.068	2.00E+11	0.000
Croydon	0.001	0.225	1.70E+09	0.010
Ealing	0.001	0.236	1.40E+09	0.010
Enfield	0.000	0.156	7.20E+09	0.000
Haringey	0.001	0.220	1.80E+09	0.010
Harrow	0.000	0.171	5.00E+09	0.000
Havering	0.000	0.163	6.10E+09	0.000
Hillingdon	0.000	0.199	2.80E+09	0.000
Hounslow	0.001	0.000	1.50E+24	0.000
Kingston-upon-Thames	0.000	0.184	3.70E+09	0.000
Merton	0.001	0.237	1.40E+09	0.010
Newham	0.001	0.257	9.90E+08	0.010
Redbridge	0.001	0.215	2.00E+09	0.000
Sutton	0.000	0.126	1.70E+10	0.000
Waltham Forest	0.000	0.203	2.60E+09	0.000
Isle of Wight	0.000	0.096	5.10E+10	0.000
Surrey	0.000	0.199	2.80E+09	0.000
West Sussex	0.000	0.103	3.90E+10	0.000
Dorset	0.001	0.244	1.20E+09	0.010
Bournemouth	0.000	0.208	2.30E+09	0.000
Poole	0.000	0.135	1.30E+10	0.000
Hampshire	0.001	0.220	1.80E+09	0.010
Portsmouth	0.000	0.142	1.10E+10	0.000
Southampton	0.000	0.213	2.10E+09	0.000
East Sussex	0.000	0.115	2.50E+10	0.000
Brighton & Hove	0.000	0.198	2.80E+09	0.000
Wiltshire	0.000	0.184	3.80E+09	0.000
Swindon	0.000	0.157	7.20E+09	0.000
Kent	0.000	0.181	4.10E+09	0.000
Medway Towns	0.000	0.175	4.60E+09	0.000
Cornwall	0.000	0.127	1.60E+10	0.000
Gloucestershire	0.000	0.110	2.90E+10	0.000
Somerset	0.001	0.223	1.70E+09	0.010
Bath & N E Somerset	0.000	0.102	4.00E+10	0.000

Variable	MC error of Beta	MC error of Beta/SE *100	Degrees of freedom	% increase in SE
Bristol	0.000	0.201	2.60E+09	0.000
North Somerset	0.000	0.126	1.70E+10	0.000
South Gloucestershire	0.001	0.259	9.50E+08	0.010
Devon	0.000	0.187	3.50E+09	0.000
Plymouth	0.000	0.117	2.30E+10	0.000
Torbay	0.000	0.180	4.10E+09	0.000
Constant	0.002	0.665	2.20E+07	0.050

†Also includes ‘vulnerable people’ client group ‡ Also include ‘substance misuse’ client group

Table 67: Imputation variance and efficiency associated with the estimate of beta for each variable in the MNL response propensity model (Outcome=Blank form)

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
Outcome=Blank form						
Physical disability†	0.001	0.000	0.001	0.011	0.010	0.999
Mental health user‡	0.002	0.000	0.002	0.002	0.002	1.000
Learning disability	0.003	0.000	0.003	0.001	0.001	1.000
18-34 years	0.003	0.000	0.003	0.002	0.002	1.000
35-44 years	0.003	0.000	0.003	0.003	0.003	1.000
45-54 years	0.003	0.000	0.003	0.002	0.002	1.000
55-64 years	0.002	0.000	0.002	0.002	0.002	1.000
65-74 years	0.002	0.000	0.002	0.002	0.002	1.000
White	0.002	0.000	0.002	0.034	0.033	0.998
Count of services	0.002	0.000	0.002	0.085	0.079	0.996
Count of services - sq	0.000	0.000	0.000	0.062	0.059	0.997
Nursing care home	0.002	0.000	0.002	0.019	0.019	0.999
Residential care home	0.002	0.000	0.002	0.034	0.033	0.998
Home care	0.001	0.000	0.001	0.052	0.050	0.998
Low level services	0.001	0.000	0.002	0.066	0.062	0.997
Direct Payment	0.002	0.000	0.002	0.034	0.033	0.998
Equipment	0.001	0.000	0.001	0.066	0.062	0.997
Northumberland	0.022	0.000	0.022	0.000	0.000	1.000
Gateshead	0.093	0.000	0.093	0.000	0.000	1.000
Newcastle upon Tyne	0.016	0.000	0.016	0.000	0.000	1.000
North Tyneside	0.021	0.000	0.021	0.000	0.000	1.000
South Tyneside	0.024	0.000	0.024	0.000	0.000	1.000
Sunderland	0.019	0.000	0.019	0.000	0.000	1.000
Hartlepool	0.021	0.000	0.021	0.000	0.000	1.000
Middlesbrough	0.039	0.000	0.039	0.000	0.000	1.000
Redcar and Cleveland	0.016	0.000	0.016	0.000	0.000	1.000
Stockton on Tees	0.026	0.000	0.026	0.000	0.000	1.000
Durham	0.011	0.000	0.011	0.000	0.000	1.000
Darlington	0.031	0.000	0.031	0.000	0.000	1.000
Barnsley	0.020	0.000	0.020	0.000	0.000	1.000
Doncaster	0.046	0.000	0.046	0.000	0.000	1.000
Rotherham	0.026	0.000	0.026	0.000	0.000	1.000
Sheffield	0.016	0.000	0.016	0.001	0.001	1.000
Bradford	0.018	0.000	0.018	0.000	0.000	1.000
Calderdale	0.020	0.000	0.020	0.000	0.000	1.000
Kirklees	0.018	0.000	0.018	0.000	0.000	1.000
Leeds	0.017	0.000	0.017	0.000	0.000	1.000
Wakefield	0.016	0.000	0.016	0.001	0.001	1.000

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
East Riding	0.021	0.000	0.021	0.000	0.000	1.000
Kingston-upon-Hull	0.508	0.000	0.508	0.000	0.000	1.000
N E Lincolnshire	0.021	0.000	0.021	0.000	0.000	1.000
North Lincolnshire	0.016	0.000	0.016	0.000	0.000	1.000
North Yorkshire	0.015	0.000	0.015	0.000	0.000	1.000
York	0.016	0.000	0.016	0.000	0.000	1.000
Bolton	0.017	0.000	0.017	0.000	0.000	1.000
Bury	1100000	0.000	1100000	0.000	0.000	1.000
Manchester	0.054	0.000	0.054	0.000	0.000	1.000
Oldham	0.017	0.000	0.017	0.001	0.001	1.000
Rochdale	0.019	0.000	0.019	0.000	0.000	1.000
Salford	0.011	0.000	0.011	0.000	0.000	1.000
Stockport	0.017	0.000	0.017	0.000	0.000	1.000
Tameside	0.016	0.000	0.016	0.000	0.000	1.000
Trafford	0.016	0.000	0.016	0.000	0.000	1.000
Wigan	0.016	0.000	0.016	0.000	0.000	1.000
Knowsley	0.025	0.000	0.025	0.000	0.000	1.000
Liverpool	0.017	0.000	0.017	0.001	0.001	1.000
Sefton	0.018	0.000	0.018	0.000	0.000	1.000
St Helens	0.014	0.000	0.014	0.000	0.000	1.000
Wirral	0.019	0.000	0.019	0.001	0.001	1.000
Halton	0.020	0.000	0.020	0.000	0.000	1.000
Warrington	0.018	0.000	0.018	0.000	0.000	1.000
Lancashire	0.022	0.000	0.023	0.000	0.000	1.000
Blackburn with Darwen	0.033	0.000	0.033	0.000	0.000	1.000
Blackpool	0.023	0.000	0.023	0.000	0.000	1.000
Cheshire East	0.010	0.000	0.010	0.000	0.000	1.000
Cheshire W. and Chester	0.012	0.000	0.012	0.000	0.000	1.000
Warwickshire	0.021	0.000	0.021	0.000	0.000	1.000
Birmingham	0.033	0.000	0.033	0.000	0.000	1.000
Coventry	0.019	0.000	0.019	0.000	0.000	1.000
Dudley	0.011	0.000	0.011	0.000	0.000	1.000
Sandwell	0.021	0.000	0.021	0.000	0.000	1.000
Solihull	0.018	0.000	0.018	0.000	0.000	1.000
Walsall	0.024	0.000	0.024	0.000	0.000	1.000
Wolverhampton	0.022	0.000	0.022	0.000	0.000	1.000
Staffordshire	0.017	0.000	0.017	0.000	0.000	1.000
Stoke-on-Trent	0.020	0.000	0.020	0.001	0.001	1.000
Herefordshire	0.018	0.000	0.018	0.003	0.003	1.000
Worcestershire	0.018	0.000	0.018	0.000	0.000	1.000
Shropshire	0.020	0.000	0.020	0.000	0.000	1.000
The Wrekin	0.019	0.000	0.019	0.000	0.000	1.000

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
Lincolnshire	0.024	0.000	0.024	0.000	0.000	1.000
Northamptonshire	0.018	0.000	0.018	0.000	0.000	1.000
Derbyshire	0.050	0.000	0.050	0.000	0.000	1.000
Derby	0.034	0.000	0.034	0.000	0.000	1.000
Leicestershire	0.045	0.000	0.045	0.000	0.000	1.000
Leicester	0.020	0.000	0.020	0.001	0.001	1.000
Rutland	0.028	0.000	0.028	0.000	0.000	1.000
Nottinghamshire	0.017	0.000	0.017	0.001	0.001	1.000
Nottingham	0.038	0.000	0.038	0.000	0.000	1.000
Hertfordshire	0.020	0.000	0.020	0.000	0.000	1.000
Norfolk	0.022	0.000	0.022	0.000	0.000	1.000
Oxfordshire	0.013	0.000	0.013	0.001	0.001	1.000
Suffolk	0.014	0.000	0.014	0.000	0.000	1.000
Luton	0.019	0.000	0.019	0.000	0.000	1.000
Buckinghamshire	0.022	0.000	0.022	0.002	0.002	1.000
Milton Keynes	0.019	0.000	0.019	0.000	0.000	1.000
Bracknell Forest	0.020	0.000	0.020	0.000	0.000	1.000
West Berkshire	0.023	0.000	0.023	0.000	0.000	1.000
Reading	0.018	0.000	0.018	0.000	0.000	1.000
Slough	0.072	0.000	0.072	0.000	0.000	1.000
Windsor & Maidenhead	0.027	0.000	0.027	0.000	0.000	1.000
Wokingham	0.027	0.000	0.027	0.000	0.000	1.000
Essex	0.017	0.000	0.017	0.001	0.001	1.000
Southend	0.030	0.000	0.030	0.000	0.000	1.000
Thurrock	0.026	0.000	0.026	0.000	0.000	1.000
Cambridgeshire	0.015	0.000	0.015	0.000	0.000	1.000
Peterborough	0.019	0.000	0.019	0.000	0.000	1.000
Bedford	0.020	0.000	0.020	0.000	0.000	1.000
Central Bedfordshire	0.028	0.000	0.028	0.000	0.000	1.000
Camden	1.009	0.000	1.009	0.000	0.000	1.000
Greenwich	0.031	0.000	0.031	0.000	0.000	1.000
Hackney	0.040	0.000	0.040	0.000	0.000	1.000
Hammersmith & Fulham	0.024	0.000	0.024	0.001	0.001	1.000
Islington	0.022	0.000	0.022	0.000	0.000	1.000
Kensington & Chelsea	0.025	0.000	0.025	0.000	0.000	1.000
Lambeth	0.022	0.000	0.022	0.001	0.001	1.000
Lewisham	0.050	0.000	0.050	0.000	0.000	1.000
Southwark	0.025	0.000	0.025	0.001	0.001	1.000
Tower Hamlets	0.034	0.000	0.034	0.001	0.001	1.000
Wandsworth	0.032	0.000	0.032	0.001	0.001	1.000
Westminster	0.022	0.000	0.022	0.000	0.000	1.000
Barking & Dagenham	0.021	0.000	0.021	0.000	0.000	1.000

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
Barnet	0.022	0.000	0.022	0.000	0.000	1.000
Bexley	0.061	0.000	0.061	0.000	0.000	1.000
Brent	0.023	0.000	0.023	0.002	0.002	1.000
Bromley	0.018	0.000	0.018	0.000	0.000	1.000
Croydon	0.014	0.000	0.014	0.001	0.001	1.000
Ealing	0.019	0.000	0.019	0.001	0.001	1.000
Enfield	0.027	0.000	0.027	0.000	0.000	1.000
Haringey	0.033	0.000	0.033	0.000	0.000	1.000
Harrow	0.031	0.000	0.031	0.000	0.000	1.000
Havering	0.020	0.000	0.020	0.000	0.000	1.000
Hillingdon	0.019	0.000	0.019	0.001	0.001	1.000
Hounslow	0.007	0.000	0.007	0.002	0.002	1.000
Kingston-upon-Thames	0.022	0.000	0.022	0.000	0.000	1.000
Merton	0.109	0.000	0.109	0.000	0.000	1.000
Newham	1.009	0.000	1.009	0.000	0.000	1.000
Redbridge	0.029	0.000	0.029	0.000	0.000	1.000
Sutton	0.022	0.000	0.022	0.000	0.000	1.000
Waltham Forest	0.021	0.000	0.021	0.000	0.000	1.000
Isle of Wight	0.032	0.000	0.032	0.000	0.000	1.000
Surrey	0.026	0.000	0.026	0.000	0.000	1.000
West Sussex	0.031	0.000	0.031	0.000	0.000	1.000
Dorset	0.038	0.000	0.038	0.000	0.000	1.000
Bournemouth	0.019	0.000	0.019	0.000	0.000	1.000
Poole	0.024	0.000	0.024	0.000	0.000	1.000
Hampshire	0.020	0.000	0.020	0.000	0.000	1.000
Portsmouth	0.035	0.000	0.035	0.000	0.000	1.000
Southampton	0.018	0.000	0.018	0.001	0.001	1.000
East Sussex	0.020	0.000	0.020	0.000	0.000	1.000
Brighton & Hove	0.042	0.000	0.042	0.000	0.000	1.000
Wiltshire	0.022	0.000	0.022	0.000	0.000	1.000
Swindon	0.037	0.000	0.037	0.000	0.000	1.000
Kent	0.032	0.000	0.032	0.000	0.000	1.000
Medway Towns	1400000	0.000	1400000	0.000	0.000	1.000
Cornwall	0.011	0.000	0.011	0.001	0.001	1.000
Gloucestershire	0.015	0.000	0.015	0.000	0.000	1.000
Somerset	0.030	0.000	0.030	0.000	0.000	1.000
Bath & N E Somerset	0.020	0.000	0.020	0.000	0.000	1.000
Bristol	0.034	0.000	0.034	0.000	0.000	1.000
North Somerset	0.038	0.000	0.038	0.000	0.000	1.000
South Gloucestershire	0.017	0.000	0.017	0.000	0.000	1.000
Devon	0.015	0.000	0.015	0.000	0.000	1.000
Plymouth	0.015	0.000	0.015	0.000	0.000	1.000

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
Torbay	0.012	0.000	0.012	0.001	0.001	1.000
Constant	0.011	0.000	0.011	0.019	0.018	0.999

†Also includes ‘vulnerable people’ client group ‡ Also include ‘substance misuse’ client group

Table 68: Imputation variance and efficiency associated with the estimate of beta for each variable in the MNL response propensity model (Outcome=Nonrespondent)

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
Outcome=Nonrespondent						
Physical disability†	0.000	0.000	0.000	0.006	0.006	1.000
Mental health user‡	0.000	0.000	0.000	0.003	0.003	1.000
Learning disability	0.001	0.000	0.001	0.003	0.003	1.000
18-34 years	0.001	0.000	0.001	0.002	0.002	1.000
35-44 years	0.001	0.000	0.001	0.002	0.002	1.000
45-54 years	0.001	0.000	0.001	0.002	0.002	1.000
55-64 years	0.001	0.000	0.001	0.003	0.003	1.000
65-74 years	0.001	0.000	0.001	0.003	0.003	1.000
White	0.000	0.000	0.000	0.025	0.025	0.999
Count of services	0.001	0.000	0.001	0.050	0.048	0.998
Count of services - sq	0.000	0.000	0.000	0.042	0.040	0.998
Nursing care home	0.001	0.000	0.001	0.010	0.010	0.999
Residential care home	0.001	0.000	0.001	0.021	0.021	0.999
Home care	0.000	0.000	0.000	0.029	0.028	0.999
Low level services	0.001	0.000	0.001	0.048	0.046	0.998
Direct Payment	0.001	0.000	0.001	0.020	0.020	0.999
Equipment	0.000	0.000	0.000	0.048	0.046	0.998
Northumberland	0.054	0.000	0.054	0.000	0.000	1.000
Gateshead	0.057	0.000	0.057	0.000	0.000	1.000
Newcastle upon Tyne	0.055	0.000	0.055	0.000	0.000	1.000
North Tyneside	0.055	0.000	0.055	0.000	0.000	1.000
South Tyneside	0.056	0.000	0.056	0.000	0.000	1.000
Sunderland	0.057	0.000	0.057	0.000	0.000	1.000
Hartlepool	0.055	0.000	0.055	0.000	0.000	1.000
Middlesbrough	0.057	0.000	0.057	0.000	0.000	1.000
Redcar and Cleveland	0.056	0.000	0.056	0.000	0.000	1.000
Stockton on Tees	0.056	0.000	0.056	0.000	0.000	1.000
Durham	2600000	0.000	2600000	0.000	0.000	1.000
Darlington	0.058	0.000	0.058	0.000	0.000	1.000
Barnsley	0.055	0.000	0.055	0.000	0.000	1.000
Doncaster	0.055	0.000	0.055	0.000	0.000	1.000
Rotherham	0.055	0.000	0.055	0.000	0.000	1.000
Sheffield	0.054	0.000	0.054	0.000	0.000	1.000
Bradford	0.054	0.000	0.054	0.000	0.000	1.000
Calderdale	0.055	0.000	0.055	0.000	0.000	1.000
Kirklees	0.057	0.000	0.057	0.000	0.000	1.000
Leeds	0.054	0.000	0.054	0.000	0.000	1.000
Wakefield	0.055	0.000	0.055	0.000	0.000	1.000

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
East Riding	0.056	0.000	0.056	0.000	0.000	1.000
Kingston-upon-Hull	0.054	0.000	0.054	0.000	0.000	1.000
N E Lincolnshire	0.056	0.000	0.056	0.000	0.000	1.000
North Lincolnshire	0.058	0.000	0.058	0.000	0.000	1.000
North Yorkshire	0.056	0.000	0.056	0.000	0.000	1.000
York	0.055	0.000	0.055	0.000	0.000	1.000
Bolton	0.053	0.000	0.053	0.000	0.000	1.000
Bury	0.054	0.000	0.054	0.000	0.000	1.000
Manchester	0.054	0.000	0.054	0.000	0.000	1.000
Oldham	0.053	0.000	0.053	0.000	0.000	1.000
Rochdale	0.055	0.000	0.055	0.000	0.000	1.000
Salford	0.053	0.000	0.053	0.000	0.000	1.000
Stockport	0.054	0.000	0.054	0.000	0.000	1.000
Tameside	0.056	0.000	0.056	0.000	0.000	1.000
Trafford	0.055	0.000	0.055	0.000	0.000	1.000
Wigan	0.056	0.000	0.056	0.000	0.000	1.000
Knowsley	0.054	0.000	0.054	0.000	0.000	1.000
Liverpool	0.055	0.000	0.055	0.000	0.000	1.000
Sefton	0.055	0.000	0.055	0.000	0.000	1.000
St Helens	0.056	0.000	0.056	0.000	0.000	1.000
Wirral	0.056	0.000	0.056	0.000	0.000	1.000
Halton	0.057	0.000	0.057	0.000	0.000	1.000
Warrington	0.055	0.000	0.055	0.000	0.000	1.000
Lancashire	0.056	0.000	0.056	0.000	0.000	1.000
Blackburn with Darwen	0.054	0.000	0.054	0.000	0.000	1.000
Blackpool	0.055	0.000	0.055	0.000	0.000	1.000
Cheshire East	10000000	0.000	10000000	0.000	0.000	1.000
Cheshire W and Chester	0.053	0.000	0.053	0.000	0.000	1.000
Warwickshire	0.055	0.000	0.055	0.000	0.000	1.000
Birmingham	0.055	0.000	0.055	0.000	0.000	1.000
Coventry	0.055	0.000	0.055	0.000	0.000	1.000
Dudley	0.052	0.000	0.052	0.000	0.000	1.000
Sandwell	0.055	0.000	0.055	0.000	0.000	1.000
Solihull	0.055	0.000	0.055	0.000	0.000	1.000
Walsall	0.056	0.000	0.056	0.000	0.000	1.000
Wolverhampton	0.056	0.000	0.056	0.000	0.000	1.000
Staffordshire	0.055	0.000	0.055	0.000	0.000	1.000
Stoke-on-Trent	0.054	0.000	0.054	0.000	0.000	1.000
Herefordshire	0.057	0.000	0.057	0.000	0.000	1.000
Worcestershire	0.056	0.000	0.056	0.000	0.000	1.000
Shropshire	0.055	0.000	0.055	0.000	0.000	1.000
The Wrekin	0.054	0.000	0.054	0.000	0.000	1.000

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
Lincolnshire	0.056	0.000	0.056	0.000	0.000	1.000
Northamptonshire	0.054	0.000	0.054	0.000	0.000	1.000
Derbyshire	0.057	0.000	0.057	0.000	0.000	1.000
Derby	0.055	0.000	0.055	0.000	0.000	1.000
Leicestershire	0.054	0.000	0.054	0.000	0.000	1.000
Leicester	0.055	0.000	0.055	0.000	0.000	1.000
Rutland	0.061	0.000	0.061	0.000	0.000	1.000
Nottinghamshire	0.055	0.000	0.055	0.000	0.000	1.000
Nottingham	0.055	0.000	0.055	0.000	0.000	1.000
Hertfordshire	0.055	0.000	0.055	0.000	0.000	1.000
Norfolk	0.056	0.000	0.056	0.000	0.000	1.000
Oxfordshire	0.053	0.000	0.053	0.000	0.000	1.000
Suffolk	0.053	0.000	0.053	0.000	0.000	1.000
Luton	0.057	0.000	0.057	0.000	0.000	1.000
Buckinghamshire	0.056	0.000	0.056	0.000	0.000	1.000
Milton Keynes	0.056	0.000	0.056	0.000	0.000	1.000
Bracknell Forest	0.057	0.000	0.057	0.000	0.000	1.000
West Berkshire	0.055	0.000	0.055	0.000	0.000	1.000
Reading	0.056	0.000	0.056	0.000	0.000	1.000
Slough	0.056	0.000	0.056	0.000	0.000	1.000
Windsor & Maidenhead	0.055	0.000	0.055	0.000	0.000	1.000
Wokingham	0.057	0.000	0.057	0.000	0.000	1.000
Essex	0.054	0.000	0.054	0.000	0.000	1.000
Southend	0.055	0.000	0.055	0.000	0.000	1.000
Thurrock	0.057	0.000	0.057	0.000	0.000	1.000
Cambridgeshire	0.054	0.000	0.054	0.000	0.000	1.000
Peterborough	0.056	0.000	0.056	0.000	0.000	1.000
Bedford	0.055	0.000	0.055	0.000	0.000	1.000
Central Bedfordshire	0.055	0.000	0.055	0.000	0.000	1.000
Camden	0.055	0.000	0.055	0.000	0.000	1.000
Greenwich	0.055	0.000	0.055	0.000	0.000	1.000
Hackney	0.056	0.000	0.056	0.000	0.000	1.000
Hammersmith & Fulham	0.055	0.000	0.055	0.000	0.000	1.000
Islington	0.054	0.000	0.054	0.000	0.000	1.000
Kensington & Chelsea	0.054	0.000	0.054	0.000	0.000	1.000
Lambeth	0.056	0.000	0.056	0.000	0.000	1.000
Lewisham	0.055	0.000	0.055	0.000	0.000	1.000
Southwark	0.055	0.000	0.055	0.000	0.000	1.000
Tower Hamlets	0.053	0.000	0.053	0.000	0.000	1.000
Wandsworth	0.055	0.000	0.055	0.000	0.000	1.000
Westminster	0.054	0.000	0.054	0.000	0.000	1.000
Barking & Dagenham	0.057	0.000	0.057	0.000	0.000	1.000

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
Barnet	0.055	0.000	0.055	0.000	0.000	1.000
Bexley	0.055	0.000	0.055	0.000	0.000	1.000
Brent	0.056	0.000	0.056	0.000	0.000	1.000
Bromley	0.055	0.000	0.055	0.000	0.000	1.000
Croydon	0.054	0.000	0.054	0.000	0.000	1.000
Ealing	0.055	0.000	0.055	0.000	0.000	1.000
Enfield	0.055	0.000	0.055	0.000	0.000	1.000
Haringey	0.054	0.000	0.054	0.000	0.000	1.000
Harrow	0.057	0.000	0.057	0.000	0.000	1.000
Havering	0.056	0.000	0.056	0.000	0.000	1.000
Hillingdon	0.057	0.000	0.057	0.000	0.000	1.000
Hounslow	2100000	0.000	2100000	0.000	0.000	1.000
Kingston-upon-Thames	0.058	0.000	0.058	0.000	0.000	1.000
Merton	0.055	0.000	0.055	0.000	0.000	1.000
Newham	0.055	0.000	0.055	0.000	0.000	1.000
Redbridge	0.058	0.000	0.058	0.000	0.000	1.000
Sutton	0.055	0.000	0.055	0.000	0.000	1.000
Waltham Forest	0.055	0.000	0.055	0.000	0.000	1.000
Isle of Wight	0.056	0.000	0.056	0.000	0.000	1.000
Surrey	0.054	0.000	0.054	0.000	0.000	1.000
West Sussex	0.055	0.000	0.055	0.000	0.000	1.000
Dorset	0.056	0.000	0.056	0.000	0.000	1.000
Bournemouth	0.054	0.000	0.054	0.000	0.000	1.000
Poole	0.056	0.000	0.056	0.000	0.000	1.000
Hampshire	0.054	0.000	0.054	0.000	0.000	1.000
Portsmouth	0.055	0.000	0.055	0.000	0.000	1.000
Southampton	0.055	0.000	0.055	0.000	0.000	1.000
East Sussex	0.055	0.000	0.055	0.000	0.000	1.000
Brighton & Hove	0.056	0.000	0.056	0.000	0.000	1.000
Wiltshire	0.055	0.000	0.055	0.000	0.000	1.000
Swindon	0.057	0.000	0.057	0.000	0.000	1.000
Kent	0.056	0.000	0.056	0.000	0.000	1.000
Medway Towns	0.055	0.000	0.055	0.000	0.000	1.000
Cornwall	0.052	0.000	0.052	0.000	0.000	1.000
Gloucestershire	0.123	0.000	0.123	0.000	0.000	1.000
Somerset	0.054	0.000	0.054	0.000	0.000	1.000
Bath & N E Somerset	0.054	0.000	0.054	0.000	0.000	1.000
Bristol	0.057	0.000	0.057	0.000	0.000	1.000
North Somerset	0.057	0.000	0.057	0.000	0.000	1.000
South Gloucestershire	0.056	0.000	0.056	0.000	0.000	1.000
Devon	0.055	0.000	0.055	0.000	0.000	1.000
Plymouth	0.054	0.000	0.054	0.000	0.000	1.000

Variable	Imputation variance			RVI	FMI	Relative efficiency
	Within	Between	Total			
Torbay	0.055	0.000	0.055	0.000	0.000	1.000
Constant	0.052	0.000	0.052	0.001	0.001	1.000

†Also includes ‘vulnerable people’ client group ‡ Also include ‘substance misuse’ client group

Appendix 8: Overdispersion Factors for the Overall Sample

Following Spiegelhalter (2005b), the degree of overdispersion in the data can be estimated by calculating an overdispersion factor, denoted ϕ, as follows, $\phi = \frac{1}{I}\sum z_i^2$, where z is the standardised Pearson residual, and I is the total number of units, in this case 149 CASSRs. Where ϕ is significantly greater than one, there is overdispersion. Table 69 shows that the overdispersion factor for each indicator here is in all cases much less than one, implying that overdispersion is not present for these indicators.

Table 69: Overdispersion factors for PIs

Indicator	CCA	IPW/CCA	CCA/MI	IPW/MI
SCRQoL PI	0.114	0.116	0.120	0.120
Satisfaction PI	0.112	0.102	0.125	0.114
Control PI	0.083	0.074	0.080	0.071
Safety PI	0.048	0.056	0.049	0.057
Information PI	0.052	0.058	0.047	0.054

Appendix 9: Additional Graphs and Tables Showing the Effect of Adjusting for Nonresponse on Performance Assessment

I noted that there is very little change in the width of the control limits, so the status changes and shifts that occur as a result of adjusting for nonresponse are in all likelihood due to a volume or 'mean' outcome effect. Following Gomes et al (2016) it is possible to decompose the observed changes in positions of CASSRs in the funnel plots into the volume and mean outcome effect and this is illustrated for all the indicators in Figure 53 to Figure 57, with graphs a to c showing the effects of MI and graphs d to f showing the effects of IPW. The volume and mean outcome effects are not easily observable on the funnel plots for either MI or IPW.

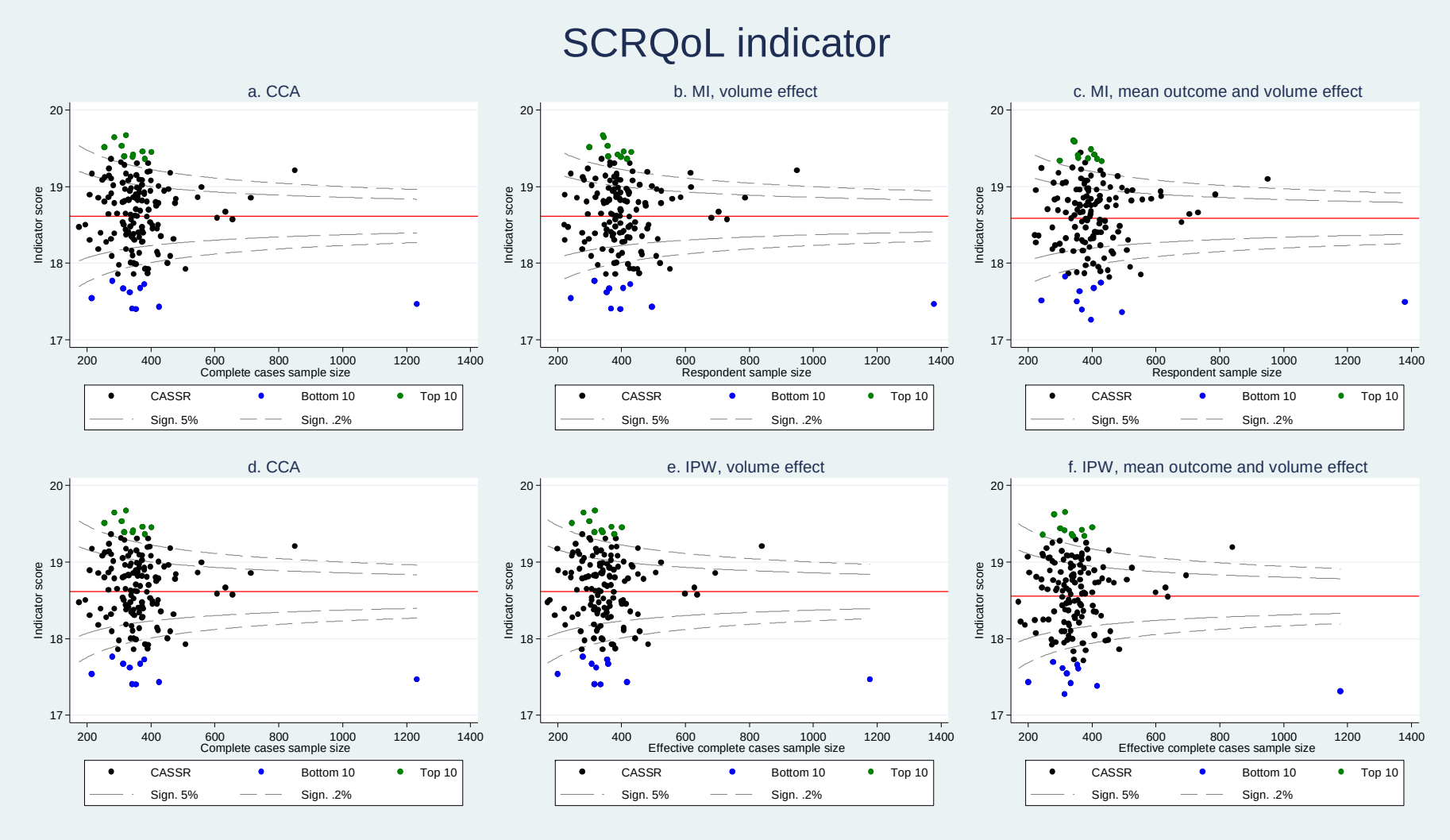

Figure 53: Funnel plots of CASSR scores on the SCRQoL indicator, showing the mean outcome and volume effects from MI and IPW

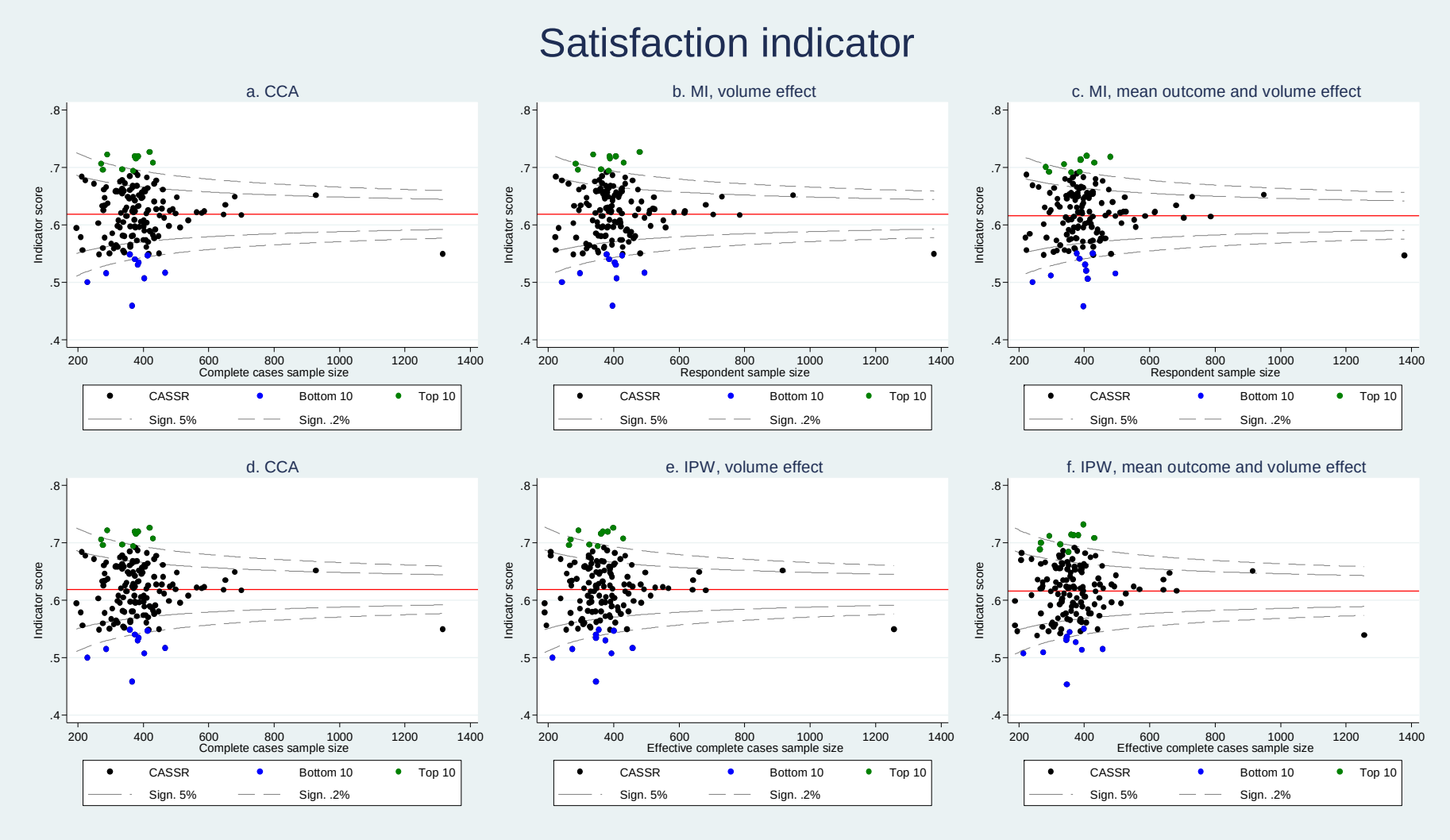

Figure 54: Funnel plots of CASSR scores on the satisfaction indicator, showing the mean outcome and volume effects from MI and IPW

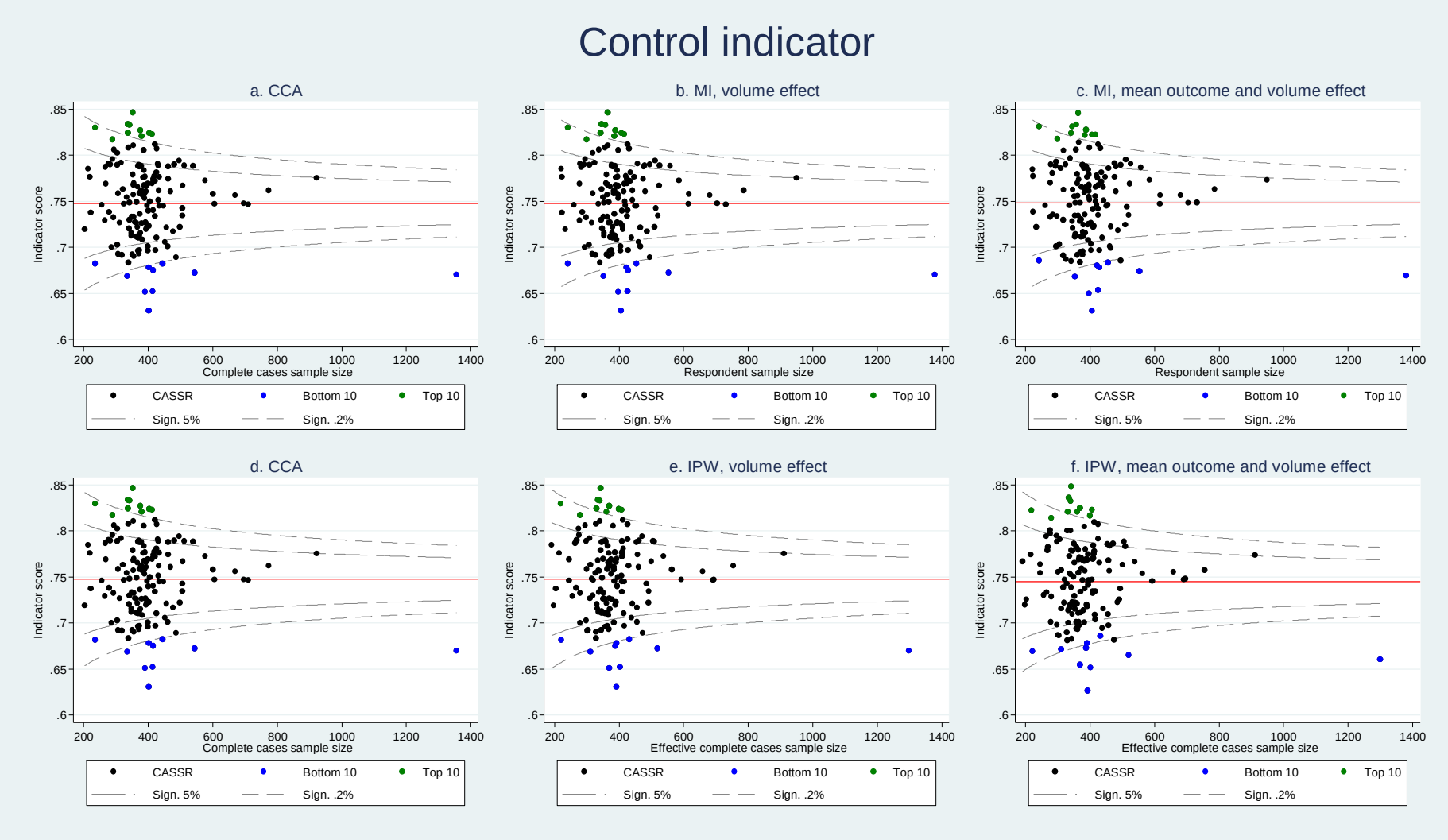

Figure 55: Funnel plots of CASSR scores on the control over daily life indicator, showing the mean outcome and volume effects from MI and IPW

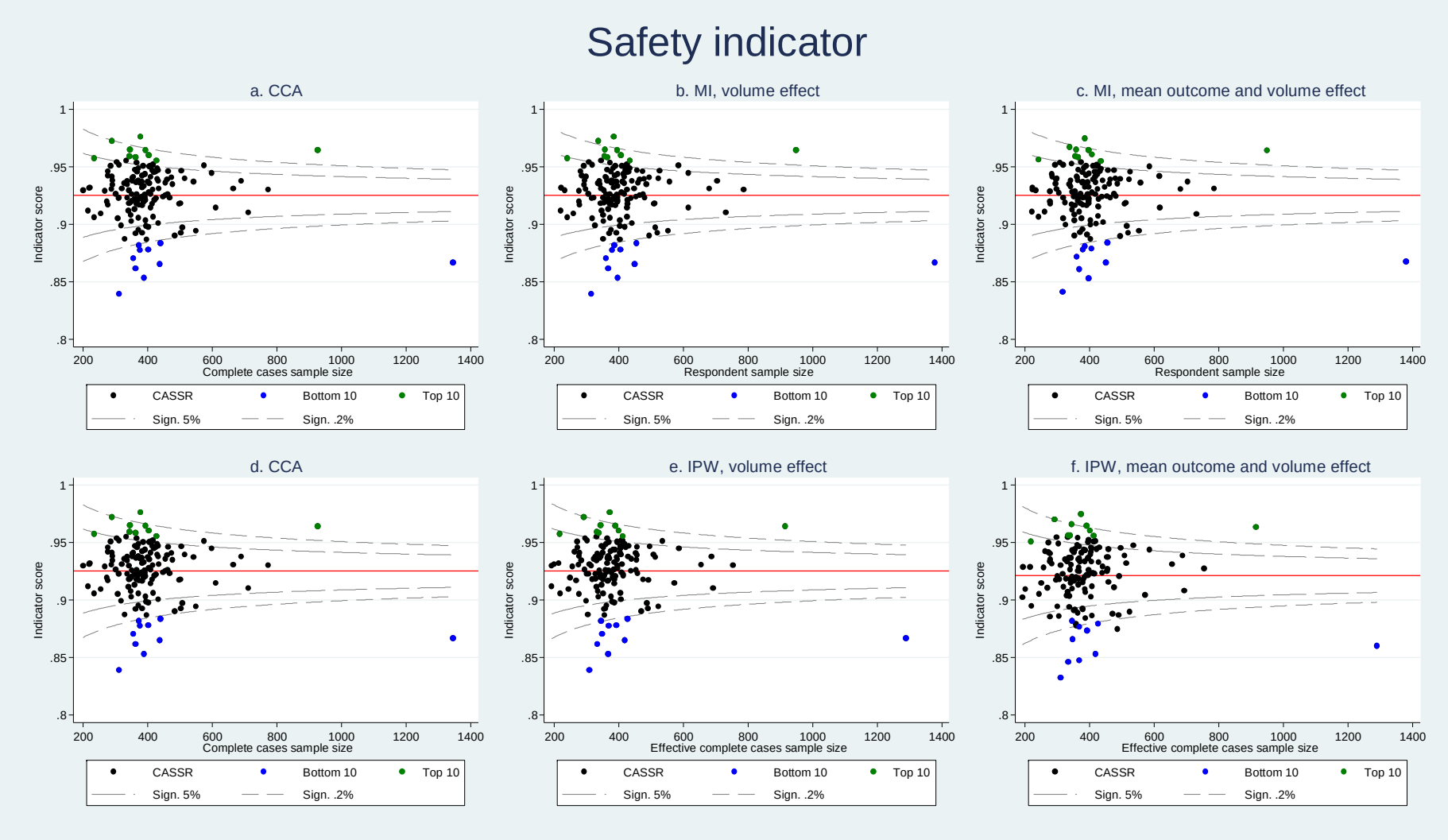

Figure 56: Funnel plots of CASSR scores on the safety indicator, showing the mean outcome and volume effects from MI and IPW

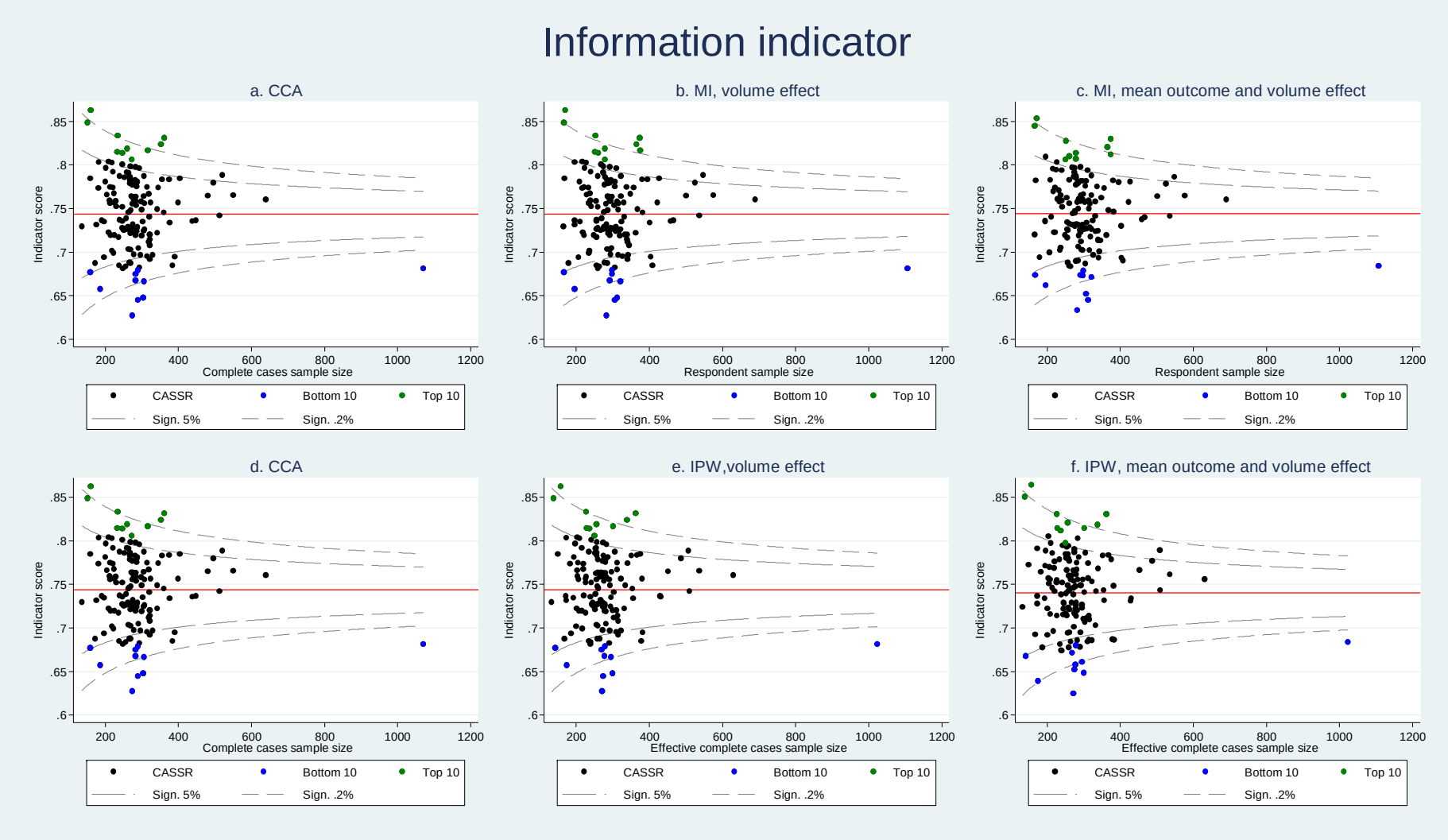

Figure 57: Funnel plots of CASSR scores on the information indicator, showing the mean outcome and volume effects from MI and IPW

Appendix 10: The Diversity of the ASCS Sample

Prior to developing the various case-mix adjustment models, I first needed to consider how to address this diversity of the ASCS sample. Where social care researchers have included multiple client groups, age groups and service settings in studies, they commonly split the sample by these characteristics at the analysis stage (Glendinning et al. 2008a). From a policy perspective, this strategy has the benefit of allowing researchers to draw conclusions that are specific to client groups, and service forms, which are often operationally separate within CASSRs. From an estimation perspective, it is often a necessity due to differences in the way data are collected for different groups, and differences in the interpretation of adjustor variables and their relationship with outcomes. Both of these problems affect the ASCS dataset.

Given these difficulties and the fact that the ASCS sample is dominated by people over the age of 65, who have a physical and/or sensory impairment and are receiving home-based or community-based services, in this analysis, I proceeded by developing separate adjustment models for the various combinations of client group, age group and care settings. Table 70 below shows the size of the sample for the sub-populations, as described by the primary client group, age and care setting. It is clear that the sub-groups are highly unbalanced. The PSD client is by far the largest, with over 70 per cent of the respondents falling into this group. The care home group is about a quarter of the size of the community-based group and older people (over 65) are about twice as numerous as adults under 65. When subdivided by these three variables, some of the sub-groups are too small for separate analysis.[80]

[80] At least that is the case when using only this dataset. It may be possible to conduct analysis on the smaller subgroups using a dataset composed of multiple years' worth of data.

Table 70: Size of the sample for analysis by location of care, primary client group and age, shown for the respondent sample with and without budget data

		Sample with budget data			Sample		
		OP	**YA**	**Total**	**OP**	**YA**	**Total**
Community	PSD	12,140	3,112	15,252	28,815	7,224	36,039
	MH	886	818	1,704	1,924	3,010	4,934
	LD	215	2,479	2,694	423	5,454	5,878
	SM	6	26	32	18	82	100
	VP	511	96	607	971	223	1,195
	Subtotal	13,758	6,531	20,289	32,151	15,993	48,146
Care home	PSD	n/a			6,618	634	7,261
	MH				1,246	557	1,804
	LD				406	2,237	2,647
	SM				15	21	36
	VP				225	23	249
	Subtotal				8,510	3,472	11,997

Legend: PSD=physically or sensorily disabled, MH=mental health client, LD=learning disability, SM=substance misuse client, VP=vulnerable person

It is also notable, although not shown in this table, that the PSD client group is the only group that has at least one respondent from each of the 149 CASSRs in at least one of the sub-group cells when further subdivided by age and care setting, or solely care setting. Even for this, the largest client group, there are a relatively small number of respondents who are aged between 18 to 64 and living in a care home. In this cell, there is not a single respondent from six of the CASSRs, but when the age groups are combined all of the CASSRs have at least one respondent from the PSD group resident in a care home. (For an unknown reason one CASSR, did not have a single sample member in the dataset identified as resident in a care home, so 148 CASSRs is the maximum possible total for the care home sample.) In developing adjustment models, it is important that the sample is as diverse as possible, i.e. has representation from all CASSRs, and is large enough to determine the relevant adjustor variables. There is a danger that models developed on the smaller sub-groups and those missing some CASSRs will be too specific to the peculiarities of the sample.

One possibility to address the small numbers problem is to collapse some of the sub-groups, for example by removing the age-group stratification and bringing some client groups together for analysis. I therefore looked at whether the distribution of the adjustor variables varied across the client groups and age groups to see whether this would be a sensible way forward. On the basis of this analysis, and some exploratory modelling, I decided to combine the vulnerable people (VP) client group with the PSD client group and remove the age group stratification from the care home sub-group.

It is not possible to combine people living in an institutional environment with those receiving services in their own homes or in the community in adjustment models, because the difference in care setting affects the interpretation of adjustor variables. While self-reported suitability of the design of home is arguably exogenous for people living in their own homes, it is clearly not so for care home residents. The design of the home is dependent on the care home's policies around the quality of the care environment. By contrast services delivering care and support to people living in their own homes have very little control (at least in the short-term) over the care environment. While social services departments (SSDs) do provide minor adaptations as part of social care packages, these will have a limited impact on the suitability of the home for the person, particularly for people with more severe disabilities. Over the longer-term social services departments (SSDs) may move people to more suitable housing (assuming they are willing) or provide Disabled Facilities Grants (DFGs) to pay for the renovation of their home for care-related needs (now as part of the Better Care Fund), but in the short-term the design of housing can clearly be seen as a factor that would affect the efficiency and effectiveness of care services and would be desirable to include in any adjustment model. Additionally, there is no intensity measure for care homes, since the budget variable reflects the average fee paid by the authority per place for a given care home and is not related to the amount of care received, so a production function model cannot be estimated with the care home sub-group.

Combining the learning disabled (LD) client group with the other groups (PSD, mental health problems, substance misuse problems and VP)[81] is problematic due to differences in the questionnaires received by the two groups. People with LD receive an

[81] The client groups are defined as follows: physical disability includes physical disability, frailty and/or temporary illness, hearing impairment, visual impairment and dual sensory loss; learning disability is the term used for intellectual impairment; substance misuse includes those with drug and / or alcohol related problems; mental health problems includes people with dementia as well as other mental health problems such as schizophrenia or depression; and vulnerable people is a catch all group to include those whose situation cannot be appropriately fitted in any of the other groups and includes asylum seekers/refugees and welfare benefits clients.

Easy Read version of the questionnaire that differs in the wording and in some cases the number of response options to the standard questionnaire. In developing the Easy Read version, the developers were guided by the principle to maintain consistency with the standard version (Malley et al. 2010). There is, however, no evidence of linguistic equivalence between the questionnaires. Any differences in the interpretation of the two alternative wordings for questions will translate into differences in the observed relationship between adjustor and outcome variables.

There are also reasons to be sceptical about including people from the remaining client groups (PSD, mental health, substance misuse and VP) in the same adjustment model, as the relationship between adjustor variables and outcomes is likely to vary across client groups. For example, measures of difficulty with ADLs are likely to be a better predictor of care needs (i.e. baseline 'risk' or outcome in the absence of services) among people with PSD than among people with mental health or substance misuse problems. In the latter two groups mood is likely to be a more important predictor of baseline risk. Age is another variable that has a heterogeneous relationship with outcome across subgroups of the sample. Since the incidence of long-term conditions rises with age, age becomes a predictor of need for social care, particularly amongst the physically disabled group and only for older age groups. The operational separation of older people's services from services for working age adults also means that older people receive different types of services to people under the age of 65, given their needs profiles, which may mean that differences by age reflect differences in how people are treated by services rather than differences in their baseline outcome state. In addition, older people are much more likely to contribute to the costs of their care (The Health and Social Care Information Centre 2012), which may affect the way they respond to the questions and therefore the relationship of other predictor variables with outcomes.

It is central to the adjustment method that the adjustment models accurately reflect the relationship between adjustor variables and outcomes. An assumption of the approach is that the coefficients from the adjustment model are equal across all population sub-groups (and performance units). If sub-groups differ substantially from the average population the model will fail to adjust adequately for these groups (Julious et al. 2001, Nicholl 2007). The greater the difference in size between sub-groups the more problematic as the average effect will reflect the effect observed for the dominant sub-group. To address the problems of differential relationships between covariates and outcomes by population sub-group, it is possible to include interactions in the models between variables describing the sub-group and adjustor variables. For example, interactions between client group and other variables, and

between age and other variables, will allow the model to reflect differences in the relationship between adjustor variables and outcomes by age and client sub-groups. Yet such models would rapidly become very complicated, which would not be desirable, and it is likely that the models would still inadequately adjust for sub-group differences (Johnson et al. 2010).

Appendix 11: Relationship Between Adjustor Variables and Performance

The relationship between each of the adjustor variables and the indicators, as well as the explanatory power is set out for each of the population sub-groups in Table 71 for the ASCOF SCRQoL indicator. There are a number of similarities across the sub-groups in terms of both the relationship adjustor variables have with the indicators and which adjustor variables make the greatest contribution to explaining differences in indicator scores. The variable capturing feelings of anxiety and depression (ANXDEP) has a fairly strong correlation with the indicator score and has strong explanatory power across all sub-groups. The high explanatory power score is explained more by the large unique contribution it makes to explaining variation in indicator scores, than by heterogeneity across areas, which varies from around four to seven per cent depending on the sub-group. By contrast, the indicators capturing reporting-related factors (i.e. how and who assisted the respondents to complete the questionnaire, and whether additional private payments were made), have in a many cases high heterogeneity (as high as 13 per cent for some), yet have low explanatory power, as they explain little variation in the outcome indicators.

The other adjustor variables, show some variation across sub-groups in their importance. The care home sub-group is the most different to the other sub-groups. With the exception of the anxiety and depression variable, none of the variables capturing underlying conditions (SPH, PAIN and ADLDIFF, FINANCES) have high explanatory power. This is despite these variables all having fairly strong correlations with the outcome indicator, and arises because both the unique contribution they make to explaining variations in the outcome indicator is low and heterogeneity in these indicators for this sub-group across areas is fairly low (zero to seven per cent). Additionally age (AGE_GP), vulnerable person (VP) and gender (MALE) have low explanatory power and make little contribution to explaining variations in the outcome indicator for this sub-group. This is despite very high heterogeneity in the age-group and vulnerable person variables across areas (19 and 53 per cent). Variations in practical help from someone living inside the care home (PH, in home), however, does have high explanatory power, due mostly to high heterogeneity across areas (around 13 per cent).

The two sub-groups of people receiving community- and home-based services are more similar to each other than the care home sub-group. In addition to the anxiety and depression variable, self-perceived health (SPH) also has high explanatory power, due to both

fairly high heterogeneity (eight to nine per cent) and high unique contribution to the variation in the outcome indicator. The other variables capturing underlying conditions (PAIN and ADLDIFF, FINANCES) all have strong correlations with the outcome indicator, but in most cases do not make a large unique contribution to explaining variation in the outcome indicator even though they show some heterogeneity across area (around seven per cent). The exception is the score of the number of ADLs completed with (at least) difficulty (ADLDIFF) which has a different pattern across the two sub-groups, making both a larger contribution to variations in the outcome indicator for the 65 and over subgroup, and having more heterogeneity for this sub-group (around seven per cent compared to zero per cent for the 18 to 65 sub-group). In the case of the ASCOF SCRQoL indicator, the ADLDIFF score (rounded to 2 s.f.s) crosses the criterion of explanatory power greater than one and therefore is included in the explanatory power model for this indicator. This is the only substantive difference in the effect of predictors between the two indicators. This suggests that the method for aggregating the different quality of life questions into a single index makes very little difference to our understanding of what impacts on differences in social care outcomes.

Another important variable in predicting variation in the outcome indicators for these two sub-groups is the self-perceived design of home question (SPHD), which has very high explanatory power. This is driven both by the high unique contribution made by this variable to explaining variation in the outcome indicators and high levels of heterogeneity across areas on this indicator (from eight to 11 per cent). As for the care home sub-group, gender (MALE) has little explanatory power and has little variation across areas, but the other three variables capturing resources and personal characteristics, i.e. practical help (PH), vulnerable person (VP) and age (AGE_GP), show very high heterogeneity across areas (from seven to 45 per cent). Despite this high heterogeneity all these variables have low explanatory power because they make little unique contribution to explaining variations in the outcome indicators for these sub-groups.

Table 71: Correlation and explanatory power of adjustor variables, with respect to ASCOF SCRQoL indicator

Variable	Care home		18 to 65		65 and over	
	Pearson's R	Explanatory power	Pearson's R	Explanatory power	Pearson's R	Explanatory power
MALE	0.007	0.181	0.010	0.000	-0.023***	0.018
AGE_GP	-0.014	0.196	-0.049**	0.022	0.051***	0.001
VP	0.005	0.012	0.010	0.093	0.007	0.026
ADLDIFF	-0.323***	0.091	-0.257***	0.000	-0.385***	**0.971**
ADLCANT	-0.283***	0.016	0.006	0.417	-0.190***	0.070
SPH	-0.394***	0.000	-0.494***	**2.842**	-0.424***	**2.880**
PAIN	-0.227***	-0.018	-0.343***	0.141	-0.219***	0.023
ANXDEP	-0.347***	**1.675**	-0.475***	**3.365**	-0.403***	**1.640**
FINANCES	-0.132***	0.079	-0.103***	0.000	-0.222***	0.104
SPHD	n/a	0.000	-0.477***	**7.724**	-0.437***	**5.654**
PH, in home	0.147***	**0.979**	0.051***	0.420	-0.024***	0.214
PH, out home	-0.146***	0.001	0.019	0.082	0.009	0.081
PROXY	-0.241***	0.147	-0.014	0.000	-0.120***	0.199
ASSIST, read	0.190***	0.033	0.030*	0.000	-0.029***	0.023
ASSIST, translate	0.041***	0.006	0.001	0.056	-0.050***	0.002
ASSIST, write	0.051***	0.004	0.040**	0.001	-0.082***	0.002
ASSIST, discuss	0.001	0.000	-0.047***	0.000	-0.078***	0.000
ASSIST, in home	0.013	0.053	-0.032**	0.000	-0.057***	0.002
ASSIST, out home	-0.137***	0.010	-0.058***	0.000	-0.119***	0.084
ASSIST, cw	0.272***	0.116	0.096***	0.000	0.075***	0.214
PRIV, own	-0.089***	0.111	-0.036**	0.037	-0.035***	0.049
PRIV, fam	-0.059***	0.000	-0.026*	0.000	-0.046***	0.000

Legend: * p<.1; ** p<.01; *** p<.001; variables in bold have an explanatory power over 1, or very close to 1

Appendix 12: Results of the Risk-Adjustment Models Estimated on the Multiply-Imputed Dataset

In this section I present the coefficient estimates for the adjustment models for the ASCOF SCRQoL indicators, estimated by CCA and on the MI dataset (see Appendix 1 and 2 for further details of the imputation method). For the models estimated on the MI dataset, I exclude cases from the estimation for which the ASCOF SCRQoL indicator was missing in the original dataset. The data in this section are presented by population subgroup and by whether the predictors included in the model are all the significant variables (SV model) or a subset of those with high explanatory power (EP model). I also separately present the OLS against the FR regression and the OLS against the RE and FE models to simplify the comparisons. All of the OLS models failed tests for normality of the error terms and homoscedasticity (confirmed on visual inspection of the data). Likewise all of the FE models fail the modified Wald test for groupwise heteroscedasticity, as implemented via the stata routine xttest3 (Baum 2000, Greene 2012). All results for these models therefore use the Huber-White sandwich estimator to correct standard errors (Huber 1967a, White 1980).

The models estimated on the multiply-imputed dataset are very similar to those estimated on the complete cases dataset (see Table 72 to Table 77). All of the coefficients have the same sign, and, although in general the coefficients are numerically different, in only a few instances is the change large enough to lead to different conclusions. For example, in the SV model for the 18 to 64 sub-group the coefficient for help translating the questionnaire (ASSIST, translate) becomes insignificant (from $p<.1$) in both the OLS and FR regressions, while the coefficient on practical help from inside the household (PH, in home) becomes more significant ($p<.1$ to $p<.01$) for the same subgroup in both the OLS and FR regressions (see Table 72). The practical help from inside the household coefficient also changes in the SV model for the 65 and over subgroup, but this time it becomes non-significant ($p<.1$ to N.S.) and only for the FR regression (see Table 74). In the care homes subgroup, the coefficients that change between the two estimation samples are those that were found to change between the regressions. Thus in the SV more, both the male dummy variable (MALE) and the age group (AGE_GP) variables change significance for both the OLS and FR regressions, with MALE becoming less significant and 55 to 64 age group generally dummy variable more significant (see Table 76). Similarly, it is the interaction term between practical help from inside the household (PH, in) and anxiety/depression

(ANXDEP) that changes in the EP model for the care home subgroup. The coefficients estimated on the imputed sample are less significant for both the OLS and FR regressions.

Table 72: Estimates for OLS and FR regression of the ASCOF SCRQoL indicator, SV model, 18 to 64 subgroup

Variable	Casewise-deleted sample (n=5,856)		Multiply-imputed sample (m=20, n=6,793)	
	OLS	**FR**	**OLS**	**FR**
	B (Robust SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (Robust SE)
Fixed effects				
SPH: good [a]	-1.061*** (0.16)	-0.419*** (0.06)	-1.11*** (0.149)	-0.448*** (0.057)
SPH: fair [a]	-2.170*** (0.157)	-0.724*** (0.057)	-2.158*** (0.147)	-0.739*** (0.054)
SPH: bad [a]	-3.483*** (0.18)	-0.970*** (0.059)	-3.48*** (0.169)	-0.987*** (0.056)
SPH: very bad [a]	-3.964*** (0.242)	-1.033*** (0.066)	-4.049*** (0.229)	-1.064*** (0.063)
SPHD	-1.092*** (0.097)	-0.260*** (0.021)	-1.141*** (0.09)	-0.271*** (0.02)
ADLDIFF	-0.069* (0.031)	-0.021** (0.007)	-0.074* (0.03)	-0.023** (0.007)
ADLDIFF*SPHD	-0.091*** (0.021)	-0.009* (0.004)	-0.086*** (0.02)	-0.008* (0.004)
FINANCE: diff [b]	-0.715*** (0.115)	-0.153*** (0.025)	-0.75*** (0.108)	-0.161*** (0.023)
FINANCE: can't [b]	-0.575*** (0.119)	-0.121*** (0.027)	-0.566*** (0.112)	-0.12*** (0.025)
ANXDEP	-1.665*** (0.077)	-0.348*** (0.016)	-1.682*** (0.074)	-0.353*** (0.016)
PH, in home	0.357* (0.151)	0.081* (0.037)	0.463** (0.142)	0.107** (0.035)
PH, out home	0.288** (0.096)	0.058** (0.021)	0.314** (0.092)	0.064** (0.02)
PH, in*ADLDIFF	0.093** (0.035)	0.016* (0.008)	0.082* (0.033)	0.013* (0.008)
MALE	-0.238** (0.089)	-0.052** (0.019)	-0.221** (0.083)	-0.049** (0.018)
ASSIST, cw	0.878*** (0.155)	0.204*** (0.037)	0.872*** (0.149)	0.202*** (0.035)
ASSIST, translate	0.320* (0.179)	0.071* (0.04)	0.138 (0.173)	0.032 (0.038)
ASSIST, write	0.221* (0.108)	0.043* (0.024)	0.215* (0.105)	0.042* (0.023)
Constant	22.299*** (0.183)	2.318*** (0.062)	22.36*** (0.17)	2.35*** (0.059)
Model statistics				
F-stat	259.98***	211.23***	296.23***	242.33***
R^2 [c]	0.431	0.431	0.435	

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: can do easily by myself; [c] R^2 for FR regression calculated as the correlation of predicted SCRQoL with observed SCRQoL.

Table 73: Estimates for OLS and FR regression of the ASCOF SCRQoL indicator, EP model, 18 to 64 subgroup

Variable	Casewise-deleted sample (n=6,513)		Multiply-imputed sample (m=20, n=6,793)	
	OLS	**FR**	**OLS**	**FR**
	B (Robust SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (Robust SE)
Fixed effects				
SPH: good [a]	-1.187*** (0.151)	-0.451*** (0.058)	-1.19*** (0.149)	-0.453*** (0.057)
SPH: fair [a]	-2.425*** (0.148)	-0.771*** (0.056)	-2.399*** (0.146)	-0.766*** (0.055)
SPH: bad [a]	-3.817*** (0.168)	-1.038*** (0.057)	-3.819*** (0.165)	-1.039*** (0.056)
SPH: very bad [a]	-4.422*** (0.227)	-1.139*** (0.063)	-4.472*** (0.226)	-1.148*** (0.063)
SPHD	-1.291*** (0.091)	-0.339*** (0.022)	-1.293*** (0.089)	-0.34*** (0.022)
ANXDEP: mod [b]	-1.290*** (0.123)	-0.397*** (0.033)	-1.277*** (0.121)	-0.395*** (0.032)
ANXDEP: extreme [b]	-2.985*** (0.251)	-0.745*** (0.051)	-3.002*** (0.252)	-0.75*** (0.051)
SPHD* ANXDEP: mod [c]	-0.264* (0.114)	0.039 (0.026)	-0.275* (0.113)	0.038 (0.026)
SPHD* ANXDEP: ext [c]	-0.573*** (0.168)	0.02 (0.034)	-0.588*** (0.167)	0.019 (0.033)
Constant	22.314*** (0.13)	2.370*** (0.055)	22.32*** (0.129)	2.37*** (0.054)
Model statistics				
F-stat	517.20***	415.93***	536.59***	427.5***
R^2 [d]	0.417	0.416	0.421	

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: not anxious or depressed; [c] Base category: SPHD*ANXDEP: not anxious or depressed; [d] R^2 for FR regression calculated as the correlation of predicted SCRQoL with observed SCRQoL.

Table 74: Estimates for OLS and FR regression of ASCOF SCRQoL indicator, significant variables model, 65 and over subgroup

Variable	Casewise-deleted sample (n=20,881)		Multiply-imputed sample (m=20, n=26,103)	
	OLS	**FR**	**OLS**	**FR**
	B (Robust SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (Robust SE)
Fixed effects				
SPH: good [a]	-0.632*** (0.097)	-0.303*** (0.041)	-0.735*** (0.086)	-0.344*** (0.037)
SPH: fair [a]	-1.624*** (0.095)	-0.603*** (0.04)	-1.673*** (0.084)	-0.63*** (0.036)
SPH: bad [a]	-2.676*** (0.109)	-0.806*** (0.041)	-2.749*** (0.097)	-0.839*** (0.037)
SPH: very bad [a]	-3.165*** (0.15)	-0.864*** (0.045)	-3.187*** (0.137)	-0.886*** (0.041)
ANXDEP	-1.350*** (0.04)	-0.304*** (0.009)	-1.344*** (0.037)	-0.305*** (0.008)
SPHD	-1.397*** (0.031)	-0.305*** (0.007)	-1.387*** (0.028)	-0.305*** (0.006)
ADLDIFF	-0.297*** (0.013)	-0.068*** (0.003)	-0.29*** (0.012)	-0.066*** (0.003)
FINANCES	-0.188*** (0.029)	-0.050*** (0.007)	-0.169*** (0.026)	-0.045*** (0.006)
PH, in home	0.156* (0.075)	0.037* (0.02)	0.126* (0.067)	0.029 (0.018)
PH, out home	0.276*** (0.052)	0.063*** (0.013)	0.291*** (0.048)	0.069*** (0.012)
PH, in *ADLDIFF	0.154*** (0.018)	0.031*** (0.004)	0.146*** (0.017)	0.030*** (0.004)
MALE	-0.204*** (0.045)	-0.047*** (0.011)	-0.25*** (0.04)	-0.059*** (0.01)
PROXY	-0.854*** (0.104)	-0.178*** (0.022)	-0.889*** (0.099)	-0.187*** (0.021)
ASSIST, cw	1.039*** (0.103)	0.272*** (0.028)	0.875*** (0.094)	0.230*** (0.026)
ASSIST, out home	-0.323*** (0.06)	-0.073*** (0.014)	-0.414*** (0.057)	-0.096*** (0.013)
ASSIST, read	0.159** (0.051)	0.033** (0.012)	0.158** (0.048)	0.033** (0.011)
PRIV, own	-0.159*** (0.042)	-0.044*** (0.01)	-0.176*** (0.038)	-0.048*** (0.009)
Constant	22.42*** (0.099)	2.377*** (0.041)	22.52*** (0.089)	2.418*** (0.037)
Model statistics				
F-stat	742.34***	668.84***	851.96***	771.31***
R^2 [b]	0.391	0.391	0.385	

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] R^2 for FR regression calculated as the correlation of predicted SCRQoL with observed SCRQoL.

Table 75: Estimates for OLS and FR regression of ASCOF SCRQoL indicator, EP model, 65 and over subgroup

Variable	Casewise-deleted sample (n=23,110)		Multiply-imputed sample (m=20, n=26,103)	
	OLS	**FR**	**OLS**	**FR**
	B (Robust SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (Robust SE)
Fixed effects				
SPH: good [a]	-0.734*** (0.091)	-0.321*** (0.039)	-0.764*** (0.086)	-0.334*** (0.037)
SPH: fair [a]	-1.730*** (0.089)	-0.609*** (0.038)	-1.745*** (0.085)	-0.618*** (0.036)
SPH: bad [a]	-2.837*** (0.104)	-0.828*** (0.039)	-2.847*** (0.098)	-0.837*** (0.037)
SPH: very bad [a]	-3.275*** (0.147)	-0.894*** (0.044)	-3.286*** (0.14)	-0.904*** (0.042)
SPHD: most [b]	-1.425*** (0.063)	-0.434*** (0.017)	-1.445*** (0.06)	-0.440*** (0.016)
SPHD: some [b]	-2.552*** (0.132)	-0.687*** (0.028)	-2.603*** (0.127)	-0.699*** (0.027)
SPHD: inappropriate [b]	-3.278*** (0.417)	-0.832*** (0.077)	-3.168*** (0.392)	-0.818*** (0.074)
ANXDEP: moderate [c]	-1.316*** (0.042)	-0.318*** (0.01)	-1.305*** (0.041)	-0.317*** (0.01)
ANXDEP: extreme [c]	-2.887*** (0.119)	-0.597*** (0.022)	-2.909*** (0.114)	-0.603*** (0.021)
ADLDIFF	-0.241*** (0.013)	-0.074*** (0.003)	-0.241*** (0.012)	-0.075*** (0.003)
ADLDIFF*SPHD: most [d]	-0.024 (0.018)	0.020*** (0.004)	-0.022 (0.018)	0.020*** (0.004)
ADLDIFF*SPHD: some [d]	-0.057* (0.03)	0.023*** (0.006)	-0.048* (0.029)	0.025*** (0.006)
ADLDIFF*SPHD: inappropriate [d]	-0.137 (0.09)	0.012 (0.016)	-0.142* (0.086)	0.012 (0.016)
Constant	22.43*** (0.084)	2.430*** (0.038)	22.47*** (0.079)	2.446*** (0.036)
Model statistics				
F-stat	1,032.44***	899.523***	1,108.36***	965.31***
R^2 [e]	0.370	0.371	0.367	

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: meets needs very well; [c] Base category: not anxious or depressed; [d] Base category: ADLDIFF*SPHD: meets needs very well; [e] R^2 for FR regression calculated as the correlation of predicted SCRQoL with observed SCRQoL.

Table 76: Estimates for OLS and FR regression of ASCOF SCRQoL indicator, SV model, care home sub-group

Variable	Casewise-deleted sample (n=5,710)		Multiply-imputed sample (m=20, n=6,863)	
	OLS	**FR**	**OLS**	**FR**
	B (Robust SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (Robust SE)
Fixed effects				
SPH	-0.820*** (0.055)	-0.230*** (0.016)	-0.841*** (0.051)	-0.237*** (0.014)
ANXDEP: moderate [a]	-1.289*** (0.09)	-0.366*** (0.025)	-1.246*** (0.084)	-0.358*** (0.023)
ANXDEP: extremely [a]	-3.052*** (0.28)	-0.680*** (0.057)	-2.984*** (0.267)	-0.670*** (0.055)
ADLDIFF^3	-0.004*** (3.59E-04)	-0.001*** (9.75E-05)	-0.005*** (3.33E-04)	-0.001*** (9.08E-05)
PH, in home	0.686*** (0.082)	0.207*** (0.026)	0.688*** (0.079)	0.210*** (0.025)
MALE	-0.326*** (0.094)	-0.086** (0.027)	-0.242** (0.084)	-0.064* (0.025)
AGE_GP: 18-34 [b]	0.057 (0.344)	0.066 (0.124)	-0.309 (0.334)	-0.066 (0.115)
AGE_GP: 35-44 [b]	-1.595*** (0.406)	-0.445*** (0.102)	-1.373*** (0.361)	-0.388*** (0.095)
AGE_GP: 45-54 [b]	-0.103 (0.241)	-0.026 (0.078)	-0.218 (0.232)	-0.066 (0.074)
AGE_GP: 55-64 [b]	-0.477* (0.217)	-0.148* (0.063)	-0.564** (0.192)	-0.173** (0.056)
AGE_GP: 65-74 [b]	-0.455** (0.15)	-0.134** (0.043)	-0.409** (0.135)	-0.120** (0.039)
AGE_GP: 75-84 [b]	-0.118 (0.097)	-0.023 (0.028)	-0.115 (0.088)	-0.022 (0.026)
PROXY	-1.051*** (0.186)	-0.235*** (0.05)	-1.086*** (0.168)	-0.244*** (0.045)
ASSIST, cw	0.720*** (0.145)	0.264*** (0.047)	0.652*** (0.126)	0.240*** (0.042)
ASSIST, out home	-0.530*** (0.142)	-0.153*** (0.042)	-0.542*** (0.124)	-0.158*** (0.038)
ASSIST, read	0.313** (0.103)	0.090** (0.029)	0.308** (0.094)	0.091** (0.027)
PRIV, own	-0.492*** (0.129)	-0.130*** (0.033)	-0.471*** (0.116)	-0.127*** (0.031)
Constant	22.13*** (0.149)	2.247*** (0.047)	22.2*** (0.13)	2.273*** (0.042)
Model statistics				
F-stat	122.98***	127.64***	143.04***	149.38***
R^{2} [c]	0.296	0.301	0.295	

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: not anxious or depressed; [b] Base category: 85 and over age group; [c] R^2 for FR regression calculated as the correlation of predicted SCRQoL with observed SCRQoL.

Table 77: Estimates for OLS and FR regression of ASCOF SCRQoL indicator, EP model, care home sub-group

Variable	Casewise-deleted sample (n=6,344)		Multiply-imputed sample (m=20, n=6,863)	
	OLS	**FR**	**OLS**	**FR**
	B (Robust SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (Robust SE)
Fixed effects				
ANXDEP: moderate [a]	-2.218*** (0.114)	-0.588*** (0.029)	-2.216*** (0.112)	-0.590*** (0.029)
ANXDEP: extremely [a]	-5.027*** (0.346)	-1.147*** (0.065)	-5.032*** (0.348)	-1.151*** (0.065)
PH in home	0.741*** (0.097)	0.261*** (0.035)	0.695*** (0.098)	0.245*** (0.035)
PH in* ANXDEP: moderate [b]	0.540** (0.184)	0.052 (0.053)	0.436* (0.188)	0.030 (0.054)
PH in*ANXDEP: extremely [b]	1.784** (0.656)	0.234* (0.14)	1.613* (0.677)	0.205 (0.141)
Constant	20.33*** (0.063)	1.712*** (0.02)	20.35*** (0.061)	1.718*** (0.02)
Model statistics				
F-stat	172.23***	187.632***	169.92***	184.82***
R^{2} [c]	0.140	0.140	0.137	

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: not anxious or depressed; [b] Base category: PH in*ANXDEP: not anxious or depressed; [c] R^2 for FR regression calculated as the correlation of predicted SCRQoL with observed SCRQoL.

The results for the risk adjustment models estimated by OLS, RE and FE regression for each subgroup and variable selection are shown in Table 78 through to Table 83. The model coefficients estimated on the multiply-imputed dataset all have effects in the anticipated direction and the results are the same in that respect as the models estimated using the complete cases only. Many of the coefficients estimated on the MI compared to the complete case sample are different, but in most cases the differences are small and the conclusions remain the same, showing evidence of convergent validity.

Nevertheless, there are some differences that are larger and some coefficients change in their significance (although, as before, not always an indication of the size of the coefficient difference). Often the coefficients that are most different between the imputed and complete case samples are those that differ most between the regression models. For example, the coefficients for self-perceived health (SPH) for the SV model estimated on the 18 to 65 subgroup show modest changes (>.05, but <.1) (see Table 78) and once again the greatest differences (in the range of 0.1 to 0.4 scale points), are found for the age group coefficients for the SV model estimated on the care home subgroup (see Table 82). This is not always the case. The coefficient are all very similar (all differences <.05) for the EP model estimated on the 18 to 65 sub-group show quite large changes despite there being quite a lot of modest changes of around 0.05 to 0.1 between the coefficients estimated by the different regression methods (see Table 79).

Table 78: Estimates for OLS, fixed and RE regression of ASCOF SCRQoL indicator, SV model, 18 to 64 sub-group

Variable	Casewise-deleted sample (n=5,856)			Multiply-imputed sample (m=20, n=6,793)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
SPH: good [a]	-1.061*** (0.16)	-1.068*** (0.203)	-1.118*** (0.162)	-1.11*** (0.149)	-1.117*** (0.191)	-1.169*** (0.144)
SPH: fair [a]	-2.170*** (0.157)	-2.172*** (0.194)	-2.202*** (0.16)	-2.158*** (0.147)	-2.161*** (0.183)	-2.194*** (0.149)
SPH: bad [a]	-3.483*** (0.18)	-3.487*** (0.209)	-3.521*** (0.181)	-3.48*** (0.169)	-3.485*** (0.197)	-3.528*** (0.172)
SPH: very bad [a]	-3.964*** (0.242)	-3.962*** (0.242)	-3.974*** (0.256)	-4.049*** (0.229)	-4.048*** (0.228)	-4.072*** (0.254)
SPHD	-1.092*** (0.097)	-1.088*** (0.096)	-1.062*** (0.101)	-1.141*** (0.09)	-1.137*** (0.089)	-1.109*** (0.09)
ADLDIFF	-0.069* (0.031)	-0.069* (0.032)	-0.072* (0.029)	-0.074* (0.03)	-0.075* (0.03)	-0.076** (0.029)
ADLDIFF*SPHD	-0.091*** (0.021)	-0.091*** (0.02)	-0.090*** (0.021)	-0.086*** (0.02)	-0.086*** (0.019)	-0.086*** (0.02)
FINANCE: diff [c]	-0.715*** (0.115)	-0.717*** (0.116)	-0.722*** (0.116)	-0.75*** (0.108)	-0.753*** (0.109)	-0.766*** (0.107)
FINANCE: can't [c]	-0.575*** (0.119)	-0.579*** (0.12)	-0.598*** (0.117)	-0.566*** (0.112)	-0.568*** (0.113)	-0.584*** (0.113)
ANXDEP	-1.665*** (0.077)	-1.663*** (0.073)	-1.661*** (0.073)	-1.682*** (0.074)	-1.681*** (0.069)	-1.677*** (0.073)
PH in home	0.357* (0.151)	0.359* (0.163)	0.372* (0.155)	0.463** (0.142)	0.464** (0.152)	0.47** (0.146)
PH out home	0.288** (0.096)	0.284** (0.096)	0.275** (0.101)	0.314** (0.092)	0.309** (0.091)	0.291** (0.098)
PH in*ADLDIFF	0.093** (0.035)	0.091** (0.035)	0.082* (0.036)	0.082* (0.033)	0.081* (0.033)	0.072* (0.033)
MALE	-0.238** (0.089)	-0.236** (0.089)	-0.226* (0.095)	-0.221** (0.083)	-0.222** (0.083)	-0.225** (0.085)
ASSIST, care worker	0.878*** (0.155)	0.880*** (0.149)	0.878*** (0.171)	0.872*** (0.149)	0.872*** (0.144)	0.861*** (0.153)
ASSIST, translate	0.320* (0.179)	0.326* (0.169)	0.313* (0.183)	0.138 (0.173)	0.142 (0.164)	0.136 (0.186)
ASSIST, write down	0.221* (0.108)	0.216* (0.107)	0.191 (0.116)	0.215* (0.105)	0.21* (0.104)	0.189* (0.11)
Constant	22.299*** (0.183)	22.310*** (0.221)	22.347*** (0.18)	22.36*** (0.17)	22.38*** (0.206)	22.42*** (0.169)
Random effects						
σ_u		0.210 (0.083)			0.195 (0.079)	
σ_e		3.289 (0.031)			3.304 (0.029)	
Model statistics						
F-stat	259.98***	191.33***	253.91***	296.23***	296.37***	292.42***
R^2	0.431		0.447	0.435		0.450

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: SPHD*SPH: very good; [c] Base category: can do easily by myself.

Table 79: Estimates for OLS, fixed and RE regression of the ASCOF SCRQoL indicator, EP model, 18 to 64 sub-group

Variable	Casewise-deleted sample (n=6,513)			Multiply-imputed sample (m=20, n=6,793)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
SPH: good [a]	-1.187*** (0.151)	-1.192*** (0.197)	-1.244*** (0.146)	-1.19*** (0.149)	-1.198*** (0.193)	-1.255*** (0.143)
SPH: fair [a]	-2.425*** (0.148)	-2.429*** (0.188)	-2.472*** (0.151)	-2.399*** (0.146)	-2.404*** (0.185)	-2.445*** (0.149)
SPH: bad [a]	-3.817*** (0.168)	-3.822*** (0.201)	-3.872*** (0.173)	-3.819*** (0.165)	-3.826*** (0.197)	-3.881*** (0.17)
SPH: very bad [a]	-4.422*** (0.227)	-4.424*** (0.229)	-4.465*** (0.254)	-4.472*** (0.226)	-4.474*** (0.226)	-4.519*** (0.251)
SPHD	-1.291*** (0.091)	-1.286*** (0.088)	-1.248*** (0.087)	-1.293*** (0.089)	-1.288*** (0.087)	-1.258*** (0.086)
ANXDEP: moderate [c]	-1.290*** (0.123)	-1.290*** (0.135)	-1.288*** (0.121)	-1.277*** (0.121)	-1.277*** (0.134)	-1.274*** (0.12)
ANXDEP: extreme [c]	-2.985*** (0.251)	-2.978*** (0.218)	-2.916*** (0.272)	-3.002*** (0.252)	-2.995*** (0.22)	-2.944*** (0.271)
SPHD* ANXDEP: moderate [d]	-0.264* (0.114)	-0.264* (0.11)	-0.264* (0.11)	-0.275* (0.113)	-0.274* (0.109)	-0.272* (0.109)
SPHD* ANXDEP: extreme [d]	-0.573*** (0.168)	-0.576*** (0.139)	-0.618*** (0.17)	-0.588*** (0.167)	-0.59*** (0.139)	-0.623*** (0.169)
Constant	22.314*** (0.13)	22.316*** (0.179)	22.318*** (0.13)	22.32*** (0.129)	22.32*** (0.175)	22.33*** (0.128)
Random effects						
σ_u		0.183 (0.089)			0.205 (0.08)	
σ_e		3.345 (0.03)			3.347 (0.029)	
Model statistics						
F-stat	517.20***	384.50***	475.57***	536.59***	526.43***	482.48***
R^2	0.417		0.432	0.421		0.436

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good; [b] Base category: SPHD*SPH: very good; [c] Base category: not anxious or depressed; [d] Base category: SPHD*ANXDEP: not anxious or depressed

Table 80: Estimates for OLS, fixed and RE regression of the ASCOF SCRQoL indicator, SV model, 65 and over sub-group

Variable	Casewise-deleted sample (n=20,881)			Multiply-imputed sample (m=20, n=26,103)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
SPH: good [a]	-0.632*** (0.097)	-0.635*** (0.118)	-0.641*** (0.094)	-0.735*** (0.086)	-0.735*** (0.104)	-0.738*** (0.085)
SPH: fair [a]	-1.624*** (0.095)	-1.626*** (0.114)	-1.631*** (0.089)	-1.673*** (0.084)	-1.678*** (0.1)	-1.685*** (0.086)
SPH: bad [a]	-2.676*** (0.109)	-2.681*** (0.124)	-2.686*** (0.108)	-2.749*** (0.097)	-2.756*** (0.109)	-2.763*** (0.102)
SPH: very bad [a]	-3.165*** (0.15)	-3.161*** (0.147)	-3.157*** (0.15)	-3.187*** (0.137)	-3.185*** (0.132)	-3.185*** (0.136)
ANXDEP	-1.350*** (0.04)	-1.345*** (0.037)	-1.342*** (0.04)	-1.344*** (0.037)	-1.337*** (0.034)	-1.333*** (0.038)
SPHD	-1.397*** (0.031)	-1.391*** (0.028)	-1.383*** (0.03)	-1.387*** (0.028)	-1.378*** (0.026)	-1.369*** (0.027)
ADLDIFF	-0.297*** (0.013)	-0.294*** (0.012)	-0.293*** (0.013)	-0.29*** (0.012)	-0.288*** (0.011)	-0.287*** (0.013)
FINANCEs	-0.188*** (0.029)	-0.182*** (0.029)	-0.178*** (0.03)	-0.169*** (0.026)	-0.168*** (0.027)	-0.168*** (0.028)
PH, in home	0.156* (0.075)	0.153* (0.08)	0.145* (0.087)	0.126* (0.067)	0.116 (0.072)	0.104 (0.083)
PH, outside home	0.276*** (0.052)	0.266*** (0.051)	0.258*** (0.056)	0.291*** (0.048)	0.274*** (0.047)	0.261*** (0.053)
PH, in *ADLDIFF	0.154*** (0.018)	0.153*** (0.018)	0.155*** (0.021)	0.146*** (0.017)	0.146*** (0.016)	0.146*** (0.019)
MALE	-0.204*** (0.045)	-0.202*** (0.044)	-0.200*** (0.047)	-0.25*** (0.04)	-0.249*** (0.04)	-0.248*** (0.041)
PROXY	-0.854*** (0.104)	-0.860*** (0.095)	-0.863*** (0.113)	-0.889*** (0.099)	-0.895*** (0.091)	-0.901*** (0.105)
ASSIST, cw	1.039*** (0.103)	1.004*** (0.1)	0.974*** (0.139)	0.875*** (0.094)	0.848*** (0.091)	0.823*** (0.097)
ASSIST, out home	-0.323*** (0.06)	-0.331*** (0.06)	-0.336*** (0.064)	-0.414*** (0.057)	-0.422*** (0.056)	-0.428*** (0.063)
ASSIST, read	0.159** (0.051)	0.153** (0.05)	0.150** (0.055)	0.158** (0.048)	0.153** (0.047)	0.149** (0.051)
PRIV, own	-0.159*** (0.042)	-0.156*** (0.042)	-0.155*** (0.042)	-0.176*** (0.038)	-0.176*** (0.038)	-0.174*** (0.038)
Constant	22.42*** (0.099)	22.41*** (0.122)	22.41*** (0.094)	22.52*** (0.089)	22.52*** (0.108)	22.53*** (0.088)
Random effects						
σ_u		0.271 (0.029)			0.246 (0.027)	
σ_e		2.878 (0.014)			2.887 (0.013)	
Model statistics						
F-stat	742.34***	597.86***	830.19***	851.96***	899.03***	932.72***
R^2	0.391		0.401	0.385		0.393

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: very good.

Table 81: Estimates for OLS, fixed and RE regression of the ASCOF SCRQoL indicator, EP model, 65 and over sub-group

Variable	Casewise-deleted sample (n=23,110)			Multiply-imputed sample (m=20, n=26,103)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
SPH: good [a]	-0.734*** (0.091)	-0.737*** (0.113)	-0.743*** (0.089)	-0.764*** (0.086)	-0.767*** (0.105)	-0.773*** (0.089)
SPH: fair [a]	-1.730*** (0.089)	-1.733*** (0.11)	-1.740*** (0.088)	-1.745*** (0.085)	-1.751*** (0.102)	-1.76*** (0.09)
SPH: bad [a]	-2.837*** (0.104)	-2.841*** (0.119)	-2.847*** (0.106)	-2.847*** (0.098)	-2.856*** (0.111)	-2.864*** (0.107)
SPH: very bad [a]	-3.275*** (0.147)	-3.275*** (0.142)	-3.276*** (0.151)	-3.286*** (0.14)	-3.289*** (0.134)	-3.292*** (0.141)
SPHD: most [b]	-1.425*** (0.063)	-1.432*** (0.068)	-1.435*** (0.056)	-1.445*** (0.06)	-1.449*** (0.065)	-1.45*** (0.055)
SPHD: some [b]	-2.552*** (0.132)	-2.532*** (0.127)	-2.512*** (0.145)	-2.603*** (0.127)	-2.581*** (0.121)	-2.563*** (0.134)
SPHD: inappropriate [b]	-3.278*** (0.417)	-3.214*** (0.36)	-3.161*** (0.398)	-3.168*** (0.392)	-3.105*** (0.335)	-3.058*** (0.381)
ANXDEP: moderate [c]	-1.316*** (0.042)	-1.314*** (0.042)	-1.313*** (0.052)	-1.305*** (0.041)	-1.303*** (0.041)	-1.302*** (0.049)
ANXDEP: extreme [c]	-2.887*** (0.119)	-2.866*** (0.096)	-2.849*** (0.12)	-2.909*** (0.114)	-2.889*** (0.093)	-2.876*** (0.11)
ADLDIFF	-0.241*** (0.013)	-0.239*** (0.013)	-0.237*** (0.013)	-0.241*** (0.012)	-0.24*** (0.012)	-0.239*** (0.013)
ADLDIFF* SPHD: most [d]	-0.024 (0.018)	-0.022 (0.018)	-0.02 (0.019)	-0.022 (0.018)	-0.019 (0.018)	-0.018 (0.019)
ADLDIFF* SPHD: some [d]	-0.057* (0.03)	-0.057* (0.028)	-0.056 (0.035)	-0.048* (0.029)	-0.047** (0.027)	-0.047 (0.032)
ADLDIFF* SPHD: inappropriate [d]	-0.137 (0.09)	-0.144* (0.071)	-0.148 (0.09)	-0.142* (0.086)	-0.148** (0.067)	-0.153** (0.088)
Constant	22.43*** (0.084)	22.42*** (0.109)	22.42*** (0.084)	22.47*** (0.079)	22.46*** (0.101)	22.46*** (0.083)
Random effects						
σ_u		0.261 (0.028)			0.259 (0.027)	
σ_e		2.924 (0.014)			2.926 (0.013)	
Model statistics						
F-stat	1,032.44***	1,024.36***	1,034.73***	1,108.36***	1,098.08***	1,188.55***
R^2	0.370	0.371	0.379	0.367		0.376

Legend: * p<.1; ** p<.01; *** p<.001; a Base category: very good; b Base category: meets needs very well; c Base category: not anxious or depressed; d Base category: ADLDIFF*SPHD: meets needs very well.

Table 82: Estimates for OLS, fixed and RE regression of the ASCOF SCRQoL indicator, SV model, care home sub-group

Variable	Casewise-deleted sample (n=5,710)			Multiply-imputed sample (m=20, n=6,863)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
SPH	-0.820*** (0.055)	-0.819*** (0.052)	-0.821*** (0.054)	-0.841*** (0.051)	-0.84*** (0.048)	-0.841*** (0.055)
ANXDEP: moderate [a]	-1.289*** (0.09)	-1.286*** (0.088)	-1.276*** (0.086)	-1.246*** (0.084)	-1.243*** (0.082)	-1.24*** (0.08)
ANXDEP: extreme [a]	-3.052*** (0.28)	-3.057*** (0.216)	-3.076*** (0.29)	-2.984*** (0.267)	-2.987*** (0.204)	-2.989*** (0.254)
ADLDIFF^3	-0.004*** (3.59E-04)	-0.005*** (2.41E-04)	-0.005***(3.55E-04)	-0.005*** (3.33E-04)	-0.005*** (3.23E-04)	-0.005*** (3.45E-04)
PH, in home	0.686*** (0.082)	0.686*** (0.087)	0.684*** (0.094)	0.688*** (0.079)	0.689*** (0.082)	0.694*** (0.087)
MALE	-0.326*** (0.094)	-0.325*** (0.094)	-0.315** (0.107)	-0.242** (0.084)	-0.241** (0.085)	-0.231* (0.092)
AGE_GP: 18-34 [b]	0.057 (0.344)	0.069 (0.4)	0.043 (0.347)	-0.309 (0.334)	-0.304 (0.368)	-0.389 (0.307)
AGE_GP: 35-44 [b]	-1.595*** (0.406)	-1.564*** (0.368)	-1.374*** (0.401)	-1.373*** (0.361)	-1.309*** (0.332)	-1.185** (0.35)
AGE_GP: 45-54 [b]	-0.103 (0.241)	-0.086 (0.266)	0.001 (0.243)	-0.218 (0.232)	-0.183 (0.244)	-0.113 (0.238)
AGE_GP: 55-64 [b]	-0.477* (0.217)	-0.460* (0.218)	-0.354 (0.226)	-0.564** (0.192)	-0.528** (0.193)	-0.447* (0.202)
AGE_GP: 65-74 [b]	-0.455** (0.15)	-0.448** (0.147)	-0.405* (0.175)	-0.409** (0.135)	-0.399** (0.134)	-0.367* (0.154)
AGE_GP: 75-84 [b]	-0.118 (0.097)	-0.117 (0.097)	-0.111 (0.101)	-0.115 (0.088)	-0.112 (0.088)	-0.102 (0.09)
PROXY	-1.051*** (0.186)	-1.044*** (0.171)	-1.006*** (0.194)	-1.086*** (0.168)	-1.07*** (0.155)	-1.03*** (0.172)
ASSIST, cw	0.720*** (0.145)	0.705*** (0.148)	0.656*** (0.158)	0.652*** (0.126)	0.627*** (0.131)	0.602*** (0.139)
ASSIST, out home	-0.530*** (0.142)	-0.530*** (0.138)	-0.527*** (0.154)	-0.542*** (0.124)	-0.537*** (0.123)	-0.517*** (0.136)
ASSIST, read	0.313** (0.103)	0.310** (0.1)	0.283* (0.112)	0.308** (0.094)	0.302** (0.092)	0.285** (0.101)
PRIV, own	-0.492*** (0.129)	-0.484*** (0.114)	-0.455** (0.138)	-0.471*** (0.116)	-0.451*** (0.106)	-0.43** (0.129)
Constant	22.13*** (0.149)	22.13*** (0.152)	22.14*** (0.147)	22.2*** (0.13)	22.20*** (0.136)	22.18*** (0.133)
Random effects						
σ_u		0.204 (0.081)			0.313 (0.056)	
σ_e		3.018 (0.029)			3.003 (0.026)	
Model statistics						
F-stat	122.98***	116.13***	110.65***	143.04***	160.39***	130.47***
R^2	0.296		0.318	0.295		0.317

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: not anxious or depressed; [b] Base category: 85 and over age group.

Table 83: Estimates for OLS, fixed and RE regression of the ASCOF SCRQoL indicator, EP model, care home sub-group

Variable	Casewise-deleted sample (n=6,344)			Multiply-imputed sample (m=20, n=6,863)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
ANXDEP: mod [a]	-2.218*** (0.114)	-2.190*** (0.105)	-2.157*** (0.107)	-2.216*** (0.112)	-2.189*** (0.105)	-2.167*** (0.104)
ANXDEP: extreme [a]	-5.027*** (0.346)	-4.988*** (0.257)	-4.932*** (0.38)	-5.032*** (0.348)	-4.99*** (0.264)	-4.948*** (0.364)
PH in home	0.741*** (0.097)	0.752*** (0.113)	0.770*** (0.105)	0.695*** (0.098)	0.704*** (0.113)	0.718*** (0.106)
PH, in*ANXDEP: mod	0.540** (0.184)	0.511** (0.189)	0.471** (0.175)	0.436* (0.188)	0.418* (0.189)	0.401* (0.174)
PH, in*ANXDEP: ext	1.784** (0.656)	1.758*** (0.495)	1.708* (0.668)	1.613* (0.677)	1.591** (0.526)	1.567* (0.692)
Constant	20.33*** (0.063)	20.30*** (0.078)	20.30*** (0.049)	20.35*** (0.061)	20.32*** (0.079)	20.33*** (0.049)
Random effects						
σ_u		0.447 (0.062)			0.493 (0.06)	
σ_e		3.293 (0.030)			3.304 (0.029)	
Model statistics						
F-stat	172.23***	188.15***	195.50***	169.92***	195.89***	172.98***
R^2	0.140		0.174	0.137		0.174

Legend: * p<.1; ** p<.01; *** p<.001; [a] Base category: not anxious or depressed; [b] Base category: PH in*ANXDEP: not anxious or depressed.

Appendix 13: Distribution of Predictions from the Risk Adjustment Models for All the PSD Sub-Groups

The relatively poor performance of all the models in predicting the extremes of the distribution is shown more clearly in Table 84 to Table 86. The mean error and mean absolute error are both largest for the categories that are farthest away from the mean of the distribution and the errors from the EP model are in all cases larger than those from the SV model. The care home sub-group has highest mean and the most skewed distribution, followed by the 65 and over sub-group, and then the 18 to 64 sub-group. It is probably for this reason that the errors are greatest for the care home sub-group at the low end of the distribution and smallest at the high end of the distribution, whereas they are greatest for the 18 to 64 sub-group at the high end of the distribution and smallest at the low end of the distribution. Across all of the models, both the RE and FE models perform better than the OLS model in predicting the extremes of the distribution, but this could be due at the upper end to predictions falling out of range more frequently for the RE and FE models. The FR models perform better the other models generally over the lower end of the distribution, but only for the SV models.

Table 84: Predictions of the models over the range of ASCOF SCRQoL, 18 to 64 sub-group

	For regions where ASCOF SCRQoL is…				
	<11	**≥11, < 14**	**≥14, < 17**	**≥17, < 20**	**≥20**
Full model	**n=442**	**n=741**	**n=1,259**	**n=1,538**	**n=1,876**
Observed mean	7.889	12.143	15.075	17.974	21.804
Predicted mean					
OLS	13.025	14.964	16.318	17.375	19.137
FRM	**12.830**	14.948	16.390	**17.439**	19.088
RE	12.845	**14.863**	**16.287**	17.409	**19.212**
FE	12.873	14.874	**16.287**	17.406	19.203
Mean error (ME)					
OLS	-5.136	-2.821	-1.243	0.599	2.668
FRM	-4.940	-2.805	-1.315	**0.535**	2.716
RE	-4.956	**-2.720**	**-1.212**	0.565	**2.593**
FE	**-4.984**	-2.731	-1.213	0.568	2.601
Mean absolute error (MAE)					
OLS	5.145	3.019	2.050	1.875	2.866
FRM	4.968	3.102	2.166	**1.830**	2.836
RE	**4.965**	**2.955**	2.036	1.865	**2.822**
FE	4.992	2.957	**2.027**	1.857	2.826
Simple model	**n=501**	**n=806**	**n=1,402**	**n=1,696**	**n=2,108**
Observed mean	7.892	12.140	15.083	17.968	21.815
Predicted mean					
OLS	13.184	15.005	16.362	17.400	19.069
FRM	13.128	14.997	16.389	**17.419**	19.053
RE	**13.021**	**14.902**	**16.330**	17.435	**19.140**
FE	13.040	14.917	16.335	17.434	19.128
Mean error (ME)					
OLS	-5.292	-2.865	-1.279	0.569	2.746
FRM	-5.236	-2.857	-1.305	0.549	2.763
RE	**-5.129**	**-2.762**	**-1.247**	**0.533**	**2.675**
FE	-5.147	-2.777	-1.252	0.535	2.688
Mean absolute error (MAE)					
OLS	5.313	3.118	2.091	1.858	2.893
FRM	5.262	3.139	2.144	1.859	2.876
RE	**5.157**	**3.058**	2.071	1.860	**2.851**
FE	5.175	3.068	**2.063**	**1.847**	2.859

Legend: values in bold are the most accurate

Table 85: Predictions over the range of ASCOF SCRQoL, 65 and over sub-group

	For regions where ASCOF SCRQoL is…				
	<11	**≥11, < 14**	**≥14, < 17**	**≥17, < 20**	**≥20**
Full model	**n=616**	**n=1,709**	**n=4,152**	**n=6,013**	**n=8,391**
Observed mean	8.468	12.252	15.140	18.035	21.710
Predicted mean					
OLS	14.309	15.791	17.007	**18.169**	19.540
FRM	**13.977**	**15.708**	17.055	18.242	19.505
RE	14.197	15.735	**16.973**	18.172	**19.574**
FE	14.212	15.744	16.978	18.172	19.569
Mean error (ME)					
OLS	-5.841	-3.540	-1.867	**-0.135**	2.170
FRM	**-5.510**	**-3.457**	-1.915	-0.208	2.205
RE	-5.730	-3.483	**-1.833**	-0.138	**2.136**
FE	-5.745	-3.493	-1.838	-0.138	2.141
Mean absolute error (MAE)					
OLS	5.842	3.581	2.223	1.528	2.361
FRM	**5.519**	**3.573**	2.373	1.529	**2.331**
RE	5.730	3.529	2.210	1.532	2.335
FE	5.745	3.537	**2.209**	**1.527**	2.337
Simple model	**n=663**	**n=1,863**	**n=4,534**	**n=6,640**	**n=9,410**
Observed mean	8.439	12.250	15.145	18.041	21.716
Predicted mean					
OLS	14.526	15.965	17.079	**18.207**	19.503
FRM	**14.412**	15.916	17.089	18.237	19.495
RE	14.421	**15.910**	**17.044**	18.210	**19.536**
FE	14.441	15.921	17.048	18.209	19.531
Mean error (ME)					
OLS	-6.087	-3.715	-1.934	**-0.165**	2.213
FRM	-5.973	-3.666	-1.944	-0.196	2.222
RE	**-5.982**	**-3.660**	**-1.900**	-0.169	**2.180**
FE	-6.002	-3.671	-1.904	-0.168	2.185
Mean absolute error (MAE)					
OLS	6.087	3.746	2.252	**1.526**	2.392
FRM	5.973	3.714	2.307	1.553	2.368
RE	**5.982**	**3.697**	**2.234**	1.535	**2.367**
FE	6.002	3.705	2.232	1.529	2.370

Legend: values in bold are the most accurate

Table 86: Predictions over the range of ASCOF SCRQoL, care home sub-group

	For regions where ASCOF SCRQoL is…				
	<11	**≥11, < 14**	**≥14, < 17**	**≥17, < 20**	**≥20**
Full model	**n=119**	**n=293**	**n=653**	**n=1,279**	**n=3,366**
Observed mean	8.782	12.188	15.147	18.105	22.027
Predicted mean					
OLS	16.750	17.332	18.196	19.113	20.323
FRM	**16.425**	**17.163**	18.182	19.181	20.326
RE	16.471	17.185	**18.091**	**19.064**	**20.385**
FE	16.492	17.211	18.108	19.069	20.376
Mean error (ME)					
OLS	-7.969	-5.144	-3.049	-1.008	1.704
FRM	**-7.643**	**-4.975**	-3.035	-1.076	1.701
RE	-7.689	-4.997	**-2.944**	**-0.959**	**1.642**
FE	-7.711	-5.024	-2.961	-0.964	1.651
Mean absolute error (MAE)					
OLS	7.969	5.144	3.142	1.662	2.054
FRM	**7.643**	**5.002**	3.231	1.790	**1.994**
RE	7.689	5.003	**3.065**	1.647	2.030
FE	7.711	5.030	3.079	**1.639**	2.028
Simple model	**n=123**	**n=317**	**n=713**	**n=1,421**	**n=3,770**
Observed mean	8.715	12.189	15.144	18.110	22.042
Predicted mean					
OLS	17.976	18.488	19.005	19.492	19.959
FRM	17.976	18.488	19.005	19.492	19.959
RE	**17.633**	**18.266**	**18.848**	**19.406**	**20.051**
FE	17.647	18.277	18.853	**19.406**	20.049
Mean error (ME)					
OLS	-9.261	-6.298	-3.861	-1.382	2.083
FRM	-9.261	-6.298	-3.861	-1.382	2.083
RE	**-8.917**	**-6.077**	**-3.703**	**-1.296**	**1.991**
FE	-8.931	-6.087	-3.708	**-1.296**	1.993
Mean absolute error (MAE)					
OLS	9.261	6.298	3.884	1.737	2.232
FRM	9.261	6.298	3.884	1.737	2.232
RE	**8.917**	**6.077**	**3.733**	1.715	**2.181**
FE	8.931	6.087	3.736	**1.709**	2.182

Legend: values in bold are the most accurate

Appendix 14: Specification of the imputation equations for the respondents to the ASCS data, in the 18 to 64 sub-group with budget data

To ensure appropriate imputation of the missing budget data I carried out multiple imputation on the subset of the data used in the production function analysis. The multiple imputation was therefore restricted to the 18 to 64 sub-group and included only those CASSRs with budget data for some cases within this sub-group. CASSRs with budget data missing for all cases were excluded from the imputation procedure. I followed broadly the same methods set out in Appendix 1 using a chained equations approach. I imputed binary variables using logistic regression (logit); ordinal variables using proportional odds models (ologit); and imputed variables with more than two nominal categories using multinomial logit models (mlogit).

The imputation models included all theoretically-relevant variables for the production function adjustment models. To ensure the plausibility of the MAR assumption, other variables that were not tested for inclusion in the adjustment model but were available in the dataset were also included in the imputation models. (Variables with more than 50 per cent of the cases missing (i.e. secondary client group and religion) were not considered for the imputation models (McKee et al. 1999, Rubin 2003).) However, due to size of the dataset and number of variables with missing data, therefore requiring imputation, a pragmatic approach had to be taken to the specification of the imputation equations to make the imputation computationally feasible. Therefore, particularly in the case of the variables imputed using mlogits, covariates were omitted for the imputation equation where the predictive significance of the variable was found to be less than 0.1. I also omitted variables that were highly collinear or perfectly predicted the outcome (e.g. in the case of helpstat).

A set of dummy variables for the CASSR were included in the imputation models (except those using mlogit due to difficulties with estimation) to capture CASSR effects. The derived variables (i.e. the ADL scales and the SCRQoL measure) were not included in the imputation equations as their derivation could not be specified in the way required by the software. However, imputation of each individual item within the scales should preserve the complexity of the relationships. Additionally not all relevant interactions could be included in the imputation models where these were specified between multi-item measures.

Due to missingness in many of the additional variables included in the imputation models to ensure the plausibility of the MAR assumption, imputation models had to be

specified for these additional variables. All covariates for the adjustment models and all additional variables were included in these models, although in the case of the mlogits a reduced set of variables were used for computational feasibility. Only variables with a p-value of 0.1 or less were included in the models.

To provide valid imputations of the service receipt variables as well as the budget variable, information on CASSR-level service provision and gross adult social care expenditure (The Information Centre for Health and Social Care 2012c, The Information Centre for Health and Social Care 2012a) and population density (Office for National Statistics 2001) were included in the imputation equations for these variables. These authority-level data were introduced to allow for shifts in the relationship between service receipt, budget and the other variables by CASSR, particularly important given some service receipt variables (meals, short-term residential care, personal budgets, professional support, equipment and other services) were completely missing within a number of CASSRs. The area cost adjustment factor was also included to control for differences in prices between areas.

Due to the missingness in the service receipt and gross adult social care expenditure and deprivation variables, these variables also had to be imputed. The continuous variables, including budget, were imputed using predictive mean matching (pmm, with ten nearest neighbours), due to the skewed nature of the distribution for these variables. Given these additional variables are not being used in the adjustment models, the outcome variables were not included in the imputation equations and only CASSR-level variables were included to mitigate excessive CASSR-level variability. All available (i.e. fully observed or in the imputation model) CASSR-level variables were used.

Appendix 15: Additional Information on the Multiple Imputation of the Data on the ASCS Respondents, in the 18 to 64 Sub-Group with Budget Data

Checking convergence of the multiple chains

Prior to imputation proper, I checked for convergence of the multiple chains, by examining plots summarising the distribution (means and standard deviations) of imputed values against iteration numbers. For this purpose I used the chainonly and savetrace options in Stata alongside the mi impute chained command. In addition I examined the behaviour of three chains, each obtained using a different set of initial values, to check the convergence and stability of the algorithm, using the add(3) option instead of chainonly. The variables did not appear to show any trends, and the three chains seem to oscillate around the same point, providing some evidence of convergence of the algorithm. Convergence appeared to be achieved around 20 iterations.

Checking the fit of the imputation models

I compared the imputed values with the observed ones for each of the continuous variables using the user-written command midiagplot (Eddings and Marchenko 2006), to check the fit of the imputation model. They demonstrate that the predictive mean matching does a relatively good job of imputing the skewed distribution of the CASSR-level aggregate service receipt, gross expenditure and budget variables.

For all other variables, to ensure the imputations produce sensible results I compared the distributional statistics calculated on the casewise-deleted sample to those calculated on the multiply-imputed sample for each variable. Table 87 shows the distributional statistics for the outcome indicators; Table 88 shows the same statistics for the risk adjustor and budget variables; and Table 89 shows the statistics for all other variables. The statistics are fairly similar across the complete cases and multiply-imputed samples, suggesting that the imputation procedure is valid.

Table 87: Comparison of distributional statistics for outcome indicators for the 18 to 64 sub-group respondent sample on casewise-deleted and multiply-imputed samples

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE

ASCOF SCRQoL	3,828	17.0	0.072	17.0	0.069
Satisfaction: Ex. Sat	4,013	1,145	28.53	1,192	28.36
Satisfaction: V. sat	4,013	1,206	30.05	1,258	29.95
Satisfaction: Q. sat	4,013	1,053	26.24	1,106	26.33
Satisfaction: Neither	4,013	346	8.62	366	8.71
Satisfaction: Q. dissat	4,013	121	3.02	129	3.07
Satisfaction: V. dissat	4,013	61	1.52	65	1.55
Satisfaction: Ex. dissat	4,013	81	2.02	86	2.04
Control: preferred	4,139	984	23.77	1,002	23.85
Control: needs met	4,139	1,726	41.70	1,749	41.64
Control: some needs	4,139	1,237	29.89	1,255	29.88
Control: high needs	4,139	192	4.64	194	4.63
Pers care: preferred	4,133	1,989	48.12	2,021	48.10
Pers care: needs met	4,133	1,741	42.12	1,769	42.11
Pers care: some needs	4,133	337	8.15	344	8.18
Pers care: high needs	4,133	66	1.60	68	1.61
Food: preferred	4,095	2,424	59.19	2,487	59.20
Food: needs met	4,095	1,250	30.53	1,282	30.52
Food: some needs	4,095	338	8.25	347	8.27
Food: high needs	4,095	83	2.03	85	2.02
Accom: preferred	4,121	2,068	50.18	2,108	50.19
Accom: needs met	4,121	1,547	37.54	1,577	37.53
Accom: some needs	4,121	425	10.31	433	10.31
Accom: high needs	4,121	81	1.97	83	1.96
Safety: preferred	4,138	1,840	44.47	1,866	44.42
Safety: needs met	4,138	1,666	40.26	1,691	40.25
Safety: some needs	4,138	413	9.98	420	10.00
Safety: high needs	4,138	219	5.29	224	5.33
Soc part: preferred	4,131	1,209	29.27	1,230	29.28
Soc part: needs met	4,131	1,397	33.82	1,420	33.79
Soc part: some needs	4,131	1,107	26.80	1,125	26.78
Soc part: high needs	4,131	418	10.12	426	10.14
Occ: preferred	4,116	810	19.68	828	19.72
Occ: needs met	4,116	1,174	28.52	1,199	28.55
Occ: some needs	4,116	1,736	42.18	1,771	42.17
Occ: high needs	4,116	396	9.62	402	9.56
Dignity: preferred	4,005	1,969	49.16	2,054	48.89
Dignity: needs met	4,005	1,270	31.71	1,330	31.65
Dignity: some needs	4,005	681	17.00	724	17.22
Dignity: high needs	4,005	85	2.12	94	2.24
Information: v. easy	4,082	659	16.14	677	16.12
Information: fairly easy	4,082	1,570	38.46	1,613	38.39
Information: fairly diff	4,082	774	18.96	798	18.98
Information: v. diff	4,082	478	11.71	493	11.74

Table 88: Comparison of distributional statistics for adjustment model covariates for the 18 to 64 sub-group respondent sample on casewise-deleted and multiply-imputed samples

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
Male	4,199	1,713	40.80	1,714	40.79
Age: 18-24	4,201	193	4.59	193	4.59
Age: 25-34	4,201	242	5.76	242	5.76
Age: 35-44	4,201	609	14.50	609	14.50
Age: 45-54	4,201	1,161	27.64	1,161	27.64
Age: 55-64	4,201	1,996	47.51	1,996	47.51
Vulnerable person	4,201	148	3.52	148	3.52
ADLs diff score	3,913	3.70	0.04	3.68	0.04
ADLs can't score	3,913	1.48	0.03	1.46	0.03
ADL, indoors: can	4,065	1,736	42.71	1,797	42.77
ADL, indoors: diff	4,065	1,693	41.65	1,746	41.56
ADL, indoors: can't	4,065	636	15.65	658	15.66
ADL, bed/chair: can	4,061	1,533	37.75	1,596	37.99
ADL, bed/chair: diff	4,061	1,598	39.35	1,648	39.22
ADL, bed/chair: can't	4,061	930	22.90	957	22.79
ADL, feed self: can	4,073	2,801	68.77	2,894	68.88
ADL, feed self: diff	4,073	922	22.64	947	22.54
ADL, feed self: can't	4,073	350	8.59	360	8.58
ADL, wash: can	4,078	1,136	27.86	1,176	27.98
ADL, wash: diff	4,078	1,436	35.21	1,477	35.17
ADL, wash: can't	4,078	1,506	36.93	1,548	36.85
ADL, un/dress: can	4,062	1,365	33.60	1,422	33.84
ADL, un/dress: diff	4,062	1,485	36.56	1,532	36.46
ADL, un/dress: can't	4,062	1,212	29.84	1,248	29.70
ADL, WC/toilet: can	4,058	2,172	53.52	2,252	53.61
ADL, WC/toilet: diff	4,058	1,095	26.98	1,132	26.95
ADL, WC/toilet: can't	4,058	791	19.49	817	19.44
ADL, face/hands: can	4,081	2,732	66.94	2,814	66.98
ADL, face/hands: diff	4,081	822	20.14	846	20.14
ADL, face/hands: can't	4,081	527	12.91	541	12.88
IADL, finance: can	4,060	1,380	33.99	1,438	34.22
IADL, finance: diff	4,060	1,102	27.14	1,141	27.15
IADL, finance: can't	4,060	1,578	38.87	1,623	38.63
SPH: very good	4,114	234	5.69	238	5.66
SPH: good	4,114	754	18.33	770	18.33
SPH: fair	4,114	1,649	40.08	1,688	40.17
SPH: bad	4,114	1,036	25.18	1,057	25.16
SPH: very bad	4,114	441	10.72	449	10.68
pain: none	4,066	727	17.88	755	17.97
pain: moderate	4,066	1,998	49.14	2,068	49.23
pain: extreme	4,066	1,341	32.98	1,378	32.80
anx/dep: none	4,050	1,385	34.20	1,443	34.36
anx/dep: moderate	4,050	2,034	50.22	2,103	50.05
anx/dep: extreme	4,050	631	15.58	655	15.59
SPHD: meets needs	4,070	1,380	33.91	1,434	34.15

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
SPHD: most needs	4,070	1,496	36.76	1,541	36.67
SPHD: some needs	4,070	891	21.89	914	21.76
SPHD: inappropriate	4,070	303	7.44	312	7.42
PH, in h'hold	4,028	2,041	50.67	2,119	50.44
PH, outside h'hold	4,028	1,492	37.04	1,559	37.10
PH, none	4,028	929	23.06	974	23.19
No assistance	3,926	1,757	44.75	1,899	45.20
Proxy	3,926	92	2.34	99	2.36
Assistance	3,926	2,077	52.90	2,203	52.44
Assist, in h'hold	3,908	884	22.62	939	22.36
Assist, out h'hold	3,908	649	16.61	700	16.66
Assist, care worker	3,908	430	11.00	462	11.00
Assist, read	3,926	1,241	31.61	1,316	31.33
Assist, translate	3,926	340	8.66	363	8.64
Assist, write	3,926	1,122	28.58	1,172	27.90
Assist, talk through	3,926	911	23.20	967	23.01
Add, own money	3,968	966	24.34	1,023	24.35
Add, family pays	3,968	236	5.95	252	5.99
Add, none	3,968	2,832	71.37	2,997	71.34
Budget (logged, base e)	3,208	8.4	0.029	8.2	0.037

Table 89: Comparison of distributional statistics for variables included to improve the MAR assumption, on casewise-deleted and multiply-imputed samples for the 18 to 64 sub-group respondent sample

	Number of complete cases	Casewise		Imputed	
		Freq / Mean	Percent / SE	Freq / Mean	Percent / SE
Home care	4,186	1,490	35.6	1,494	35.6
Day care	4,174	477	11.4	482	11.5
Meals	4,078	99	2.4	102	2.4
Direct Payment	4,135	26	0.6	28	0.7
Professional support	4,112	1,146	27.9	1,168	27.8
Short-term residential	3,888	658	16.9	697	16.6
Equipment	4,077	1,382	33.9	1,421	33.8
Other services	4,005	272	6.8	287	6.8
Ethnicity: white	4,140	3,429	82.8	3,480	82.8
Ethnicity: mixed	4,140	42	1.0	43	1.0
Ethnicity: black	4,140	319	7.7	322	7.7
Ethnicity: Asian	4,140	262	6.3	266	6.3
Ethnicity: other	4,140	88	2.1	89	2.1
QoL: so good	4,137	135	3.3	137	3.3
QoL: very good	4,137	674	16.3	685	16.3
QoL: good	4,137	1,015	24.5	1,030	24.5
QoL: alright	4,137	1,573	38.0	1,596	38.0
QoL: bad	4,137	469	11.3	477	11.4
QoL: very bad	4,137	172	4.2	175	4.2
QoL: so bad	4,137	99	2.4	101	2.4
SP help effect: better	4,033	2,031	50.4	2,106	50.1
SP help effect: none	4,033	1,100	27.3	1,146	27.3
SP help effect: sl worse	4,033	782	19.4	821	19.5
SP help effect: worse	4,033	120	3.0	128	3.0
Outdoor access: all places	4,038	968	24.0	1,008	24.0
Outdoor access: difficult	4,038	1,727	42.8	1,794	42.7
Outdoor access: unable	4,038	969	24.0	1,008	24.0
Outdoor access: not leave	4,038	374	9.3	391	9.3

Checking sufficiency of the number of imputations

Following imputation, I carried out checks to ensure the number of imputations was sufficient for reporting the outcome indicators and key variables used in the adjustment models, using the methods set out in Appendix 2.

Sufficiency of the number of imputations for reporting mean statistics

Table 90 shows the estimates of the Monte Carlo error (MC error) for the mean for each of the outcome variables for the respondent sample. Table 91 breaks down the variance into its between- and within-imputation components and provides estimates of the FMI and the relative efficiency of the estimates. The statistics in both tables imply that 30 imputations are sufficient for reporting sample means for the outcome indicators. Specifically, the MC error of the mean is less than ten per cent of the standard error of the mean, the relative efficiency of 30 imputations compared to an infinite number of imputations is over 99% and the FMI is <0.3. The smallest degrees of freedom and largest increase in the SE are found for the satisfaction and information indicators, but statistics for these suggest that 30 imputations is more than sufficient.

Table 90: Monte Carlo estimates of the mean for all outcome indicators and effect of imputation on the standard error of the mean (18 to 64 sub-group with budget data)

Variable	MC error of mean	MC error of mean/SE *100	Degrees of freedom	% increase in SE
ASCOF SCRQoL	0.001	1.468	4,143	0.34
Satisfaction indicator	0.001	4.219	2,801	2.88
Control indicator	0.000	1.859	4,086	0.54
Safety indicator	0.000	1.900	4,078	0.56
Information indicator	0.000	2.420	3,937	0.92

Table 91: Imputation variance and efficiency associated with the mean for each outcome indicator (18 to 64 sub-group with budget data)

	Imputation variance					Relative efficiency
Variable	Within	Between	Total	RVI	FMI	
ASCOF SCRQoL	0.005	0.000	0.005	0.007	0.007	1.000
Satisfaction indicator	0.000	0.000	0.000	0.058	0.055	0.998
Control indicator	0.000	0.000	0.000	0.011	0.011	1.000
Safety indicator	0.000	0.000	0.000	0.011	0.011	1.000
Information indicator	0.000	0.000	0.000	0.018	0.018	0.999

Sufficiency of the number of imputations for reporting case-mix adjustment models

I estimated the MC error for the mean of each model parameter for each of the case-mix adjustment models estimated on the 18 to 64 sub-group with budget data. I also calculated the FMI and the relative efficiency of the estimates, using the Stata routines as previously discussed. This analysis suggested that 30 imputations were sufficient for reporting all the model coefficients. Specifically, the MC error of the coefficient mean is less than ten per cent of the standard error of the mean, the relative efficiency of 30 imputations compared to an infinite number of imputations is over 99% and the FMI is <0.3 for all coefficients for the risk-adjustment models. For the production function models the efficiency is similar very high for all coefficients and the FMI is <0.3 for all variables, except the budget variable for the FE estimator with the EP specification where the FMI is 0.31. For all of the production function models the FMI is fairly high for the budget variable being around .26 to .28. This suggests that 30 imputations are sufficient for using the estimates from these models.

Appendix 16: Results of the Risk-Adjustment and Production Function Models Estimated on the Multiply-Imputed Dataset

In this section I present the coefficient estimates for the production function and risk-adjustment models for the ASCOF SCRQoL indicators, estimated by CCA and on the MI dataset (see Appendix 14 and 15 for further details of the imputation method). For the models estimated on the multiply-imputed dataset, I exclude cases from the estimation for which the ASCOF SCRQoL indicator was missing in the original dataset. I show the estimates from the three different statistical estimators and the two different strategies for selection of covariates. The data in this section are presented by model type (production function or risk-adjustment) and by whether the predictors included in the model are all the significant variables (SV model), the simplified set of significant variables (SSV model) or a subset of those with high explanatory power (EP model).

All of the OLS models failed tests for normality of the error terms and homoscedasticity (confirmed on visual inspection of the data). Likewise all of the FE models fail the modified Wald test for groupwise heteroscedasticity, as implemented via the stata routine xttest3 (Baum 2000, Greene 2012). All results for these models therefore use the Huber-White sandwich estimator to correct standard errors (Huber 1967a, White 1980).

As Table 92 through Table 97 show, the models estimated on the multiply-imputed dataset are very similar to those estimated on the complete cases dataset.

Table 92: Estimates for ASCOF SCRQoL indicator risk adjustment models, SV model

Variable	Casewise-deleted sample (n=2,516)			Multiply-imputed sample (m=30, n=3,828)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
SPH: good [b]	-1.143*** (0.249)	-1.160*** (0.318)	-1.247*** (0.263)	-1.146*** (0.203)	-1.151*** (0.264)	-1.179*** (0.2)
SPH: fair [b]	-2.215*** (0.243)	-2.225*** (0.307)	-2.268*** (0.271)	-2.114*** (0.201)	-2.117*** (0.256)	-2.143*** (0.208)
SPH: bad [b]	-3.188*** (0.279)	-3.197*** (0.332)	-3.229*** (0.278)	-3.315*** (0.232)	-3.319*** (0.276)	-3.343*** (0.237)
SPH: very bad [b]	-3.981*** (0.369)	-3.989*** (0.377)	-4.003*** (0.404)	-4.031*** (0.311)	-4.03*** (0.317)	-4.059*** (0.334)
SPHD	-1.023*** (0.163)	-1.035*** (0.157)	-1.067*** (0.163)	-1.116*** (0.125)	-1.11*** (0.123)	-1.084*** (0.117)
ADLDIFF	-0.018 (0.048)	-0.02 (0.048)	-0.028 (0.043)	-0.045 (0.041)	-0.045 (0.041)	-0.047 (0.04)
ADLDIFF*SPHD	-0.102** (0.034)	-0.099** (0.032)	-0.089* (0.035)	-0.09** (0.027)	-0.089** (0.026)	-0.087** (0.027)
FINANCES: diff [c]	-0.622*** (0.182)	-0.625*** (0.184)	-0.647*** (0.177)	-0.683*** (0.148)	-0.686*** (0.149)	-0.692*** (0.136)
FINANCES: can't [c]	-0.742*** (0.183)	-0.751*** (0.186)	-0.805*** (0.181)	-0.641*** (0.151)	-0.643*** (0.154)	-0.66*** (0.149)
ANXDEP	-1.607*** (0.121)	-1.613*** (0.113)	-1.643*** (0.132)	-1.668*** (0.101)	-1.666*** (0.092)	-1.664*** (0.104)
PH in home	0.404* (0.241)	0.407 (0.263)	0.419 (0.26)	0.653** (0.192)	0.654** (0.208)	0.658** (0.202)
PH out home	0.273* (0.148)	0.266* (0.147)	0.236 (0.16)	0.289* (0.124)	0.282* (0.123)	0.261* (0.139)
PH in*ADLDIFF	0.07 (0.054)	0.065 (0.055)	0.05 (0.061)	0.045 (0.045)	0.043 (0.046)	0.03 (0.048)
MALE	-0.171 (0.138)	-0.168 (0.137)	-0.161 (0.163)	-0.203* (0.113)	-0.2* (0.112)	-0.19 (0.115)
ASSIST, care worker	0.689** (0.226)	0.680** (0.214)	0.646* (0.256)	0.698*** (0.2)	0.692*** (0.192)	0.668** (0.198)
ASSIST, translate	0.291 (0.255)	0.285 (0.24)	0.26 (0.277)	0.083 (0.218)	0.092 (0.207)	0.107 (0.217)
ASSIST, write down	0.105 (0.165)	0.11 (0.163)	0.129 (0.195)	0.039 (0.142)	0.039 (0.141)	0.046 (0.15)
Constant	22.20*** (0.284)	22.24*** (0.346)	22.38*** (0.264)	22.27*** (0.228)	22.29*** (0.283)	22.308*** (0.212)
Random effects						
σ_u		0.283 (0.120)			0.235 (0.095)	
σ_e		3.325 (0.048)			3.352 (0.039)	
Model statistics						
F-stat	96.18***	76.05***	149.04***	160.47***	159.87***	196.9***
R^2	0.405		0.428	0.426		0.441

Legend: * p<.1; ** p<.01; *** p<.001

Table 93: Estimates for ASCOF SCRQoL indicator production function models, SV model

Variable	Casewise-deleted sample (n=2,516)			Multiply-imputed sample (m=30, n=3,828)		
	OLS, PF	**RE, PF**	**FE, PF**	**OLS, PF**	**RE, PF**	**FE, PF**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
Budget (ln)	0.098* (0.042)	0.098* (0.042)	0.094* (0.046)	0.126** (0.039)	0.126** (0.04)	0.135** (0.045)
Area cost adjustment [a]	-2.853** (1.063)	-2.887** (1.097)	n/a	-2.85** (0.894)	-2.84** (0.904)	n/a
SPH: good [b]	-1.097*** (0.249)	-1.109*** (0.318)	-1.217*** (0.261)	-1.103*** (0.204)	-1.106*** (0.264)	-1.147*** (0.2)
SPH: fair [b]	-2.155*** (0.244)	-2.163*** (0.307)	-2.225*** (0.273)	-2.046*** (0.201)	-2.049*** (0.256)	-2.088*** (0.212)
SPH: bad [b]	-3.103*** (0.28)	-3.111*** (0.333)	-3.159*** (0.278)	-3.216*** (0.234)	-3.218*** (0.277)	-3.241*** (0.241)
SPH: very bad [b]	-3.860*** (0.372)	-3.869*** (0.378)	-3.923*** (0.415)	-3.9*** (0.313)	-3.903*** (0.318)	-3.957*** (0.345)
SPHD	-1.002*** (0.162)	-1.012*** (0.157)	-1.066*** (0.163)	-1.09*** (0.125)	-1.089*** (0.123)	-1.078*** (0.117)
ADLDIFF	-0.038 (0.049)	-0.039 (0.049)	-0.047 (0.044)	-0.068 (0.042)	-0.068* (0.041)	-0.072* (0.04)
ADLDIFF*SPHD	-0.100** (0.034)	-0.098** (0.032)	-0.088* (0.035)	-0.088** (0.027)	-0.088** (0.026)	-0.086** (0.028)
FINANCES: diff [c]	-0.631*** (0.181)	-0.632*** (0.184)	-0.651*** (0.178)	-0.697*** (0.147)	-0.698*** (0.148)	-0.703*** (0.137)
FINANCES: can't [c]	-0.794*** (0.184)	-0.796*** (0.186)	-0.835*** (0.181)	-0.694*** (0.152)	-0.694*** (0.154)	-0.703*** (0.15)
ANXDEP	-1.586*** (0.121)	-1.591*** (0.113)	-1.630*** (0.131)	-1.649*** (0.101)	-1.649*** (0.092)	-1.653*** (0.104)
PH in home	0.428* (0.241)	0.43 (0.263)	0.444* (0.261)	0.677*** (0.193)	0.68** (0.208)	0.706** (0.205)
PH out home	0.24 (0.148)	0.24 (0.147)	0.239 (0.161)	0.254* (0.125)	0.255* (0.123)	0.269* (0.14)
PH in*ADLDIFF	0.068 (0.054)	0.066 (0.055)	0.054 (0.061)	0.045 (0.045)	0.045 (0.045)	0.034 (0.048)
MALE	-0.16 (0.137)	-0.159 (0.137)	-0.159 (0.162)	-0.184 (0.113)	-0.183 (0.113)	-0.175 (0.114)
ASSIST, care worker	0.639** (0.227)	0.634** (0.214)	0.605* (0.257)	0.636** (0.201)	0.633** (0.193)	0.598** (0.201)
ASSIST, translate	0.32 (0.254)	0.313 (0.24)	0.267 (0.277)	0.122 (0.219)	0.125 (0.207)	0.124 (0.219)
ASSIST, write down	0.076 (0.166)	0.08 (0.163)	0.101 (0.197)	0.007 (0.143)	0.007 (0.141)	0.014 (0.15)
Constant	24.38*** (1.193)	24.45*** (1.255)	21.61*** (0.481)	24.24*** (1.021)	24.23*** (1.049)	21.20*** (0.466)
Random effects						
σ_u		0.214 (0.142)			0.149 (0.136)	
σ_e		3.321 (0.047)			3.346 (0.039)	
Model statistics						
F-stat	87.80***	68.68***	141.40***	144.21***	144.01***	184.90***
R^2	0.408		0.429	0.430		0.444

Legend: * p<.1; ** p<.01; *** p<.001

Table 94: Estimates for ASCOF SCRQoL indicator risk adjustment models, Simplified SV model

Variable	Casewise-deleted sample (n=2,517)			Multiply-imputed sample (m=30, n=3,828)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
ADLDIFF	-0.070* (0.03)	-0.071* (0.031)	-0.077* (0.03)	-0.100*** (0.025)	-0.101*** (0.026)	-0.108*** (0.029)
FINANCE: diff [b]	-0.583** (0.179)	-0.587** (0.183)	-0.608*** (0.177)	-0.646*** (0.144)	-0.65*** (0.147)	-0.658*** (0.133)
FINANCE: can't [b]	-0.592*** (0.166)	-0.606*** (0.173)	-0.672*** (0.164)	-0.569*** (0.138)	-0.572*** (0.143)	-0.59*** (0.141)
SPH	-0.991*** (0.079)	-0.990*** (0.077)	-0.980*** (0.085)	-1.024*** (0.067)	-1.024*** (0.065)	-1.027*** (0.073)
SPHD: most [c]	-1.646*** (0.154)	-1.654*** (0.163)	-1.670*** (0.151)	-1.691*** (0.129)	-1.686*** (0.135)	-1.644*** (0.132)
SPHD: some [c]	-3.012*** (0.2)	-3.012*** (0.195)	-2.981*** (0.182)	-3.089*** (0.165)	-3.072*** (0.161)	-3.011*** (0.167)
SPHD: inappropriate [c]	-4.316*** (0.373)	-4.315*** (0.289)	-4.297*** (0.426)	-4.341*** (0.295)	-4.308*** (0.238)	-4.188*** (0.329)
ANXDEP: moderate [d]	-1.402*** (0.147)	-1.412*** (0.155)	-1.438*** (0.153)	-1.336*** (0.122)	-1.334*** (0.129)	-1.327*** (0.125)
ANXDEP: extreme [d]	-3.306*** (0.273)	-3.318*** (0.236)	-3.371*** (0.294)	-3.509*** (0.223)	-3.503*** (0.192)	-3.496*** (0.222)
PH in home	0.613*** (0.143)	0.600*** (0.142)	0.557*** (0.15)	0.72*** (0.118)	0.713*** (0.117)	0.676*** (0.128)
PROXY	-1.002* (0.436)	-0.953* (0.454)	-0.742 (0.455)	-0.584 (0.381)	-0.566 (0.389)	-0.467 (0.398)
ASSIST, care worker	0.682** (0.218)	0.677** (0.208)	0.663** (0.246)	0.667** (0.193)	0.664*** (0.187)	0.654** (0.197)
Constant	22.29*** (0.185)	22.33*** (0.207)	22.39*** (0.184)	22.39*** (0.149)	22.41*** (0.168)	22.41*** (0.158)
Random effects						
σ_u		0.310 (0.114)			0.247 (0.093)	
σ_e		3.331 (0.048)			3.355 (0.039)	
Model statistics						
F-stat	133.46***	106.86***	168.58***	221.35***	225.60***	226.16***
R^2	0.402		0.427	0.425		0.440

Legend: * p<.1; ** p<.01; *** p<.001

Table 95: Estimates for ASCOF SCRQoL indicator production function models, Simplified SV model

Variable	Casewise-deleted sample (n=2,517)			Multiply-imputed sample (m=30, n=3,828)		
	OLS, PF	**RE, PF**	**FE, PF**	**OLS, PF**	**RE, PF**	**FE, PF**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
Budget (ln)	0.098* (0.042)	0.098* (0.042)	0.094* (0.046)	0.126** (0.039)	0.127** (0.039)	0.133** (0.045)
Area cost adjustment [a]	-3.044** (1.057)	-3.087** (1.11)	n/a	-2.986** (0.886)	-2.975** (0.9)	#N/A
ADLDIFF	-0.090** (0.031)	-0.090** (0.032)	-0.093** (0.031)	-0.123*** (0.026)	-0.123*** (0.027)	-0.13*** (0.031)
FINANCE: diff [b]	-0.597*** (0.178)	-0.598** (0.183)	-0.616*** (0.178)	-0.664*** (0.144)	-0.665*** (0.146)	-0.672*** (0.134)
FINANCE: can't [b]	-0.654*** (0.168)	-0.659*** (0.174)	-0.710*** (0.164)	-0.63*** (0.139)	-0.63*** (0.143)	-0.64*** (0.141)
SPH	-0.963*** (0.08)	-0.964*** (0.077)	-0.960*** (0.088)	-0.992*** (0.068)	-0.993*** (0.065)	-0.998*** (0.076)
SPHD: most [c]	-1.613*** (0.154)	-1.623*** (0.163)	-1.662*** (0.151)	-1.661*** (0.129)	-1.661*** (0.134)	-1.627*** (0.132)
SPHD: some [c]	-2.932*** (0.201)	-2.940*** (0.196)	-2.963*** (0.18)	-3.01*** (0.166)	-3.008*** (0.161)	-2.976*** (0.166)
SPHD: inappropriate [c]	-4.240*** (0.374)	-4.248*** (0.289)	-4.288*** (0.423)	-4.247*** (0.296)	-4.241*** (0.238)	-4.168*** (0.328)
ANXDEP: moderate [d]	-1.396*** (0.147)	-1.403*** (0.154)	-1.431*** (0.153)	-1.335*** (0.121)	-1.334*** (0.128)	-1.328*** (0.126)
ANXDEP: extreme [d]	-3.251*** (0.274)	-3.263*** (0.237)	-3.340*** (0.293)	-3.46*** (0.223)	-3.459*** (0.191)	-3.463*** (0.221)
PH in home	0.642*** (0.146)	0.635*** (0.144)	0.597*** (0.154)	0.755*** (0.121)	0.755*** (0.119)	0.734*** (0.132)
PROXY	-1.022* (0.44)	-0.988* (0.453)	-0.744 (0.452)	-0.608 (0.383)	-0.6 (0.388)	-0.485 (0.398)
ASSIST, care worker	0.630** (0.219)	0.627** (0.209)	0.614* (0.248)	0.605** (0.195)	0.602** (0.188)	0.576** (0.2)
Constant	24.67*** (1.166)	24.75*** (1.235)	21.61*** (0.448)	24.49*** (0.992)	24.48*** (1.017)	21.31*** (0.412)
Random effects						
σ_u		0.237 (0.132)			0.152 (0.134)	
σ_e		3.328 (0.048)			3.349 (0.039)	
Model statistics						
F-stat	116.66***	93.23***	156.33***	189.78***	193.76***	203.25***
R^2	0.405		0.428	0.429		0.443

Legend: * p<.1; ** p<.01; *** p<.001

Table 96: Estimates for ASCOF SCRQoL indicator risk adjustment models, EP model

Variable	Casewise-deleted sample (n=2,819)			Multiply-imputed sample (m=30, n=3,828)		
	OLS	**RE**	**FE**	**OLS**	**RE**	**FE**
	B (Robust SE)	*B* (SE)	*B* (Robust SE)	*B* (Robust SE)	*B* (SE)	*B* (Robust SE)
Fixed effects						
SPH: good [a]	-1.258*** (0.228)	-1.267*** (0.303)	-1.348*** (0.225)	-1.233*** (0.2)	-1.241*** (0.266)	-1.277*** (0.191)
SPH: fair [a]	-2.458*** (0.224)	-2.463*** (0.293)	-2.503*** (0.255)	-2.375*** (0.195)	-2.381*** (0.256)	-2.417*** (0.202)
SPH: bad [a]	-3.550*** (0.256)	-3.552*** (0.314)	-3.581*** (0.269)	-3.686*** (0.223)	-3.694*** (0.274)	-3.732*** (0.234)
SPH: very bad [a]	-4.411*** (0.343)	-4.418*** (0.352)	-4.462*** (0.384)	-4.482*** (0.301)	-4.487*** (0.31)	-4.54*** (0.317)
SPHD	-1.286*** (0.141)	-1.288*** (0.133)	-1.291*** (0.125)	-1.338*** (0.122)	-1.335*** (0.117)	-1.316*** (0.104)
ANXDEP: mod [b]	-1.173*** (0.189)	-1.183*** (0.207)	-1.232*** (0.191)	-1.272*** (0.164)	-1.281*** (0.182)	-1.304*** (0.162)
ANXDEP: extreme [b]	-2.659*** (0.399)	-2.673*** (0.336)	-2.737*** (0.452)	-2.781*** (0.351)	-2.778*** (0.295)	-2.752*** (0.394)
SPHD* ANXDEP: mod [c]	-0.203 (0.177)	-0.199 (0.167)	-0.166 (0.16)	-0.122 (0.153)	-0.113 (0.146)	-0.089 (0.136)
SPHD* ANXDEP: ext [c]	-0.637* (0.26)	-0.630** (0.21)	-0.607* (0.275)	-0.672** (0.223)	-0.669*** (0.183)	-0.681** (0.232)
Constant	22.21*** (0.196)	22.23*** (0.275)	22.28*** (0.189)	22.25*** (0.172)	22.27*** (0.242)	22.27*** (0.17)
Random effects						
σ_u		0.241 (0.124)			0.258 (0.094)	
σ_e		3.369 (0.045)			3.384 (0.039)	
Model statistics						
F-stat	198.6***	155.45***	237.48***	290.44***	289.66***	296.18***
R^2 [d]	0.395		0.416	0.415		0.431

Legend: * p<.1; ** p<.01; *** p<.001

Table 97: Estimates for ASCOF SCRQoL indicator production function models, EP model

Variable	Casewise-deleted sample (n=2,819)			Multiply-imputed sample (m=30, n=3,828)		
	OLS, PF *B* (Robust SE)	**RE, PF** *B* (SE)	**FE, PF** *B* (Robust SE)	**OLS, PF** *B* (Robust SE)	**RE, PF** *B* (SE)	**FE, PF** *B* (Robust SE)
Fixed effects						
Budget (ln)	0.045 (0.037)	0.046 (0.038)	0.051 (0.042)	0.052 (0.036)	0.052 (0.037)	0.056 (0.041)
Area cost adjustment [a]	-3.053** (1.001)	-3.079** (1.011)	n/a	-3.101*** (0.889)	-3.093** (0.911)	n/a
SPH: good [b]	-1.222*** (0.228)	-1.228*** (0.303)	-1.331*** (0.227)	-1.205*** (0.2)	-1.21*** (0.266)	-1.262*** (0.193)
SPH: fair [b]	-2.414*** (0.225)	-2.418*** (0.293)	-2.480*** (0.258)	-2.339*** (0.196)	-2.344*** (0.257)	-2.394*** (0.206)
SPH: bad [b]	-3.502*** (0.257)	-3.505*** (0.314)	-3.549*** (0.273)	-3.65*** (0.224)	-3.655*** (0.274)	-3.696*** (0.238)
SPH: very bad [b]	-4.355*** (0.344)	-4.361*** (0.353)	-4.439*** (0.389)	-4.432*** (0.302)	-4.439*** (0.31)	-4.518*** (0.321)
SPHD	-1.282*** (0.141)	-1.284*** (0.133)	-1.296*** (0.125)	-1.329*** (0.122)	-1.33*** (0.117)	-1.321*** (0.104)
ANXDEP: mod [d]	-1.195*** (0.189)	-1.199*** (0.206)	-1.236*** (0.191)	-1.288*** (0.165)	-1.291*** (0.182)	-1.312*** (0.163)
ANXDEP: extreme [d]	-2.665*** (0.401)	-2.671*** (0.336)	-2.728*** (0.453)	-2.781*** (0.352)	-2.779*** (0.294)	-2.748*** (0.395)
SPHD* ANXDEP: mod [e]	-0.181 (0.178)	-0.179 (0.166)	-0.159 (0.159)	-0.104 (0.152)	-0.101 (0.146)	-0.082 (0.137)
SPHD* ANXDEP: ext [e]	-0.602* (0.26)	-0.601** (0.21)	-0.600* (0.275)	-0.647** (0.223)	-0.648*** (0.183)	-0.675** (0.233)
Constant	25.00*** (1.106)	25.03*** (1.139)	21.83*** (0.438)	25.04*** (0.992)	25.04*** (1.034)	21.78*** (0.41)
Random effects						
σ_u		0.157 (0.168)			0.165 (0.131)	
σ_e		3.367 (0.045)			3.383 (0.039)	
Model statistics						
F-stat	165.84***	128.13***	213.77***	234.20***	234.60***	254.82***
R^2	0.398		0.416	0.415		0.431

Legend: * p<.1; ** p<.01; *** p<.001

Appendix 17: Distribution of Predictions from Risk-Adjustment and Production Function Models

The relatively poor performance of all the models in predicting the extremes of the distribution is shown more clearly in Table 98. The mean error and mean absolute error are both largest for the categories that are farthest away from the mean of the distribution. There is no clear 'best' regression approach for the EP models, but for the SV and Simplified SV models, the production function model estimated by FE regression appears to be the best by a small margin. All predictions are also within range for this estimation sample, which was not the case for the larger sample used in Chapter 6.

Table 98: Predictions over the range of ASCOF SCRQoL for risk adjustment and production function models

	For regions where ASCOF SCRQoL is…				
	<11	**≥11, < 14**	**≥14, < 17**	**≥17, < 20**	**≥20**
SV model	**n=189**	**n=308**	**n=521**	**n=702**	**n=796**
Observed mean	7.974	12.172	15.106	18.010	21.824
Predicted mean					
OLS	13.127	15.202	16.544	17.401	19.024
RE	12.912	15.050	16.509	17.458	19.106
FE	12.921	15.044	16.499	17.458	19.113
OLS, PF	13.102	15.186	16.530	17.413	19.035
RE, PF	12.930	15.060	16.501	17.462	19.100
FE, PF	**12.908**	**15.042**	**16.491**	**17.463**	**19.118**
Mean error (ME)					
OLS	-5.153	-3.030	-1.438	0.609	2.800
RE	-4.939	-2.878	-1.403	0.552	2.718
FE	-4.948	-2.871	-1.394	0.552	2.711
OLS, PF	-5.129	-3.014	-1.424	0.597	2.790
RE, PF	-4.957	-2.888	-1.395	0.548	2.724
FE, PF	**-4.935**	**-2.870**	**-1.385**	**0.547**	**2.706**
Mean absolute error (MAE)					
OLS	5.165	3.189	2.081	1.879	2.970
RE	4.954	3.095	2.070	1.867	2.900
FE	4.962	3.084	2.064	1.868	2.898
OLS, PF	5.143	3.191	2.074	1.884	2.958
RE, PF	4.973	3.103	**2.063**	**1.856**	2.901
FE, PF	**4.952**	**3.089**	2.066	1.865	**2.893**
Simplified SV model	**n=189**	**n=308**	**n=521**	**n=702**	**n=797**
Observed mean	7.974	12.172	15.106	18.010	21.827
Predicted mean					
OLS	13.205	15.212	16.524	17.387	19.033
RE	12.984	15.057	16.489	17.444	19.117
FE	12.978	15.052	16.485	17.447	19.122
OLS, PF	13.174	15.196	16.511	17.399	19.044
RE, PF	13.001	15.068	16.482	17.448	19.111
FE, PF	**12.966**	**15.049**	**16.477**	**17.452**	**19.127**
Mean error (ME)					
OLS	-5.231	-3.040	-1.418	0.623	2.793
RE	-5.011	-2.885	-1.384	0.566	2.709
FE	-5.005	-2.880	-1.379	0.563	2.705
OLS, PF	-5.201	-3.024	-1.405	0.611	2.782
RE, PF	-5.027	-2.896	-1.377	0.562	2.716
FE, PF	**-4.993**	**-2.877**	**-1.371**	**0.558**	**2.700**
Mean absolute error (MAE)					
OLS	5.244	3.153	2.073	1.893	2.980
RE	5.029	3.052	2.060	1.884	2.909
FE	5.020	**3.048**	2.057	1.885	2.908
OLS, PF	5.216	3.159	2.066	1.897	2.961
RE, PF	5.047	3.060	**2.053**	**1.871**	2.909
FE, PF	**5.011**	3.050	2.058	1.882	**2.901**

	For regions where ASCOF SCRQoL is…				
	<11	**≥11, < 14**	**≥14, < 17**	**≥17, < 20**	**≥20**
EP model	**n=216**	**n=332**	**n=586**	**n=776**	**n=909**
Observed mean	7.963	12.154	15.121	17.994	21.814
Predicted mean					
OLS	13.224	15.232	16.591	17.434	18.969
RE	13.033	15.089	16.555	**17.476**	19.055
FE	**13.040**	15.098	16.556	17.473	19.052
OLS, PF	13.202	15.223	16.590	17.435	18.978
RE, PF	13.067	15.101	16.553	17.472	19.047
FE, PF	**13.040**	**15.097**	**16.552**	17.471	**19.056**
Mean error (ME)					
OLS	-5.261	-3.078	-1.470	0.559	2.845
RE	**-5.070**	**-2.935**	-1.434	**0.518**	2.759
FE	-5.077	-2.944	-1.434	0.521	2.762
OLS, PF	-5.239	-3.069	-1.468	0.558	2.836
RE, PF	-5.104	-2.947	-1.432	0.522	2.767
FE, PF	-5.077	-2.944	**-1.431**	0.522	**2.758**
Mean absolute error (MAE)					
OLS	5.285	3.276	2.104	1.848	2.976
RE	**5.117**	**3.185**	2.107	1.849	**2.910**
FE	5.123	3.189	2.102	**1.843**	2.914
OLS, PF	5.270	3.277	2.102	1.855	2.968
RE, PF	5.149	3.187	**2.099**	**1.843**	2.915
FE, PF	5.126	3.192	2.103	1.845	**2.910**

Legend: values in bold are the most accurate

Appendix 18: Effect of Adjustment on Performance Assessment Using ASCOF SCRQoL Indicators

The correlation statistics for the ASCOF SCRQoL PI for CASSRs estimated using different regression methods are shown for the three sub-groups in Table 99 to Table 101 for the complete case analysis. The correlation statistics for the ASCOF SCRQoL PI for CASSRs estimated using different methods for indirect standardisation are shown for the three sub-groups in Table 102 to Table 104 for the complete case analysis. Correlations are very high, being above 0.99 in all cases. Kendall's tau is also very high, with values >.90.

Table 99: Correlation statistics between adjusted indicators†, estimated using different regression methods, care home sub-group (n=148)

Complete case analysis	1a. OLS regression	1b. FR regression	2a. FE regression	2b. RE regression
Pearson's R^2 (p-value)				
1a. OLS regression		1.000	1.000	1.000
1b. FR regression	0.993		1.000	1.000
2a. FE regression	1.000	0.993		1.000
2b. RE regression	1.000	0.993	1.000	
Rho (p-value)				
1a. OLS regression		1.000	1.000	1.000
1b. FR regression	0.991		1.000	1.000
2a. FE regression	0.999	0.991		1.000
2b. RE regression	1.000	0.991	1.000	
Tau (p-value)				
1a. OLS regression		1.000	0.996	0.997
1b. FR regression	0.923		0.996	0.997
2a. FE regression	0.985	0.925		0.998
2b. RE regression	0.997	0.922	0.988	

Legend: † Indicators 1a, 1b, 2a and 2b estimated using the error approach i.e. observed – expected; lower quadrant SV models, upper quadrant EP models

Table 100: Correlation statistics between adjusted indicators† for SV and EP covariate sets, estimated using different models, 65 and over sub-group (n=149)

Complete case analysis	1a. OLS regression	1b. FR regression	2a. FE regression	2b. RE regression
Pearson's R^2 (p-value)				
1a. OLS regression		0.999	1.000	1.000
1b. FR regression	0.997		0.999	0.999
2a. FE regression	1.000	0.997		1.000
2b. RE regression	1.000	0.997	1.000	
Rho (p-value)				
1a. OLS regression		0.998	1.000	1.000
1b. FR regression	0.997		0.998	0.998
2a. FE regression	1.000	0.997		1.000
2b. RE regression	1.000	0.997	1.000	
Tau (p-value)				
1a. OLS regression		0.970	0.994	0.997
1b. FR regression	0.955		0.969	0.969
2a. FE regression	0.991	0.954		0.997
2b. RE regression	0.995	0.956	0.996	

Legend: † Indicators 1a, 1b, 2a and 2b estimated using the error approach i.e. observed – expected; lower quadrant SV models, upper quadrant EP models

Table 101: Correlation statistics between adjusted indicators† generated using different regression methods, 18 to 64 sub-group (n=149)

Complete case analysis	1a. OLS regression	1b. FR regression	2a. FE regression	2b. RE regression
Pearson's R^2 (p-value)				
1a. OLS regression		0.998	1.000	1.000
1b. FR regression	0.995		0.998	0.998
2a. FE regression	1.000	0.995		1.000
2b. RE regression	1.000	0.995	1.000	
Rho (p-value)				
1a. OLS regression		0.998	1.000	1.000
1b. FR regression	0.995		0.998	0.998
2a. FE regression	1.000	0.995		1.000
2b. RE regression	1.000	0.995	1.000	
Tau (p-value)				
1a. OLS regression		0.966	0.994	1.000
1b. FR regression	0.941		0.965	0.965
2a. FE regression	0.989	0.944		0.994
2b. RE regression	0.999	0.942	0.990	

Legend: † Indicators 1a, 1b, 2a and 2b estimated using the error approach i.e. observed – expected; lower quadrant SV models, upper quadrant EP models

Table 102: Correlation statistics between adjusted indicators†, estimated using different methods for generating the indicator, care home sub-group (n=148)

Complete case analysis	Average ratio	Individual ratio	Error
Pearson's R^2 (p-value)			
Average ratio		0.998	1.000
Individual ratio	0.994		0.997
Error	1.000	0.994	
Rho (p-value)			
Average ratio		0.997	1.000
Individual ratio	0.991		0.997
Error	1.000	0.991	
Tau (p-value)			
Average ratio		0.958	0.996
Individual ratio	0.930		0.958
Error	0.991	0.930	

Legend: † Indicators based on OLS only; lower quadrant SV models, upper quadrant EP models

Table 103: Correlation statistics between adjusted indicators†, estimated using different methods for generating the indicator, 65 and over sub-group (n=149)

Complete case analysis	Average ratio	Individual ratio	Error
Pearson's R^2 (p-value)			
Average ratio		0.995	1.000
Individual ratio	0.995		0.995
Error	1.000	0.995	
Rho (p-value)			
Average ratio		0.993	1.000
Individual ratio	0.995		0.994
Error	1.000	0.995	
Tau (p-value)			
Average ratio		0.937	0.992
Individual ratio	0.944		0.937
Error	0.992	0.945	

Legend: † Indicators based on OLS only, except EB error; lower quadrant SV models, upper quadrant EP models

Table 104: Correlation statistics between adjusted indicators†, estimated using different methods for generating the indicator, 18 to 64 sub-group (n=149)

Complete case analysis	Average ratio	Individual ratio	Error
Pearson's R^2 (p-value)			
Average ratio		0.983	0.999
Individual ratio	0.986		0.982
Error	0.999	0.984	
Rho (p-value)			
Average ratio		0.979	1.000
Individual ratio	0.981		0.978
Error	1.000	0.981	
Tau (p-value)			
Average ratio		0.886	0.987
Individual ratio	0.892		0.882
Error	0.987	0.888	

Legend: † Indicators based on OLS, except for Eb error; lower quadrant SV models, upper quadrant EP models

The correlation statistics for the ASCOF SCRQoL PI for CASSRs estimated using different methods for indirect standardisation are shown in Table 105 for the complete case analysis. Correlation statistics for the ASCOF SCRQoL PI for CASSRs estimated using different model specificatons are shown in Table 106. The correlation statistics for the ASCOF SCRQoL PI for CASSRs estimated using different regression methods are shown in Table 107 for the complete case analysis.

Table 105: Correlation statistics between adjusted indicators†, with indicators generated using different methods, partial 18 to 64 sub-group with budget data (n=83)

Complete case analysis	Average ratio	Individual ratio	Error
Pearson's R^2 (p-value)			
Average ratio		0.982	0.999
Individual ratio	0.987		0.982
Error	0.997	0.985	
Rho (p-value)			
Average ratio		0.974	0.999
Individual ratio	0.983		0.975
Error	0.996	0.980	
Tau (p-value)			
Average ratio		0.885	0.979
Individual ratio	0.907		0.885
Error	0.961	0.889	

Legend: † Indicators based on OLS, PF; lower quadrant Simplified SV models, upper quadrant EP models

Table 106: Correlation statistics between adjusted indicators† from SV, SSV and EP models, partial 18 to 64 sub-group with budget data (n=83)

Indicator	Covariate comp	Pearson's R^2 (p-value)	Rho (p-value)	Tau (p-value)	% pairs change order
Complete case analysis					
1a. OLS	SV v SSV	0.994	0.993	0.936	3.2%
regression	SV v EP	0.939	0.935	0.790	10.5%
	SSV v EP	0.940	0.933	0.785	10.8%
1b. FE regression	SV v SSV	0.995	0.994	0.942	2.9%
	SV v EP	0.939	0.937	0.800	10.0%
	SSV v EP	0.939	0.934	0.789	10.6%
1c. RE regression	SV v SSV	0.995	0.993	0.938	3.1%
	SV v EP	0.939	0.936	0.791	10.5%
	SSV v EP	0.940	0.934	0.787	10.6%
2a. OLS	SV v SSV	0.994	0.992	0.935	3.3%
production	SV v EP	0.929	0.924	0.772	11.4%
function	SSV v EP	0.929	0.918	0.758	12.1%
2b. FE production	SV v SSV	0.995	0.994	0.944	2.8%
function	SV v EP	0.936	0.931	0.789	10.6%
	SSV v EP	0.936	0.930	0.784	10.8%
2c. RE production	SV v SSV	0.994	0.993	0.935	3.3%
function	SV v EP	0.929	0.925	0.773	11.4%
	SSV v EP	0.929	0.918	0.757	12.2%
Multiply-imputed sample					
1a. OLS	SV v SSV	0.995	0.992	0.936	3.2%
regression	SV v EP	0.982	0.974	0.869	6.6%
	SSV v EP	0.982	0.975	0.868	6.6%
1b. FE regression	SV v SSV	0.996	0.992	0.934	3.3%
	SV v EP	0.982	0.975	0.868	6.6%
	SSV v EP	0.982	0.973	0.864	6.8%
1c. RE regression	SV v SSV	0.995	0.992	0.936	3.2%
	SV v EP	0.982	0.975	0.868	6.6%
	SSV v EP	0.982	0.974	0.867	6.7%
2a. OLS	SV v SSV	0.995	0.993	0.937	3.2%
production	SV v EP	0.972	0.958	0.848	7.6%
function	SSV v EP	0.970	0.957	0.838	8.1%
2b. FE production	SV v SSV	0.996	0.992	0.936	3.2%
function	SV v EP	0.975	0.961	0.846	7.7%
	SSV v EP	0.974	0.962	0.840	8.0%
2c. RE production	SV v SSV	0.995	0.993	0.935	3.2%
function	SV v EP	0.972	0.957	0.847	7.7%
	SSV v EP	0.970	0.957	0.838	8.1%

Legend: † Indicators 1, 2a and 2b estimated using the error approach i.e. observed – expected.

Table 107: Correlation statistics between adjusted indicators† generated using different regression methods, CCA sample, partial 18 to 64 sub-group with budget data (n=83)

Complete case analysis	1a. OLS	1b. FE	1c. RE	2a. OLS, PF	2b. FE, PF	2c. RE, PF
Pearson's R^2 (p-value)						
1a. OLS		1.000	1.000	0.951	1.000	0.950
1b. FE	1.000		1.000	0.951	1.000	0.950
1c. RE	1.000	1.000		0.951	1.000	0.950
2a. OLS, PF	0.951	0.951	0.951		0.950	1.000
2b. FE, PF	1.000	1.000	1.000	0.951		0.949
2c. RE, PF	0.950	0.949	0.949	1.000	0.950	
Rho (p-value)						
1a. OLS		1.000	1.000	0.935	1.000	0.933
1b. FE	0.999		1.000	0.935	1.000	0.933
1c. RE	1.000	1.000		0.935	1.000	0.933
2a. OLS, PF	0.943	0.941	0.943		0.934	1.000
2b. FE, PF	0.998	0.999	0.999	0.943		0.932
2c. RE, PF	0.940	0.939	0.940	1.000	0.941	
Tau (p-value)						
1a. OLS		0.993	0.998	0.820	0.992	0.816
1b. FE	0.987		0.994	0.817	0.999	0.814
1c. RE	0.995	0.990		0.820	0.993	0.816
2a. OLS, PF	0.823	0.820	0.825		0.817	0.997
2b. FE, PF	0.977	0.987	0.979	0.825		0.814
2c. RE, PF	0.820	0.817	0.822	0.997	0.822	

Legend: † Indicators 1a, 1b, 2a and 2b estimated using the error approach i.e. observed – expected; lower quadrant Simplified SV models, upper quadrant EP models