The London School of Economics and Political Science

Understanding adult social care using Large Language Models with administrative records

Sam Rickman

A thesis submitted to the Department of Health Policy of the London School of Economics and Political Science for the degree of Doctor of Philosophy, London, November 2024

Statement of authorship

I certify that the thesis I have presented for examination for the 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). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. 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 82,042 words.

Acknowledgements

First, I would like to thank my supervisors, Jose-Luis Fernandez and Juliette Malley. Jose-Luis, without whom there would have been no project, no data, and no thesis, provided steadfast support, including creating the conditions for me to focus on my thesis in the critical months before submission. Juliette's thoughtful feedback and detailed on the numerous flaws in my many drafts were invaluable. Together, my supervisors' insights and perceptive comments made the methods more robust, the arguments clearer and (I hope) the thesis more coherent and comprehensible. Their time, expertise, and generosity strengthened all aspects of this work, and especially the two papers they co-authored. I am deeply grateful for their guidance throughout this journey.

I also owe a great debt of thanks to Uche Osuagwu, the local authority's Public Health Data Manager, for his tireless dedication to managing the difficult process of data extraction and quality. His efforts ensured that the administrative data used in this research was comprehensive - and our weekly meetings about data quality were more fun than they sound. I am also grateful to Ivor Bucker and William Wood at the Intelligence Solutions for London team, who made vital contributions to data pseudonymisation and Information Governance, without which the data could not have left the local authority's servers. From CPEC, I was enormously helped by Hannah Kendrick's generosity and effort in helping to establish inter-rater reliability among human annotators, Nic Brimblecombe's valuable advice in the final months of this work, and Anji Mehta, who arranged many crucial resources, particularly the GPU-equipped PC that allowed me to run the analysis. I would also like to thank my examiners, Fiona Steele and Paul Clarkson, for their careful reading of my thesis and their insightful questions and comments during the viva and in their written reports. I am grateful for their time and thoughtful engagement with my work.

I would like to acknowledge the individuals whose lives are reflected in the local authority administrative records used in this research. While these records were not created for research, there would have been no thesis without the records and those they describe, and I hope this work might form part of an incremental contribution to improving the way data is used, and ultimately decision-making, in social care services. I also extend my gratitude to the social care practitioners whose detailed documentation made this research possible. I would also like to thank the funding bodies that supported this research: the NIHR Policy Research

Unit in Adult Social Care, the Applied Research Collaboration (ARC) North Thames, and NHS Digital.

I am deeply grateful to the developers and communities behind the open-source software that made this research possible. While there are too many libraries to list individually, I am particularly indebted to the authors of the R and Python languages, the Python `pandas`, `numpy`, `spacy`, `nltk`, and `scikit-learn` libraries, and particularly the Python Hugging Face `transformers` package which provided a consistent and intuitive interface for working with large language models, and the R `data.table`, `kableExtra` and `ggplot2` packages. These tools, alongside Quarto and LaTeX which enabled me to typeset this document and the papers, were fundamental to my analyses and presentation of results. My gratitude also extends to Explosion AI for their generosity in providing a free research licence for their proprietary annotation software, Prodigy. This tool was an essential resource in this research and its availability greatly facilitated the annotation process. I am also especially grateful to Terry Therneau, author of the R `survival` package, for replying to my emails, and for incorporating in an updated version of his package (an improved version of) my code for expressing a formula for stratification by multiple levels in a multi-state Cox Proportional Hazards model.

On a personal note, starting a PhD during a global pandemic was far from easy, and I am deeply grateful to my family for their unwavering support throughout this journey. My mother-in-law, Agnes Munday, provided invaluable help during this time, for which I am immensely thankful. My parents, Gill and Dan, offered an immeasurable amount of emotional and practical support that sustained me through the most challenging period, and without which I would not have reached this stage. I am forever grateful to my wife, Eloise for her patience, and for her insight during my interminable discussions about my findings. She has taken on so much during this process. Her love and encouragement have been a constant source of strength and completing this thesis would not have been possible without her. Finally, my daughter, Frida, for whom this thesis has taken up nearly half her life, has been an endless source of emotional support. Her admirable tolerance of my frequent unavailability and her fresh perspectives and theories — on topics related to this work, like what fairness is, and everything else — have meant more to me than I can express and been a constant source of joy through it all. I am profoundly grateful to all my family.

Preface: personal motivation

When I began my career in social work, I expected that it would mainly involve face-to-face work. However, I quickly found that most of my time consisted of recording. I spent hours each day documenting every conversation, assessment, and intervention. I wrote down why I had made every decision, and defended why I had not taken actions that I might have. Yet despite the time invested in creating these records, it seemed to me that no one was drawing on this mountain of data to understand the bigger picture — what was working, what was not, and how services could be improved.

In my first week working in an adult community social work team, I had to tell a young man, paralysed in a motorcycle accent, that the Independent Living Fund which supported him to live at home was to be shut down. This experience was typical of my time in public services. I had finished my undergraduate degree in the summer of 2008, in the midst of a global financial crisis that had a profound effect on public services internationally [1]. I worked with older adults, as well as individuals with learning disabilities, mental health conditions, and substance use issues in their own homes, care homes, hospitals, and prisons. Between 2010 and 2017, real-terms funding for adult social care declined by approximately 8%, and while spending has increased slightly since then, it remains around 2% lower than in 2010 [2, 3], despite the growing population of people with care needs [4]. Since the 2008 financial crisis, both in England and internationally, there has been a trend towards higher eligibility thresholds and a reduction in the range of services provided [see e.g. 5, 6, 7].

Overall funding is set nationally, but decisions about how to allocate limited resources fall to "street-level" workers [8]. My colleagues and I were responsible for those choices — and for their consequences. It was during these years that I became increasingly frustrated with the lack of information on the impact of funding decisions. We were constantly being asked to do more with less, but no one seemed to know how these cuts were affecting the people we were supposed to be helping. Was cutting one service really saving money, or just shifting costs to another part of the system, like health or housing? It was painfully clear to me that we were making decisions without knowing what the impact would be, because that evidence was not always available.

By the time I became a social work manager, I was acutely aware that despite the vast amounts of documentation we produced, we struggled to find meaningful insights when they were most needed. Social workers in England spend upwards of 60% of their time reading and writing case notes in electronic systems [9]. I was responsible for reviewing years' worth of documentation to make decisions about case allocation and prioritisation, as well as supervising the new notes generated by a team of social workers who recorded every telephone call, home visit and care plan. We were drowning in paperwork, yet we lacked the information to make decisions the next time the budget was reduced.

This frustration led me to consider whether technology could help make sense of these vast records. Could computational methods extract meaningful insights from the unstructured data we generated? I was not alone in this line of thinking; as I worked on this thesis, large language models began to be adopted in social care practice — not just to understand records but to generate them [10]. This thesis represents my effort to evaluate the potential of these tools — incorporate their development as I have been writing it — within the context of social care.

Abstract

This thesis explores Large Language Models (LLMs) for addressing two related challenges in adult social care: the scarcity of information about care recipients with the highest needs, and the significant administrative workload on practitioners documenting service delivery. Using the administrative social care records for 3,046 older adults who were receiving care in a London local authority between 2016 and 2020, the thesis evaluates LLMs for improving access to unstructured data in care records, and for reducing the administrative burden through automation.

The research demonstrates that LLMs can extract valuable information from rich, free text social care case notes — an important development, as survey data often fails to capture those with the highest needs. LLMs can generate a structured indicator of loneliness among older adults, outperforming traditional natural language processing methods. Using this indicator in a survival analysis finds loneliness is a significant predictor of care home entry. These results suggest that, when combined with existing statistical methods, LLMs can improve our understanding of the factors influencing service use in social care.

In addressing LLMs for reducing the administrative workload, the thesis evaluates gender bias in LLMs used for automating documentation tasks, such as generating or summarising case notes. This reveals meaningful differences in gender bias among state-of-the-art language models, emphasising the need for rigorous evaluation when integrating LLMs into social care practice.

As one of the first studies to explore LLMs in social care, this research concludes that LLMs are useful for improving access to information in administrative records that is otherwise unavailable. However, this requires a substantial investment of time and expertise. LLMs hold promise for reducing administrative burdens in social care, but their implementation must be approached cautiously. Ethical and practical considerations around accuracy and bias are essential to ensure these innovative models are used effectively and equitably in social care.

Preliminaries

Statement of co-authored work

Chapter 6 and Chapter 7 were co-authored with my supervisors, Jose-Luis Fernandez and Juliette Malley, of the Care Policy and Evaluation Centre (CPEC) at LSE. As lead author of both chapters, I was primarily responsible for the conceptualisation, including the research questions, choice of analytical approaches, data preparation, code development, methodological design, and drafting. For Chapter 6, I prioritised the area for free text extraction, independently selected and implemented the language models, chose the programming languages and framework, wrote the software, and performed the analysis and evaluation, with oversight from my supervisors on other methodological aspects such as construct validity. For Chapter 7, I explored and evaluated multiple methodological approaches in discussion with my supervisors before selecting and implementing the most appropriate method. I led the design, execution, and interpretation of findings and prepared the original draft of both chapters. Both my supervisors provided valuable input on the research questions, methodology, and the content and structure of drafts. The submitted papers are a collaborative effort.

Current status of research papers

This thesis includes three research papers, all of which had progressed through initial editorial assessment at their respective journals but were not yet published at the time of submission in November 2024. However, two papers had been accepted by the time I prepared the final version of this thesis in March 2025:

1. Rickman, S., Fernandez, J., and Malley, J. (2025). Understanding patterns of loneliness in older long-term care users using natural language processing with free text case notes. *PLOS ONE*. ISSN 1932-6203. DOI: 10.1371/journal.pone.0319745. Accepted for publication February 2025. The version included in Chapter 6 closely reflects the version accepted for publication.
2. Rickman, S., Fernandez, J., and Malley, J. (2025). Loneliness as a risk factor for time to care home entry for older adults receiving community care. *Innovation in Aging*. ISSN 2399-5300. DOI: 10.1093/geroni/igaf010.

Accepted for publication in January 2025. The version included in Chapter 7 is very similar to the version acceptd for publication.

3. Rickman, S. (2024). *Gender bias in large language models in long-term care.* Submitted to *BMC Medical Informatics and Decision Making* in September 2024. It passed initial editorial assessment on 10th October 2024 and a manuscript similar to Chapter 8 remains under peer review in March 2025.

Modified versions of Chapter 6, Chapter 7 and Chapter 8 have been approved by the Department of Health and Social Care (DHSC) as Care Policy and Evaluation Centre (CPEC) Working Papers, and a preprint of Chapter 8 is publicly available [11].

Statement on reproducible research

This thesis is a research project which used individual-level, administrative records from a local authority. The Care Policy and Evaluation Centre (CPEC) at the London School of Economics (LSE) acted as a data processor, rather than data controller, and does not have permission to publish or share these records. Nevertheless, I have tried to ensure that the research is as reproducible as possible and have published two GitHub repositories:

1. Chapter 6: The GitHub repository includes the final classification model and synthetic data, enabling the code to run. It can be applied to large volumes of free text to classify notes as indicating loneliness or social isolation. The repository is permanently archived on Zenodo [12].

2. Chapter 8: The GitHub repository contains the code and synthetic data necessary to replicate the entire analysis. By following the provided instructions, the analysis can be reproduced using either synthetic data or custom data. The repository also extends the analysis to include OpenAI's ChatGPT. It is permanently archived on Zenodo [13].

Table of contents

List of Figures

List of Tables

List of abbreviations

ADLs	Activities of Daily Living
ADASS	Association of Directors of Adult Social Services
AI	Artificial Intelligence
ANOVA	Analysis of Variance
ARC	Applied Research Collaboration
ASC-FR	Adult Social Care Activity and Finance Report
ASC	Adult social care
ASCII	American Standard Code for Information Interchange
ASCS	Adult Social Care Survey
ASR	Automated speech recognition
BART	Bidirectional Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CAF	Common Assessment Framework
CAG	Confidentiality Advisory Group
CES	Center for Epidemiologic Studies (often as CES-D for depression scale)
CI	Confidence Interval
CLD	Client Level Data
CPEC	Care Policy and Evaluation Centre
CRediT	Comprehensive Taxonomy for Authorship Attribution
CSU	Commissioning Support Unit
DHSC	Department of Health and Social Care
DPA	Data Protection Act 2018
DPIA	Data protection impact assessment
DPs	Direct payments
EHRs	Electronic Health Records
ELSA	English Longitudinal Study of Ageing
EU	European Union
FDA	Food and Drug Administration
FN	False negative
FP	False positive
FRS	Family Resources Survey
GB	Gigabyte
GDP	Gross domestic product
GDPR	General Data Protection Regulation
GEE	Generalised Estimating Equations
GloVe	Global Vectors for Word Representation
GPs	General practitioners
GPT	Generative pre-trained transformer
GPU	Graphics Processing Unit
GRUs	Gated recurrent units

GVIF	Generalised variance inflation factor
HES	Hospital episode statistics
HHS	Health and Human Services
HIPAA	Health Insurance Portability and Accountability Act
HR	Hazard ratio
HRS	Health and Retirement Study
HSE	Health Survey for England
IADLs	Instrumental Activities of Daily Living
ICT	Information and communication technology
ID	Identification
IT	Information technology
IV	Instrumental variable
JSON	JavaScript Object Notation
LA	Local authority
LDA	Linear Discriminant Allocation
LGA	Local Government Association
LLMs	Large language models
LOTI	London Office of Technology and Innovation
LSE	London School of Economics and Political Science
LSTM	Long and Short-Term Memory
LTC	Long-term care
MHRA	Medicines and Healthcare products Regulatory Agency
NA	Not applicable (missing data)
NFA	No fixed abode
NHS	National Health Service
NICE	National Institute for Health and Care Excellence
NIHR	National Institute for Health Research
NLP	Natural language processing
NLTK	Natural Language Toolkit
NN	Neural network
NULL	Null value or non-existent
OECD	Organisation for Economic Co-operation and Development
ONS	Office for National Statistics
OT	Occupational Therapist
OTs	Occupational therapists
POW	Production of Welfare
PRU	Policy research unit
PSR	Primary support reason
PSSRU	Personal Social Services Research Unit
QDA	Quadratic Discriminant Analysis
QQ	Quantile-quantile
RCT	Randomised Control Trial
RNNs	Recurrent neural networks
RoBERTa	Robustly Optimised Bidirectional Encoder Representations from Transformers
ROUGE	Recall Oriented Understudy for the Gisting Evaluation
SALT	Short and Long Term Support

SAP	Single Assessment Process
SARs	Safeguarding Adults Reviews
SD	Standard Deviation
SDoH	Social Determinants of Health
SES	Socio-economic status
SIL	Social isolation and loneliness
SNI	Social needs index
SW	Social worker
tf-idf	Term-frequency, inverse document frequency
TN	True negative
TP	True positive
UCLA	University of California, Los Angeles
UK	United Kingdom
UKDA	UK Data Archive
US	United States
USD	US Dollars
WEF	World Economic Forum
WHO	World Health Organization
YOB	Year of birth

1 Introduction: information scarcity and administrative burden in social care

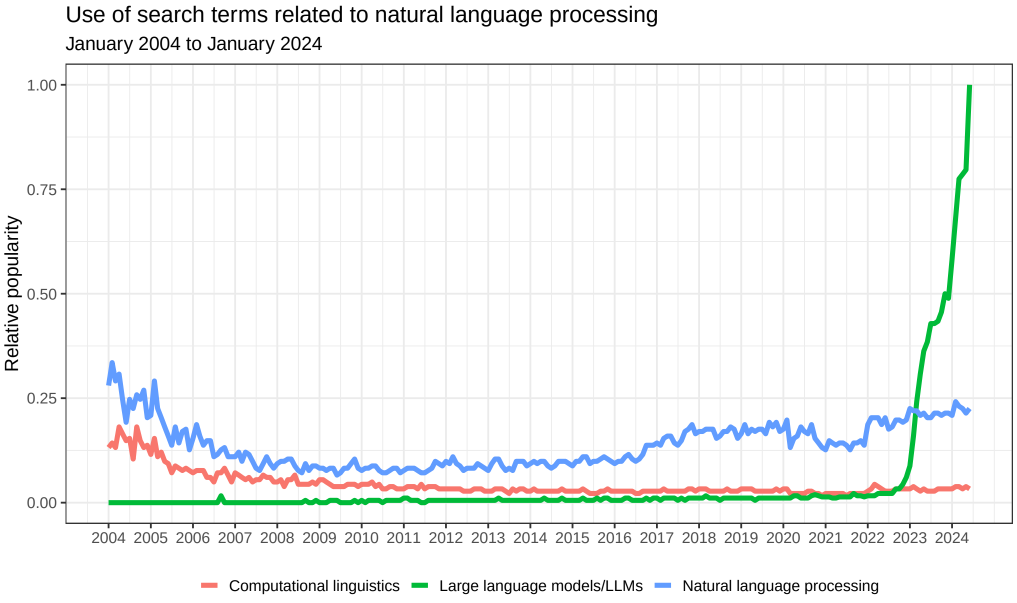
This thesis addresses two related challenges in adult social care: the scarcity of useful information about people receiving care and the heavy administrative workload on staff documenting service delivery. I explore how Large Language Models (LLMs)[1] can be used to address these issues. In July 2024, the London Office of Technology and Innovation (LOTI) published a report entitled *Opportunities for AI in Adult Social Care Services*, outlining potential applications of LLMs in social care. These include providing case summaries, automating documentation, predictive modelling, generating accessible documentation, and developing chatbots for public access to care information [14]. As of November 2024, several local councils in England have piloted or adopted LLMs for these purposes [15, 10].

The common thread among these proposed uses — and the focus of this thesis — is improving decision-making by enhancing access to information held in electronic records. Given the wide range of potential applications, this work concentrates on two key areas. Firstly, I investigate whether free text administrative records can generate insights into how changes in care provision affect people's lives. Local authorities hold vast amounts of free text data on care users that largely remain unanalysed [16]. This thesis examines whether language models can interpret millions of words recorded over thousands of hours to improve our understanding of how to deliver adult social care services more effectively.

Secondly, this thesis is about the use of LLMs in social care practice. This topic has developed during my time researching this area. When I conceived of this thesis in 2019, I used the term *natural language processing* (NLP) to refer to computational analysis of free text. I viewed NLP as a type of statistical method, and did not imagine that language models would be used by social care practitioners. Google search trends data indicates that the term *Large Language Model* (LLM) was not in use until October 2020 [17], and only became widely used after the release of ChatGPT in November 2022, as set out in Figure 1.1. Since then, LLMs have been adopted in England to reduce the administrative burden in adult social

[1] I define the term Large Language Models in Chapter 3. However, for the purpose of this introduction, I use the term LLM to refer to pre-trained, general-purpose language models released in 2019 or later.

care through generating summaries of case notes or writing them automatically [15, 10]. It is an enticing prospect that technology might deliver efficiency savings, freeing up time for social workers and cutting public sector costs, without reducing services. But the accuracy of these summaries is unknown, and there are no established methods for assessing it. Additionally, LLMs can reproduce bias present in their training data [18].



Data from Google Trends

Figure 1.1: Use of search terms related to natural language processing

This thesis examines the emerging use of LLMs in adult social care, focusing on two key areas: research and practice. From the research perspective, it explores how LLMs can extract valuable information from free text records to improve our understanding of care provision and its effects on people's lives. In terms of care practice, it investigates LLMs used to reduce the heavy administrative workload by automating documentation tasks, such as generating summaries of case notes. This introductory chapter outlines the pressing need in social care for better evidence, and more efficient ways to handle the administrative burden. The potential role of LLMs in addressing both challenges is explored in the chapters that follow.

1.1 The need for better social care evidence

Adult social care, internationally often called long-term care, refers to support for people who need help with daily activities such as washing, dressing, and eating. This thesis focuses on the English, statutory adult social care system, where local authorities assess needs and deliver services to around 660,000 people in England, of whom almost 400,000 are aged over 65 [19]. The English context provides a valuable setting for studying data in social care, with recent and future policy initiatives emphasising expanding and improving data use. During the Covid-19 pandemic, health and care data sharing increased in England and internationally [20]. However, a review by the Department of Health and Social Care (DHSC) in December 2021 highlighted significant gaps in the available data on social care service users [21], prompting a push for greater investment in data collection and infrastructure in adult social care [22]. In 2022, DHSC published *Data Saves Lives* [23], outlining £300 million in funding and plans to collect pseudonymised client-level data (CLD) from local authorities. Since March 2024, data from CLD has been published in aggregate form [19], reflecting the ongoing policy focus on using data to enhance care quality and promote independent living [24].

The English context is suited to exploring the use of LLMs in particular, with targeted funding for social care innovation [25] and local councils already using LLMs to try to reduce workloads and support decision-making [15]. An Autumn 2024 survey by the Association of Directors of Adult Social Services (ADASS) indicated that senior local authority managers ranked funding for enhanced digital tools, including AI, as their top priority for improving services [26]. The administrative dataset used in this thesis comes from an English local authority, and comparisons are made with other English data sources, such as surveys, to assess validity and generalisability. While the focus is on English data, the findings have broader international significance. Issues like the administrative burden on care practitioners, or the potential of routinely collected data to address policy questions about care user needs, are global concerns.

Davies and Fernandez [27] wrote that social care decision-makers, "have no evidence about what service packages on average produce what outcomes for persons with different risks and needs" (p369). Understanding the impact of care services requires more than just counting the hours of care delivered; it involves looking at needs, demographics, and individual circumstances [28, 29, 30]. However, the secondary datasets available to researchers for analysing these questions are surveys and nationally aggregated statistics, both of which have limitations [31]. Surveys often exclude those with the highest needs, such as those who lack capacity to participate or those living in care homes, leading to an under-representation of those most dependent on social care services [31, 32, 33]. Survey attrition, particularly among those with declining health, further complicates the reliability of this data for understanding long-term care dynamics such as factors affecting the risk of care home entry [34]. Furthermore, even among statutory care users who are represented in surveys, participants tend to report lower care needs than those seen in administrative data (as I show in Chapter 5).

The other published source of information about older adults using social care services in England comes from statistics published by the Department for Health and Social Care (DHSC) and NHS England [31]. While these sources provide valuable insights, they are typically aggregated at the national level and do not offer the individual-level detail needed for in-depth analysis. Although the individual-level dataset CLD is expected to become available soon in England, the information it will include is minimal, and currently there is insufficient data for assessing how well care services are working [35].

1.2 The administrative burden in social care

Alongside the relative paucity of information for evaluation of social care, social care practitioners face challenges from the volume of documentation required to deliver and manage services. Administrative data, the collection of documents that organisations use to record decisions and activities [36], typically includes individual-level, time-variant information on care needs, planning, and service delivery [37]. In social care, this type of data has been routinely captured in electronic systems since the late 1990s [38]. While this thesis focuses on records collected by local authorities in England for assessing eligibility and managing services, similar systems are used internationally, with most OECD countries employing electronic needs assessments to determine eligibility and record care planning and delivery [39]. This administrative burden is largely due to the need for comprehensive, detailed records [40], but recording, updating, and reviewing case notes takes up the majority of social workers' time [9, 41].

1.2.1 Documentation burden

The issue of documentation burden is not unique to social care. High demands for record-keeping are evident across professions that involve public interaction, professional accountability, and legal oversight. In healthcare, a time-use study found that physicians spent almost twice as much time completing electronic health records (EHRs) as they did seeing patients [42]. Similarly, police officers report frustration with the growing volume of paperwork required for transparency, audit, and accountability, often at the expense of time spent on active policing [43, 44, 45]. Education scholars describe expectations around recording and oversight increasing the administrative burden on academic staff, redirecting time and energy away from teaching and research [46, 47].

In social care, the documentation burden is acute. A study of social workers in England revealed that 57 of 60 participants estimated they spent more than half their time on documentation, while time-use logs found the actual proportion exceeded 60% [9]. Similar findings have been reported internationally, with 78% of Scottish social workers citing a high administrative workload [41] and French social workers identifying record-keeping as their most time-consuming activity [48].

The introduction of electronic health and social care records in England in the 1990s aimed to standardise documentation and broaden its scope [49, 50, 51]. Policy initiatives in the next decade, such as the Single Assessment Process (SAP) [52] and the Common Assessment Framework (CAF) [53], aimed to develop a more consistent and comprehensive assessment process for older people across health and social care [54]. This resulted in the widespread adoption of assessment forms comprised of large numbers of structured fields [55, 56], whose scope was further expanded after the introduction of the Care Act 2014 [57]. Such forms allowed for the quick generation of statistical reports for performance management of workers and monitoring population need [58, 59]. However, as more data was required, structured digital systems were criticised for increasing bureaucracy and reducing time for care management [60, 61].

Additionally, computerised checklists designed for managerial oversight can cause tension with policy goals such as person-centred care [62, 63, 54]. The 2011 Munro Review of child protection social work in the UK recommended prioritising the needs of frontline workers over creating management information when designing ICT-based systems [64], and others have since argued that structured fields in administrative records should be reduced to relieve the data entry burden on practitioners [65].

A 2024 scoping review of good practice in digital social care records found that social care recording systems include narrative free text, preferred by practitioners, and structured fields for audit and performance management [37]. Recording information in structured forms can take longer than entering the same information in free text [66], and social workers have expressed concerns that completing forms detracts from meaningful engagement during face-to-face meetings [67]. This affects morale and turnover; time spent recording is the most significant factor causing social workers to question their future in the profession [67]. While digitising long-term care records remains a policy goal in the UK [22], concerns about the time burden of structured forms are reflected in the UK government's 2023 adult social care data strategy, which stresses that data collection should impose "minimal burden" to "free up staff time for direct care" [68].

Recent shifts in social work practice appear to be leading to an increase in the volume of free text data. Social workers have reported that structured assessments can focus narrowly on functional abilities rather than capturing uncertainty, nuance or the full complexity of social and emotional needs [69, 70, 71, 67]. There has been a move towards strengths-based approaches in adult social care in England, focused on interventions to prevent or reduce care needs before a formal statutory care assessment is completed [72, 73, 74]. Supporters argue that these approaches lower costs, improve care, and lead to case management systems with more free text and fewer structured data fields [75]. However, critics contend that the increased use of free text may not actually reduce the administrative burden [73] and that shifting away from traditional care management risks masking reductions in care, as individuals with unmet needs may never receive a formal assessment [76, 77]. Regardless of the debate, documentation consumes a substantial portion of care management time, with free text recording representing a considerable share of this effort.

1.2.2 Data overload

Although there is a literature on the principles, benefits and costs of social work recording [e.g. 78, 40, 79], there is no published data on the volume of free text in social work records. The dataset used in this thesis sheds some light on the scale of the information recorded. There are 114.4 million total words of free text recorded for the 3,046 individuals in the dataset used in this thesis, who are a subset of older people receiving social care in one local authority between 2015 and 2020. The median number of words in case records for an individual in the administrative dataset used in this thesis was 29,650 and the mean is 37,568. To put this into context, a typical social worker managing around 40 cases [41] might need to sift through around 1.5 million words of free text at any given time. As workers regularly close cases and take on new ones, they constantly need to familiarise themselves with large amounts of information.

While it is impractical, and not usually necessary, for workers to read all this data, a significant amount of time is spent gathering and retrieving information [14, 15]. Recording systems include features to flag important information, but these can be error-prone. Manually set alerts can mean key information is missed, or excessive alerts can lead to “alert fatigue” [80, 81]. While focusing on the most recent records is often manageable, a meta-analysis of Safeguarding Adults Reviews (SARs) in England shows that avoidable harm has occurred when workers were unaware of important details within their records [82].

In healthcare, the term “information overload” is common, but Pohl [83] argues that “data overload” is a more accurate description, as it refers to too much uncategorised data and too little usable information. A meta-analysis of electronic health records found that large volumes of recorded information can become a barrier to providing care [84]. Excessive data, scattered across different systems, may never be retrieved, obstructing effective care [85, 69]. Interfaces which prioritise relevant data improve comprehension and reduce cognitive burden and errors [86, 87, 88].

1.3 Research question: evaluating LLMs for improving decision-making in social care

This thesis explores the overarching questions: Can LLMs improve decision-making in social care by increasing access to information held in care records, and what are the potential challenges associated with their use in generating, summarising and interpreting these records? Social care services face significant problems associated with recording, accessing and using the vast amounts of free text data in administrative records. The growing recognition of the need for better social care data highlights the potential of large, mostly untapped sources like administrative records. This thesis seeks to make the case for addressing gaps in knowledge through the use of computational methods to extract information from free text records, with the ultimate goal of improving both the evaluation and administration of adult social care.

Given the implementation of LLMs and their potential to ease data overload, there is a strong rationale for quantifying their effectiveness in social care settings. This includes assessing their effectiveness in reducing the risks associated with missed or misinterpreted information, identifying new risks created by inaccurate information or bias, and determining whether such issues can be mitigated. It is essential to rigorously assess how accurately LLMs can extract and summarise information from care records. By examining the accuracy of LLMs for extracting information from free text care records, the thesis aims to determine whether they can support decision-making processes.

However, LLMs have a tendency to replicate and even amplify biases present in the data they are trained on, so accuracy alone is not a sufficient condition for the use of LLMs in social care. LLMs learn from vast amounts of text data, which often contain historical biases, stereotypes, and discriminatory language. As a result, the models can internalise these biases and reflect them in their outputs. Bender et al. [18] describe LLMs as “stochastic parrots”, based on their tendency to regurgitate content to which they have been exposed. In social care, the reproduction of such biases is particularly worrying. For example, biased language in case summaries could influence social worker perceptions and decisions, potentially leading to unequal treatment of service users based on gender, race, or other characteristics.

Addressing these ethical issues is challenging. While companies developing LLMs can reduce overtly hateful or abusive language through careful selection of training data and refining models [89], subtler forms of bias can be harder to detect and address [90]. Linguistic differences in how groups are described can reinforce social inequalities [91]. A key part of this thesis is investigating bias in LLMs used in social care settings. By systematically assessing the outputs of LLMs, the thesis aims to find out whether they treat people differently based on gender, which could have serious implications for fairness and equality in care provision.

Moreover, this thesis highlights the need to develop evaluation metrics that can effectively assess both the benefits and potential risks associated with LLMs. The importance of transparency and fairness in algorithms is increasingly recognised in policies and regulations, such as the European Union’s AI Act [92], which stresses the need for oversight mechanisms to reduce risks associated with AI systems. A key thread of this thesis is that while there are advantages of LLMs, there are also drawbacks. If there is a desire for algorithmic fairness, it is essential to establish metrics that can effectively measure both the benefits and the potential risks, such as bias or inaccuracies, posed by these models.

1.4 Data used in the thesis

In this thesis, I use administrative records from a London local authority for all adults who were aged 65 years or over by 31st August 2020, and who had been receiving adult social care services in the community for at least a year at some point since 1st January 2016. These records are typically compiled by social workers, care managers, and occupational therapists employed by local authorities, as part of their professional duty to assess, plan, and monitor the care provided.

Local authorities in England are responsible for organising social care support for their populations. Unlike acute medical care, adult social care is often delivered continuously from eligibility until the end of a person's life [93]. Adult social care services in England encompass a range of support, including domiciliary care, residential and nursing care, and direct payments that allow individuals to arrange their own care [94]. In England, under the Care Act 2014 and related eligibility criteria, every person seeking publicly funded care undergoes an assessment to determine their level of need, which includes the ability to perform activities of daily living (ADLs) and instrumental activities of daily living (IADLs). The documentation collected in the course of needs assessment and service delivery serves multiple purposes: identifying and describing care needs, monitoring care delivery and costs, providing accountability in the case of legal challenges or complaints, demonstrating alignment with professional values, and supporting individuals' access to their own information [40, 67].

While the study draws on records from a single local authority, the questions it explores have a wider relevance. Although the specific local authority is not named here, to protect the anonymity of individual records, I also extensively use survey data throughout the thesis to test the validity of the findings, and demonstrate that the local authority is not an outlier, sharing demographic similarities with other English local authorities. Furthermore, the challenges faced in England — such as the administrative burden in care, targeting support effectively within the constraints of public funding, and understanding the relationship between care provision and outcomes — are common to many countries' long-term care systems [39]. The purpose of this analysis is to inform policy and practice beyond

England, addressing global questions about the use of data for decision-making in long-term care.

1.5 Structure of thesis

This thesis is structured into nine chapters, including this introduction. Chapter 2 makes the case for using administrative records to evaluate social care services, with a particular emphasis on free text. The chapter outlines the data required for evaluation, discusses existing data sources in England, and identifies gaps in current survey and national data collections. Finally, it highlights the opportunities and challenges of using administrative data, focusing on issues such as scalability, access, and reliability.

Chapter 3 examines the role of LLMs in social care in the context of large volumes of free text records. The chapter begins by defining LLMs and reviewing them in contrast to traditional natural language processing (NLP) methods. It then establishes a theoretical framework for understanding the process of transforming unstructured text data into useful information. Finally, the chapter explores how LLMs can be used in social care, including their potential for improving information available for predictive modelling, and for automating and summarising documentation. Given the rapid development of LLMs during the course of this thesis, I cannot address every possible use case. However, this chapter argues that if LLMs are to play a role in public services, it is essential that they be evaluated for accuracy, bias, and ethical considerations.

Chapter 4 details the specific dataset used in this thesis. This chapter has three aims. Firstly, it outlines the demographic characteristics of individuals receiving services, highlighting patterns in age, ethnicity, and support needs. Secondly, it demonstrates the robustness of the findings of the thesis through a detailed examination of data quality, such as handling missing information and inconsistencies in service data. Finally, it discusses the process of transforming administrative records into a format suitable for research, emphasising the work involved in preparing this data for analysis, which is a key finding of the thesis.

Chapter 5 explores more deeply the available survey data. It contains an empirical comparison of administrative data with survey data, to argue that while surveys are invaluable for population-level research, they have significant limitations when capturing information about under-represented groups such as statutory care users or those with the highest needs. This makes the case for using administrative records.

I then present the three research papers which explore different aspects of how LLMs can be used to improve decision-making in social care.

1. **Using Large Language Models to extract loneliness from free text social care notes**

   Chapter 6 explores the use of LLMs to identify loneliness and social isolation in older adults using publicly funded long-term care services. The study focuses on 1.1 million free text case notes from the administrative records of a London council, covering 3,046 older adults. I compare the effectiveness of LLMs against traditional NLP methods, finding that LLMs are much more accurate. This supports the broader conclusions of this thesis, highlighting the new possibilities that LLMs can bring to social care applications.

2. **Loneliness as a risk factor for care home entry**

   Chapter 7 examines the impact of loneliness on time until care home entry, using the loneliness metric extracted with an LLM in the first paper. The study controls for other factors like demographic characteristics, and uses a survival model with competing risks to account for the possibility of death before entering a residential or nursing care facility. The policy implications relate not only to the impact of loneliness on care home entry, but also to the use of LLMs to extract data for evaluating services, used in conjunction with structured data and established econometric methods.

3. **Evaluating gender bias in LLMs for summarising long-term care notes**

   Chapter 8 evaluates gender bias in summarising long-term care records using state-of-the-art, open-source LLMs released in 2024: Meta's Llama 3 and Google's Gemma. The study involved creating two versions of case notes for older adults, which vary only by the gender of the individual receiving care. The summaries generated by Llama 3 and Gemma were then compared to those produced by older benchmark models, T5 and BART, to assess any gender-based variations in the summaries.

Chapter 9 reflects on the research objectives and findings of the thesis, exploring the opportunities LLMs present for improving understanding of services while emphasising the importance of evaluating risks and benefits when using LLMs in social care research and practice. It also discusses the rapid development of LLMs during the research period and implications for future studies, including challenges related to accuracy, bias, and regulation. Finally, it addresses the balance between innovation and regulatory oversight in adopting LLMs for social care, reflecting on the future uses of LLMs with administrative records, considering potential opportunities as well as practical, technical and ethical challenges.

2 Why use administrative records for evaluating social care?

The 2024 report by the London Office of Technology and Innovation (LOTI) on artificial intelligence in social care highlights the potential of LLMs to extract valuable information from free text administrative records for use in predictive models and service evaluations [14]. Despite this promising outlook, there are significant barriers to using administrative records, including difficulties in data access, issues of scalability, and the high transaction costs associated with processing unstructured data. Nevertheless, free text administrative records contain information that is not available from other sources, offering the potential for rich insights into the needs and characteristics of people receiving social care. In this chapter, I will outline the data required to evaluate social care services effectively. Following that, I will describe the currently available data and its gaps. Finally, I will provide an overview of the opportunities and challenges associated with using administrative data, emphasising how advancements like LLMs could address some of these challenges.

2.1 What information is required to evaluate social care services?

In this section, I discuss the key indicators needed to evaluate social care services, describing the existing sources of evidence and the gaps in these sources. Additionally, I examine how different types of data contribute to our understanding of social care needs and outcomes. The purpose of this is to emphasise the importance of information about individual needs and characteristics which are found within administrative records.

2.1.1 Production of Welfare

The Production of Welfare (POW) approach is a theoretical framework which helps identify the key groups of factors which determine social care outcomes and their expected interrelationship [28], as illustrated in Figure 2.1.

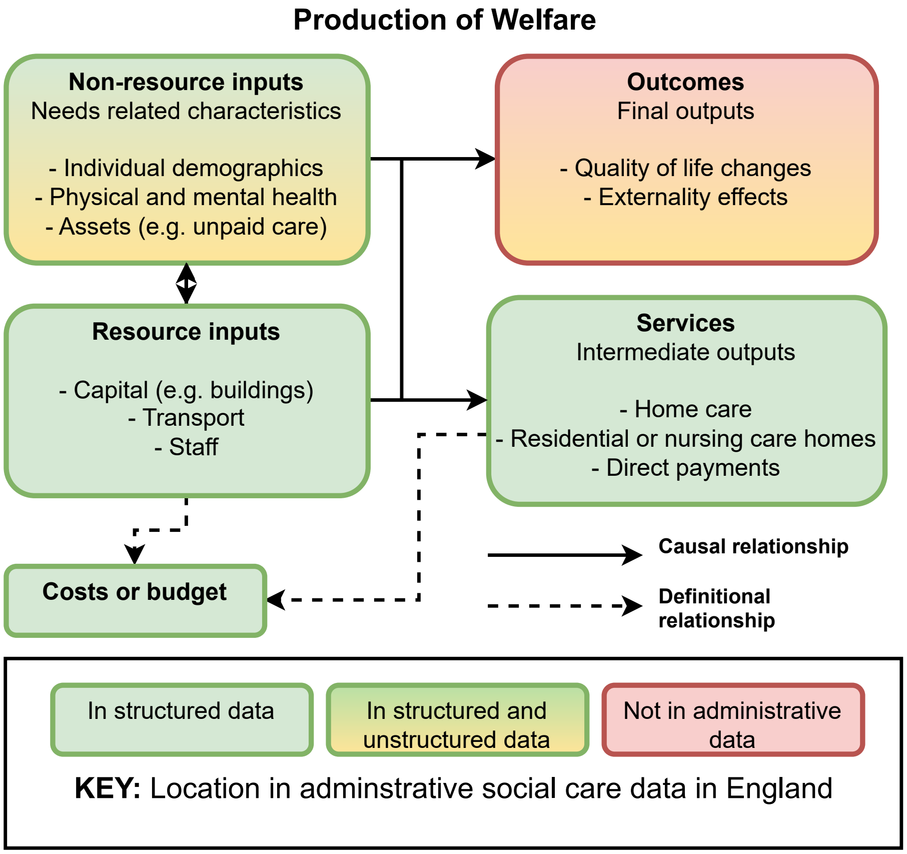

Production of Welfare
Non-resource inputs
Needs related characteristics
- Individual demographics
- Physical and mental health
- Assets (e.g. unpaid care)
Outcomes
Final outputs
- Quality of life changes
- Externality effects
Resource inputs
- Capital (e.g. buildings)
- Transport
- Staff
Services
Intermediate outputs
- Home care
- Residential or nursing care homes
- Direct payments
Costs or budget
Causal relationship
Definitional relationship
In structured data
In structured and unstructured data
Not in administrative data
KEY: Location in adminstrative social care data in England

Figure 2.1: Production of Welfare (adapted from Knapp [28])

I have adapted the figure from Knapp [28] to try to illustrate whether data might be expected to be included in administrative data. This is necessarily an oversimplification as there is no single format for administrative data, and even within individual local authorities the type of data held changes over time. I discuss this in detail in Chapter 4. However, the general point is that there is a considerable amount of the data relevant to evaluating social care services within local authority records, both within structured forms and free text. There is also some information, such as direct measures of quality of life, which one would not generally expect to be contained within administrative records.

In social work literature, need is often defined in terms of expressed demand for care, or a functional impairment related to a specific activity, e.g. an inability to dress independently [95, 96]. However, the POW framework stresses the importance of considering a wider range of need-related characteristics associated with (or mediating) the potential to benefit from social care support. These include functional impairments i.e. ability to complete activities of daily living, personal characteristics, such as ethnicity and gender, and wider factors, including social networks and environmental factors such as housing conditions and local area deprivation. The circumstances of individuals receiving care, which are extensively documented in free text records, are paramount to outcomes [97].

2.1.2 The importance of needs and demographics

The importance of controlling for needs when establishing the effects of resource inputs on outcomes can be demonstrated with an example adapted from Davies and Fernandez [29]. I generate some outcomes data for a group of individuals with high needs, and another whose needs are low. The equation given for the data generation process is,

$$\text{Outcomes}_i = \text{intercept}_i + \text{effectiveness} \cdot (\mathcal{N}(0,1) + \text{inputs}_i)$$

Where

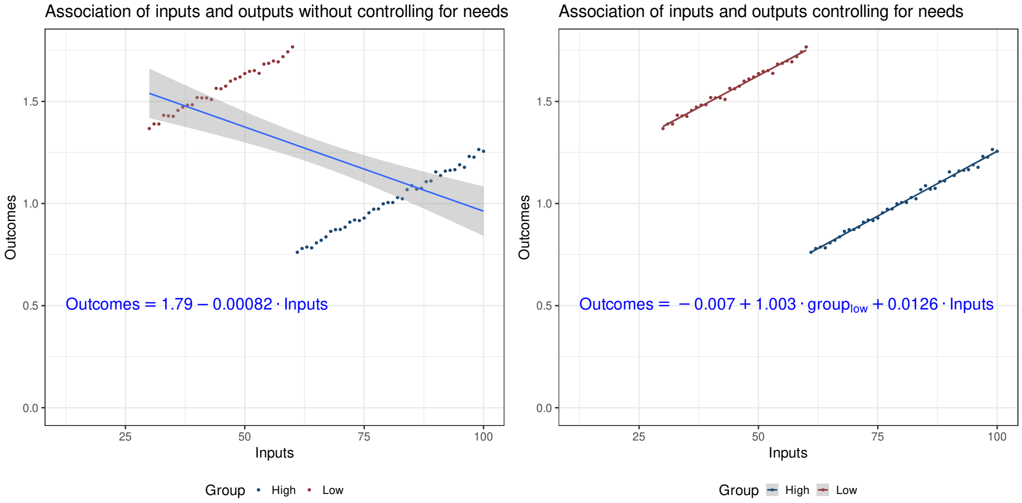
$$\text{intercept}_i = \begin{cases} 0 & \text{if inputs}_i > \tau \\ 1 & \text{if inputs}_i \leq \tau \end{cases}$$

τ is a threshold indicating whether an individual is in the high or needs low group, reflecting that individuals in the lower needs group receive a lesser quantity of services, but that they may have better outcomes related to lower levels of disability. I set $\tau = 60$, effectiveness $= 0.0125$ and inputs $\in [30, 100] \cap \mathbb{Z}$. For reproducibility, I added the random noise defined in $\mathcal{N}(0,1)$ using the R `rnorm()` function immediately after setting the random seed to 1 using `set.seed(1)` [98]. I then ran two regression models, specified as follows:

$$\text{Outcomes}_i = \beta_0 + \beta_1 \cdot \text{inputs}_i + \epsilon_i \qquad \text{(Model 1)}$$

$$\text{Outcomes}_i = \beta_0 + \beta_1 \cdot \text{inputs}_i + \beta_2 \cdot \text{group}_i + \epsilon_i \qquad \text{(Model 2)}$$

A well-specified model should estimate that $\beta_1 \approxeq 0.0125$. However, Model 1, which does not include the needs group, finds a negative (and significant at $\alpha = 0.05$) association between inputs and outcomes, as set out in Figure 2.2a. Model 2, which includes needs, estimates $\beta = 0.0126$, as set out in Figure 2.2b. While this data is contrived, it closely follows a similar specification presented in Davies and Fernandez [29], which was based on the principle that the circumstances and characteristics of individuals have a greater impact on outcomes than quantity of service received. Given this concern about confounding by indication [99], it is crucial that researchers have a broad understanding of individual characteristics when using observational data to assess the effectiveness of interventions in the care system [29, 30, 100].



(a) Model 1

(b) Model 2

Figure 2.2: Comparison of model with and without needs

I summarise in Figure 2.3 key areas of need relevant to the analysis of long-term care using the framework proposed by Abdi et al. [101] for classifying the care and support needs of older adults living with chronic conditions. I include in this figure where such information might be found in administrative data. Structured records primarily include information on functional ability and services received, detailed in Chapter 4. Free text data contains other salient information such as interpersonal relationships, environmental factors and interactions with care professionals. Information not covered by a specific form will be recorded in case notes such as action plans created in monthly supervision meetings, records of case conferences or contact with the person receiving services, family members or other professionals such as care agencies, GPs or mental health services [102].

I outline the contents of free text data in Chapter 4, but the exact composition of unstructured data in case records is not known, as most of it has never been analysed [16].

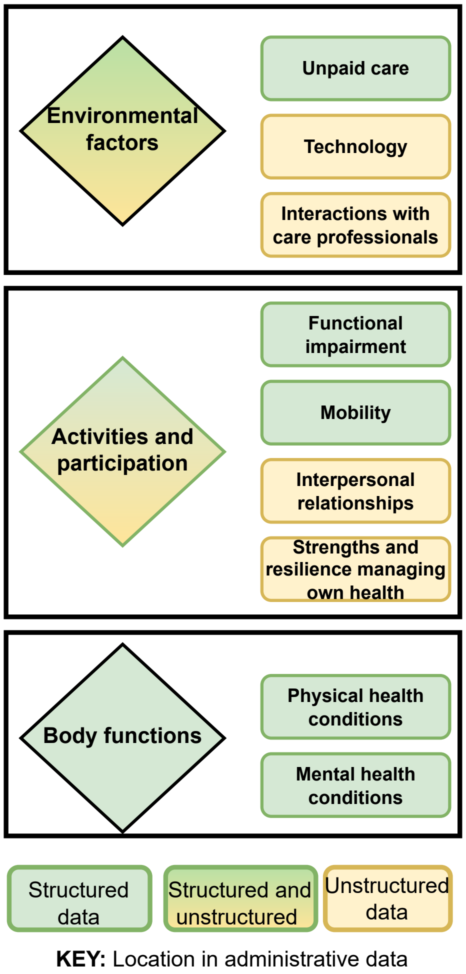

Environmental factors
Unpaid care
Technology
Interactions with care professionals
Activities and participation
Functional impairment
Mobility
Interpersonal relationships
Strengths and resilience managing own health
Body functions
Physical health conditions
Mental health conditions
Structured data
Structured and unstructured
Unstructured data
KEY: Location in administrative data

Figure 2.3: Key characteristics of older long-term care users

2.2 Available social care data sources in England

Once it is accepted that information about the needs, demographics and characteristics of individuals receiving social care services is required to evaluate those services, the question is where such information is available. A central argument of this thesis is that, despite significant barriers to accessing, cleaning, and analysing administrative data, the benefits are substantial. This data not only contains unique information unavailable elsewhere but also covers large populations that would be impractical to recruit in trials or surveys. The dataset that I use in this thesis relates to older people receiving publicly funded adult social care in a London local authority. I outline here the other available data sources for this group, i.e. older adults receiving statutory care in England.

The 2015 scoping review of adult social care data in England by King and Wittenberg [31] is the most recent, comprehensive review of sources of information for English long-term care users. There have been new releases, updates and amendments to all the datasets included since 2015. These have been quite substantial in the case of aggregated national service use data and some surveys, which have added new questions or even modules that are relevant to social care. In this section I will provide an updated summary of the description of the key data sources highlighted in King and Wittenberg [31]. I set out below that aggregated statistics and survey data have many advantages, but also limitations which create the rationale for finding new data sources, such as administrative data.

2.2.1 Survey data

Surveys are commonly used to assess population-level needs, but have limitations when it comes to representing older adults with significant social care needs such as publicly funded care users. Individuals with fewer social and health resources may be less likely to respond to surveys, and the majority of participants in English surveys do not have substantial social care requirements [103]. Under-representation occurs because individuals lacking the capacity to consent to participate — such as those with severe cognitive impairments — are often excluded from research [33], as seen in the Health Survey for England [31]. Even when initial inclusion is possible, participants whose needs increase may be lost to follow-up in longitudinal studies like the English Longitudinal Study of Ageing (ELSA) [32, 34]. Additionally, people residing in care homes are excluded by design from the Family Resources Survey (FRS) and Understanding Society [31, 32, 104, 105], further limiting the representation of those with the highest needs. Surveys can also face issues of under-reporting, where respondents may not accurately disclose their true levels of need [106]. There has been little analysis in the literature examining whether survey participants are representative of social care users. However, I explore in Chapter 5 evidence suggesting that individuals who report receiving statutory care services in surveys are not representative of publicly funded care users, but tend to have systematically lower needs than those reflected in administrative data. The implication for research and policy is that administrative records may provide a more appropriate and comprehensive source of information for understanding the needs and experiences of individuals who use statutory social care services.

2.2.2 National data collections

The other main sources of information about older people who use adult social care services in England are published by the Department for Health and Social Care (DHSC) or NHS England [31]. I include below an updated summary of the national datasets set out in King and Wittenberg [31], as well as new or soon-to-be-published datasets. As in the case of survey data, such sources contain useful information. The main shortcoming is that current data is not published at individual level, and while there will soon be an individual-level dataset, so far the information included within it is limited.

2.2.2.1 Existing national data

National data about adult social care is published by the Department for Health and Social Care. In particular, the Short- and Long-Term Support (SALT) Adult Social Care Activity and Finance Report (ASC-FR) includes information on the number of people receiving social care and related expenditure in England [107]. However, these statistics are aggregated at local authority level. There is not currently a comprehensive, individual-level, time-variant national database for adult social care users comparable to those in health such as Hospital Episode Statistics (HES) [108]. Data in ASC-FR and SALT are banded into age groups of over and under 65, and generally not split by gender.

Data from SALT indicates that there were 226,990 older people receiving publicly-funded care in England in 2022/23. This is around 2.1% of the total population of over 65s in England. However, it is relatively unusual for people in their sixties to need long-term care. It is likely that the proportion of the population receiving statutory care of, for example, over 85s is considerably higher. The mean age in the administrative dataset used in this thesis is around 82, which is close to the mean age for older people receiving care in ELSA and Understanding Society (81 and 80, respectively) [109, 110]. However, it is not possible to use the national data such as SALT to establish the mean age for older people receiving care, as individual-level age is not included. The same applies to other demographic characteristics such as ethnicity and gender, which are in some tables in SALT reported at aggregate level, while in others are excluded entirely. Functional ability is not reported in SALT.

There are also aggregate level statistics published monthly under the Community Services Statistics [111], relating to individuals receiving care in NHS-funded community services, which might be used to capture a picture of demographics and needs in an area. These contain local authority level aggregate numbers of referrals by age band (children, adult or older adult), ethnicity and gender, but also do not contain functional ability. While these datasets are useful for an overview of some care provision, they are not suitable for answering questions about the impact of care needs on service use.

There is individual-level, functional data published in the Adult Social Care Survey (ASCS) dataset published by NHS England [112]. While this survey provides individual-level data, the recruitment process is not representative of care users, and it does not contain weights or other means to account for under-representation of subgroups by age, gender, ethnicity or functional ability [113, 114]. Furthermore, the majority of this data contains questions relating to care outcomes, such as satisfaction with the home environment or care services. It does contain self-reported ADL questions, but this cannot be linked with service receipt data to quantify the impact of needs on service use. Finally, ASCS is cross-sectional, with no way to link individuals over time. There are other national data sources published, such as workforce data, safeguarding concerns and applications for deprivations of liberty. However, these do not contain individual-level functional ability and are not intended to capture the needs of adult social care services.

2.2.2.2 Upcoming national data: Client Level Data

In February 2023, DHSC published *Care data matters: a roadmap for better adult social care data* [68]. This set out plans to implement Client Level Data (CLD), a national collection of pseudonymised, individual-level social care data from all local authorities. Collection of data from local authorities started on a quarterly basis in July 2023. The move from aggregate to individual level information increases opportunities for research. Aggregate statistics have been published since March 2024 [19], but at the time of writing in November 2024, individual-level data is not publicly accessible. However, access to individual-level CLD would not yet make research using administrative records redundant. A technical specification [115] for the data sets out that while it includes demographic information such as gender, ethnicity and age, and service cost information, it does not yet include functional information about ability to complete ADLs and IADLs, or other needs-related information such as cognition, housing or social circumstances. If one wishes to examine, for example, whether loneliness contributes to care home entry (as in Chapter 7), this would require individual-level data on loneliness. It is also necessary to control for functional ability as, for example, it would be expected that loneliness would be correlated with disability. Neither loneliness nor functional ability are included in the CLD schema, so it would not currently be possible to answer such questions with this data.

While CLD represents progress, it is not yet comparable with administrative datasets in health such as HES [116], as the specification includes a minimal subset of the data recorded by local authorities. This seems unlikely to change imminently, owing to the practical challenges in national aggregation of individual-level adult social data at administrative level. The Autumn 2024 ADASS survey identified challenges in standardising information across local authorities, with less than half of councils stating they agreed that the information in CLD accurately reflects their area [26].

2.2.3 Administrative records

Administrative records capture a wide range of information on service use and individual needs. They are less likely than surveys to suffer from limitations such as sampling bias or missing data. Furthermore, social care records contain much more information about factors relevant to care than general population surveys. They also contain free text fields for recording salient information not captured by structured fields. This section explores the types and amounts of data available within these records, highlighting how they complement survey data by offering detailed, longitudinal, and in some cases continuous insights into social care users. I will outline the strengths and limitations of both structured and unstructured administrative data, showing how administrative records, and in particular free text, can fill gaps left by traditional data collection methods.

2.2.3.1 What are administrative records?

Administrative records are collected in the course of exercising statutory duties. In England, the Care Act 2014 s.9 established a duty to carry out a needs assessment of adults who may be eligible for care (although a similar duty existed prior to this under the NHS and Community Care Act 1990 s.47). There is no legally mandated format for the recording of such interventions, although local authorities tend to use an assessment form. Local authority social care data can be recorded by a range of individuals, including social workers, care managers and occupational therapists. Figure 2.4 outlines the assessment and care planning process as described in the Care Act 2014 guidance [94].

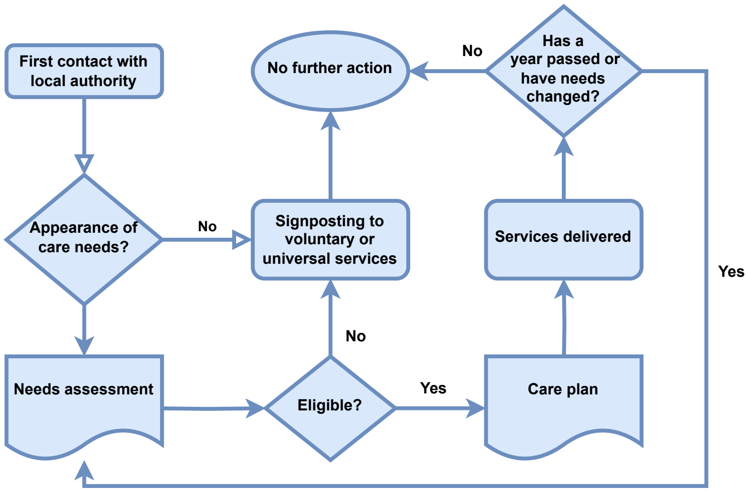
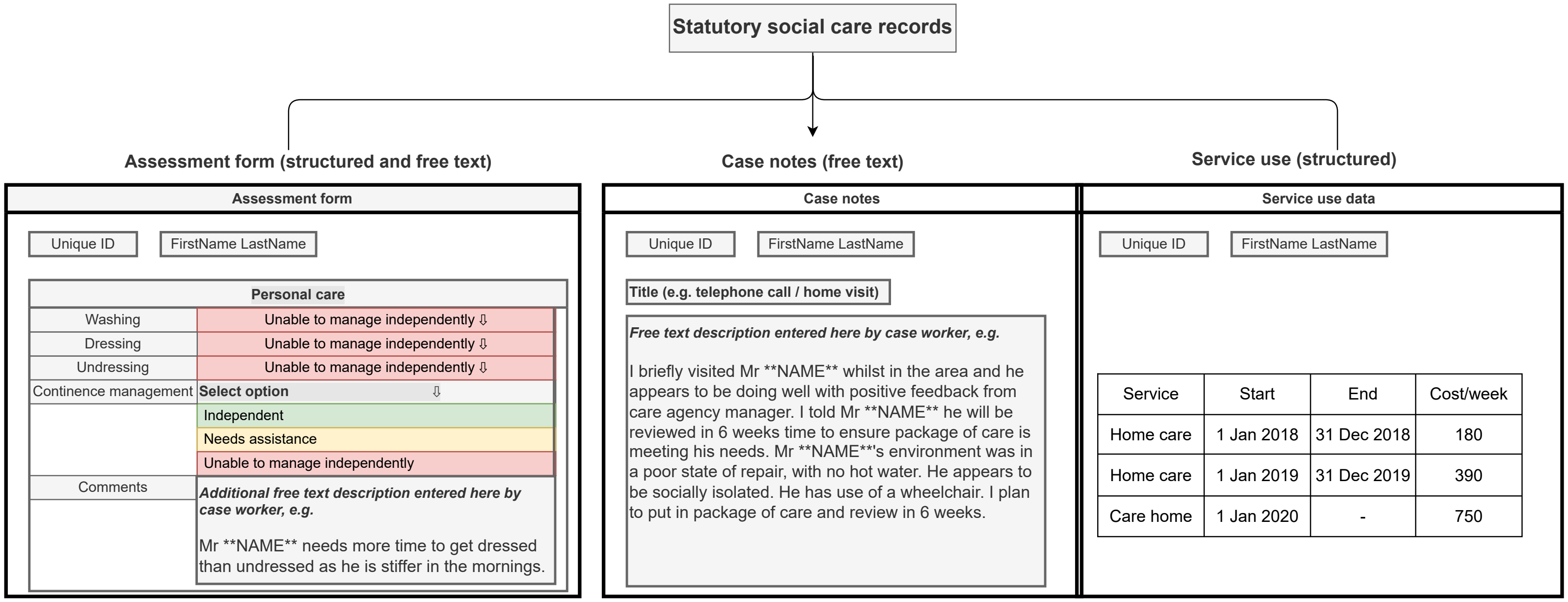

First contact with local authority
No further action
No
Has a year passed or have needs changed?
Appearance of care needs?
No
Signposting to voluntary or universal services
Services delivered
Yes
No
Needs assessment
Eligible?
Yes
Care plan

Figure 2.4: Local authority social care assessment process

Information is recorded using a combination of structured data fields and free text. Structured data is information that is trivially machine-readable, such as a binary gender field, or a numeric field recording age [117]. Free text fields can appear within structured assessment forms, or in separate sections of case management systems to record information not covered by structured forms ("case notes"). While the Care Act 2014 mandates the eligibility criteria, it does not specify how data should be recorded, and the type, format and structure of data collected can vary temporally within local authorities, and between local authorities. I include in Figure 2.5 examples of structured and unstructured data in adult social care case notes.



Additional free text description entered here by case worker, e.g.

Mr **NAME** needs more time to get dressed than undressed as he is stiffer in the mornings.

Free text description entered here by case worker, e.g.

I briefly visited Mr **NAME** whilst in the area and he appears to be doing well with positive feedback from care agency manager. I told Mr **NAME** he will be reviewed in 6 weeks time to ensure package of care is meeting his needs. Mr **NAME**'s environment was in a poor state of repair, with no hot water. He appears to be socially isolated. He has use of a wheelchair. I plan to put in package of care and review in 6 weeks.

Service	Start	End	Cost/week
Home care	1 Jan 2018	31 Dec 2018	180
Home care	1 Jan 2019	31 Dec 2019	390
Care home	1 Jan 2020	-	750

Figure 2.5: Structured and unstructured data in adult social care case notes

2.2.3.2 The purpose of administrative records

Administrative social care records serve multiple purposes, each reflecting distinct demands placed on social workers and care managers (unqualified workers responsible for assessing need and arranging care). In her comprehensive book examining the role of social care records, O'Rourke [67] identifies three primary purposes of administrative records: functionality, accountability and values.

Functionally, social care records must support the administration and evaluation of service delivery. For example, operationally, needs assessments keep track of disability over time, and care plans specify how services should be delivered. Case notes can also provide detailed instructions for continuity where cases are closed or transferred. Social care records are also used for performance monitoring, providing statistical data about individuals receiving care or care managers' performance. Accountability refers to the use of records in disputes or investigations, where they must provide an audit trail of decision-making to protect workers and their organisations against legal challenges or complaints. Value demands encompass both professional social work values, and the knowledge that individuals have access to their own records. This may entail recording information in a way that reflects the individual's strengths, needs, and preferences, rather than focusing solely on deficits or risks [72, 73].

These incentives can overlap. For example, the functional purpose of recording care needs and how they are met can provide accountability in the case of legal challenge, and information for statistical returns. However, they can also conflict. O'Rourke [67] describes the balancing act for workers who feel compelled to exaggerate dependency to make a case for funding, while simultaneously wishing to accentuate the strengths of the person they are describing, who has access to their own records. It is important to understand these incentives and where they might lead to inaccurate information being recorded, or relevant information being omitted. I discuss the approach to evaluating the reliability of the administrative records used in this thesis in Chapter 4.

2.2.3.3 Research access to administrative records

Administrative records contain detailed, personal information about people receiving care. Individuals have a reasonable expectation that data collected for their personal care and support will remain confidential and be used solely for the purposes for which it was collected. In England, the Data Protection Act

2018, which ratified the EU General Data Protection Regulation (GDPR), sets out stringent requirements for the processing of personal data. These regulations mandate that personal data must be processed lawfully, fairly, and transparently, and collected for specified, explicit, and legitimate purposes. Furthermore, individuals have specific rights under these laws, including the right to be informed about how their data is used, the right to access their data, and the right to object to certain types of processing. In the context of social care, this means that service users expect their sensitive information to be handled confidentially and in accordance with local authority privacy notices.

Any secondary use of this data, such as for research purposes, requires careful consideration to ensure compliance with legal obligations and to maintain public trust. One major obstacle for researchers is difficulty in obtaining the data, which is often compounded by public concerns over privacy and data sharing. Projects can be abandoned for information governance reasons. For example, the *care.data* data sharing initiative to join up records across hospitals and the community was cancelled, after opposition from the public about the lack of clear communication about the use of individual data and how to opt out [118].

Furthermore, permission to access this data can be subject to delays, which hampers timely research and analysis [119]. As a result, studies tend to be limited to data from a single local authority or, at most, a handful of councils, making it challenging to generalise findings across different regions due to the variability in data collection and recording practices among local authorities [see e.g. 120, 119]. Additionally, free text information presents a particular challenge, as it is often excluded from information-sharing agreements due to concerns about the potential for containing identifiable information, including that of third parties who have not consented to the sharing of their data [121]. In this study, software was developed to pseudonymise free text data. I discuss this further in Chapter 4.

2.2.3.4 Scalability of research with administrative records

Currently, the analysis of administrative social care data for establishing the impact of needs on service use in England must begin at a local council level, as there is no nationally mandated structure for social care forms. There is little recent evidence on the variation in the type of social care assessment forms used. In England in the 1990s there was wide variability in the format and content of forms [122, 123]. There have since been significant efforts to address this, particularly with the introduction of the Single Assessment Process (SAP) in 2004, which aimed to introduce a more standardised assessment process for older people across health and social care [52] and the Common Assessment Framework (CAF) introduced in 2009 [53]. Both policies aimed to ensure that older people's needs were assessed comprehensively, and to reduce duplication in the assessment process by improving information sharing between health and social care. The focus of these policies was through increasing partnership working between local authority adult social care departments and relevant health authorities. However, implementation of SAP varied across the country [54]. Local authorities appeared to prioritise the use of standardised assessment forms over other aspects of SAP [124] and by 2010, there appeared to have been some convergence, with more than half of local authorities using one of two nationally-accredited forms. Nevertheless, tensions between central and local government resulted in differences in implementation of SAP between areas [56, 124].

The SAP and CAF were superseded by The Care Act 2014, which defined new, national eligibility criteria for adult social care, and did not prescribe any specific forms for recording this information [94]. However, there remain barriers to local authorities designing their own forms and systems. Councils must demonstrate that they meet their statutory duties, and any council who designs rather than purchases forms assumes a degree of legal risk. Furthermore, there are technical costs associated with implementing IT systems, which it may be more economical for councils to share. The result of this is that most local authorities purchase forms and IT systems from private sector suppliers. There is little public information on which systems are used. A 2019 report by the Local Government Association (LGA) into social care data found the market was dominated by two suppliers, with 75% of councils in England using their systems (though not necessarily the same assessment forms) [125].

While the dominance of two providers might suggest there would be consistency across councils, structured data can be challenging to standardise within local authorities, as well as between them. For example, a paper reviewing the assessment and care planning process in Birmingham City Council found that there were six forms involved at the time of the study [120]. Four of these were different types of needs assessment forms, which might record the individual's functional ability with different questions. Additionally, it appeared that despite a codified assessment and care planning process, individuals could appear at stages in the middle of the process without having been recorded earlier. Similarly, care users might regularly be recorded as requiring a follow-up action at one stage of the progress, with no record of that action taking place.

The 2019 LGA report noted that at that time there appeared to be a trend towards diversification in the market. There is evidence that there has been innovation in the way that social work records are collected since 2019 [15, 126], although it is not clear the extent to which the market share of the two providers has changed. A 2024 report by the DHSC into streamlining social care assessments in England named only two providers of social care IT services, who were the same as mentioned in the 2019 LGA report [127]. Nevertheless, even if providers of databases remain the same, the changes in forms used by local authorities over time create challenges in harmonising administrative data. As well as new IT systems, forms can be updated because of legislation, reorganisation of internal processes or policy changes such as the decision to move from local to nationally accredited SAP forms [128], or the more recent shift to a strengths-based approach [72]. This can be seen in the data used in this thesis, which use the same database of needs assessment data recorded between 2010 and 2020. Chapter 4 describes the nine assessment forms used over this time, each of which has multiple versions, owing to changes in structured data triggered by updates to policies, local priorities or the law.

While structured data is machine-readable, temporal changes in forms and recording practices mean that the process of translating it into information can be complex. A systematic review of the use of local authority data for research found widespread challenges in how data are coded [16]. For example, a 2015 study found there are "considerable resources and time" required to clean adult social care data to the quality that it can be used for research purposes, requiring decisions made by researchers and tacit knowledge of individuals who collect these data [119]. I discuss this process with the dataset used in this thesis in Chapter 4.

The use of different assessment forms over time can also introduce problems with missing data. For example, in the administrative dataset used in this thesis, a needs assessment form that was temporarily used in late 2014, just before the introduction of new forms associated with the Care Act, did not include questions relating to memory or orientation. While this means there can be missing data in administrative records, where this is related to the form that was in use at the time, it is not as problematic as missing data correlated with individual need, such as in survey data.

Data cleaning needs to be duplicated for every additional structured field introduced into a research dataset. Given the range of data recorded across a variety of forms and IT systems used by councils [125], this would quickly become highly resource-intensive for all 152 councils in England. Despite this limitation, as I show in Chapter 5, the number of statutory social care users within a single local authority database is much higher than the number in longitudinal, national surveys. For example, the data in this study from one London local authority contains 3,046 care users. A sample size of thousands of individuals can offer far more statistical power than the numbers of social care users available in surveys, making it possible to compare subgroups of the population, for example on the basis of ethnicity, that would not be possible with survey data.

Structured long-term care data requires substantial time and expertise to translate into useful information. However, the result can be information that is simply unavailable from surveys and nationally aggregated datasets. For example, administrative records contain continuous service receipt data, rather than repeated observations at survey waves. Furthermore, as set out in Figure 2.4, in England there is a legal duty to update functional information if there is a change in care needs. If this duty is met, it should allow for more accurate recording of changes in functional ability over time than surveys.

2.2.3.5 Quality of administrative data

Administrative data is not recorded for research purposes, and its use in research requires an understanding of the context in which it is recorded, the incentives of those creating the records, and the fact that the reliability of information can vary widely; some fields may be highly accurate, while others may be inconsistent or unreliable. In social care in England, there are limitations in the identification of health needs in social care assessments. Davies and Fernandez [27] found that care managers' perception of the needs and characteristics of individuals differed from those of care users. Challis et al. [129] compared the needs assessments of social work care managers with those conducted by specialist clinicians, finding care managers missed a significant number of conditions, and cognitive impairment in particular was under-recorded. Clarkson et al. [130] found that depression was systematically under-recorded in social care assessments.

Clarkson et al. [55] found that — while recording improved after the introduction of the SAP — there remained widespread under-identification in social care needs assessment of medical diagnoses and disagreement between care managers' assessment of functional ability and those of the individual. For dressing, bathing and using the toilet, Cohen's κ was 0.27, 0.24 and 0.46. This measure is bounded between -1 and 1, with 0 representing agreement no better than random chance, and $\kappa \leq 0.4$ representing poor agreement [131]. However, it is challenging to interpret the extent to which inconsistencies between professional assessment and self-reported measures reflect inaccuracy in social care records, as individuals can under-report their own health needs. Stoye and Zaranko [106] compared self-reported medical conditions in ELSA with information from hospital records and found omissions in both, though administrative data was more accurate. I address this in Chapter 5, which compares self-reported and observed cognitive function in ELSA, finding that there is not a strong association between the two measures.

There is limited other literature on the accurate identification of functional needs, such as ability to complete ADLs, in administrative data. Spiers et al. [132] is a systematic review of social care supply on healthcare utilisation. They find 12 studies using administrative data examining the relationship between the availability of social care support and healthcare utilisation, seven of which are from the UK. However, six use national data returns, which it is not possible to disaggregate to assess. The seventh study, Bardsley et al. [133], uses individual-level local authority records. However, there is no triangulation with other data sources capturing need, and it is not possible from the published data to undertake such a secondary analysis.

Challenges in the reliability of administrative records are not unique to social care. General medical practitioners systematically under-identify dementia [134] and depression [135]. Research into recording of health diagnoses in medical records in the US found that the correct primary diagnosis was recorded for 57% of visits [136]. Similarly to social care [55, 130], specificity in health records tends to be high, ranging from 0.9 to 0.99, with false diagnoses recorded quite rarely [137]. Conversely, the sensitivity for clinical conditions was lower (between 0.75 and 0.96), with diagnoses not always recorded. A study in the Netherlands found that correlation between self-reported and actual functional ability was 0.2, and argued that this difference was so large that the two constructs should be treated as complementary but separate measures [138]. Children's social care records also contain missing information [139]. This is not necessarily a barrier to the use of administrative data in research, but it means that those using such data must have sufficient understanding of how it is recorded to assess which parts of the data are of sufficient quality [140].

Inaccuracy in data is of particular concern where, as in the case of under-recording, it is systematic rather than stochastic [141]. Incentives in social care recording can both create and mitigate the opportunities for such bias. In adult social databases in England, service use information is often linked to payment systems, so must be recorded accurately for providers to be paid [125]. This is reassuring as providers are incentivised to ensure that every service is logged. It is possible there could be a slight upwards bias to costs, if there exist unscrupulous providers who are less likely to report recording errors resulting in overpayments. Nevertheless, if one assumes that local authorities would notice significant overpayments, and providers the converse, service use data is likely relatively reliable.

In contrast, there are various incentives that may lead to needs being under-reported in social care administrative data. Social work has a "blame culture" [142]. Records can be scrutinised if an adverse event occurs and workers may omit information to avoid accountability [143, 67], or because they are aware that the services required cannot be provided. For instance, if a social worker knows that they are unlikely to be able to provide a social inclusion service, they might choose not to document that an individual is lonely. This might occur because the worker wishes to avoid recording unmet needs, which could increase stress and legal risks [144]. Alternatively, it may be that recording this information may be seen as an unnecessary data collection effort [141] given that no action is likely to be taken.

Another incentive may be that workers over-report a person's needs to meet eligibility criteria for certain services or funding [67]. The Care Act 2014 eligibility criteria require that individuals must meet specific thresholds or are ineligible for all publicly funded services. In close cases, social workers might inflate the severity of a client's condition to ensure they qualify. This practice can be driven by a genuine desire to secure the necessary resources for clients, but it may lead to inaccuracies in the data. This phenomenon has been observed in various contexts, where the financial incentives or performance metrics tied to specific outcomes prompt workers to record data in ways that are not entirely accurate [145, 146, 147].

While this is concerning, there are reasons to be reassured. Social care records are subject to supervision and audit, which are effective mechanisms for reducing inaccuracy [136]. Fraudulent recording incurs significant legal risk to professional registration and job security [148]. These incentives are most likely to affect recording in borderline cases, where there is room for interpretation. Social care incentives are not likely to impact recording as much as payments, and even in a situation with financial incentives for physicians the correlation between recorded structured data and actual need was very high ($r \geq 0.98$) [146]. There is no reason to assume that the issue in adult social care is more widespread. However, it should be considered and accounted for. There is likely to be a distribution of fidelity in administrative records, with some fields more accurate than others. I discuss in detail the way that I established the reliability of the administrative dataset used in this thesis in Chapter 4.

Finally, the other approach for managing this issue is through the interpretation of results derived from administrative records. An approach used in Minchin et al. [147] estimates the effect of an intervention on the *documented* quality of care. The same approach can be taken when extracting information from social care records. For example, when extracting loneliness from free text, rather than assuming that this necessarily measures real loneliness, it can be understood as a measure of whether loneliness is recorded. While this might appear to side step the issue, it is useful in policy and practice terms. Chapter 7 finds that, controlling for needs-related factors, the measure of loneliness extracted from free text at the time of the first assessment predicts time until care home entry. Such a finding could be integrated into case management systems to highlight the risk of care home entry based on what is recorded in case notes.

2.2.3.6 Free text administrative data

One of the most significant advantages of administrative records is the wealth of free text data they contain. Unlike structured data, which is confined to predefined fields and categories, free text allows practitioners to record nuanced observations, contextual details, and personal narratives that offer a richer understanding of an individual's circumstances. This unstructured data can capture the complexities of social care needs, interventions, and outcomes in a way that structured data often cannot [69, 70, 71, 67]. For instance, case notes and assessment summaries may include detailed accounts of a person's daily challenges, social relationships, and emotional well-being — information that is invaluable for comprehensive evaluations of care services. Researchers in the US have used natural language processing with free text children's social work records to automatically identify abuse [149, 150]. Another study uses US data to categorise, from free text, types of intervention delivered by social workers, such as care coordination, financial planning or referral to community services [151]. In health, researchers have extracted information about social determinants of health from free text data [152, 153, 154].

Furthermore, the reliability of administrative data can be particularly improved by including free text, expanding the data available for establishing internal consistency. A study which combined free text records with structured fields increased the positive predictive value of allergic drug reaction from 46% to 86% [155]. Similarly, a study predicting length of stay found the addition of free text notes into the model significantly improved model performance over using structured data [156].

Yet there have been no published papers using free text social care data in England. This is partly because the very richness of free text data also presents challenges. The unstructured nature of this information makes it difficult to extract and analyse systematically. It can be challenging to handle the variability in language, terminology, and recording styles found across different practitioners and time periods. Additionally, the sheer volume of free text — often amounting to millions of words — means that manual analysis is impractical. Without effective tools to process and interpret this data, much of its potential value remains untapped. In the next chapter, I describe the recent developments in LLMs and the potential for their use for insight into the information contained within free text administrative data in social care.

3 Large Language Models for understanding social care

This chapter outlines uses of LLMs with administrative records to enhance understanding of social care and inform decision-making. The chapter covers definitions of LLMs, their theoretical advantages over traditional NLP methods, and potential applications in social care, before discussing practical considerations. I begin by defining LLMs and contrasting them with traditional natural language processing (NLP) methods used with administrative care records. I discuss the advantages and limitations of previous NLP methods when applied to free text care records, followed by an overview of the emergence of LLMs and their improved capacity for processing text in care records. Next, I present a theoretical framework distinguishing data, information, and knowledge, to clarify the benefits of language models for transforming raw data into useful information. Finally, I examine the potential of LLMs as a method for addressing the core challenges of extracting meaningful information from large volumes of unstructured data, and reducing the manual effort needed to document social care. This covers LLMs for evaluating social care services using administrative records, including their roles in predictive modelling, understanding service use, and reducing the administrative burden through automation. This chapter highlights the opportunities provided by LLMs, underscoring the importance of evaluating their accuracy and biases to assess their suitability in social care settings.

3.1 What are large language models?

In this section, I will define the term Large Language Model (LLM), giving a brief overview of how NLP methods were used with administrative care records prior to LLMs, and explain how LLMs are distinguished from earlier NLP models. Although using LLMs with long-term care records is relatively new, using NLP to extract information from administrative health records is more established [see e.g. 157, 152, 158, 159]. Until the mid-late 2010s, NLP models were trained for specific tasks in certain domains, such as identifying the presence of medical conditions in free text records [160]. However, such models were "brittle and sensitive to slight changes in the data distribution and task specification" [161]. Furthermore, it was challenging to create models for administrative care records, which often include domain-specific acronyms and terminology, and typographic errors [159].

Development of NLP models to parse such data requires large volumes of it which, for information governance reasons, are not generally available [162, 160].

The development of LLMs has been driven through increases in computing power, larger datasets and the development of new machine learning methods [163]. These models are considered large because of the vast number of parameters — numeric weights in computational neural networks — they contain relative to earlier models, improving their capacity to learn, represent, and classify complex patterns in language. While earlier NLP methods also used neural networks to model language, I describe below how LLMs are distinguished by their scale and a model architecture designed to represent text in context-specific ways [164]. LLMs also differ from earlier language models in that they tend to be general purpose rather than domain-specific, and through their ability to generate coherent text [163]. Since the explosion of pre-trained, general-purpose LLMs in 2022, where increased performance can be achieved by scaling the model size rather than through large volumes of domain-specific data [165], there have been far more opportunities for the use of language models in long-term care.

I include in this section examples of relevant research using smaller language models, such as dictionary-based approaches, count-based methods, and pre-trained word vector models, describing the strengths and limitations of such approaches. This forms the argument that LLMs are more suitable than earlier NLP methods for addressing specific challenges in administrative care records, because of their improved capacity to accurately represent context-specific information.

3.1.1 Dictionary-based approaches

Dictionary-based natural language processing aims to assess the presence of concepts by defining lists of words or phrases. Text is pre-processed, which means removing stop words (such as conjunctions, articles and prepositions), correcting common spelling errors and undertaking context-specific abbreviation expansion [162]. Words are lemmatised, or restored to a base form. For example, “walks”, “walked” and “walking” will all be replaced with the lemmatised form, walk [166]. After this, text is matched to a pre-defined list of words and related phrases. For example, Perera et al. [158] established housing status from mental health records by searching for the terms “homeless”, “NFA” and “No Fixed Abode”. Negation detection aims to ensure that phrases such as “she is not diabetic” are not miscategorised as indicating diabetes [167].

However, there are drawbacks to dictionary-based approaches. It can be difficult to classify sentences such as, “She has nowhere to live”, “He used to be homeless”, or, “Since she spoke to the homeless person’s unit she has been housed”. Negation detection is challenging in dictionary-based approaches, leading to false positives and false negatives [168]. Additionally, negative results of a dictionary-based approach can be hard to interpret. In Chapter 6, where I establish whether a worker has recorded whether an individual is lonely, I considered using a dictionary. However, unlike approaches which rely on counts of words, the performance of a dictionary-based approach depends on the rules defined. If such an approach cannot identify whether loneliness is recorded, it does not necessarily mean the method is inappropriate. It may mean that the dictionary rules need to be improved.

3.1.2 Count-based approaches

An alternative method to dictionary-based approaches are count-based approaches, also known as a Bag of Words (BoW) model [169]. In a BoW model, the text in each document is treated as a collection (or “bag”) of words, disregarding grammar and word order. The model creates a vocabulary from all the unique words in the text corpus and then represents each document as a vector. The representation of all documents is a large, sparse matrix where the columns are the set of all words in the corpus (i.e. all documents), each document is represented by a row, and the value of each field in each row indicates the frequency of the relevant word in the sentence [170]. I set out an example of a corpus of three documents in Table 3.1, with the resulting BoW vectors in Table 3.2.

Table 3.1: Bag of Words sentences

No.	Sentence
1	Mr Smith refuses to go to the day centre after he eats.
2	My Smith refuses to eat after the day centre.
3	He does not want support with social inclusion.

Such approaches create a word vector representing each sentence, which can be used with a variety of algorithms, depending on the downstream task. For example, in Chapter 6, I compare LLMs for identifying loneliness in case notes with

Table 3.2: Bag of Words vectors

centre	day	eat	refuse	smith	inclusion	social	support	want
1	1	1	1	1	0	0	0	0
1	1	1	1	1	0	0	0	0
0	0	0	0	0	1	1	1	1

the labelled BoW representation used as an input to binary classification algorithms such as logistic regression, random forest and a sequential neural network. I also use a slightly modified BoW representation called term-frequency inverse-document frequency (tf-idf), which weights each term in inverse proportion to its frequency [171] to reduce the impact of words which appear in many documents, such as “care”. The equation used for this weighting is given in Section 11.1.

BoW models are a simple way to represent text which can be effective. However, the simplicity has drawbacks. Word order is important in natural language. Malkin et al. [172] give the example of the sentences, “The scared mouse chased the hungry cat”, and “The hungry cat chased the scared mouse”, both of which would be represented in the same way by a BoW model. The same issue can be seen with the first and second sentences in Table 3.2, which have the same representation but different meanings. Additionally, the sparsity and high dimensionality of the matrices created by BoW can pose technical problems for classification algorithms [173]. To mitigate this, similar pre-processing methods are applied as with dictionary-based approaches, such as lemmatisation and removal of stop words, and sometimes post-processing such as dimensionality reduction methods. However, such processing can degrade meaning [169]. Furthermore, sentences with semantically similar meanings, but different words, will be mapped to vectors that may have no overlapping features (i.e. columns) in a BoW representation, leading to a loss of semantic understanding. This can be seen with second and third sentences in Table 3.2, which are close in meaning but are mapped to orthogonal vectors. This limitation means that BoW models often fail to capture the deeper context and relationships between words in a sentence, making them less effective for tasks requiring nuanced language interpretation [173]. As a result, more sophisticated approaches, such as dense word embeddings or contextual representations, are often preferred for complex NLP tasks. I outline these methods, with examples, in the next sections.

3.1.3 Dense word embeddings

Dense word embeddings were created to resolve the problems of information loss from pre-processing, the computational limitations of large, sparse vectors and the issue of semantically similar sentences having different vector representations. In 2013, Mikolov et al. [174] created Word2Vec, a model that revolutionised natural language processing by creating dense vector representations of words, known as word embeddings [175]. Unlike the BoW model, which treats words as independent and isolated entities, Word2Vec captures meaning based on distributional semantics, perhaps best described by JM Firth as the principle that, “You shall know a word by the company it keeps” [176].

The key innovation of Word2Vec lies in its ability to map words with similar meanings to vectors that are close in the high-dimensional vector space, and also for the relationship between words to be encoded through these distances. I provide examples of this below. To achieve these encodings, the Word2Vec model employs a neural network to generate word embeddings by learning patterns from large corpora of free text. Specifically, Word2Vec uses two main architectures: Continuous Bag of Words, which predicts a target word based on its surrounding context, and Skip-gram, which predicts the surrounding context based on a central word [174]. These training methods allow Word2Vec to create a language representation model that captures a wide range of word associations, making the resulting embeddings highly effective for various NLP tasks [177].

The semantic meanings encoded in vector space are demonstrated in Mikolov et al. [178], which describes how (unlike in count-based models) the words “king” and “queen” appear close to each other in vector space. Furthermore, in their representation, the relationship between words is encoded through vector offsets, such that $(king - man) \approxeq (woman - queen) \approxeq (uncle - aunt)$. This ability to encode both semantic similarity and relational patterns in word vectors was transformative compared to count-based NLP models, as grouping words with similar meanings together simplified the decision boundaries for classification algorithms.

I give in Figure 3.1 an example of this approach to language representation using the pre-trained Global Vectors for Word Representation (GloVe) [179], which like Word2Vec creates dense word embeddings to represent semantic meaning. As the GloVe vectors are 300-dimensional, I used Truncated Singular Value Decomposition (a dimensionality-reduction method) [180] to create a 2-dimensional representation of the vectors for the purpose of the plot. Although this involves some information loss, it is evident from visual inspection of the figure how a decision boundary might be drawn by a classifier to distinguish relevant words.

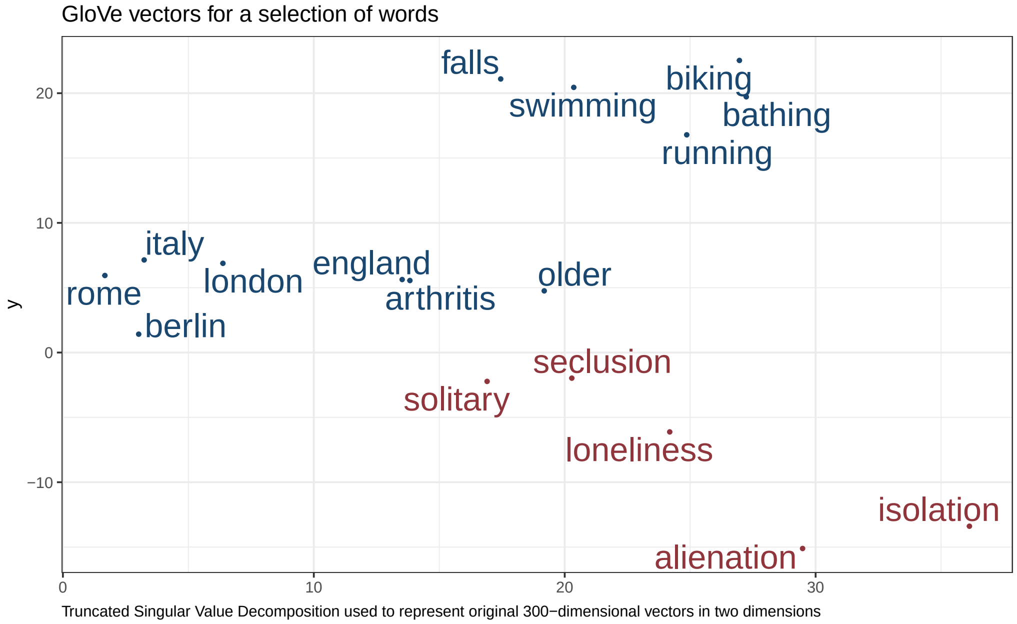


Figure 3.1: GloVe embeddings

However, while these embeddings represented a significant leap forward, they still have drawbacks. Firstly, they are fixed, meaning that once trained they do not adapt to new contexts or tasks without retraining, which can be a limitation if the models are applied in domains where a word may be used in a different sense to its use in the training data [181]. Secondly, pre-trained word embeddings like Word2Vec and GloVe assign a single vector representation to each word, which can lead to ambiguity when a word has multiple meanings (polysemy). A canonical example of this is the word “bank”, which can refer to a financial institution or the side of a river, but pre-trained embeddings do not distinguish between these senses, potentially causing confusion in downstream tasks [182]. This limitation is especially problematic in cases where meaning is highly context-dependent. For instance, the word “social” varies significantly in relevance to loneliness in care records depending on its context — e.g. “social worker” versus “social isolation”. Polysemy is a limitation of models such as GloVe and Word2Vec which assign a single vector to each word, conflating differences in meaning. I illustrate this in Figure 3.2, where “social” has one representation despite distinct potential meanings.

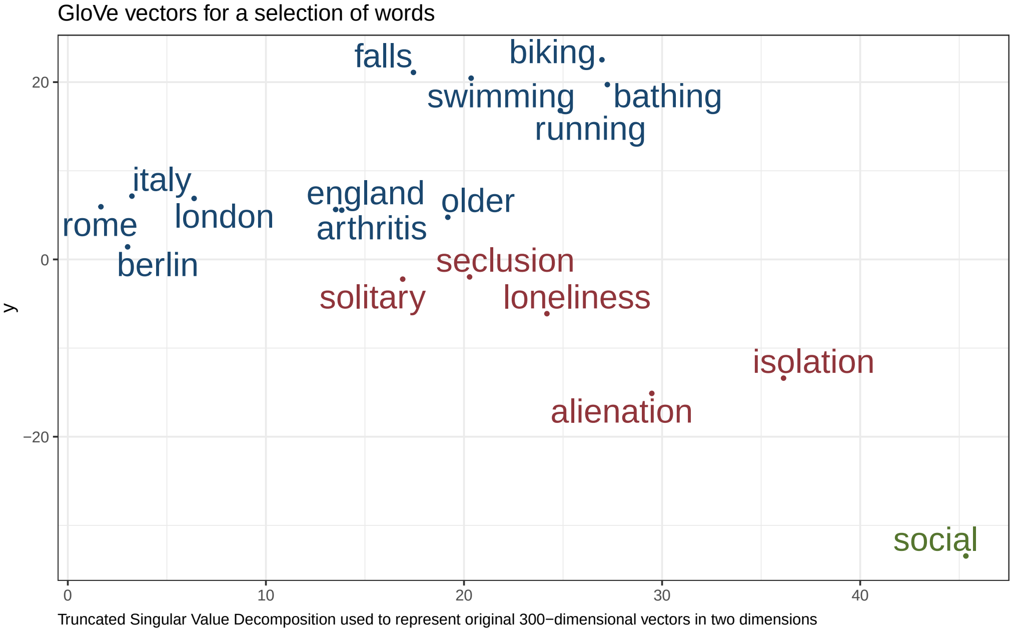


Figure 3.2: GloVe embeddings with "social"

Another limitation of models like Word2Vec is that they generate embeddings at word- rather than sentence-level, which poses challenges when dealing with sentences of varying lengths. Most classification models require a consistent input size, so sentences of different lengths must be standardised. One common approach is padding, where shorter sentences are extended with placeholder values to match the length of the longest sentence. While padding can improve accuracy, it often comes at the expense of performance, especially when datasets include many short sentences alongside very long ones [183]. Another approach is pooling, such as averaging the word embeddings in a sentence across each dimension to create a fixed-size sentence representation [184]. Pooling is efficient and sometimes effective but can lead to the loss of important information [185]. More recent language models address these limitations by generating context-specific vector representations at the sentence or document level.

3.1.4 Transformer-based models and LLMs

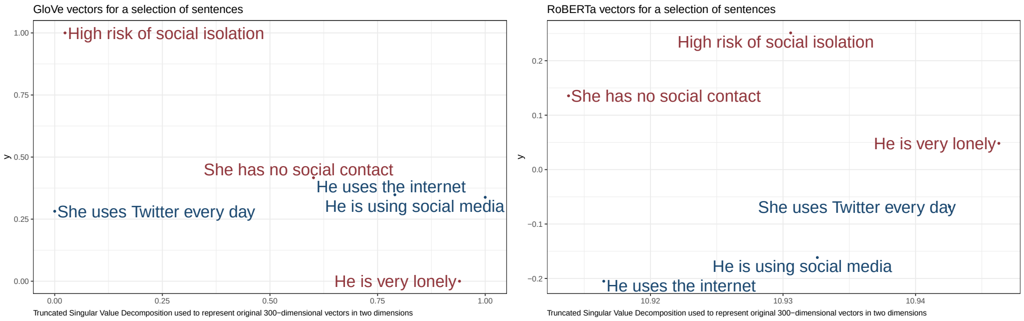
Contextualised word embeddings, created to address the issue of polysemy and fixed word representations, marked a significant advancement in natural language processing. This allowed for more sophisticated approaches, seen in models like

InferSent or Universal Sentence Encoder, which aim to create an embedding capturing the meaning of an entire sentence rather than individual words [186]. The Bidirectional Encoder Representations from Transformers (BERT) model, introduced in 2018, significantly advanced state-of-the-art performance across multiple language model metrics [187]. Transformers are a type of neural network architecture which allow models to compute contextual relationships between words, and their use is one of the key distinguishing features of LLMs [164]. BERT, an early LLM, considered the full context of a word by looking at both the words before and after it in a sentence, allowing it to generate different embeddings for the same word depending on its surrounding context [188].

This means models like BERT contain parameters not just for training word vectors to an optimal mapping of words in vector space, but also for inference, allowing them to generate context-dependent word embeddings dynamically. In contrast, a pre-trained word vector model like Word2Vec uses a neural network during training, which contain parameters (weights and biases) but once trained, Word2Vec essentially functions as a large lookup table, requiring minimal computational power to fetch the vector for a particular word.

There were earlier context-aware models, prior to BERT, such as those based on traditional Recurrent Neural Networks (RNNs). However, the context of a word may be influenced by ranges longer than just the immediately surrounding ones. RNNs worked well over a few words but were limited in their ability to consider longer sequences due to challenges like vanishing gradients — where the gradients used to update model weights during training decay exponentially, preventing the model from effectively learning long-term dependencies — and the need to retain too much information over extended sequences [189, 190]. Prior to models like BERT, Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) improved upon this by introducing mechanisms to better manage information over longer sequences, which mitigated vanishing gradient issues [191]. While these architectures could remember important information over extended sequences, such networks were autoregressive, with the state at time t dependent on the state at $t-1$, limiting the possibility to run them in parallel and preventing efficient processing of very long texts [192]. In 2017, the transformer neural network architecture was developed, which could retain information over long contexts and run efficiently in parallel [192], and this was followed by the release of BERT in 2018 [187]. The large version of BERT contained 340 million parameters [193]. These parameters determine how the model processes input text into vectors. Models with more parameters require greater computational resources [194].

The capacity to create context-dependent vectors was a major step forward, as it allowed BERT to better represent the nuanced meanings of words and phrases in a way that static embeddings like Word2Vec or GloVe could not, and it eliminated the need for a fixed context window, enabling long range dependencies to be encoded in vectors. In Chapter 6 I use Robustly Optimised Bidirectional Encoder Representations from Transformers (RoBERTa) [195], a model which uses the same architecture as BERT but is trained on more data. I present in Figure 3.3 a comparison of the sentence vector representation of a selection of sentences using GloVe and RoBERTa. The GloVe sentence vectors are a mean of the word vectors across 300 dimensions, which is an efficient and often effective method of creating sentence representations [185]. The RoBERTa vectors are the 768-dimensional embedding of the `[CLS]` token, which indicates the start of the sentence, and is encoded to represent the semantic meaning of the sentence [195]. In Figure 3.3 I have again reduced the dimensionality to two, which leads to some information loss. Nevertheless, it appears that RoBERTa is able to distinguish “social media” and “social isolation” more effectively than a model like GloVe, which cannot account for polysemy. This is a strong case for using such models.



(a) GloVe sentence vectors

(b) RoBERTa sentence vectors

Figure 3.3: Sentence vectors

However, despite the significant advantages of BERT and its derivatives such as RoBERTa, it works as it was primarily designed: as a classification model. Its architecture is well-suited for tasks where the goal is to predict labels or categories based on input text, as it excels at encoding a context-dependent representation of words within a sentence. However, its design does not naturally lend itself to generative tasks like summarisation or text completion. While BERT can be fine-tuned for a variety of downstream tasks, its core architecture is optimised for representing rather than generating language [187].

3.1.5 Generative LLMs

The limitations of models like BERT in generative tasks led to the development of models specifically designed for text generation, such as the Generative Pre-trained Transformer (GPT) models from OpenAI, Meta's Llama series and Google's Gemini and Gemma models [196, 197, 198]. There is no technical reason that such models must be larger than BERT derivatives. Open AI's first GPT model had 110 million parameters, while the larger versions of the BERT models had 340 million parameters and RoBERTa 355 million [193]. However, in practice, the models that have been successful for general purpose text generation have been very large relative to BERT. GPT-2 had 1.5 billion parameters and this trend continues, with 2024 generative models eclipsing the size of BERT-based models. For example, Google's Gemma has 7 billion parameters [199], and Meta's Llama 3 model was released in 8 billion and 70 billion parameter versions [200].

Unlike BERT, which focuses on representing words in vector space, generative models are designed to write coherent, contextually appropriate text based on a given prompt. The architecture of such models is designed to generate text, rather than map it to a vector, and such models are suited to tasks such as creating or summarising documentation [161].

3.2 Theoretical framework: data, information and knowledge

The focus of this thesis is evaluating LLMs for resolving information challenges in social care. The thread that unifies the chapters is the transformation of raw data into meaningful information, and subsequently deriving knowledge from this information. I distinguish data, information and knowledge in this thesis using the theoretical approach found in information sciences. In particular, I use the definition in Meadow and Yuan [201]:

1. *Data*: a set of symbols with little or no meaning to a recipient.
2. *Information*: is a set of symbols that does have meaning or significance to their recipient
3. *Knowledge*: the accumulation and integration of information received and processed by a recipient.

I present in Figure 3.4 an example of free text data using this framework.

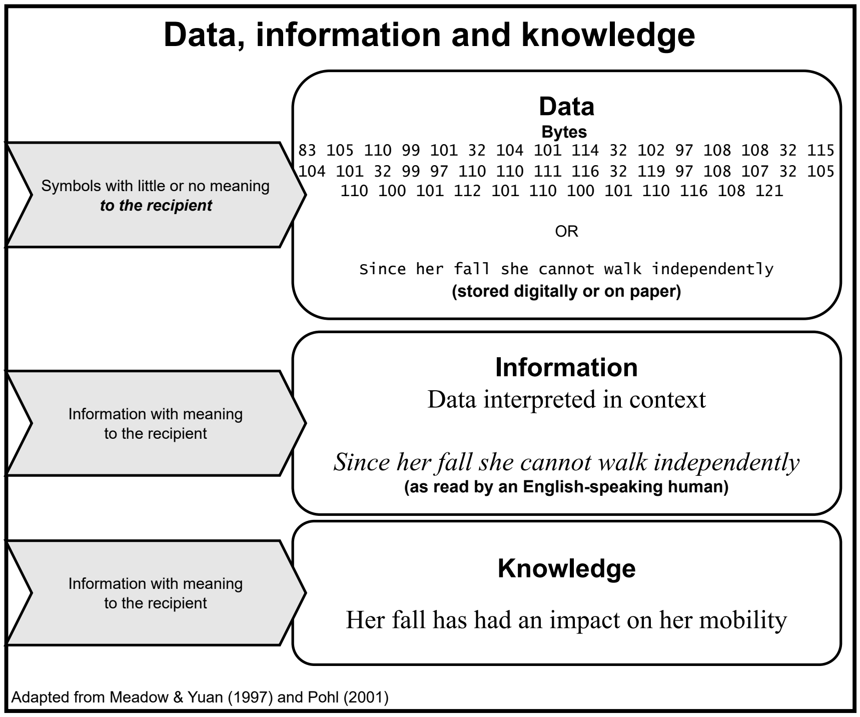

Data, information and knowledge
Symbols with little or no meaning
to the recipient
Data
Bytes
83 105 110 99 101 32 104 101 114 32 102 97 108 108 32 115
104 101 32 99 97 110 110 111 116 32 119 97 108 107 32 105
110 100 101 112 101 110 100 101 110 116 108 121
OR
Since her fall she cannot walk independently
(stored digitally or on paper)
Information with meaning
to the recipient
Information
Data interpreted in context
Since her fall she cannot walk independently
(as read by an English-speaking human)
Information with meaning
to the recipient
Knowledge
Her fall has had an impact on her mobility
Adapted from Meadow & Yuan (1997) and Pohl (2001)

Figure 3.4: Data, information and knowledge

It is quite straightforward that the representation of text as bytes would be considered data, but perhaps harder to immediately see how the text, “Since her fall she cannot walk independently”, could be considered either data or information, depending on whether it is stored electronically or read by a human. Such distinctions are generally made in the domain of computer science, because humans automatically and subconsciously apply their knowledge to interpret data within a specific context [83]. Context is required to convert data into meaningful information [83]. Bytes representing text do not have a meaning when stored on disk. The text would similarly have little value as information if placed in front of a human who could not read English [201]. It is only through context-specific interpretation that data becomes information. This awareness of context is the application of existing knowledge to information [202].

This distinction between text data and text information is perhaps easier to illustrate with a document containing one million words. In this case, storing the data is straightforward, but extracting the information it contains may not be. The data used in this thesis, administrative records of older people receiving long-term care in one local authority, contains 114.4 million words of free text. Such data sources provide an opportunity to extract information, but simply reading it all is not a practical approach. Consequently, the vast majority of unstructured administrative data has never been analysed [16].

Structured data, such as a field representing gender or age, is encoded in a format where it is already mapped to meaning. This makes it relatively straightforward to extract and summarise from structured data, for example, the proportion of women or the median age. However, there is no corresponding, unambiguous way to map a word to meaning. As the cognitive scientist Colin Cherry sets out,

> The full meaning of a word does not appear until it is placed in its context, and the context may serve an extremely subtle function... And even then the "meaning" will depend upon the listener, upon the speaker, upon their entire experience of the language, upon their knowledge of one another, and upon the whole situation. Words do not "mean things" in a one-to-one relation.

— [203, p.10]

This framework for understanding data, information, and knowledge highlights the inherent challenges in extracting meaningful insights from unstructured text. In social care, where large amounts of free text administrative records are collected, this distinction becomes critically important. The difficulty lies not only in the sheer volume of data but also in the nuanced interpretation required to transform this data into actionable knowledge. Traditional NLP methods fall short in addressing these challenges due to the complexity and contextual dependencies of language. There is an opportunity to use advanced computational approaches, such as LLMs, to overcome these obstacles. However, it is essential to critically evaluate the accuracy and potential biases of these models to determine whether they are appropriate in social care settings.

3.3 Theory to practice: LLMs in social care

The administrative records in social care are rich with data but require sophisticated methods to extract useful information and build knowledge. Increasingly, LLMs may be able to transform large volumes of unstructured free text data into formats which are easier for humans to glean information from, such as structured data or brief, free text summaries. The format of the output of LLMs will differ depending on its purpose. To create data for researchers to use in a regression model, it might be appropriate to use a discriminative LLM, which is trained to map input text into structured data [204]. This is the approach taken in Chapter 6, where a binary indicator of loneliness is derived from free text. I then use this structured data in a survival analysis to model time until care home entry based on loneliness and other demographic and needs-related factors, which comprises Chapter 7. Alternatively, to create data to inform social workers' daily practice, it might be appropriate to use a generative LLM, i.e. a model whose output is new sequences of text [205]. Such models can be used for transforming large volumes of text data into brief summaries, and I evaluate the gender bias of such models in Chapter 8. I give an overview of the distinction between generative and discriminative models in Figure 3.5.

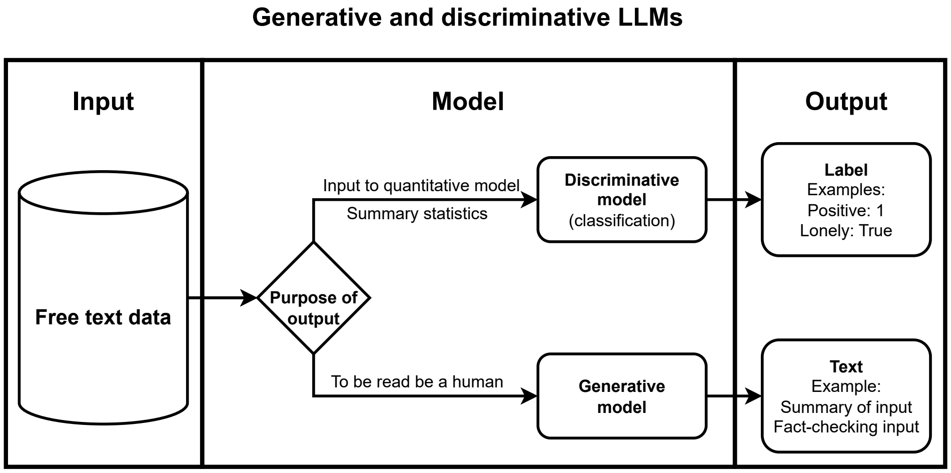


Figure 3.5: Generative and discriminative LLMs

3.3.1 LLMs in social care practice

The 2024 LOTI report *Opportunities for AI in Adult Social Care Services* notes that there are opportunities for LLMs to be used to generate documentation, and to generate summaries to allow faster decision-making [14]. The report lists five potential uses for LLMs in social care [14].

1. Providing case summaries for workers.
2. Automatic transcription of meeting notes.
3. Predictive forecasting to allow for early intervention in care.

4. Generating accessible, easy-read documents for adults with learning disabilities.
5. Development of a chatbot that allows care users to input a search term and be provided with care options.

In this section, I will discuss the use of LLMs in social care practice, focusing in particular on the first two points, summarising existing records, and generating new documentation, and highlighting both the potential benefits and the concerns regarding accuracy and bias.

3.3.1.1 Summarising existing records

Given the problem of data overload in health and care, the opportunity for using LLMs to improve practice is clear. LLMs have been used to process large volumes of care records and provide concise summaries [206, 207]. A study in which physicians evaluated LLM and human-generated summaries for completeness, correctness and conciseness found LLM versions were as good as (45%) or better than (36%) most medical expert summaries, though 19% were worse. In particular, the LLM summaries were more concise and complete, with fewer misinterpretations, inaccuracies and hallucinations (i.e. the inclusion of information that was not in the original text) than medical experts [208]. High quality, relevant records are associated with improved quality of care [209, 210]. However, accuracy has not been comprehensively determined or even defined in social care.

At the time of writing in November 2024, LLMs are starting being used to address data overload. One English council has identified that they have two major problems with *data sprawl* [85] in children's social care:

> 1. *Time-consuming searches: Finding the information needed often requires trawling through multiple sources, delaying timely intervention for vulnerable children.*
> 2. *Limited access to historical data: Crucial insights can be buried within years of casework data, making it difficult to draw on past experiences to improve outcomes.*

— [15]

This local authority implemented a chatbot based on a GPT model LLM, which searches through existing data and generates summaries of relevant information. This is a new project which was nominated for an innovation award in 2024 [211], and the use of such products is not currently widespread. However, there appears to be a will, or at least an acceptance, that LLMs will be used to ease some documentation pressures in care. A survey of medical practitioners found that 91% anticipate the LLMs will be used to summarise health records and 71% anticipate they will reduce workloads [212]. A literature review found that care professionals see ChatGPT as a valuable tool for creating summaries of free text notes, despite concerns about transparency, bias, privacy and accuracy [213, 214, 215].

3.3.1.2 Generating new records

The LOTI report also outlines how LLMs can be used to reduce the administrative burden in social care by automating the transcription of meeting notes between social workers and service users [14]. The 114 million words of notes in the data used in this thesis would have taken a considerable amount of time to record, and this is reflective of the general administrative burden in health and social care [9, 42, 41, 48]. The proposed solution involves using automatic speech recognition models to transcribe audio recordings of meetings, which are then summarised into case notes by LLMs, allowing workers to focus on service delivery during interactions. After the meeting, the worker can validate and make minor corrections to the transcription. The LLM could be further tailored to extract key actions, summaries, and statutory information required for reporting. By integrating this functionality with existing social care systems, the report suggests that time spent on administrative tasks could be significantly reduced.

A 2024 survey of healthcare practitioners found that 95% anticipate that LLMs will be used to automatically generate health records [212], perhaps because such products already exist in health and social care [126, 15]. Google is offering a healthcare product that works with recordings of clinician-patient conversations to “instantly convert data into drafts of medical notes, which physicians then review and finalise before they’re transferred in real time to the hospital’s electronic health record” [216]. A feature on the use of Microsoft’s LLM-based Copilot on the UK Local Government Association (LGA) website states, “social care workers experience a significant reduction in administrative burden, freeing them to dedicate more time to direct client care” [15]. A similar product offered by a competitor states it saves one day per week of worker time [217]. One local authority that has introduced LLM generated case notes reports “significant time savings in completing case notes and assessments, with an average reduction of 50 to 60 per cent”. [15]

Although there are no published statistics about how many councils are using similar products, at the time of writing in November 2024, there are nine local authorities in England whose website contains a privacy notice stating that LLMs may be used to process their social care data [218, 219, 220, 221, 222, 223, 224, 225, 226]. A further three local authorities are featured using LLMs in social care on the LGA website [15]. Other local authorities mention on their websites that they use LLMs, either explicitly for adult social care [e.g. 227, 228] or more generally that LLMs are integrated into their MS Office suite [e.g. 229, 230, 231, 232, 233, 234, 235]. In a June 2024 LGA survey of English local authorities, 43% of respondents saw the greatest opportunities for AI in health and adult social care [p.32 15].

3.3.1.3 Efficiency savings from LLMs in social care practice

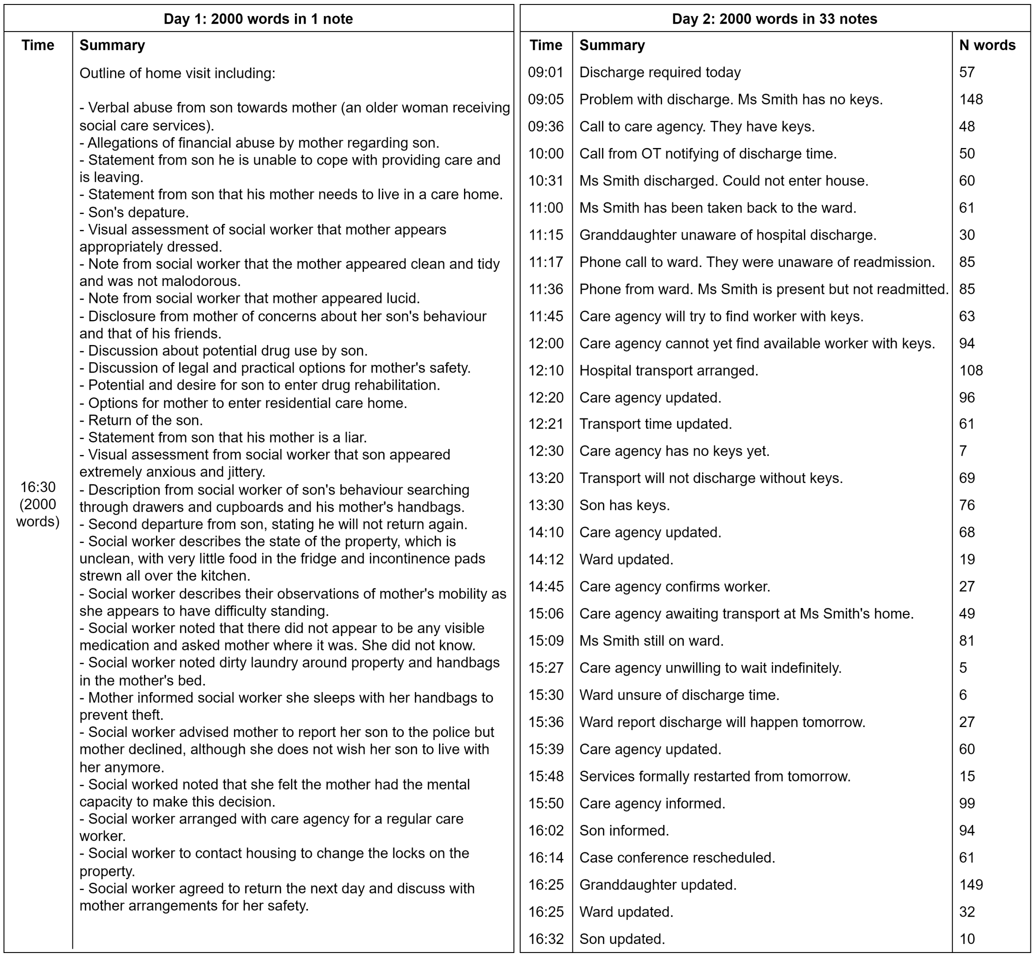
The potential for LLMs to generate efficiency savings through reducing documentation pressures is a key argument for their adoption and is cited in both academic literature and in reports from local authorities adopting them [e.g. 212, 15]. This is particularly appealing in social care, where demand is increasing and budgets are highly constrained [26, 4]. However, the scale of these potential savings remains uncertain and requires careful examination.

In the financial year 2022/23, net current expenditure on adult social care in England was £20.2 billion, with approximately 10% allocated to social work activities such as needs assessment, care planning, and safeguarding [3]. Figure 9.2 illustrates the proportion of the adult social care budget spent on such activities, which has remained relatively stable at around £1.9–2 billion annually in real terms since 2015 [3]. While some suppliers claim that their LLM products can save up to eight hours of administrative time per worker per week [e.g. 217], there is not yet an impartial assessment of the time saved, or the quality of the documentation produced. Extrapolating supplier claims to England's £2 billion annual expenditure on social work activities suggests possible savings of £450 million.[1] However, such estimates depend on the extent to which documentation

[1] This is higher than the spend on social workers in England calculated from workforce data and unit costs. Skills for Care report that there were 15,600 filled, full time equivalent local authority social worker posts in 2021/22 [236], which multiplied by a unit cost of £85,174 per social worker in the same year [237] is around £1.33 billion. However, the Skills for Care data only includes qualified social work posts, and not the large but unknown number of unqualified care managers exercising social work functions [238]. It is more appropriate to use the ASC-FR figure to count the cost of assessment and care planning, regardless of whether this is undertaken by qualified social workers or non-qualified staff.

tasks can be automated effectively.

The capacity of LLMs to deliver these savings varies depending on the nature of the tasks involved. For example, summarising long narrative notes, such as detailed accounts of home visits, may lend itself to automation. However, such notes often include observations or professional judgements that cannot be automated. Case notes can contain information that may not be available to an LLM model. Alternatively, there may be instances where there are a significant number of words recorded, yet limited opportunity to save time. To demonstrate these points, I include in Figure 3.6 two examples of case notes from the administrative dataset where there were around 2,000 words recorded in a day. In the first example, all the words are in one long description of a home visit. In the second example, the words are spread across 33 case notes.



Day 1: 2000 words in 1 note

Time	Summary
16:30 (2000 words)	Outline of home visit including: - Verbal abuse from son towards mother (an older woman receiving social care services). - Allegations of financial abuse by mother regarding son. - Statement from son he is unable to cope with providing care and is leaving. - Statement from son that his mother needs to live in a care home. - Son's depature. - Visual assessment of social worker that mother appears appropriately dressed. - Note from social worker that the mother appeared clean and tidy and was not malodorous. - Note from social worker that mother appeared lucid. - Disclosure from mother of concerns about her son's behaviour and that of his friends. - Discussion about potential drug use by son. - Discussion of legal and practical options for mother's safety. - Potential and desire for son to enter drug rehabilitation. - Options for mother to enter residential care home. - Return of the son. - Statement from son that his mother is a liar. - Visual assessment from social worker that son appeared extremely anxious and jittery. - Description from social worker of son's behaviour searching through drawers and cupboards and his mother's handbags. - Second departure from son, stating he will not return again. - Social worker describes the state of the property, which is unclean, with very little food in the fridge and incontinence pads strewn all over the kitchen. - Social worker describes their observations of mother's mobility as she appears to have difficulty standing. - Social worker noted that there did not appear to be any visible medication and asked mother where it was. She did not know. - Social worker noted dirty laundry around property and handbags in the mother's bed. - Mother informed social worker she sleeps with her handbags to prevent theft. - Social worker advised mother to report her son to the police but mother declined, although she does not wish her son to live with her anymore. - Social worked noted that she felt the mother had the mental capacity to make this decision. - Social worker arranged with care agency for a regular care worker. - Social worker to contact housing to change the locks on the property. - Social worker agreed to return the next day and discuss with mother arrangements for her safety.

Day 2: 2000 words in 33 notes

Time	Summary	N words
09:01	Discharge required today	57
09:05	Problem with discharge. Ms Smith has no keys.	148
09:36	Call to care agency. They have keys.	48
10:00	Call from OT notifying of discharge time.	50
10:31	Ms Smith discharged. Could not enter house.	60
11:00	Ms Smith has been taken back to the ward.	61
11:15	Granddaughter unaware of hospital discharge.	30
11:17	Phone call to ward. They were unaware of readmission.	85
11:36	Phone from ward. Ms Smith is present but not readmitted.	85
11:45	Care agency will try to find worker with keys.	63
12:00	Care agency cannot yet find available worker with keys.	94
12:10	Hospital transport arranged.	108
12:20	Care agency updated.	96
12:21	Transport time updated.	61
12:30	Care agency has no keys yet.	7
13:20	Transport will not discharge without keys.	69
13:30	Son has keys.	76
14:10	Care agency updated.	68
14:12	Ward updated.	19
14:45	Care agency confirms worker.	27
15:06	Care agency awaiting transport at Ms Smith's home.	49
15:09	Ms Smith still on ward.	81
15:27	Care agency unwilling to wait indefinitely.	5
15:30	Ward unsure of discharge time.	6
15:36	Ward report discharge will happen tomorrow.	27
15:39	Care agency updated.	60
15:48	Services formally restarted from tomorrow.	15
15:50	Care agency informed.	99
16:02	Son informed.	94
16:14	Case conference rescheduled.	61
16:25	Granddaughter updated.	149
16:25	Ward updated.	32
16:32	Son updated.	10

Figure 3.6: Examples of case notes

It seems plausible that a single, long note which is a record of a visit would be a good example of a case note that could be automatically summarised from a transcript of the audio in considerably less time than it might have taken to write. However, closer examination of the note reveals that while an LLM might generate a note more quickly, it could not be equivalent as the note includes content that a summarisation model would be unable to generate. Firstly, there are observations that cannot be deduced from an audio transcript. For example, the social worker remarks that the mother was clean and not malodorous, the son appeared jittery, incontinence pads were strewn across the kitchen and there was very little food in the fridge. Secondly, the worker not only describes the conversation with the mother but also makes a professional assessment of her capacity to make decisions about her own safety. This is more than a summary. It serves to discharge a legal duty which, under the Mental Capacity Act 2005, cannot be met by an AI model. Furthermore, any such assessment by an AI model may constitute automated-decision making which is prohibited under the UK General Data Protection Regulation (GDPR).

The second example is quite different, but also seems to provide an example of a case where the potential for time-saving by LLMs may be limited. The 2000 words written in case notes on this day were spread across 33 notes over a period of over seven hours, and the maximum number of words in a single case note was 149. Each note records a brief phone call arranging one aspect of a discharge. For example, coordinating with the ward about hospital transport, the care agency about when their next visit will be, or with the son about access to the property. While it is technically possible for LLMs to record the phone calls and write up these notes, it seems unlikely that the majority of the time would have been taken by the transcribing of the calls rather than assessing the issues which need to be resolved prior to discharge, deciding who to contact and speaking with them. These two, quite different examples both demonstrate that there may be limitations to the benefits of generative LLMs for creating case notes.

It is not yet established how much time LLMs products might save, or how comparable their output is to that generated by humans. While it is certainly true that there are large volumes of case notes that take a long time to generate, it is not clear what proportion of these notes could be automatically generated. Substituting handwritten notes with automatically generated ones is likely to save time, but would not constitute an efficiency saving unless it were established that the documentation created was of comparable quality [239]. Realising potential savings depends not only on the appropriateness of LLMs for different documentation tasks, but also on the quality and accuracy of their outputs, which I explore in the next section. A comprehensive evaluation of potential efficiency savings, including factors such as adoption, scalability, and organisational readiness [240], is beyond the scope of this thesis, which focuses on specific aspects of the use of LLMs in social care. However, this remains an important area for future research and I return to it in Chapter 9.

3.3.1.4 Accuracy of LLMs used in care practice

The effectiveness of AI products to create or retrieve information from case notes is yet to be determined in social care. The established metrics for quantifying the accuracy of generated texts arise from machine translation, and work by comparing an LLM generated text to a reference text created by a human. This can be done lexically. For example, Bilingual Evaluation Understudy (BLEU) compares the overlap of n-grams (phrases of n or more words) in reference and generated texts and is considered a reliable metric [241]. Alternatively, Recall Oriented Understudy for the Gisting Evaluation (ROUGE) compares overlaps of n-grams, word sequences and word pairs [241]. As neither approach takes account of synonyms, often the Bidirectional Encoder Representations from Transformers (BERT) score is used, which measures semantic similarity through the cosine distance of the reference and generated text in vector space [241]. However, with medical records, Van Veen et al. [208] found that the correlation with human judgement is not higher than 0.25 for any of these methods.

It is not yet established whether these metrics are sufficient for evaluating accuracy of summarisation of social care records. However, both the performance of such metrics in healthcare and issues inherent in the way that the metrics are designed might suggest it is likely that new metrics will need to be developed. I demonstrate in Table 3.3 some examples of typical sentences that one might find in social care case notes.

Table 3.3: Reference sentences

Sentence No.	Sentence
1	She is very disabled and at risk of falls and abuse from her son.
2	He lives with his daughter and she provides all care.
3	She lives on her own and is lonely.
4	He has advanced dementia and cannot be left alone.
5	Grade 4 bed sore needs dressing changed 4x daily

I set out in Table 3.4 two possible generated versions of each sentence. Version *a.* is lexically similar, but semantically different, either omitting key information or contradicting the original sentence. In contrast, version *b.* is semantically similar but rephrased using different terms.

Table 3.4: Sentence summaries

Sentence	
1. She is very disabled and at risk of falls and abuse from her son.	
a	She is very disabled and at risk of falls and lives with her son.
b	Needs assistance with all ADLs. Safeguarding concerns re son. Falls risk.
2. He lives with his daughter and she provides all care.	
a	He lives with his daughter and she provides no care
b	Daughter shares the flat and gives informal assistance with ADLs
3. She lives on her own and is lonely.	
a	She lives on her own terms and is not lonely.
b	Resides alone. Socially isolated.
4. He has advanced dementia and cannot be left alone.	
a	He has mild dementia and can be left alone.
b	His Alzheimer's is severe and he must be accompanied at all times.
5. Grade 4 bed sore needs dressing changed 4x daily	
a	Grade 1 pressure sore needs dressing changed 1x daily
b	Extremely severe necrotic pressure ulcer requires QDS intervention

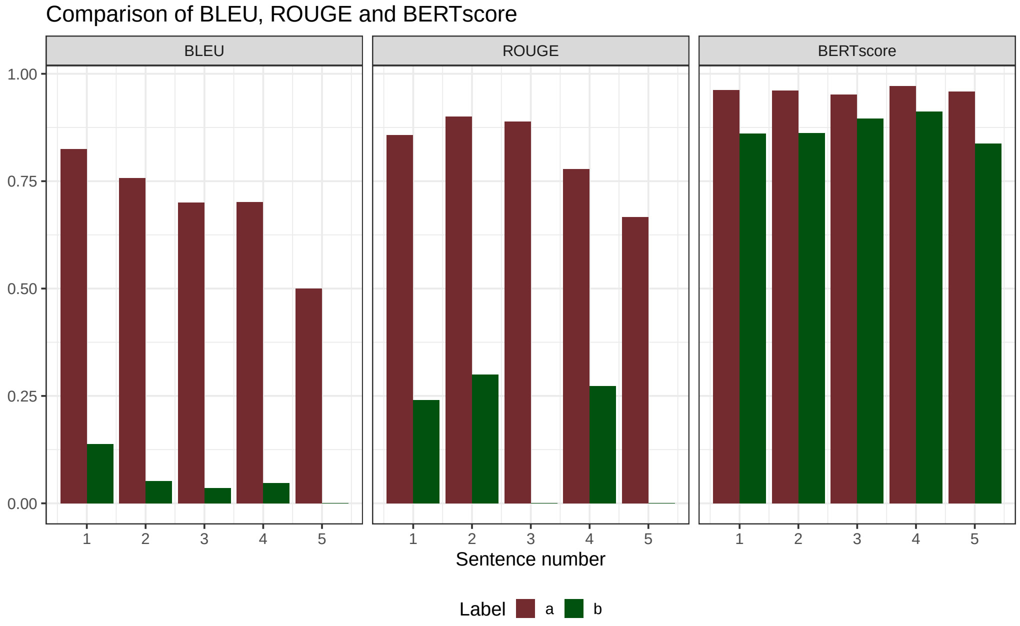


Figure 3.7: Rouge, BLEU and BERTscore

Figure 3.7 presents the ROUGE, BLEU, and BERTscore metrics for each version of the sentences. BLEU and ROUGE, which rely solely on lexical similarity, consistently rate the first version (which omits key information or reverses meaning) as closer to the reference sentence than the second version (which is lexically different but semantically similar). The second version frequently receives very low scores due to insufficient n-gram overlaps. While BERTscore captures some semantic similarity between the second version and the reference sentence, it still ranks the first version as more similar. This indicates that these metrics may not be appropriate for evaluating the accuracy of summaries. For instance, in the first example, the sentence omitting the risk of abuse scores highly across all three metrics.

Developing a suitable evaluation metric for free text summaries in social care presents unique challenges. The omission of even a few highly salient words, such as those indicating a risk of abuse, might have severe consequences. While Van Veen et al. [208] found that LLM-generated summaries of text records could outperform human summaries overall, it will be important to establish what the implications would be in the one-fifth of instances where LLM summaries were judged to be worse than those of professionals. Additionally, the study was limited by practical challenges, such as the length of doctor-patient conversation transcripts, which made some assessments of accuracy infeasible. Similarly, the size and complexity of social care data, combined with the lack of suitable evaluation metrics, make it impractical for me to assess summarisation accuracy within the scope of this thesis alongside the evaluation of bias and focus on LLMs for classification and generating inputs for statistical modelling. However, evaluating accuracy is central to the adoption of LLMs in social care, and I revisit it in Chapter 9.

3.3.1.5 Bias of LLMs used in care practice

Another important concern regarding the use of LLMs in care is the potential for bias [242]. Bias in LLMs refers to undesirable, systematic differences in outputs across demographic characteristics, particularly those that disadvantage groups who are marginalised or legally protected [243]. Such biases can be related to gender, race, ethnicity, socioeconomic status, or other characteristics.

Bias in AI products which generate social care documentation can exist in either of its components. Firstly, digital sound files are generated through audio recording of interactions with users of care services, and automatic speech recognition

(ASR) models are then used to create a transcript of these conversations [207]. Subsequently, a summarisation LLM is used to synthesise this transcript into a case note. There are racial disparities in ASR models, with transcription errors disproportionately affecting ethnic minorities [244]. Furthermore, ASR products are prone to hallucination, where gaps in speech are filled with fabricated content [245]. This is particularly concerning in a social care context, as harmful hallucinations were more likely to occur where individuals have aphasia, which is a symptom commonly associated with conditions such as stroke or dementia, which often lead to social care needs. Secondly, the summarisation LLM may exhibit bias such as the omission of relevant information or the inclusion of extraneous or even false information. Bias in LLMs used for relieving the administrative burden in social care could impact decision-making, as well as erode public trust in the use of such products in social care services.

The rationale for the research question in Chapter 8, which evaluates gender bias in state-of-the-art LLMs used to summarise social care notes, was driven by the need for careful evaluation of bias in these models. While the potential for bias spans various social dimensions, this thesis focuses on gender bias for several reasons. Firstly, gender disparities in social care are well-documented, with women representing the majority of care recipients, unpaid carers and care workers [246, 247, 248, 249, 107]. Decisions made about social care disproportionately affect women [250], and evaluating gender bias in LLMs aligns with these broader equity concerns. Secondly, gender bias in LLMs continues to be a focus of ongoing research, and it has been observed in LLMs used in other domains, such as hiring algorithms and language translation [251, 252]. Finally, as gender is a prominent and easily identifiable characteristic in social care case notes, it provides a clear entry point for analysing how biases might manifest in LLM outputs.

Although this thesis narrows its focus to gender bias, the same methodological and conceptual challenges apply to evaluating biases across other dimensions, such as race or socioeconomic status. These broader questions, alongside the issues of accuracy and applicability, remain essential to the responsible deployment of LLMs in social care and are revisited in Chapter 9. The same concerns apply to the use of LLMs extracting structured information from administrative records to support the evaluation of social care services, as I discuss in the next section.

3.3.2 LLMs for evaluating social care services

Administrative social care records contain contemporaneous, detailed information about the characteristics of people receiving social care services. However, large volumes of this data are stored as free text. Traditional NLP methods of extracting information from free text are often limited [253]. LLMs may offer more potential for success through their improved capacity to understand natural language, making it possible to identify and structure relevant information from complex and heterogeneous datasets.

It is difficult to quantify the amount of information stored in free text administrative records, though in Chapter 4 I give an overview of the 114.4 million total words of free text recorded for the 3,046 individuals included in this thesis. While size of the dataset is easiest to measure, the amount of information contained in these words will depend on the heterogeneity, noise and complexity of the data [254]. The range of the sources of the information recorded, the richness of the text, and how the documents relate to one another contributes to the amount of information contained [255, 256]. In social care, text may be gathered from several sources, such as the person receiving care, medical professionals or family members, and this variety is not measured in bytes. Conversely, unstructured data can contain noise, which increases size without a corresponding increase in information. Noise could be created through spelling error, poor grammar or transcription error which render text data impossible to decipher [257, 258]. There might also be repetition. An example seen in the administrative dataset is that occasionally a worker may copy and paste emails into case records, choosing to include the entire thread each time they send or receive a new message, inflating the size of the data without increasing the information contained. Finally, the amount of information contained in data also depends on its complexity, i.e. the amount of work required to extract information from data. Gandomi and Haider [254] describe the idea of data complexity as *the need to connect, match, cleanse and transform data* (p139).

Structured and unstructured data both require cleaning. I outline the process with structured data in Chapter 4. The process of extracting information from unstructured data is a central theme of this thesis. While the precise amount of data is difficult to quantify, 114.4 million words is too much for an individual to read. Meadow and Yuan [201] define data as *potential information.* In these terms, if LLMs can increase the potential for extracting information from administrative data over traditional NLP models, it has the capacity to improve understanding of the needs, interventions and outcomes of social care users.

3.3.2.1 Predictive modelling and early intervention

The LOTI report highlights predictive modelling as one of the most promising applications of LLMs in social care [14]. Specifically, it suggests using LLMs to extract information for use in machine learning models to forecast the likelihood of a person being hospitalised within a given timeframe. The proposed system would generate a prioritised list of users at risk, providing workers with risk scores to guide triaging decisions. For instance, when logging into a case management system, social workers would be able to see which individuals are at high risk of hospitalisation in the next six months. A model would integrate structured, needs-related metrics such as mobility, prior hospitalisations, and care visit frequencies with free text data about the needs and circumstances of care users. The idea of such products is that support could be targeted on this basis, though this would require further work into whether there are effective interventions to reduce risk of hospitalisation in those identified.

LLMs may be a useful tool to extract the information from the large quantity of data in administrative records, as they offer the potential to extract meaningful insights from free text, filling the gaps left by survey data and structured administrative data. I examine the capacity of LLMs to extract information from free text administrative records in Chapter 6, where I also compare their accuracy against traditional NLP models. However, while NLP and LLMs are widely used with health data [e.g. 259, 260, 261, 152, 262, 153, 154], I am aware of only a handful of papers which use similar methods with social work or long-term care data [149, 150, 151]. The analysis in Chapter 6 is, to my knowledge, the first paper using LLMs to extract characteristics of care users from administrative social care records for the purpose of understanding the relationship between needs and service use, and the first study to extract any information from free text social care data in England.

3.3.2.2 Using LLM-extracted data for understanding service use

One purpose of extracting structured from unstructured data is to use it to evaluate social care services. An important area for evaluation relates to potential effects of changes in the types of care services provided. Internationally, the increased pressure on public funds of an ageing population has led to a retrenchment in community care services, with more focus on personal care and nursing tasks, and services for loneliness and social needs shifted to unpaid carers [see 5, 263, 264, 6].

However, there may be unintended consequences of this approach. Social inclusion services such as day centres can lead to a reduction in loneliness [265], a factor associated with higher risk of care home admission for older people observed in surveys [266, 267]. Despite this, there has been limited analysis of how reducing the availability of these services for those with the highest needs — particularly individuals receiving publicly funded care — affects the likelihood of care home entry. This informs the research question in Chapter 7, which explores loneliness as a risk factor for care home entry among older social care users using structured data extracted in Chapter 6. While this approach can highlight a strong association between loneliness and earlier care home entry, it cannot establish whether reducing loneliness reduces this risk. Nonetheless, these types of analyses demonstrate how information extracted with LLMs can deepen understanding of the relationships between needs and service use, providing valuable insights for policymakers and paving the way for further research into the consequences of service changes.

3.4 The importance of data quality

While the potential for LLMs to transform unstructured administrative records into meaningful insights is significant, their effectiveness is inherently tied to the quality and reliability of the underlying data. Before we can confidently apply LLMs to extract information and evaluate social care services, it is crucial to evaluate the extent to which the data they are being applied to is robust. In the following chapter, I turn my attention to the dataset used in this thesis, detailing the processes involved in acquiring, pseudonymising, and preparing the data for analysis. By thoroughly examining the steps taken to address challenges such as data quality, completeness, and consistency, I aim to provide a clear understanding of the strengths and limitations of the data used in the subsequent analyses. This rigorous assessment not only underscores the importance of domain knowledge when using LLMs for social care research, but also demonstrates the reliability of the findings.

4 Establishing the reliability of administrative data

In this chapter, I examine the data used in this thesis, focusing on the processes involved in acquiring, pseudonymising, and preparing the dataset for analysis. I detail the steps taken to ensure data quality, including addressing challenges such as missing information, over-redaction of free text, and inconsistencies in service cost and needs assessment data. By documenting these processes, I aim to provide a clear understanding of the dataset's structure and reliability, which serves to demonstrate the robustness of the findings in the analyses presented in subsequent chapters. This chapter also illustrates the substantial amount of complex and iterative work required to assess data quality and transform administrative records into a format suitable for research, which is a key finding of the thesis. The nature of this process underscores the importance of understanding the intricacies of data collection and processing in social care. I also describe in this chapter the demographic characteristics of individuals receiving services, highlighting patterns in age, ethnicity, and support needs.

4.1 Data pseudonymisation and egress

The dataset used in this thesis comprises the pseudonymised, local authority administrative records for older adults using adult social care in a London local authority. This group was defined as adults aged at least 65 years on the 31st August 2020 who had been in receipt of long-term (defined as at least one year) community adult social care services during the period 1st January 2016 - 31st August 2020. This included 3,046 individuals in this local authority.

4.1.1 Information governance and ethics

The project partners were the Care Policy and Evaluation Centre (CPEC) at the London School of Economics and Political Science (LSE), the local authority and the local NHS Commissioning Support Unit (CSU). These organisations made a joint application to the NHS Confidentiality Advisory Group (CAG) for permission to use free text administrative social care records and received approval in August 2020 (application reference number 20/CAG/0043). This approval was renewed yearly. A Data Processing Agreement was signed between CPEC and the local authority, who remained the data controllers, while CPEC acted as data processors. The free text data was pseudonymised by the local authority, then passed to the CSU for further removal of any remaining identifiable information, such as NHS numbers, before being sent to CPEC. The data flow process is illustrated in Figure 4.1. The first data extract was sent using this process in August 2021.

This study uses secondary data from administrative records, which were pseudonymised prior to egress to remove identifiable personal information (e.g., names, addresses, NHS numbers, and other unique identifiers). This process is described in the next section. According to the UK General Data Protection Regulation (GDPR), processing of these data was conducted under the legal basis of legitimate interests, which does not require individual opt-in consent. A Data Processing Impact Assessment was completed, and the details of the project were made publicly available via a Privacy Notice on the local authority's website, with local opt-out options provided. Ethics approval for the project was granted by the LSE Personal Social Services Research Unit's ethics committee on 30th May 2019, in compliance with the LSE's Research Ethics Policy.

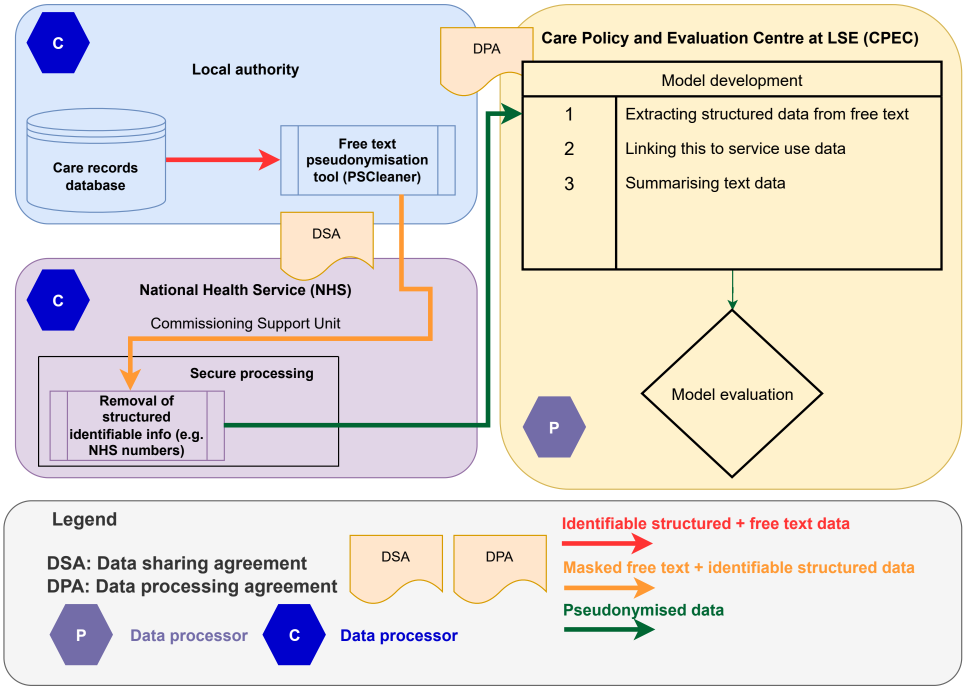

C
Local authority
DPA
Care records database
Free text pseudonymisation tool (PSCleaner)
DSA
C
National Health Service (NHS)
Commissioning Support Unit
Secure processing
Removal of structured identifiable info (e.g. NHS numbers)
Care Policy and Evaluation Centre at LSE (CPEC)
Model development
1 Extracting structured data from free text
2 Linking this to service use data
3 Summarising text data
Model evaluation
P
Legend
DSA: Data sharing agreement
DPA: Data processing agreement
DSA
DPA
P Data processor
C Data processor
Identifiable structured + free text data
Masked free text + identifiable structured data
Pseudonymised data

Figure 4.1: Data flow and data sharing agreements

4.1.2 Development of the free text pseudonymisation tool

The free text data was pseudonymised using *PSCleaner* [268], an open-source tool developed by the CSU specifically for this project. This tool was designed to remove personally identifiable information, such as names, telephone numbers, locations, dates, and currency from text data. Development was an iterative process. Prior to the extraction of the data, I received an early version of the tool and its source code in September 2020, and wrote some tests to ensure the tool's reliability.[1] During this process, I identified some words that were incorrectly redacted, including relevant terms like "hygiene," "enjoy," and "discomfort." I shared this list with the developers, and we arranged a meeting to discuss the issue. I pointed out the area in the code where it appeared these false positives were occurring. The developers subsequently updated the software, and these words were no longer redacted.

After this issue was resolved, the local authority conducted further testing of the tool in October 2020 on a sample of 13 individual records. They manually verified that these records contained no personal information before sharing them. My assessment confirmed that while no personally identifiable information remained, some instances of over-redaction persisted. Specifically, responses to structured questions, which could not contain free text, were unnecessarily redacted. I raised this with the local authority, who reviewed the affected fields. It became apparent that non-dictionary words beginning with capital letters were being misidentified as names. For example, responses such as *SuppLiv* (supported living), *Shelt* (sheltered housing), *Nurs* (nursing care home), and *Res* (residential care home) were all redacted. Once this issue was highlighted, the developers updated the code again. Additionally, the local authority decided not to apply the pseudonymisation software to responses where free text could not be entered. We continued with further testing and established a fortnightly meeting schedule with the local authority and CSU from December 2020 to August 2021. The purpose of these meetings was to ensure all stakeholders were confident that the pseudonymisation software was fully refined and to discuss data egress procedures. The successful development of the tool was a sustained and collaborative process, relying on ongoing commitment and engagement from all parties over several months to refine and implement the solution.

[1] The tests were written in Python and TypeScript [269], as the software used regular expressions, which can behave inconsistently across different programming languages due to variations in syntax and implementation [270]. For details of the code, see the GitHub repository [268].

4.2 Reliability of structured data

I received the first data file in August 2021, marking the beginning of a collaborative process of assessing data quality which continued until July 2022. Over the course of this period, I engaged in an iterative cycle of identifying anomalies, providing feedback, and receiving clarifications or additional data as needed. This section outlines that process to demonstrate the robustness of the findings derived from this dataset. The main components of the data received from the local authority are detailed in Figure 4.2. This figure also includes the dates each file was received, highlighting the iterative nature of the data extraction process needed to resolve issues like missing records and metadata.

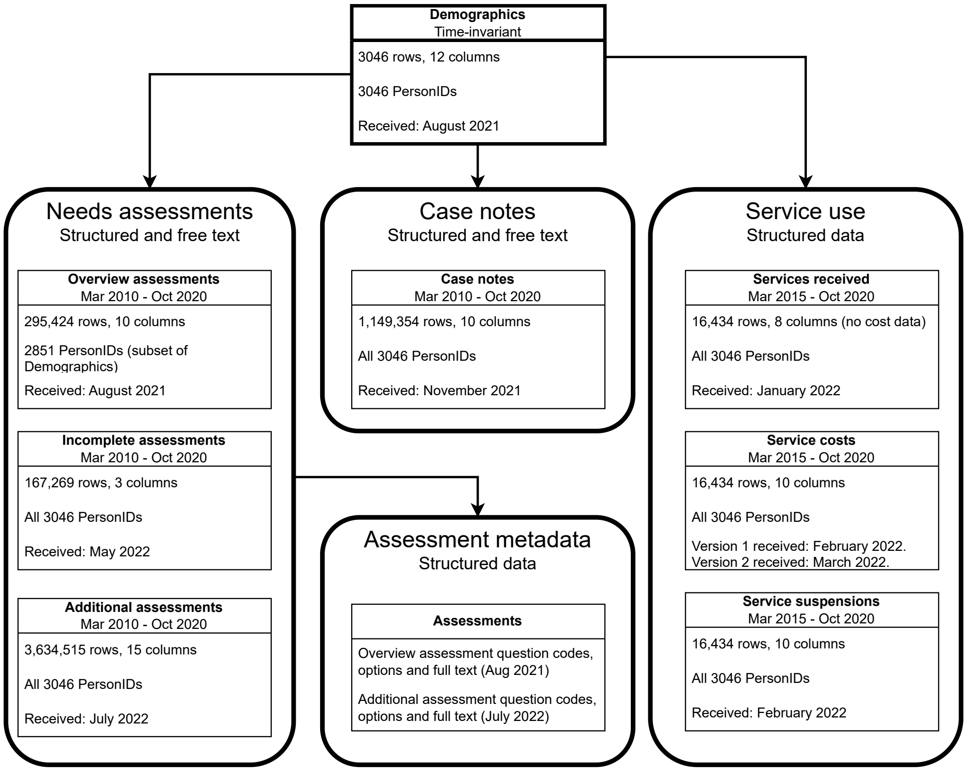


Figure 4.2: Outline of data received

A query was developed to identify all individuals in the borough who were aged 65 years and over by 31st August 2020 and who had been receiving adult social care services in the community for at least a year at some point since 1st January 2016. This resulted in a cohort of 3,046 people, included in the *Demographics* file received in August 2021, which contained information on sex, year of birth, year of death (if applicable), and ethnicity.

4.2.1 Needs assessment data

The first file extracted contained needs assessment data for 2,851 individuals. Upon reviewing the data, I noted that each of the 295,424 rows corresponded to the response to a specific assessment question for a given PersonID on a particular date. The needs assessment form template I had been sent comprised 49 questions, so I initially expected 49 rows per person for each assessment date. However, I discovered many instances where only two or three responses were recorded, and others where between 30 and 52 questions had been answered, as illustrated in Figure 4.3.

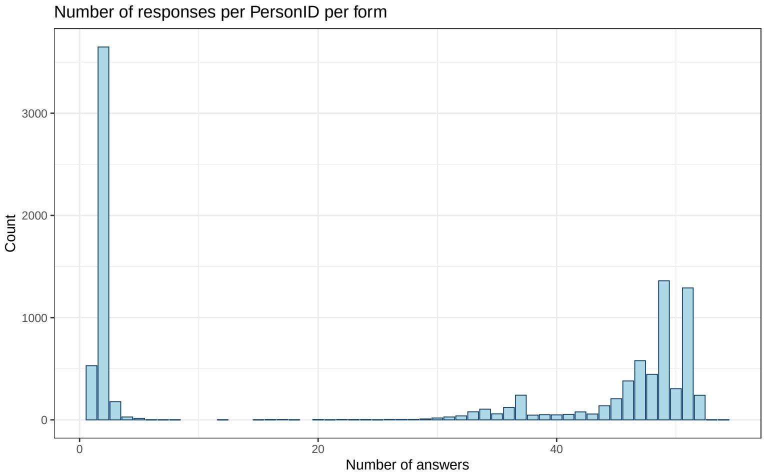


Figure 4.3: Number of responses per PersonID per form

Through my weekly meetings with the council, I was able to determine that there had been several minor changes to the assessment form over the period in question. This explained the clustering of responses around 49, corresponding to periods when the form had a different number of questions. Lower numbers of responses, particularly very low numbers, likely represented forms that had been started but not completed. While it was straightforward to identify forms with fewer than, for example, 40 responses, this presented a potential issue. If a form was incomplete, the responses should not be considered accurate and ought to be excluded from the analysis. However, determining whether a form was incomplete based solely on the number of responses was challenging. As the number of fields changed over time, it was difficult to establish whether forms with fewer than the maximum fields were unfinished, particularly as answers from previous forms were in some cases automatically carried through when a new form was generated. Furthermore, a form could have all fields filled in, yet still not be finalised. I communicated these concerns to the local authority, and in August 2021, they provided the *Incomplete Assessments* file. This additional data allowed me to exclude assessments that had not been finalised from the analysis. The distribution of the number of responses per PersonID per date after this exclusion process is shown in Figure 4.4. This reflects the varying number of questions in different versions of the assessment form over time. There were no longer any forms with fewer than 46 fields included, as the data only contains complete forms.

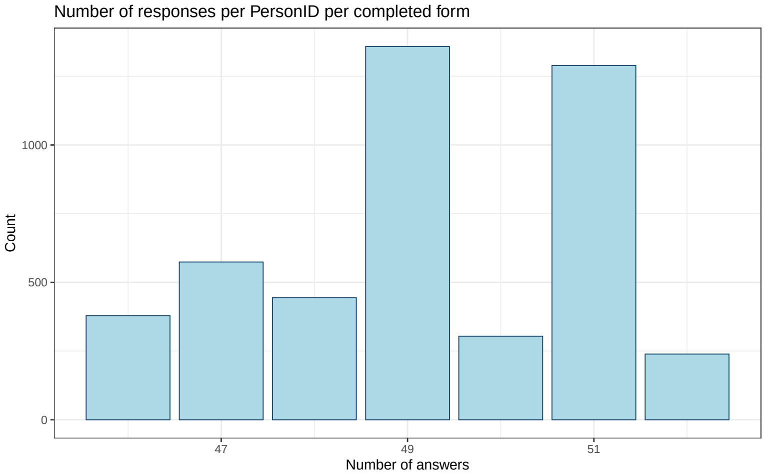


Figure 4.4: Number of responses per PersonID per completed form

Another significant issue with the data at this stage was that all individuals receiving a service should have undergone a needs assessment, so it was unclear why the data only included a subset of individuals: 2,851 out of 3,046. Upon closer examination of the service use data, it became evident that some data was missing. I will outline this process below.

4.2.2 Service use data

I initially received the service use data in January 2022. This dataset contained a row per PersonID for each service, along with the start date, end date, and type of service (e.g., direct payment, residential care, day centre). However, it did not include cost data, which was essential for the analysis, particularly in the analysis of loneliness on care home entry, where costs following the initial assessment were included as a covariate. The local authority was highly motivated to extract all relevant data throughout the process. After I provided a rationale for the necessity of cost data, their Information Governance lead confirmed that it was within the original project scope, and the cost data was sent in February 2022.

Upon reviewing this data, while the overall distribution of costs appeared correct, some questions arose regarding specific aspects of the data. There were ‘NULL’ values, indicating a missing or undefined entry, raising questions about cost data completeness. Notably, some services recorded with ‘NULL’ costs were unlikely to be without cost, as outlined in Table 4.1. Additionally, the cost data contained only a single numeric field, making it unclear whether this figure represented the cost to the local authority or the individual receiving the service.

Table 4.1: Services with NULL costs

Service	N
Community - Homecare	363
Community - Equipment etc	83
Reablement	72
Permanent nursing placement	34
Permanent residential placement	26
Community - Day care	25

Further discussions with the local authority revealed that the database actually contained four cost fields, but only one had been extracted. A new version of the data was provided in March 2022, along with the internal algorithm used to determine which of the four cost fields should take precedence. After applying this algorithm to the data, the only services with NULL or zero weekly costs were those where this was appropriate, such as equipment, health-funded services, and reablement (which are block-funded rather than individually-costed). Costs were now associated with all residential, nursing, and domiciliary care services. These adjustments ensured that service cost data was comprehensive.

4.2.3 Extraction of all needs assessment data

After receiving time-variant cost data, I was able to investigate the issue of the missing assessment data. I plotted the trajectories of cost over time, overlaid with the dates of each assessment. The Care Act assessment process mandates that there should be an updated assessment whenever there is a change in need. However, it was clear that there were significant changes in provision of services which were not preceded by an assessment. I include in Figure 4.5 a sample of nine such cases, with assessment dates overlaid on to the cost trajectories. I discussed with the local authority that the extracted assessment forms seemed to be a subset of all needs assessment. This would explain why there were some individuals with no needs assessments at all, and others with changes in service but no corresponding record of a change in need.

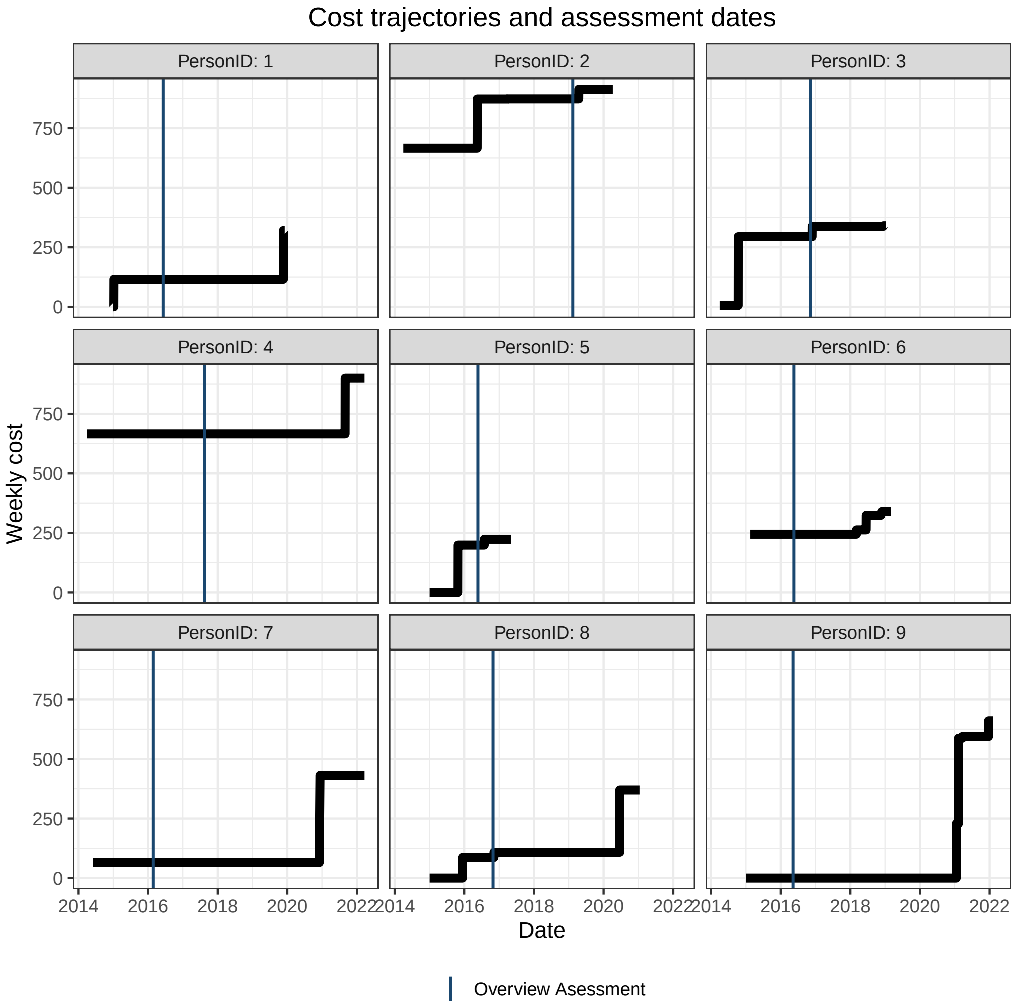


Figure 4.5: Cost trajectories and assessment dates (original forms)

In response to this, the local authority generated an extract containing all completed forms for the individuals in question, as the original extract of assessment forms had only included one form. Although this was the primary assessment form used during the period of observation, several other assessment forms were also in use. The use of different forms was not related to need but to internal processes. If an assessment and care plan were already in place, generally a review form would be completed, though this contained almost identical fields to an assessment form. Additionally, the primary assessment form was introduced shortly after the Care Act 2014 came into force in April 2015, and there were other assessment forms used during the period of observation. While there were nine assessment forms in total, there were multiple versions of each form over time, as questions may be added, or drop-down response options changed (though the majority of these changes are relatively minor). I include in Table 4.2 a breakdown of the versions of each form in the administrative dataset used in this thesis.

Table 4.2: Versions of assessment forms in the administrative dataset

Form Name	N	%	Versions	Form Type
Assessment Form 1	9099	38.2	20	Assessment
Assessment Form 2	3697	15.5	15	Assessment
Assessment Form 3	3146	13.2	3	Assessment
Assessment Form 4	2284	9.6	9	Assessment
Review Form 1	1735	7.3	15	Review
Assessment Form 5	1194	5.0	13	Assessment
Assessment Form 6	1162	4.9	9	Assessment
Assessment Form 7	789	3.3	13	Assessment
Review Form 2	722	3.0	3	Review
Total	**23828**	**100.0**	**100**	**-**

Although it happens to add up to 100, the number of versions of forms is a count and not a percentage.

I include in Figure 4.6 the previous care trajectories, now with the additional relevant forms overlaid. After producing these plots and discussing them with the local authority, it became evident that the additional assessment forms were relevant, as they were capturing need prior to a change in service provision. In July 2022, the local authority extracted and provided this additional data, which included around 3.6 million rows of assessment questions, encompassing all 3,046 PersonIDs. This is labelled as *Additional Assessments* in Figure 4.2.

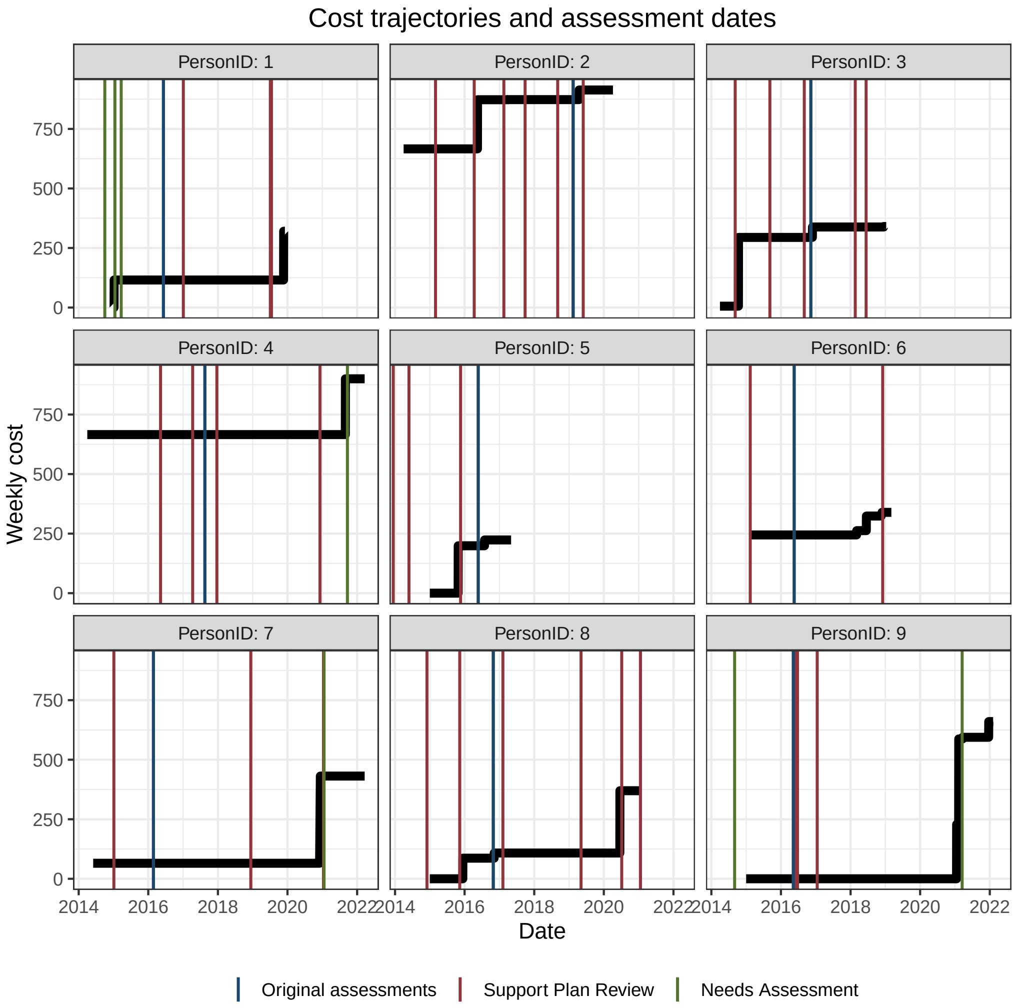


Figure 4.6: Cost trajectories and assessment dates (additional forms)

4.2.4 Establishing reliability of structured fields

Once it was established that the coverage of the data was comprehensive, I needed to ascertain which fields were reliable. In this section I describe this process, which consists of comparing the distribution of characteristics like age and ethnicity to the expected distribution from national datasets. I also compare the responses in related fields for the same individuals. The conclusion of this is that functional ability fields recorded in the needs assessment forms appear to be reliable: the distribution is as expected, they are updated regularly, and internally consistent. However, not all information recorded in administrative records is reliable. In particular, the Primary Support Reason field (PSR) does not appear to reflect expected levels of need, and I do not use it in the subsequent analysis in this thesis.

4.2.4.1 Age

The study is focused on older adults who are long-term users of adult social care. The definition of this group is: adults aged 65 years and over by the 31st August 2020 who were in receipt of adult social care services in the community for at least a year at some point since 1st January 2016. This included 3,046 individuals in this local authority. The median age of all individuals during the period of observation was 81 years, with a distribution as set out in Figure 3. The demographics data provided year of birth and year of death, so the mean is an average of each age rounded to the nearest year. Individual age in 2020 was calculated as 2020 − year of birth, except in cases where the person had died before 2020, in which case I used year of death − year of birth.

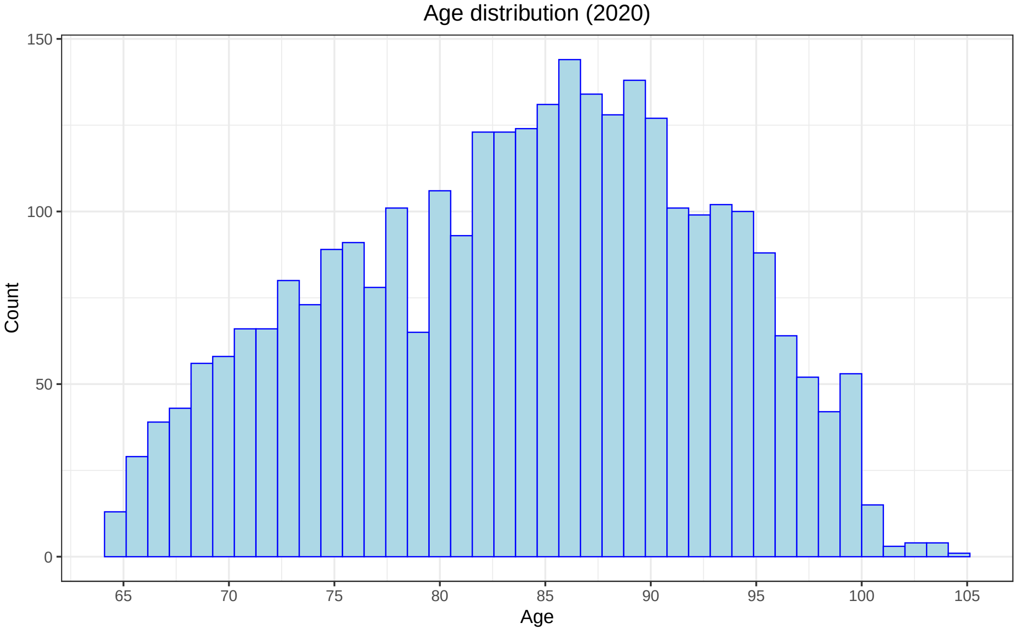
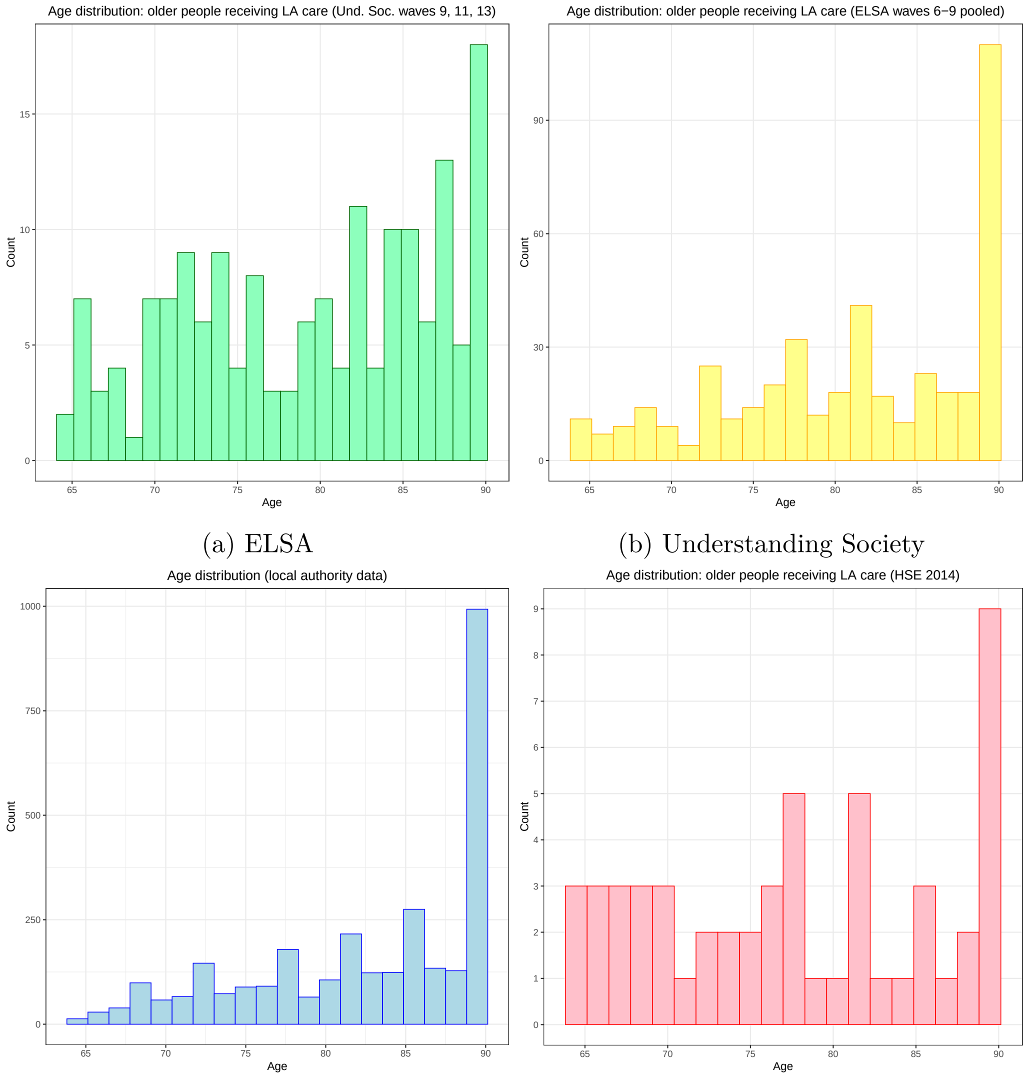


Figure 4.7: Age distribution

In Figure 4.8, I present a comparison of the age distribution for individuals aged 65 and over who report receiving local authority care, using survey data. I focus on the 2014 wave of the Health Survey for England (HSE), as subsequent waves report age in five-year bands. In both HSE and the English Longitudinal Study of Ageing

(ELSA), individuals over 90 are grouped together as aged 90. For consistency, I apply the same grouping to the administrative data and Understanding Society. With individuals over 90 banded together, the mean ages are 78 in HSE, 79 in Understanding Society, 81 in ELSA, and 82 in the administrative data. ELSA and HSE show a higher proportion of people at the lower end of the age range, which may reflect the issue discussed in the introduction: as people age and their needs increase, they are less likely to participate in surveys. However, since the totals for these groups are relatively low in absolute terms, it is difficult to draw firm conclusions. Overall, the age distribution in the administrative data aligns broadly with expectations based on survey data, with the distributions appearing relatively similar and the largest group being those aged 90 and over.



(a) ELSA

(b) Understanding Society

(c) Administrative data

(d) HSE

Figure 4.8: Comparison of age distribution with survey data

4.2.4.2 Ethnicity

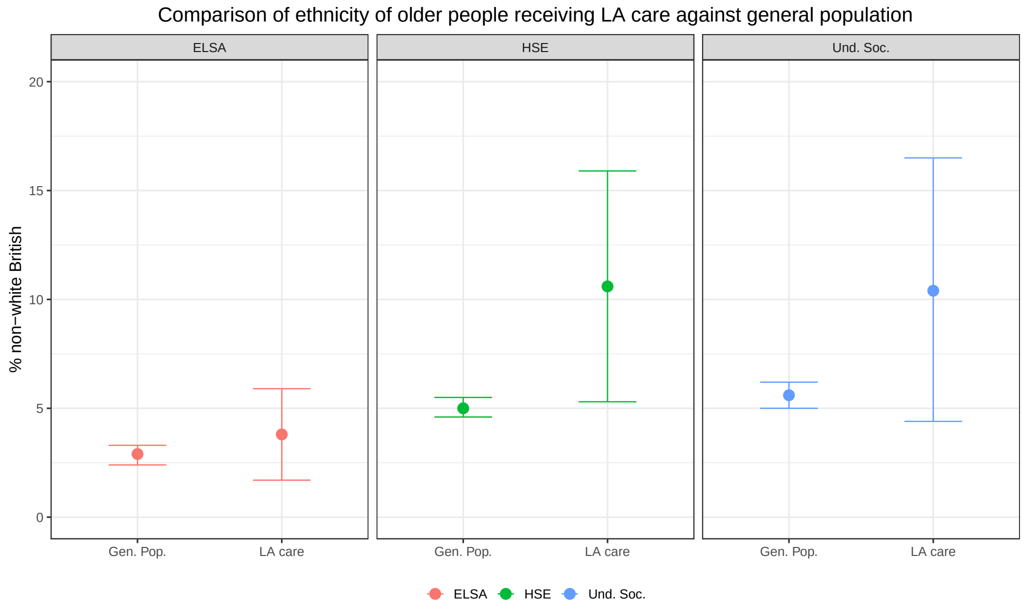
Table 4.3 shows the breakdown of ethnicity by gender in the administrative data. Almost half (47.8%) of those receiving services are recorded as White British. It is worth noting that there is significantly more ethnic diversity among those receiving statutory adult social care compared to the borough's general population, where 79.6% of individuals aged 50 and over were White British in 2021. In that same year, there were 12,700 people over 50 from minority ethnic backgrounds in the local area, although a detailed breakdown by age, local authority, and ethnicity is not available [271]. While the 1,590 people who were not White British cannot be directly compared to this figure — as they represent a flow of individuals over five years — it appears that individuals from minority ethnic backgrounds are more likely to receive statutory adult social care.

Table 4.3: Breakdown of ethnicity by gender (administrative data)

Ethnicity	F	M	Total	%
White British	881	575	1456	47.8
White other	402	227	629	20.7
Black	280	179	459	15.1
Not known	189	132	321	10.5
Asian	94	53	147	4.8
Mixed	21	13	34	1.1
Total	**1867**	**1179**	**3046**	**100.0**

Note: The raw data does not contain categories for Black, Asian or mixed British.

Given that this is a London borough, we would expect to see a higher proportion of individuals from ethnic minorities compared to surveys covering the whole of England, so it is not appropriate to compare absolute proportions with national survey data. However, in Figure 4.9, I compare the proportion of non-White British individuals in surveys who report receiving local authority care with the proportion of non-White British people in the general population aged over 65. Although the confidence intervals are wide due to the small numbers of people receiving local authority care, the general trend remains consistent: a higher proportion of individuals from ethnic minorities receive local authority care. This trend supports the reliability of the ethnicity data in the dataset. In Chapter 5, I also compare the proportion by ethnicity with the Adult Social Care Survey (ASCS) data, finding the proportion of individuals from ethnic minorities in this dataset is very close to that recorded in ASCS, which supports the same conclusion.

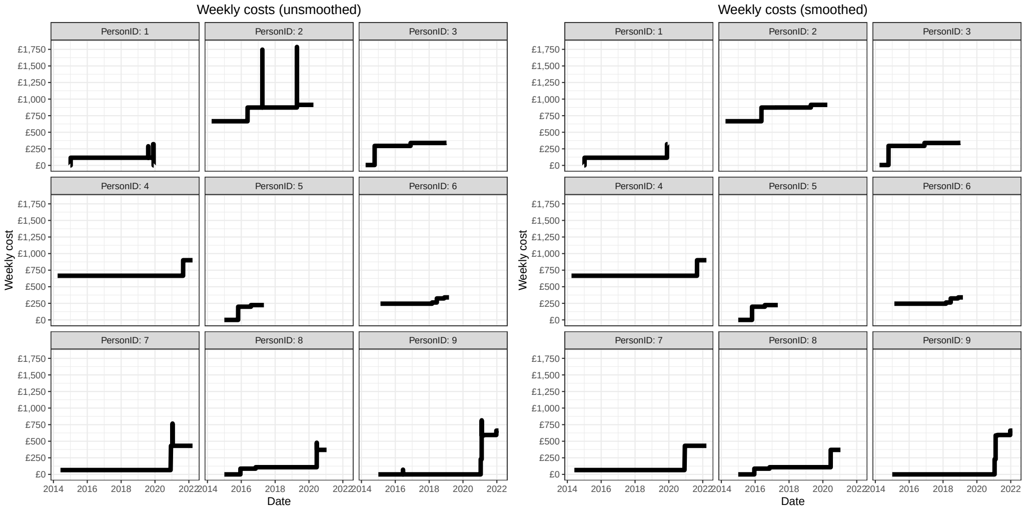


ELSA: Waves 6−9. Understanding Society: Waves 9, 11, 13. Health Survey for England: 2015−2019

Figure 4.9: Comparison of ethnicity of older people receiving LA care

4.2.4.3 Service costs: data cleaning/cost smoothing

I received service cost data in the format of a list of services provided for each PersonID, with each row containing start and end dates and the service cost. This table contained 16,434 rows. I transformed this into a daily panel format, with a row for each service received by each PersonID for every day that they received a service. In this format the data comprised 6,372,162 rows, and enabled straightforward calculation of the total cost of services per person per day by summing all services provided on that day. However, this approach led to a recurring issue where service costs appeared to briefly double. I set out in Figure 4.10a an example of this across a sample of nine cost trajectories. PersonID 2 appears to jump from requiring around £850 to £1700 a day of services in 2017 and 2019. This individual resided in the same nursing care home from 2016 onwards. However, when the weekly price was subject to a yearly contractual uplift, the service was formally ended, and a new service was created. The end date of the previous service and start date of the new service were the same, leading to this apparent doubling of cost for one day. This occurs frequently, and we see the same effect for PersonIDs 1, 7, 8 and 9. I resolved this by assuming that where such spikes occurred at a time that a service ended and a new one began, this reflected the data entry process rather than an actual doubling of services received, and the true cost that should be used is the cost of the new service. The result of the same trajectories after this can be seen in Figure 4.10b.



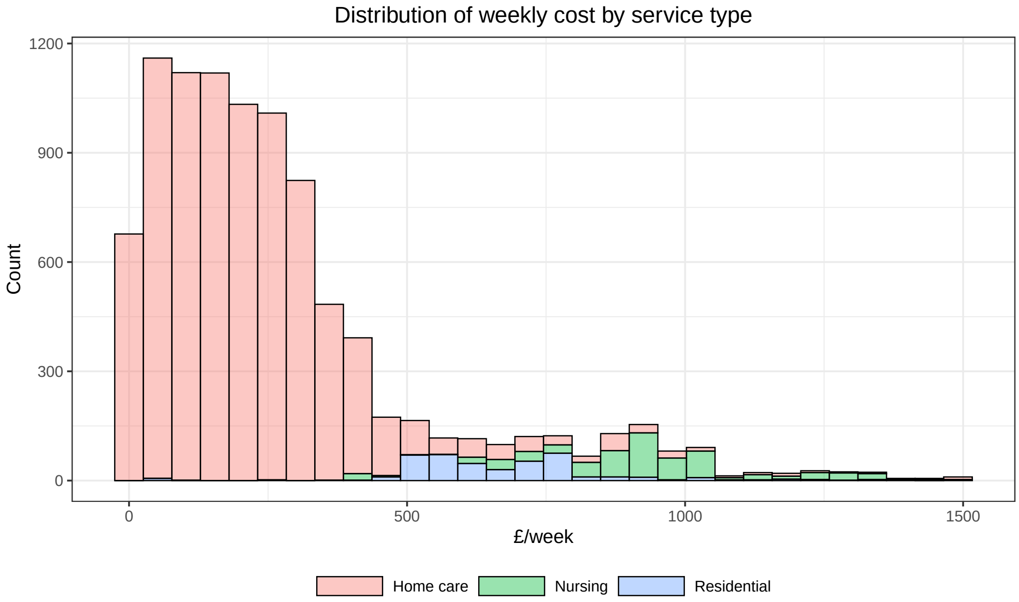
(a) Service cost over time (unsmoothed)

(b) Service cost over time (smoothed)

Figure 4.10: Service costs: before and after smoothing

4.2.4.4 Service costs: distribution of costs

I present in Figure 4.11 the distribution of cost for each service type. The plot appears as I expected. Cost data should be right-skewed, with the distribution of community home care packages dropping off at the point it becomes less expensive to meet the person's needs in a residential care home. I also expected to see the distribution of nursing care to the right of residential care, as it reflects individuals with higher needs. There were a handful of outliers, such as around £200 per week residential or nursing care costs in some years, which I was able to confirm with the local authority were errors and were straightforward to remove.



Note: A small number of services with costs > £1500 per week are excluded from this plot

Figure 4.11: Distribution of service costs

The data includes service-level costs for all services but does not provide information on the number of units delivered. This means it is not possible to determine the unit cost for services that can be provided in varying amounts, such as home care (which might be delivered for different numbers of hours per week) or day care. However, it is possible to determine unit costs for nursing and residential care, as individuals require exactly one unit (i.e. bed). I compared these costs with the Unit Costs of Health and Social Care from 2016 to 2020 [272, 273, 274, 275, 276]. To account for higher land and labour costs in London, I multiplied the Unit Cost by the mean London multiplier of 1.203 provided in the unit cost data over these years. The results, shown in Figure 4.12, indicate that the administrative costs are very close to the expected unit costs. Additionally, the costs of residential and nursing care in the administrative data make sense relative to each other, with nursing care being more expensive than residential care. Both also show a general trend of increasing by around 3-5% per year, which suggests that the cost data is reliable.

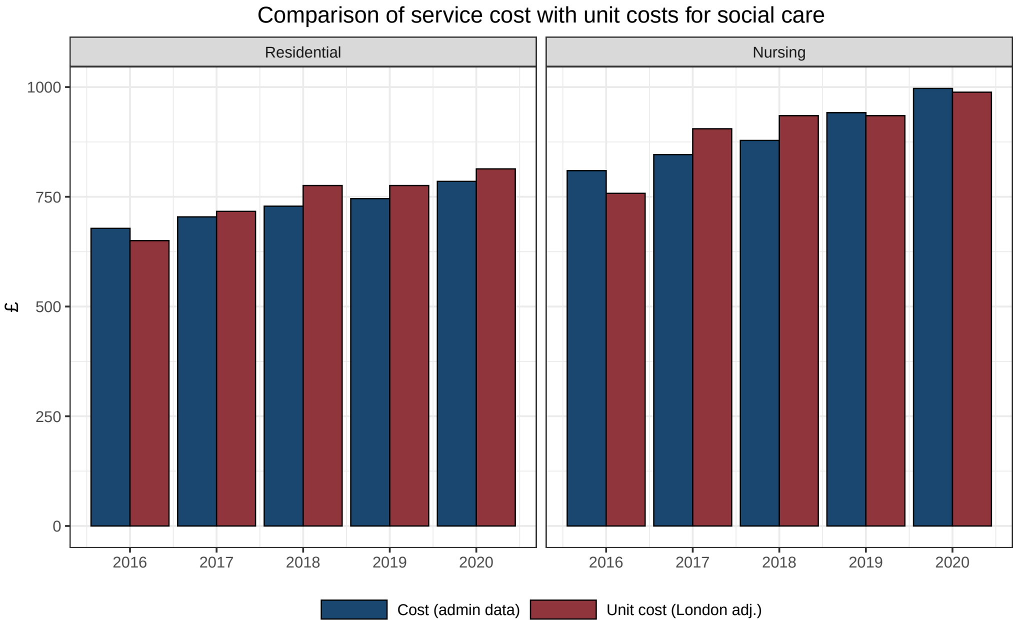
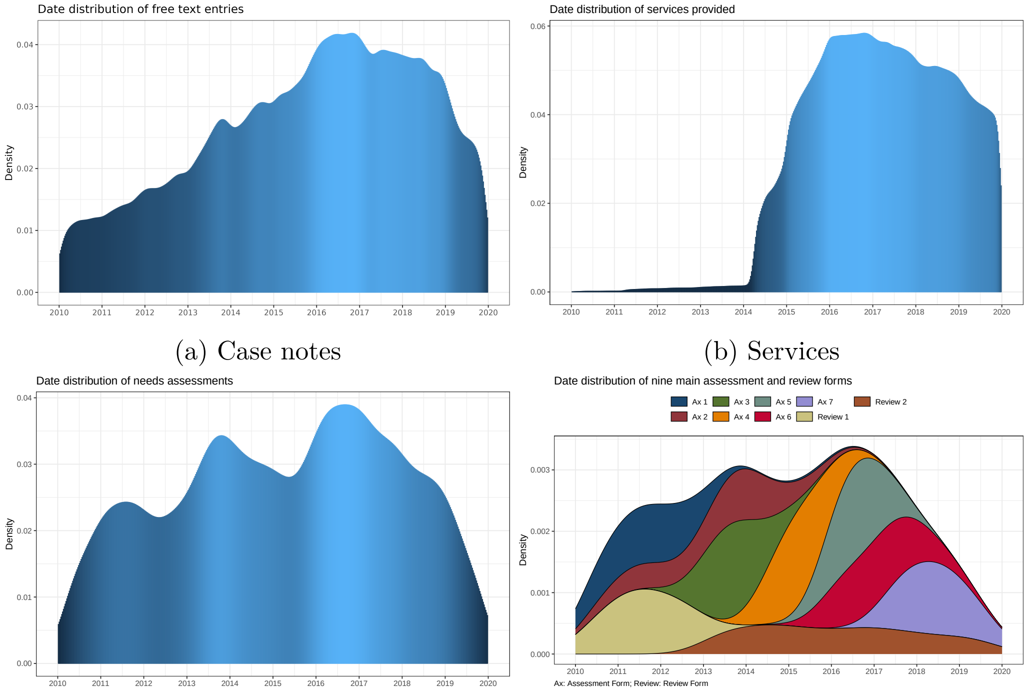


Figure 4.12: Comparison of administrative service costs with unit costs for social care

4.2.4.5 Date distribution of records

In Figure 4.13, I illustrate the date distribution of records across each dataset. There are certain artefacts in the data, influenced by how the cohort was defined and the data extraction process, which must be considered during analysis. The cohort selection query required individuals to have been receiving services since at least 1st January 2016 and to have been receiving services for at least one year at some stage the period until August 2020. The expected decline in assessments, notes, and services from 2019 reflects that individuals who began receiving services after August 2019 could not have been included in the cohort, as they could not meet the one-year service receipt requirement. The extract includes all needs assessments and case notes for the 3,046 individuals in the cohort, which explains the gradual decline observed prior to 2016 in Figure 4.13a and Figure 4.13c. However, Figure 4.13b presents a notably different pattern, with a steep drop-off before 2016. This difference arises because, unlike assessments and case notes, where all relevant records were included for each identified individual, services which started before 2016 were only included if still being delivered as of January 2016. This means that an individual who started receiving services on 1st January 2015 and stopped receiving any services on 1st January 2016 would be included in the data. However, if they had received a service from 31st December 2010 to 31st December 2015, at which point they stopped receiving any services, they would not be included. Unfortunately, this means that while the exported records contain case notes from 2008, and assessment date from 2010, this cannot be linked to all services received before 2016, and in particular there is very little service data prior to 2015. As a result, in Chapter 7, which focused on loneliness at the time that services began, I was unable to determine whether individuals who had assessments prior to 2015 had received services at this time. This limited the identification of the initial receipt of care date from the 3,046 individuals in the cohort to a subset of 1,011 people.



(a) Case notes

(b) Services

(c) Needs assessments

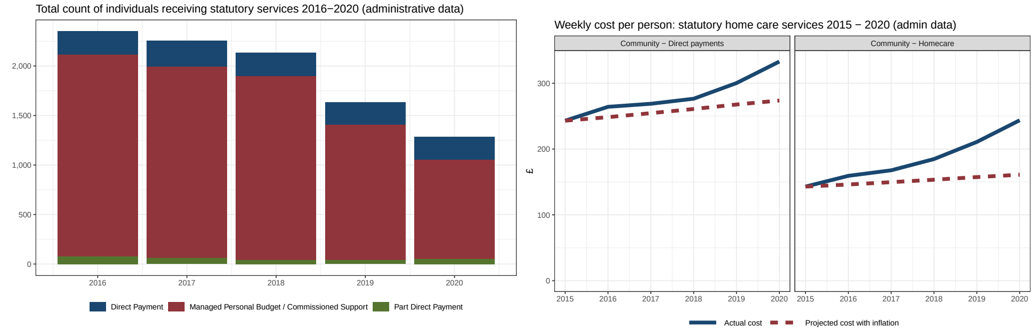
(d) Needs assessments (by form)

Figure 4.13: Date distribution of extracted records

The plots also show that the number of case notes, services, and assessments peaked in 2016-17. This was caused by changes in statutory requirements and internal processes within the local authority. The Care Act 2014, which came into force in March 2015, played a key role. Discussions with the local authority revealed that after the Act's implementation, there was a period of staff training focused on the new eligibility criteria, followed by the introduction of a new assessment form in November 2015 (labelled *Ax 5* in Figure 4.13d). This was part of a broader effort to ensure everyone was assessed under the new criteria [57].

The figure also highlights that the use of nine different assessment forms over time was largely due to temporal changes in policy about which forms to use.

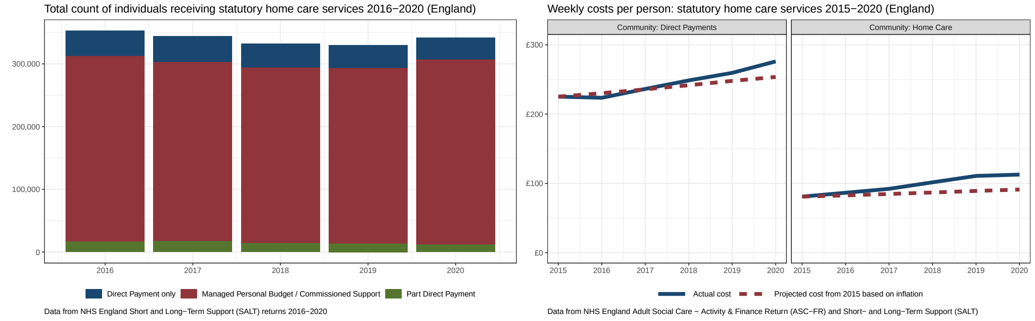
While the number of services provided decreases over time, the mean cost of these services increases, as shown in Figure 4.14b. In this figure, I compare the actual rise in weekly costs since 2015 with the hypothetical increase if costs had risen solely by the health rate of inflation [277]. The fact that actual costs are higher, despite fewer individuals receiving services, suggests a tightening of eligibility criteria during the observation period, i.e. fewer people are receiving services, but those who do have higher needs.



(a) Needs assessments: new PersonIDs per year

(b) Mean weekly cost by service type

Figure 4.14: Yearly changes in assessments and service costs (administrative data)



Data from NHS England Short and Long−Term Support (SALT) returns 2016−2020

Data from NHS England Adult Social Care − Activity & Finance Return (ASC−FR) and Short− and Long−Term Support (SALT)

(a) Total individuals receiving care (England)

(b) Total yearly gross current expenditure (England)

Figure 4.15: Trend in statutory home care costs and recipients (England)

The trend in the local authority aligns with the overall pattern in England, as shown in Figure 4.15. Specifically, Figure 4.15a illustrates that the number of adults aged over 65 receiving home care in England decreased from 2015-2019, which is consistent with the administrative data. Although there was an increase nationally in 2020, the administrative cohort definition excludes individuals who began receiving care after August 2019, so this increase would not be reflected in the administrative data. Figure 4.15b further demonstrates that while the number of individuals receiving home care decreased, average spending per person increased from 2015-2020, and that this was a real terms rise, higher than would be expected from inflation alone. This mirrors the trend observed in the administrative dataset. The absolute figures are slightly higher in the administrative data, but this is not surprising, as the cohort is limited to individuals receiving services for at least one year.[2] Additionally, London services would be expected to cost more than the average across England. I show in Figure 4.14a the count of individuals receiving care services in the administrative dataset. In 2016, 2017 and 2018, we see similar patterns in the administrative data of individuals receiving services, and proportions of direct payments, to those in SALT set out in Figure 4.15. The trends in the numbers of individuals included in the plots are very different in 2019 and 2020. However, this is to be expected, as individuals who started to receive care during or after the final quarter of 2019 could not be included in the cohort, as they could not fulfil the condition of using care services for at least a year by 31st August 2020.

4.2.4.6 Functional ability

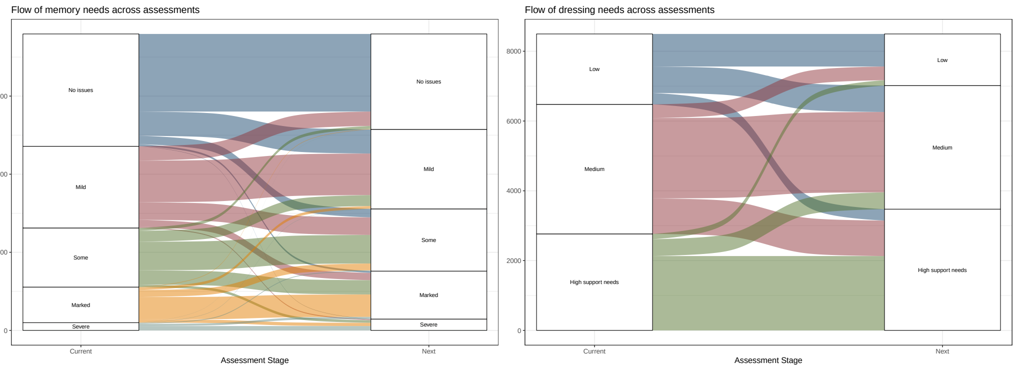
One challenge in establishing the reliability of functional ability data by comparing it to survey data is that surveys are not representative of the needs of older people receiving statutory social care services, a key argument in this thesis (see Chapter 5). However, it is possible to assess reliability by comparing the needs recorded in one assessment to those recorded in subsequent assessments.

Typically, needs of older people receiving long-term care would either remain stable or deteriorate over time [278, 279]. For instance, it would be highly unusual for an individual identified as having severe memory issues in one assessment to show no issues in the next. While such changes might occasionally occur — for example, if an assessment had been undertaken while the individual had been acutely confused while in hospital — these instances are exceptional. Similarly,

[2] The SALT definition of individuals receiving long-term support *encompasses any service or support which is provided with the intention of maintaining quality of life for an individual on an ongoing basis, which has been allocated on the basis of eligibility criteria/ policies (i.e. an assessment of need has taken place), and which is subject to regular review.* There is no minimum period of time for support to qualify as long-term.

physical needs, such as the ability to dress independently, are generally not expected to show significant improvement in older adults receiving long-term social care (rather than short-term services focused on improvement such as reablement), though minor improvements might occur in some cases. We would expect to see some cases of individuals moving from having very low needs to very high needs, which might occur after an event such as a fall or stroke, but if the local authority is fulfilling its duties to assess needs every time they change, we would hope that most deterioration would be incremental.

I do not perform a statistical test to assess reliability of assessment in functional needs, as the literature indicates that function is generally expected to decline, there is no specific hypothesis about the distribution of how often needs might improve, deteriorate, or remain stable in statutory care users. Instead, I present in Figure 4.16 the flow of needs between assessments for memory and dressing. Broadly, the results indicate that needs tend to remain the same or deteriorate over time. While there are occasional instances of improvement, these are usually modest, such as a shift from “Mild” memory issues to “No issues”. There is some sudden decline, but this is quite unusual. Cases where individuals with marked or severe memory issues subsequently exhibit no needs are extremely rare. Similarly, individuals with high support needs for dressing are rarely recorded as having low support needs in subsequent assessments. This pattern demonstrates consistency and plausibility in how needs are documented over time, supporting the reliability of the functional data recorded in needs assessments. However, not all needs-related fields appear to be as reliable, as I discuss in the next section.



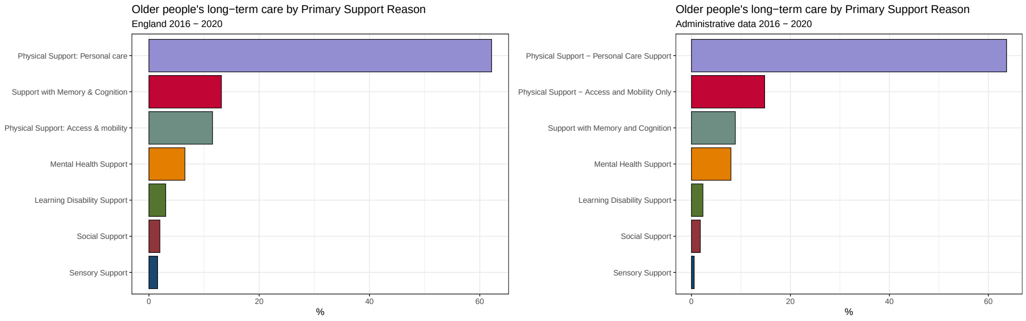
(a) Memory

(b) Administrative data

Figure 4.16: Flow of needs across assessments

4.2.4.7 Primary Support Reason

The needs-related data primarily is derived from the needs assessment forms. However, there is also a field called Primary Support Reason (PSR), which is stored in the *Demographics* file. This is because it is not updated at the time of needs assessment. Instead, it is created at the same time as an individual's case file, and can be updated at any time. The PSR data in the administrative dataset is largely consistent with the national data from SALT [107], as set out in Figure 4.17. In this sense it initially appears reliable.



(a) England

(b) Administrative data

Figure 4.17: Older people receiving care by Primary Support Reason

However, the PSR field is not always consistent with the needs assessment data. Of the 3,046 individuals in the administrative dataset, 2,495 have only one PSR recorded, 2,110 of which are for physical support. I present in Figure 4.18 the functional memory needs recorded at successive needs assessments for these 2,110 people. It is clear that there are significant changes in their memory needs over time. At the first assessment, almost 50% have no memory needs. However, by sixth or greater assessment, this figure is under 25%. We see corresponding increases in proportions of individuals with some memory needs, with more individuals having marked or severe memory issues by the sixth assessment than those with no memory issues. This is unsurprising, as we would expect the needs of some older adults who have had repeated care needs assessments to have deteriorated. However, despite this recorded deterioration, the PSR was not updated for any of these individuals whose memory worsened.

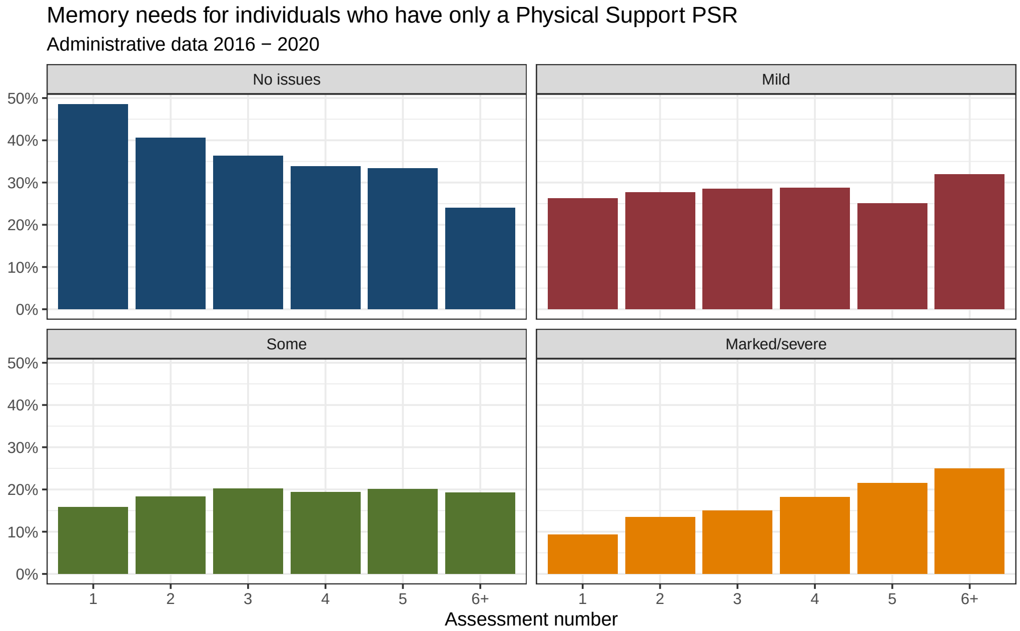


Figure 4.18: Memory needs by PSR

We see a similar phenomenon when examining the PSR of individuals who have marked/severe memory needs and reside in care homes. Dementia often contributes to care home entry, as individuals who are very physically disabled but not cognitively impaired can often live safely at home with regular care visits, while those with impaired cognition may be unable to be left safely between visits. Around 70% of older adults in care homes in England are estimated to have dementia [280, 281, 282]. However, SALT data indicates that over the 2016-2020 period, only 22% of older adults receiving long-term care services in residential or nursing care homes had a Support with Memory and Cognition PSR [107]. In the administrative dataset, the figure is 17.5%. It is possible in the administrative data to compare this breakdown by functional memory ability as recorded in a needs assessment. I set out this comparison in Table 4.4.

Table 4.4: Comparison of PSR for individuals in care homes (SALT and administrative data)

	England		Administrative data			
	All care homes		All care homes		Marked/severe memory & care home	
PSR	N	%	N	%	N	%
Physical Support: Personal care	596225	54.7	674	56.0	655	56.7
Support with Memory & Cognition	244380	22.4	211	17.5	207	17.9
Mental Health Support	99570	9.1	116	9.6	79	6.8
Physical Support: Access & mobility	87280	8.0	158	13.1	176	15.2
Learning Disability Support	33465	3.1	23	1.9	15	1.3
Social Support	15508	1.4	18	1.5	22	1.9
Sensory Support	13625	1.2	3	0.2	2	0.2

It is difficult to directly compare across three different constructs, i.e. Primary Support Reason, a diagnosis of dementia and having marked cognitive impairment. Additionally, it may be challenging for workers to determine a "primary" support reason for individuals with significant physical and cognitive impairments. Nevertheless, this analysis again indicates that the PSR field is not updated as frequently as functional needs, and that the memory PSR is much lower than would be expected given the prevalence of dementia in adults in care homes, while memory recorded in needs assessment appears more responsive to changes in need, as it declines over repeated observations. An explanation for this may be that the PSR is completed when an individual initially approaches a local authority, and is not updated, even as needs change, perhaps because it is not part of the needs assessment form. I do not use the PSR field in the analysis and instead use the functional assessments recorded in the needs assessment forms. This highlights the importance of understanding the process by which data is entered, and the caution that must be used when reporting on and interpreting fields in the data.

4.2.5 Harmonisation of functional ability across forms

The needs assessment forms contained functional needs assessment data regarding assistance required with using the toilet, washing, dressing, mobility, shopping, meal preparation and staying safe in the home. They also ask whether the person being assessed lives alone and receives unpaid care. The analysis included nine assessment forms, each with multiple versions of questions over time. ADL question responses were often consistent across forms, with a total of three sets of possible answers. One set of potential response codes was $\{0, 1, 2, 3, 4\}$, another $\{0, 2, 3, 4, 5\}$ and the final one $\{1, 2, 3, 4, 5, \mathrm{N}, \mathrm{U}, \mathrm{X}\}$. The assessment responses only contained response codes but not the question text or the meaning of each response. I requested the metadata associated with each question and response and received it in July 2022, and was able to join it to the relevant tables. The meaning of each response is set out in Table 4.5.

Table 4.5: ADL questions and responses

Form number	Question text	Response code	Response text	Response combined
1	Using the toilet/managing continence	0	Fully independent (no need) / Not applicable	Low/no needs
1	Using the toilet/managing continence	1	Largely independent (some difficulty/pain)	Low/no needs
1	Using the toilet/managing continence	2	Partial independence (sometimes help/prompting)	Moderate
1	Using the toilet/managing continence	3	Limited independence (always help/prompting)	High
1	Using the toilet/managing continence	4	High support needs (cannot undertake at all)	High
1	Using the toilet/managing continence	5	Very high support needs (needs support of two)	High
2	Your situation	0	Little or no difficulty/risk (can manage alone)	Low/no needs
2	Your situation	2	Significant difficulty/risk (sometimes needs help)	Moderate
2	Your situation	3	Significant difficulty/risk (always needs help)	High
2	Your situation	4	Unable to manage - needs one other to undertake	High
2	Your situation	5	Unable to manage - needs two others to undertake	High
3	Using the toilet/managing continence	0	Fully independent	Low/no needs
3	Using the toilet/managing continence	1	Largely independent	Low/no needs
3	Using the toilet/managing continence	2	Partial independence	Moderate
3	Using the toilet/managing continence	3	Limited independence	High
3	Using the toilet/managing continence	4	High support needs	High
3	Using the toilet/managing continence	5	Very high support needs	High
3	Using the toilet/managing continence	N	Not applicable	Low/no needs
3	Using the toilet/managing continence	U	Not assessed	Missing
3	Using the toilet/managing continence	X	Not known	Missing

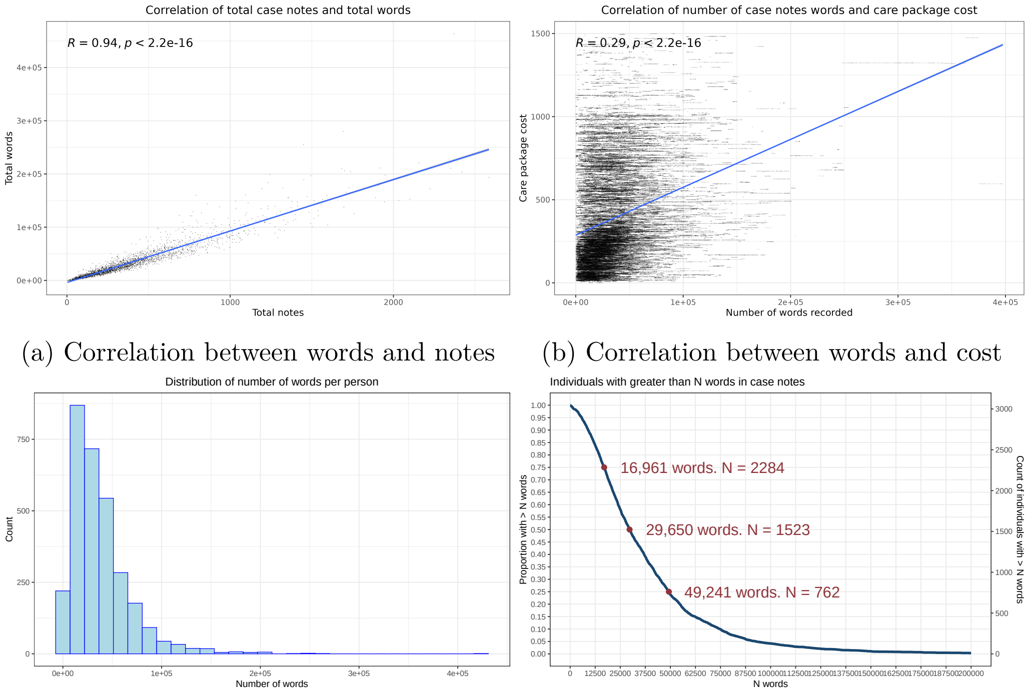
For Chapter 7, where these responses were included as covariates in a regression, the responses needed to be comparable. Additionally, the five categories were not all meaningfully different. I grouped the ADL answers into three categories: *Low/No needs*, *Moderate* or *High*. I manually mapped each response to the desired category based on the response text rather than the numeric value, as set out in Table 4.5, and repeated the process across all ADLs. This highlights the issue raised in Witham et al. [119] about the significant investment of time and domain knowledge required to aggregate even structured administrative data into the format required for research.

4.3 Reliability of free text data

4.3.1 Completeness of free text data

I include in Figure 4.19 a summary of information relating to case notes. Figure 4.19a shows the strong correlation ($r = 0.94$) between the total number of case notes recorded and the total number of words recorded for each PersonID. This is a reassuring sign that there were no systemic issues with the extraction (and also provides the rationale for only including one of these as a covariate in Chapter 7). Figure 4.19b shows the correlation between the number of case notes and the cost of the care package. There is a significant correlation of 0.29. There is no benchmark figure in the literature to compare this against. However, a positive correlation is reassuring, indicating that higher numbers of case notes are somewhat associated with higher care package costs. We would expect the number of case notes to be associated with total cost, as the act of gathering information for an assessment and care plan will usually generate case notes recording home visits, contacts with family or professionals. It is also reassuring that there is variance in the total cost that is unexplained by number of case notes, as there are many social work contacts that will generate notes but not costs, such as annual reviews with no changes to care, cases with complex family dynamics, or safeguarding investigations. A very high correlation between number of notes and care package costs would indicate that the quantity of services could be modelled by counting case notes alone, which might undermine the case for using complex language models to extract information from the content of free text data.

Figure 4.19c, illustrates the distribution of the number of words per person. There is a long tail, showing a small number of individuals with extremely high word counts. The distribution has a mean of 37,568 and median of 29,650 words per person. This suggests that while a few cases involve extensive documentation, possibly reflecting more complex care needs or more detailed record-keeping practices for certain individuals, average cases still contain a considerable amount of free text. I set this out in Figure 4.19d, which shows the proportion of people whose case notes contain more than the number of words on the x-axis. This shows that, for example, around 25% of people have around 50,000 or more words of case notes. This indicates that even outside the most extreme cases, a substantial amount of information is being recorded, which has implications for the time and resources required for case management and review.



(a) Correlation between words and notes

(b) Correlation between words and cost

(c) Words per person

(d) Individuals with more than N words

Figure 4.19: Free text figures

Establishing the reliability of free text fields is different to the structured fields. While it is possible to create quantitative metrics, such as word counts or the number of case notes per person, these metrics can highlight problems with data quality, and provide insight into the volume of the data, but they do not provide even a superficial description of its content. Unlike fields quantifying functional ability, where values can be compared and analysed, free text content is inherently unstructured, making simple comparisons difficult. The true value of free text lies in the information it contains, but this information must first be extracted and processed before it can be used in a meaningful way. This challenge of extracting information from this data forms the basis of much of this thesis.

4.3.2 Quantity of free text

There are 783,252,474 bytes of free text needs assessment and case note data for the individuals in this study, and 261,429,895 bytes of structured needs assessment data. I include in Figure 4.20 a rough calculation of the number of words recorded per worker per day in the data used in this thesis. This is calculated as,

$$Words_{worker} = \frac{Words_{day}}{N_{SW} + N_{OT}}$$

Where $Words_{day}$ is the total number of words recorded per day, and N_{SW} and N_{SW} are, respectively, the total number of local authority social workers and occupational therapists working with older people. This figure should be taken as an approximate estimate as, while the number of staff by job role is published at local authority level [283], the proportion who work with older people is published at national level [284]. Furthermore, it only includes words recorded in administrative records for this subset of older adults, and excludes those who receive services for less than a year. It also does not include all forms of administrative recording such as emails, instant messaging and handwritten notes. The figure includes free text in needs assessment forms but not structured data selected from drop down menus, or text entered into other forms such as for safeguarding investigations. It excludes non-qualified workers who may be entering notes whose numbers are not easy to measure. The total declines towards the end of the period partly due to the way the cohort was constructed, and partly as it appears that fewer people were receiving services over time. Nevertheless, these rough estimates indicate that the average worker in this local authority generally added around 1000 - 1500 words of notes to their recording systems every day for the individuals in this cohort in the subset of forms captured in the data.

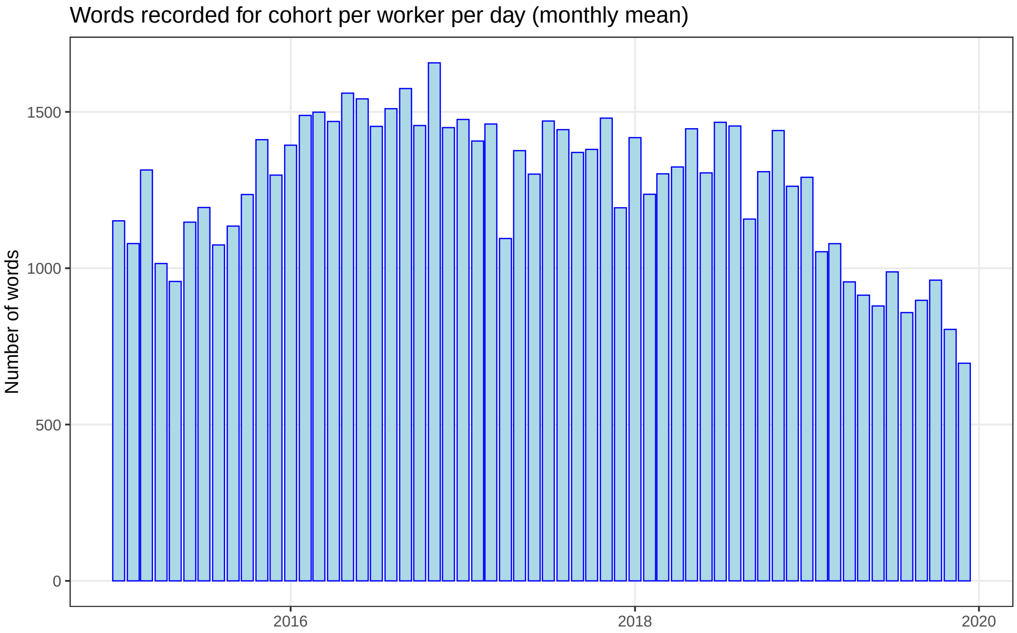


Figure 4.20: Words recorded per worker per day

4.3.3 Content of free text data

While the content of structured data is determined by its schema, free text records are not so well-defined. There is little published literature on the content of free text social care records, so I present in Figure 4.21 a breakdown of the topics covered in the case note titles that appear most frequently in the administrative data used in this thesis. These titles comprise 63,911,860 words (72% of case note words) over 804,000 entries (77% of all notes). I have grouped semantically similar titles such as *Telephone call* and *Phone call*. The titles do not always illuminate the nature of the content. The top title, *Case update* could cover a range of topics. The same applies to other frequent titles such as *Telephone call*, *Contact* and *Record of email*. However, the titles in Figure 4.21 do provide some insight. I present in Table 4.6 a breakdown of the topics by category. 37.6% of notes are related to case management, which includes case updates, case summaries, screenings, allocations and reviews. 20.6% are contacts, such as phone calls and emails and home visits, although such notes tend to be longer, and comprise 25.3% of words. A similar proportion contain records of professional assessments by Occupational Therapists (OTs), physiotherapists and sensory workers. Case notes also contain many notes about variations in service provision such as service suspensions or restarts. Interestingly, there is a relatively small proportion of urgent actions, with only 5.8% of notes or 5.6% of words including safeguarding (allegations of abuse and neglect) and “no replies”, where an individual does not answer the door to care workers. This shows one of the limitations of these descriptive statistics. It would be imprudent to assume that this means these are a relatively small part of the role, as social workers inevitably focus on high risk cases even where the probability of the risk occurring is low [285], and managing risk of abuse and neglect can be very time-consuming and stressful [286]. As outlined in Chapter 3, more complex methods than word counts are required to extract meaningful information from unstructured data. These methods are used in the subsequent chapters.

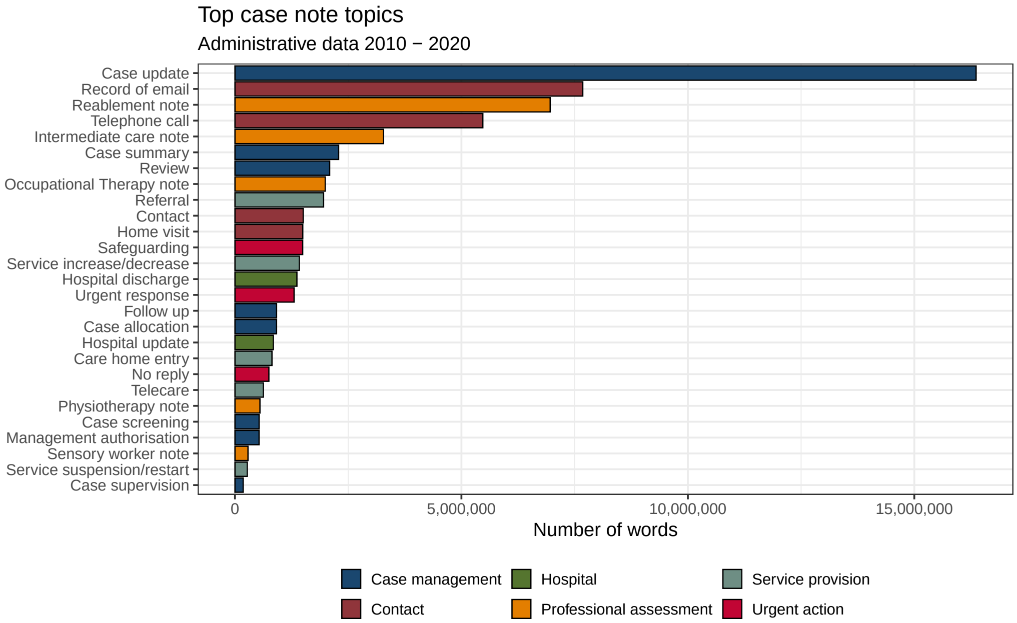


Figure 4.21: Top 100 case notes topics

Table 4.6: Categories of case note titles

Category	N Notes	%	N Words	%
Case management	302,354	37.6	23,825,547	37.3
Contact	165,623	20.6	16,156,482	25.3
Professional assessment	161,146	20.0	13,072,461	20.5
Service provision	88,186	11.0	5,093,557	8.0
Urgent action	46,353	5.8	3,548,869	5.6
Hospital	40,788	5.1	2,214,944	3.5

Free text is found in needs assessment forms, as well as case notes. I present in Figure 4.22 the breakdown of the number of words contained in the 100 most frequent free text assessment questions in the administrative data used in this thesis. The text in response to these questions comprises 28,059,980 words, which is 86.2% of the total free text in assessment responses (as opposed to case notes). Many of the top 100 questions are versions of the same question on different forms, and I was able to aggregate these 100 questions into 28 categories, which I present in the figure. The *Summary of needs* tend to be the first question in the needs assessment, and are usually written in the form of a pen portrait describing the individual's needs, circumstances and changes since the last assessment.[3] There is a considerable amount of text about a wide range of topics, some of which are covered in structured data, such as personal care and nutrition, and some which are not, such as social networks, managing finances and the views of the person receiving care, their family and other professionals.

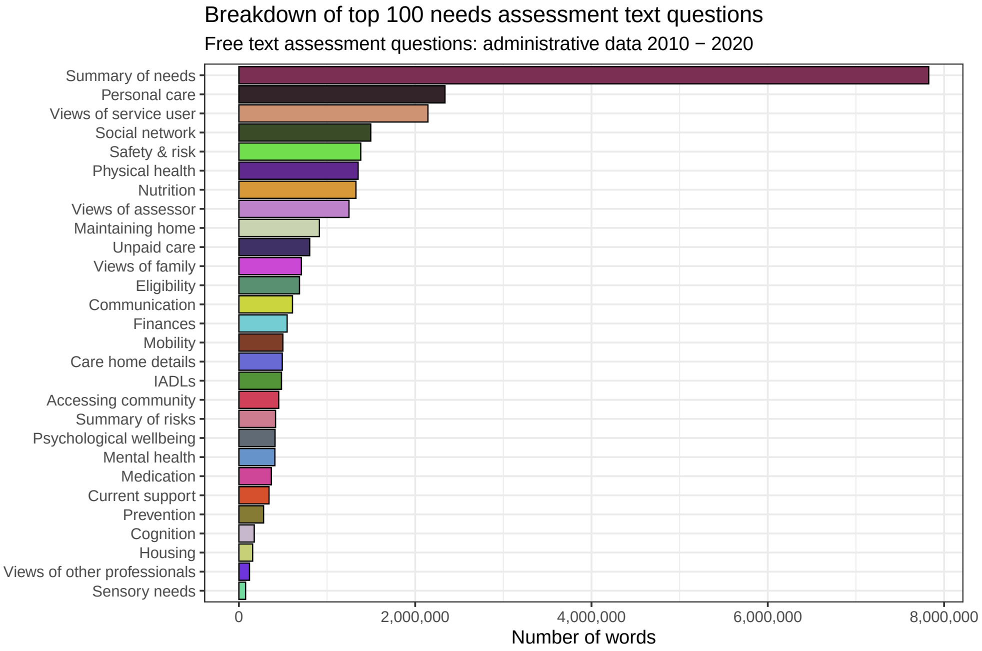


Figure 4.22: Breakdown of top 100 needs assessment questions

[3] It is these responses that I use in Chapter 8, to assess bias in LLM summarisation models.

4.4 Establishing the foundation for comparison with survey data

This chapter has detailed the processes involved in acquiring, pseudonymising, and preparing the administrative data, demonstrating its reliability and robustness for providing insights into the needs and experiences of social care users. However, it also describes the significant practical obstacles and technical complexity of transforming raw administrative records into a format suitable for research. In the next chapter, I will compare the administrative data presented here with data about statutory care users contained within surveys. This comparison will underscore that challenges inherent in survey methodologies — such as attrition, under-reporting, and exclusion of certain populations — lead to incomplete capturing of the needs of statutory social care users. By bridging the gap from the reliability of this administrative dataset to the limitations of surveys, I make the case that despite the practical obstacles inherent in using routinely-collected data, administrative records play a critical role in understanding social care and informing effective policy and practice.

5 Limitations of survey data for evaluating social care in England

5.1 Introduction

Understanding the needs of older people who use statutory social care services is crucial for developing effective policies and interventions. This chapter investigates the use of surveys for accurately capturing the needs of statutory social care users in England. Using comparisons between survey and administrative data, I examine specific issues such as attrition, under-reporting, and methodological constraints in survey design. Population surveys are a common tool for gathering data on social care needs, yet there is evidence they may not adequately represent individuals with the greatest care needs. The overwhelming majority of participants in English population surveys do not have significant social care needs (defined as inability to complete three or more activities of daily living [ADLs]) [103]. It is not clear, however, whether those individuals included in surveys who receive statutory care services are a small but representative sample, or have systematically lower needs than individuals in administrative data. This is the question I explore in this chapter.

I focus on key datasets that provide individual-level information about care recipients, including functional ability to complete ADLs, receipt of unpaid care, and use of statutory care services. These datasets include the Family Resources Survey (FRS) [287], Health Survey for England (HSE) [288], English Longitudinal Study of Ageing (ELSA) [109], Understanding Society [110], and the Adult Social Care Survey (ASCS) [112]. While there are many other surveys, such as the Whitehall study of UK civil servants [289] or UK Biobank [290], I exclude them here as they are less relevant to the focus of this thesis.[1] I also exclude the census and the ONS Longitudinal Study [293] due to their infrequent data collection and lack of information about functional ability. Surveys have strengths over administrative data, such as fewer restrictions on accessing data, and include individuals who do not need social care or purchase it privately, but I do not focus on these aspects in this chapter as they are less relevant to the studies discussed in this thesis.

[1] Additionally, it is established that large-scale voluntary surveys such as UK Biobank have limitations around selection bias towards healthy participants [291, 292].

5.1.1 Challenges in capturing the needs of statutory care users

Several methodological issues in survey design and data collection contribute to the under-representation of individuals with the highest care needs. One primary reason is the intentional exclusion of certain populations. For instance, individuals who enter care homes are excluded from FRS [104] and Understanding Society [105]. Additionally, those with higher needs are frequently omitted from research because they lack the capacity to consent to participate [33], as seen in HSE [31].

Participants whose needs increase may be lost to follow-up, as observed in ELSA [32]. This attrition exacerbates the under-representation of individuals with high care needs. Attrition is systematically related to health outcomes and socioeconomic status. There is attrition bias in ELSA associated with impaired cognition, and this may lead to underestimating the needs of populations most relevant to policymakers [294, 295]. Further examinations of longitudinal surveys in Japan and the US have found that attrition is associated with declining functional ability and social isolation [296, 297].

Under-reporting in surveys is another significant concern. Individuals may under-report their needs due to factors such as social desirability bias or misunderstanding of survey questions [298]. A study comparing self-reported medical diagnoses in ELSA linked with Hospital Episode Statistics (HES) data found that more than half of respondents did not report serious conditions they had been diagnosed with in a hospital within the previous two years [106]. Under-reporting was particularly associated with impaired cognition and was more common among men. Studies have also found low correlation between self-reported and observed measures of functional ability in the US and the Netherlands [138, 299].

5.1.2 Limitations of survey design

There are inherent limitations in cross-sectional studies (such as ASCS, HSE and FRS) for understanding social care users as it is not possible to measure the progression of needs over time. Longitudinal studies, such as ELSA and Understanding Society, allow for individual needs and circumstances to be captured at repeated intervals. Furthermore, longitudinal studies can distinguish between ageing effects and cohort effects, and control for time-invariant unobserved characteristics [300, 301]. However, ELSA and Understanding Society have limitations in the context of evaluating social care. Participants with high needs may not enroll or be lost to follow-up owing directly to declining health and impaired cognition, or factors correlated with this, such as socioeconomic status [34, 302, 295, 294]. Attrition is especially concerning when studying factors associated with cognitive function in older populations, as cognitive ability is linked to continued participation, potentially skewing results in ways difficult to adjust for through reweighting [295]. Additionally, capturing significant changes in care provision, such as care home entry, is problematic even in longitudinal surveys. When care home entry is recorded, it often occurs between waves. While interval-censoring survival models can manage this, approximations can increase standard errors and even lead to biased results [303].

While these methodological challenges pose significant problems for researchers investigating issues related to high-need populations, attrition is a persistent issue in panel studies and there are various approaches available to manage it, such as multiple imputation, or Bayesian methods that estimate missing values as additional parameters [304]. Survey weights are also commonly used to address attrition, although their effectiveness is sensitive to how the weights are calculated [305]. Despite these methods, limitations remain, particularly when unobserved variables contribute to dropout and cannot be directly accounted for [306, 307]. The extent to which these methods are appropriate for addressing attrition found in statutory care users will depend on the association between social care needs and survey dropout. However, the relationship between the association between survey attrition and functional social care needs in general, and factors such as cognition in statutory social care users in particular, have not been analysed in English surveys.

5.1.3 Research question: how well can surveys capture statutory social care use?

In light of these methodological challenges, I explore: To what extent can survey data accurately capture statutory social care use and the needs of high-need populations? It would be easier to use survey data for the analyses in Chapter 6 and Chapter 7, where I use data extracted with a LLM from free text case notes in a regression model to assess whether loneliness is associated with time to care home entry. Indeed, a similar analysis in 2018 using ELSA data established that older people who are lonely are at greater risk of care home entry [267]. However, the concerns about the extent to which those with higher needs are captured in surveys raises questions about the generalisability of such findings to statutory care users. In this section, I compare survey data with administrative records to assess whether English survey data is suitable for research into older people's use of publicly funded care.

5.2 Methods

5.2.1 Comparing surveys and administrative data

I initially compare the sample sizes of older people who report receiving statutory care across data sources, contrasting the administrative dataset used in this thesis with data from ELSA, Understanding Society, HSE, FRS, and ASCS. I include the most recent waves of each survey that contain social care data. In ELSA, the social care module is introduced in wave 6 (2013/14), so I include waves 6 to 9. For Understanding Society, social care questions are asked every other wave from 2018, so I use waves 9, 11, and 13. I use the most recent available waves of the HSE, which are from 2014 to 2019, and yearly ASCS data from 2014/15 to 2022/2023.

I then present a comparison of functional ability between individuals reportedly receiving statutory care in HSE, ELSA, Understanding Society, ASCS, and those in the administrative data used in this thesis. I exclude FRS because it does not contain questions about the ability to complete activities of daily living (ADLs). To ensure the comparison is between similar groups, I use only the subset of individuals in the administrative dataset who are living in the community (rather than in care homes) and I use needs from the time of initial assessment. I only use fields which I determined were reliable as set out in Chapter 4. I only include those individuals in surveys who report that they are in receipt of local authority care. As the absolute numbers of statutory care users in population surveys are relatively low, I pool the data across survey waves.

To assess the extent to which default survey weights mitigate the issue of under-representation, I weight survey responses using the weights provided with each survey. For ELSA and Understanding Society, I use the longitudinal weights, and for HSE, I use the cross-sectional weights. I create confidence intervals using the R `survey` package. For the administrative data, I create confidence intervals clustering by Person ID with equal weights for each individual. I do not include ASCS in this part of the analysis as it is not published with weights. I discuss below whether it would be possible to create social care survey weights from the available information.

5.2.2 Local authority level comparison between administrative records and ASCS

The ASCS dataset, unlike other surveys, includes the respondent's local authority. I analyse whether the differences in reported needs between the administrative data and survey data are due to unique characteristics of the local authority whose data is used in this thesis by looking at the distribution of support needs across local authorities using ASCS data. In this analysis, I focus on key functional and demographic indicators included in ASCS, such as support with dressing, receipt of unpaid care, gender, and ethnicity. I create estimates of these characteristics by local authority and compare them with results from ELSA, Understanding Society, HSE, and the administrative records used in this thesis. I plot these distributions to visually assess how much needs vary across local authorities and compare them with the proportions from national surveys and the administrative dataset. This helps identify whether the differences between the administrative data and survey data come from unique local factors or reflect wider discrepancies across data sources. While I cannot identify by name the specific local authority whose data is used in this thesis, I plot the distribution of these needs across local authorities using the ASCS data (removing two councils with very small populations, the City of London and Isles of Scilly), and highlight the position of the sample local authority in this distribution.

As ASCS contains around 200 older people's responses per year for the local authority in question, I can also directly compare it with the administrative data used in this thesis. I do this by using ASCS to estimate the mean level of needs among older individuals in the same local authority captured in the administrative records. I compare key indicators of functional support that appear in both the ASCS and administrative data, including needs for assistance with dressing, toileting, and demographic characteristics. I use the R `survey` package to apply the survey design, giving equal weight to all responses as there are no weights provided with the ASCS data. For each indicator, I first filter the ASCS data to exclude missing or non-disclosed values and then apply the survey design to obtain the mean proportion of individuals who report care needs by local authority, also generating confidence intervals to provide an estimate of the precision of these means. I then compare these weighted survey estimates with the equivalent mean levels recorded in the administrative data for the same local authority. I consider the administrative records to be the population data in this case, and thus do not present them with confidence intervals.

5.2.3 Measuring under-reporting in surveys

While it is possible to measure under-reporting by linking survey data to administrative datasets, I cannot use the data in this thesis for this purpose, as it is from one local authority and is unlikely to contain many ELSA respondents. However, under-reporting in surveys can also be identified by assessing the internal consistency of several responses. For example, self-reported data on unpaid care within ELSA can be examined through mismatches between respondents' and spouses' answers to whether care is provided for a partner [308]. I use a similar approach to examine under-reporting of functional cognitive ability in ELSA.

As most functional questions in surveys are self-reported, it is difficult to decompose any differences in the functional ability of those in surveys with individuals in administrative records, to establish how much is under-reporting and how much sample composition. However, ELSA waves 1 to 9 include an immediate word recall question (`imrc`). Respondents are read a list of ten words and then asked to recall them immediately to the interviewer, receiving a score from 0 to 10, indicating the number of words they can remember [109]. Immediate word recall is an effective test for cognitive impairment [309]. ELSA waves 1-4 and 7-9 also include the question, *How would you rate your memory at the present time?*, with possible responses: *Excellent, Very good, Good, Fair* and *Poor* (`slfmem`).

To establish the association of functional ability with self-reporting, I look at the association between performance on an immediate word recall task and self-reported memory. I use a χ^2 test to assess whether there is an association between immediate recall and self-assessed memory. Since the null hypothesis of no association seemed unlikely given the intuitive relationship between these variables, I also use Cramér's V to quantify the strength of any association [310]. I also use the R `survey` package [311] to calculate confidence intervals for self-reported memory among ELSA respondents who can recall zero words, reweighting using the longitudinal weights provided with ELSA clustered by unique person ID, to indicate the precision of these estimates.

I use the same approach to explore the differences between functional ability and self-reporting in statutory care users in particular using data from ELSA waves

6 to 9, by comparing the distribution of immediate word recall scores among individuals reporting that they receive local authority care with those of other individuals, and by comparing this with the distribution for self-reported memory. This comparison cannot be made among statutory care users in Understanding Society, as the immediate recall question is only included in wave 3, and receipt of local authority care from wave 7 onwards. Similarly, the Health Survey for England and the Family Resources Survey contain a memory question but no functional test of memory. ASCS does not include any questions about memory or cognition.

5.2.4 Measuring the relationship between needs and survey attrition

To investigate the impact of cognitive function on attrition in longitudinal surveys, I conduct an analysis using data from ELSA and Understanding Society. The aim is to assess whether participants with lower cognitive function are more likely to be lost to follow-up in subsequent survey waves, which could contribute to under-representation of individuals with higher needs in survey data. Participants are considered lost to follow-up if they do not participate in the subsequent survey wave and have not died between waves. Participants who continue to participate in the next survey wave are included as the comparison group. I use the same immediate word recall and self-reported memory variables as above in ELSA. To measure self-reported memory issues in Understanding Society, I used a binary variable indicating whether participants experienced memory problems, `cgsrmem2_dv`. The immediate word recall question in Understanding Society is `cgwri_dv`.

By comparing the proportions of participants lost to follow-up in each cognitive function group, I aim to identify whether higher attrition rates are associated with cognitive function or self-reported cognitive function. I conduct χ^2 tests to assess whether the differences in attrition rates between groups are statistically significant. A significant difference in the rate of loss to follow-up between groups would suggest that cognitive abilities have an impact on participation in ELSA and Understanding Society over time. I examine this association by comparing the proportions of participants lost to follow-up depending on both observed cognitive function (immediate word recall) and self-reported memory issues. Standardised residuals are calculated to understand the magnitude of the differences between observed and expected frequencies. A significance level of $\alpha = 0.05$ is used to determine statistical significance.

5.3 Results

5.3.1 Sample sizes of statutory social care users

The vast majority of individuals in the included English surveys, excluding ASCS, report they are not receiving statutory social care, as set out in Table 5.1. Around 3.6% of individuals aged 65+ in England are in receipt of statutory social care in 2024.[2] However, the proportion of individuals who report receiving social care in surveys (presented in parentheses in Table 5.1) is lower than this. The exception to this is ASCS, where of respondents are receiving statutory social care services. While ASCS is an extremely useful source of data, it has other limitations, which I discuss in the subsequent sections.

Table 5.1: Number of older people receiving statutory care per survey wave

Year	ELSA	Und. Soc.	HSE	FRS	ASCS
	N (%)	N (%)	N (%)	N (%)	N (%)
2013	78 (0.9%)				
2014			54 (0.5%)		38,963 (100%)
2015	111 (1.1%)		36 (0.3%)		39,869 (100%)
2016			31 (0.3%)		39,061 (100%)
2017	135 (1.3%)		23 (0.2%)		33,269 (100%)
2018		90 (0.2%)	21 (0.2%)	87 (2%)	35,764 (100%)
2019	99 (1.2%)		26 (0.3%)	110 (2.6%)	32,401 (100%)
2020		46 (0.1%)		40 (2.1%)	3,188 (100%)
2021				61 (1.7%)	29,954 (100%)
2022		31 (0.1%)		94 (1.6%)	29,946 (100%)
2023					30,584 (100%)

ELSA: English Longitudinal Study of Ageing, Und. Soc.: Understanding Society, HSE: Health Survey for England, FRS: Family Resources Survey. ASCS: Adult Social Care Survey. Understanding Society asks social care questions every other wave.

[2] This is based on 394,000 people aged 65+ receiving statutory care and a population of 10.8 million aged 65+ [19, 312].

5.3.2 Comparison of functional ability (unweighted)

I present in Figure 5.1 a comparison between administrative and unweighted survey data for key demographic and functional variables among social care recipients, highlighting the substantial differences in reported need levels. Across demographic factors, such as gender and living alone, the rates in administrative and survey data are relatively similar.[3] The difference in ethnicity is expected as we are comparing a London local authority with a national picture. However, across functional ability, individuals in administrative data have far higher needs than the comparable cohort included in surveys.

Table 7.1 contains the data in the plot in tabular form, including the total individuals as well as percentages.[4] It shows that, for example, 52% of individuals in administrative data require support with using the toilet. However, the self-reported rates among older statutory care users in survey data range from 19% - 38%. There are similar patterns with memory impairment and requiring support to get dressed. The table includes the sample size (N) as well as the percent, which indicates another advantage of administrative data. Although we are comparing a snapshot from one local authority with survey data over 5 years, the number in each category in the administrative dataset is generally larger than survey data by at least a factor of four, and sometimes as large as 20. The exception to this is ASCS, which has a much larger sample size than the administrative data.

[3] I use unweighted survey responses as the purpose of this plot is to demonstrate the lower numbers of individuals with self-reported needs in surveys, using raw counts. However, I examine the impact of weighting below in Figure 5.2. It does not make a meaningful difference to the proportion of individuals with self-reported care needs.

[4] Table 7.1 and Figure 5.2 do not include data from ASCS, HSE or Understanding Society on awareness of risk or support required with meal preparation, which are not recorded in these surveys.

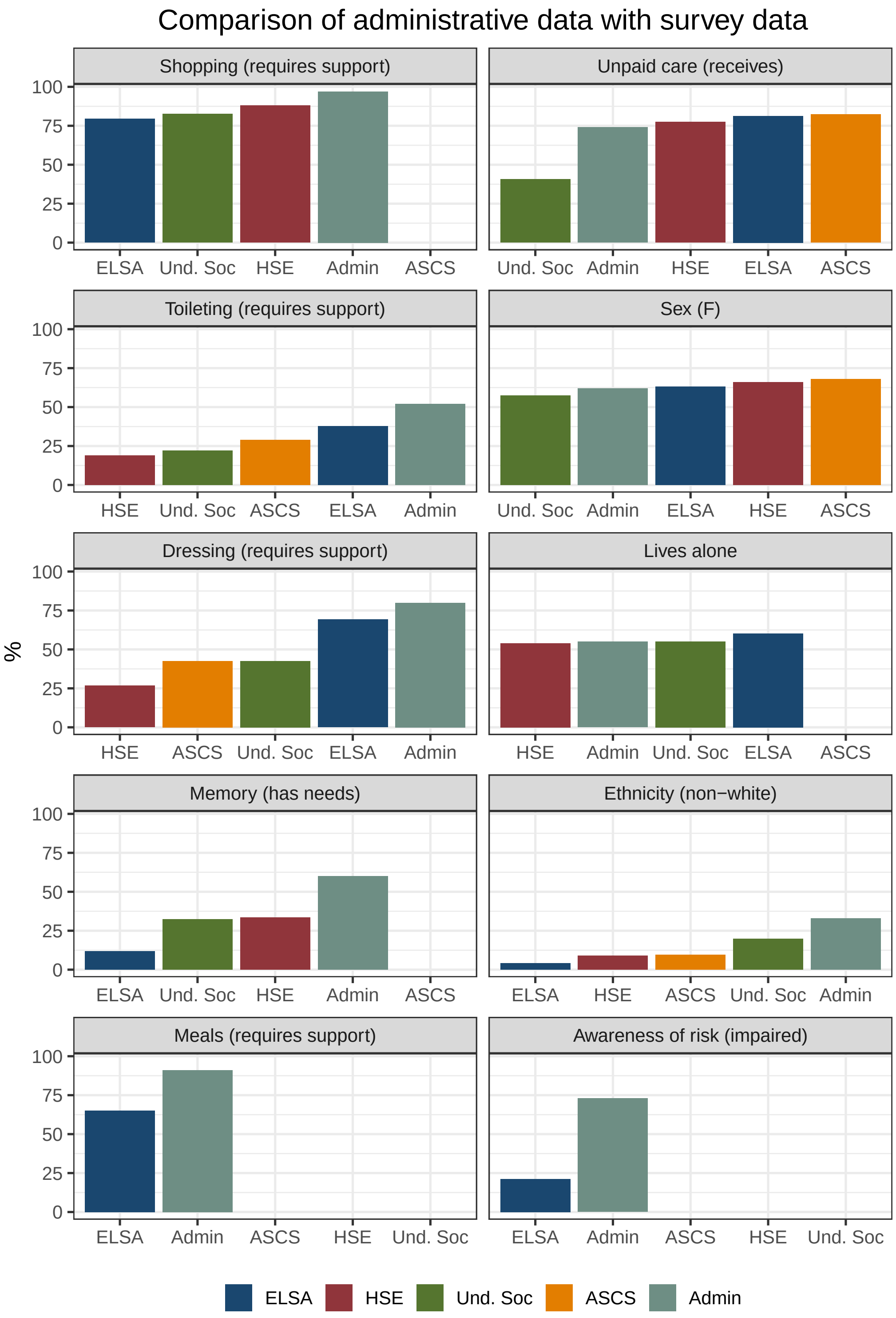


Figure 5.1: Comparison of administrative data with survey data (unweighted)

Table 5.2: Comparison of needs and demographics between survey and administrative data

	Admin		ELSA		HSE		Und. Soc.		ASCS	
	N (%)	N Uniq	N (%)	N Uniq	N (%)	N Uniq	N (%)	N Uniq	N (%)	N Uniq
Ethnicity (non-white)	364 (33%)	364	18 (4%)	15	17 (9%)		33 (20%)	32	28,408 (9.5)	
Toileting (requires support)	570 (52%)	570	160 (38%)	132	36 (19%)		37 (22%)	36	87,673 (29)	
Lives alone	608 (55%)	608	255 (60%)	203	103 (54%)		92 (55%)	84		
Memory (has needs)	664 (60%)	664	50 (12%)	49	64 (34%)		54 (32%)	52		
Sex (F)	686 (62%)	686	267 (63%)	208	126 (66%)		96 (58%)	90	212,756 (68)	
Awareness of risk (impaired)	806 (73%)	806	90 (21%)	81						
Unpaid care (receives)	819 (74%)	819	344 (81%)	277	148 (77%)		68 (41%)	65	245,332 (82.4)	
Dressing (requires support)	877 (80%)	877	293 (69%)	233	51 (27%)		71 (42%)	67	128,401 (42.5)	
Meals (requires support)	998 (91%)	998	276 (65%)	219						
Shopping (requires support)	1066 (97%)	1066	336 (79%)	272	168 (88%)		138 (83%)	129		

N Uniq is number of unique individuals (as data is pooled). This is not known for HSE and ASCS which are cross-sectional.

5.3.3 Comparison of functional ability (weighted)

We can assess the extent to which default survey weights mitigate this discrepancy in needs by re-weighting the raw responses used to create the plots in Figure 5.1. This is set out in Figure 5.2. There are some small differences between the weighted comparison in Figure 5.2 and the unweighted one in Figure 5.1, but weighting does not ultimately change the overall story, with individuals in surveys appearing demographically similar to those in administrative data across gender, living alone and unpaid care, but reporting substantially higher levels of functional ability. I do not present ASCS in Figure 5.2 as it does not include weights. I discuss below whether weights could be calculated for ASCS and other surveys.

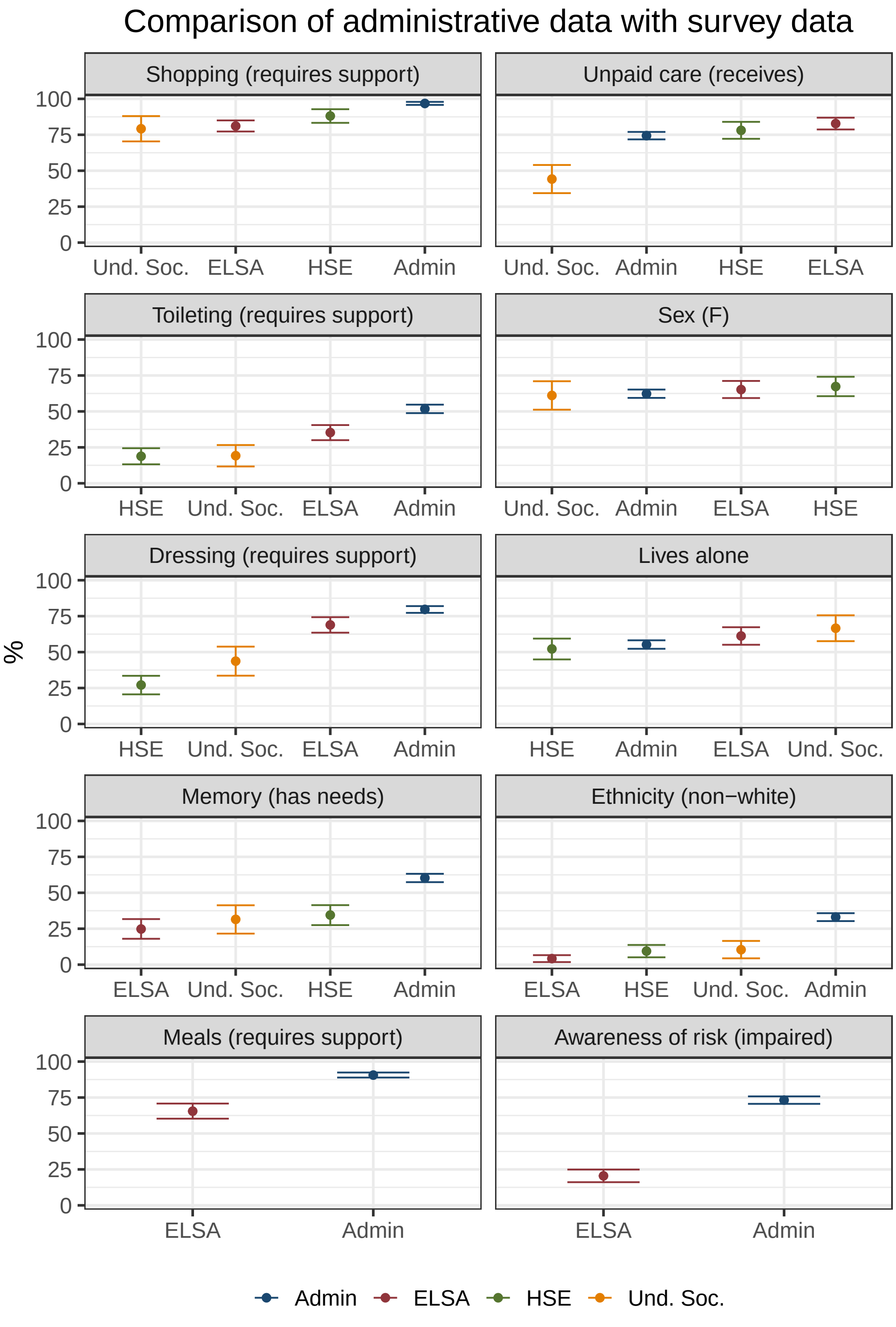
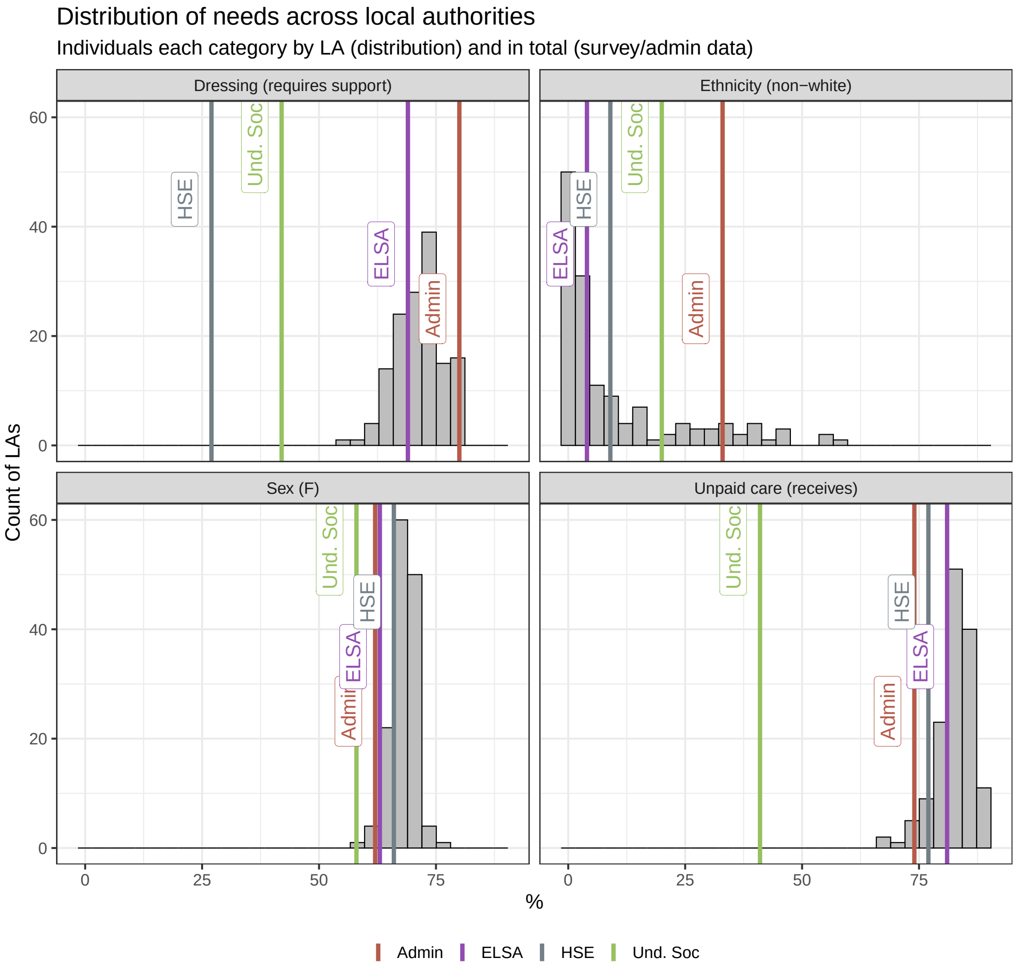


Figure 5.2: Comparison of administrative data with survey data (weighted)

5.3.4 The impact of local variation on differences between administrative and survey data

Statutory care services have a national eligibility criteria set out in the Care Act 2014. Nevertheless, demographic factors and differences in implementation of the eligibility rules can yield differences between local authorities. However, the differences in needs between the administrative data and survey data cannot be attributed to the fact that it comes from a particularly unusual local authority. In Figure 5.3, I present the distribution of ASCS responses across local authorities for support needed with dressing, receipt of unpaid care, sex, and ethnicity. I have overlaid these distributions with the proportions from ELSA, Understanding Society, HSE, and the administrative data. The variation in the demographic characteristics (sex and ethnicity) across local authorities is wide enough to explain the discrepancies observed, but this is not always the case with the needs-related characteristics.



Distribution from pooled Adult Social Care Survey 2014−2022.
Vertical lines show the proportions in each category by data source.

Figure 5.3: Distribution of needs across local authorities

There is a notably wide distribution in ethnicity in Figure 5.3, with some local authorities being almost exclusively white and others having a majority non-white population. This is the only category exhibiting such a pronounced pattern. The percentage of women in the population is particularly consistent across local authorities, with results from other surveys and administrative data generally clustering around the average. The distribution of those receiving unpaid care is similarly consistent. However, the percentage of individuals requiring support with dressing varies between 52% and 78%, with a median of 67%. The interquartile range for support with dressing across local authorities is 7.7%, indicating relatively limited variation compared to the rates reported in HSE or Understanding Society, where the percentage requiring support with dressing is 25-40% lower than the ASCS figure. In contrast, the administrative data shows a rate of 80% needing support, higher than any individual local authority in ASCS. This disparity might be due to the recruitment and strategy used by ASCS, which I discuss below.

It is not the case that the sample local authority has unusually high needs. In fact, it falls within the lowest 15% for England in ASCS in terms of the proportion of individuals needing support with dressing, using the toilet and in receipt of unpaid care. I set this out in Figure 5.4.

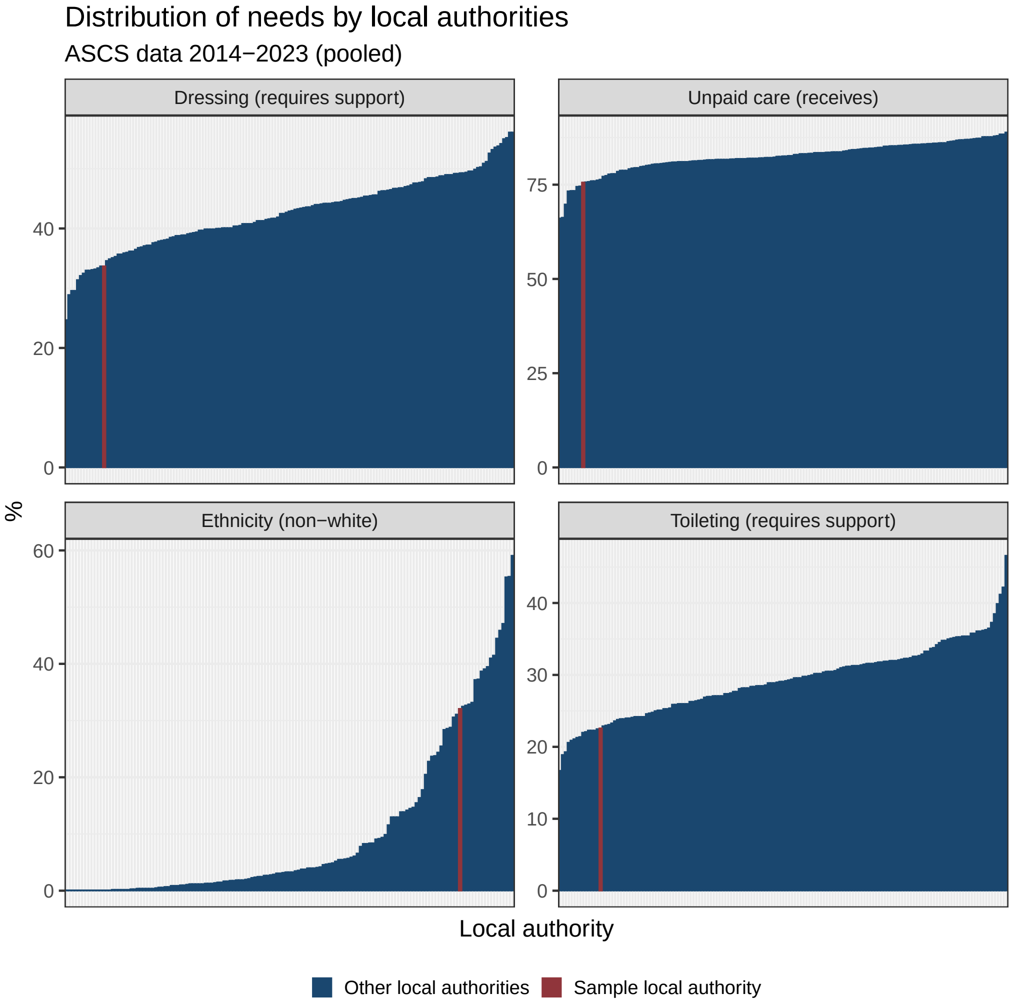


Figure 5.4: Distribution of needs in sample local authority

However, the needs in this local authority may be higher than they appear in ASCS. It is notable in Figure 5.4 that the local authority is ranked towards the lower end of the distribution for the percentage of individuals who require support with dressing. However, in Figure 5.3, the results from the administrative records of the same local authority indicate that it is at the upper end of the distribution. To investigate this tension, I set out in Figure 5.5 the comparison between the needs reported in the relevant local authority in ASCS and those recorded in that authority's administrative records. Once again, we see the pattern that demographic characteristics are very similar in both data sources, but needs-related factors appear significantly higher in administrative records than the survey data. In particular, ASCS data indicates that in this local authority 22.6% of individuals require support with using the toilet (0.95 CI 20.6% - 24.6%), and 31.3% with dressing (0.95 CI 29.2 - 33.7%). The values recorded in the administrative records are far outside these confidence intervals, at 52% and 80% respectively. This may partly reflect under-reporting or sampling bias in ASCS. I explore these phenomena in the next sections.

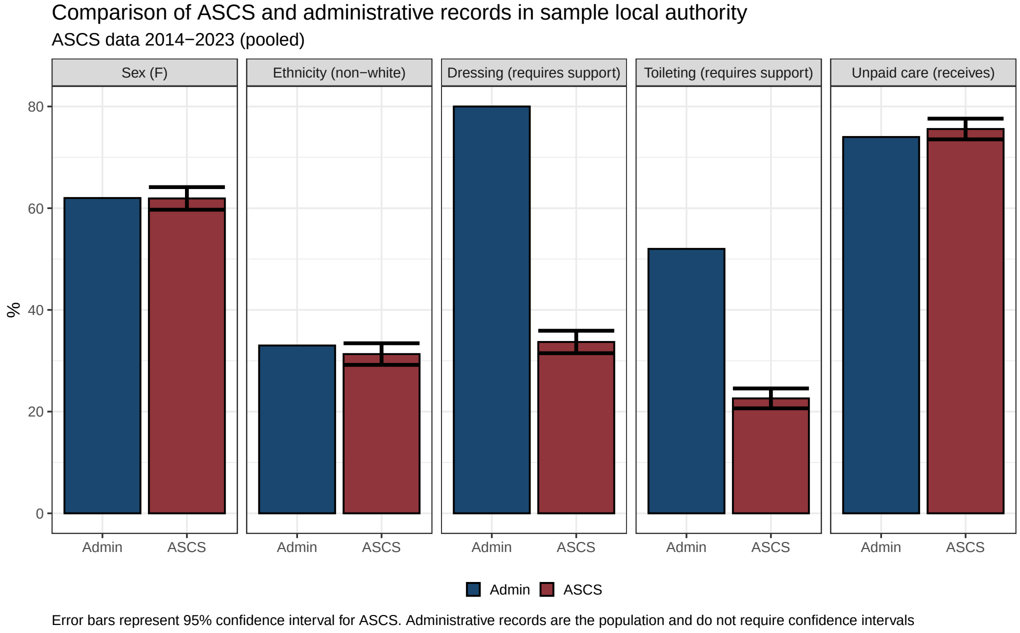
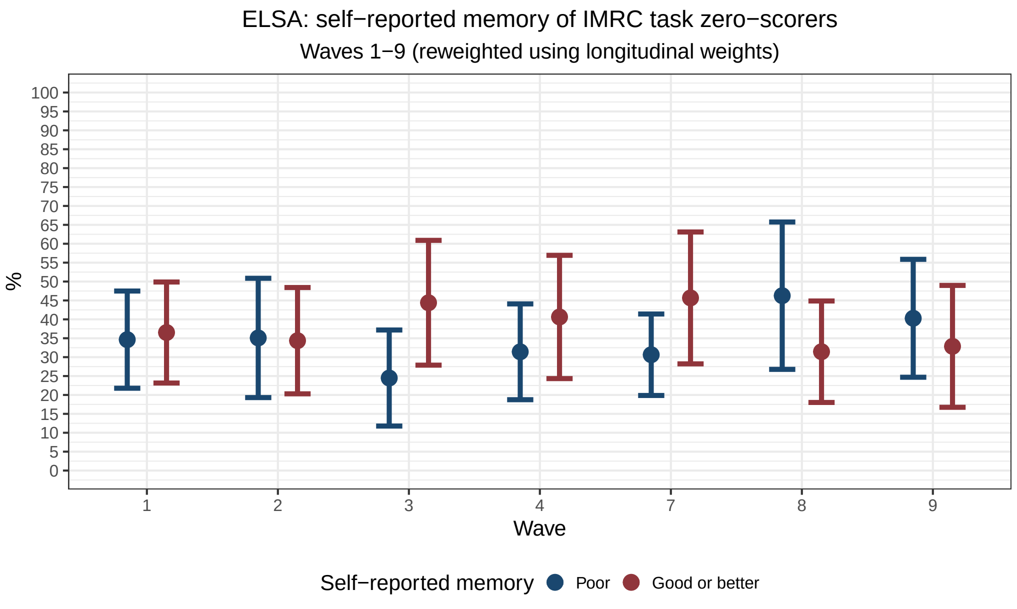


Figure 5.5: Comparison of ASCS and administrative records in sample local authority

5.3.5 The impact of under-reporting on differences between administrative and survey data

I compare responses to the ELSA questions about self-reported memory and immediate word recall to investigate under-reporting of memory issues. A χ^2 test indicated that there is an association between self-reported memory and immediate word recall ($p < 0.001$). However, this is perhaps not the most appropriate test, as it would be surprising if the null hypothesis (i.e. that there is no association) were true. The Cramér's V between the two questions is 0.115. This provides a measure of association for categorical data, with values ranging from 0 to 1, with scores in the range 0.1 - 0.2 indicating a weak association [313]. This suggests that while some relationship exists between immediate recall and self-assessed memory, immediate recall performance alone does not strongly predict self-perception of memory.

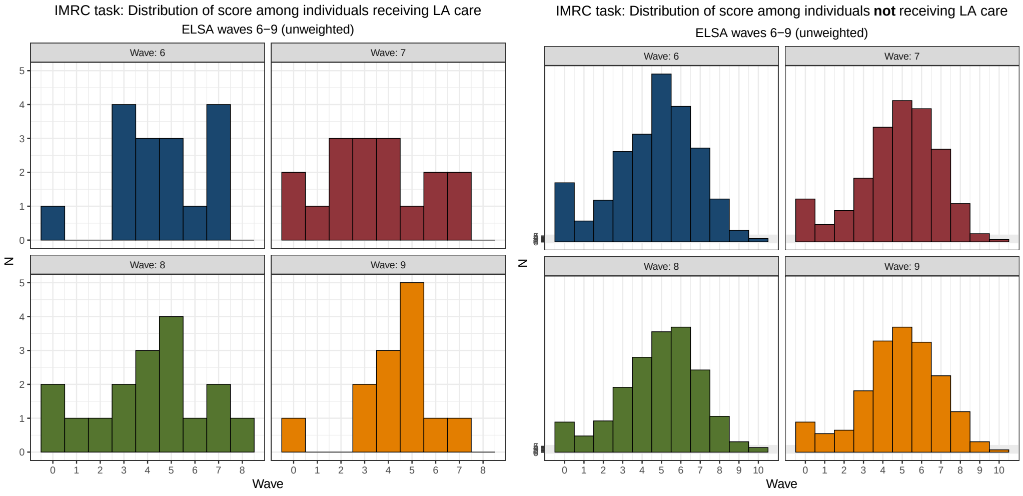
The weakness of the relationship is particularly evident among individuals who can recall zero out of 10 words immediately after they were read out. There are 599 respondents with an immediate recall with a score of zero pooled across ELSA waves containing both questions (waves 1-4 and 7-9). Around 34% of such individuals reported that their memory was poor. This is set out per wave in Figure 5.6, where it is compared against the number of people reporting their memory is *Good* or better. Until wave 7, more respondents who scored zero in immediate recall reported good than poor memory. This trend has somewhat reversed in waves 8-9, but it remains the case in later waves that around a third of those who score zero report that they have good or better memory. The confidence intervals are quite large, as the raw number of individuals per wave is relatively low (the largest is 41 in wave 7). Nevertheless, it is evident from these figures that there is a large overlap in the confidence intervals for reporting good memory and poor memory among individuals who could not immediately recall any words, indicating that self-reported memory may not be a good proxy for cognitive impairment, particularly among those who are most impaired.



Note: self−reported memory was omitted from ELSA waves 5 and 6

Figure 5.6: Self-reported memory among ELSA respondents scoring 0 on the immediate word recall test

I set out in Figure 5.7 a comparison of the distribution of immediate recall scores between individuals receiving local authority care and others from wave 6 onwards, when ELSA introduced the social care module. It does not appear that the majority of people receiving local authority care have highly impaired memory.



Note: Receipt of LA care was added to ELSA in wave 6

(a)

Note: Receipt of LA care was added to ELSA in wave 6

(b)

Figure 5.7: Self-reported memory among ELSA respondents scoring 0 on the immediate word recall test

There were 63 individuals receiving LA care who responded to this question in ELSA, which was too few for a wave-by-wave comparison of all scores. However, I grouped scores into *Low* (0-2), *Medium* (4-6) and *High* (7-10) and performed a χ^2 test of the observed distribution of memory recall scores from individuals receiving local authority care against the expected distribution of scores from all others, pooled across waves. This resulted in a p-value of 0.33, indicating that we should not reject the null hypothesis that there is no association between memory score and receipt of social care services. It is possible that with a larger sample of statutory care users there would be more of an effect, as visual examination of the plots suggests the people who score zero may be over-represented in statutory care users, particularly in wave 7. Nevertheless, there is no strong evidence that there are differences in functional memory ability between those receiving statutory care and others. We would expect people receiving local authority care to have higher needs than the general population, both because of the evidence from administrative records and the minimum eligibility thresholds for statutory care. This indicates that while under-estimation of one's own needs caused by self-reporting needs is an issue, it does not entirely explain the apparently lower needs of individuals receiving statutory care in surveys than administrative data. The other possible explanation for the individuals with care needs not having significantly greater memory issues is the impact of need on survey participation.

5.3.6 Correlation between needs and survey attrition

It is notable in Table 7.1 that the mean number of *unique* individuals with care needs as a proportion of the total individuals with care needs pooled across waves is around 78% in ELSA and 95% in Understanding Society. This indicates that individuals with higher needs do not frequently appear in more than one wave. To examine this in more detail, I set out in Table 5.3 the number of people in ELSA who are lost to follow-up (excluding those who die between waves), compared against those who appear in the next survey wave, based on both self-reported memory issues and immediate word recall score. Around a third of individuals with an immediate recall score of zero are lost to follow up, compared with just under a fifth of individuals who score higher. This is a highly significant difference, with $p < 0.001$ in a χ^2 test, and standardised residuals around 31 standard deviations from the expected result if the distributions were equal. However, when stratified by self-reported memory score, the proportion of those who state their memory is 'Poor' that are lost to follow up (19.3%) is very close to the proportion who state their Memory is ‘Fair’ or better (20.3%) ($p = 0.38$ in a χ^2 test). I include in Table 5.4 the same table for Understanding Society, based only on wave 3 data, which was the only wave with the immediate recall question, and using the derived binary variable indicating self-reported memory issues. Again, those who score zero on the recall task are likelier to be lost to follow up (20.7% vs 14.8%, $p < 0.001$), while we do not see the same difference based on self-reported memory (13.5% vs 13.2%, $p = 0.36$). Immediate word recall is evidently a greater predictor of attrition than self-reported memory in ELSA and Understanding Society.

Table 5.3: Individuals in ELSA lost to follow up based on memory

	Word recall score					Self-report				
Status	Memory recall > 0		Memory recall = 0		p	No memory issues		Memory issues		p
	N	%	N	%	<0.001	N	%	N	%	0.38
Lost to follow up	3263	19	257	33.3	***	3040	20	522	19.3	
Remained in survey	13916	81	515	66.7	***	12133	80	2185	80.7	

ELSA: English Longitudinal Study of Ageing waves 6-9. Self-reported memory variable: slfmem. Immediate word recall variable: imrc

Table 5.4: Individuals in Understanding Society lost to follow up based on memory

	Word recall score					Self-report				
Status	Memory recall > 0		Memory recall = 0		p	No memory issues		Memory issues		p
	N	%	N	%	<0.001	N	%	N	%	0.36
Lost to follow up	7204	14.8	119	20.7	***	3934	13.5	2158	13.2	
Remained in survey	41621	85.2	455	79.3	***	25179	86.5	14180	86.8	

Understanding Society waves 3-4. Self-reported memory variable: cgsrmem2_dv. Immediate word recall variable: cgwri_dv

5.4 Discussion and limitations

These comparisons between survey and administrative data indicate that individuals in surveys report lower needs than would be expected from administrative records, that the association between self-reported memory and measures of cognitive ability are statistically significant but not as strong as expected, and that there is an association between cognitive impairment and survey attrition. However, the administrative data comes from a single local authority, which may limit the generalisability of the findings to other regions or the country as a whole. Differences in local policies, demographics, and service provision practices can affect the applicability of the results to other settings. It is impossible to be certain about the extent to which one can extrapolate from trends observed with one local authority’s data. Nevertheless, it appears from this analysis that the higher levels of impairment recorded in the administrative data cannot be attributed to particularly high needs in that local authority. Additionally, while there is limited literature comparing administrative social care records to survey data, other research using administrative data appears to find they are higher. A study using data from 10 authorities in England and Wales in the 1990s found that a sample cohort of care users had higher needs than national data would indicate [27, 314]. Similarly, a table in a report assessing the impact of the Care Act using administrative records shows that 1,616 older people sampled from administrative data across 32 local authorities found workers reported that 46% of individuals required support to use the toilet [57]. While this is not directly comparable to the figures in Table 7.1 as only around 65% of participants were older people, it is higher than is indicated by survey data.

Another limitation is that, while administrative data provides detailed records, it may contain errors or inconsistencies due to variations in how information is recorded by different staff members. Missing data, over-redaction of sensitive information, or changes in data collection methods over time can impact the reliability of the findings. While the quality of administrative data used here has been assessed (I set out in Chapter 4 the mechanism by which I established the reliability of the data used in this chapter), and the literature highlights systematic sampling and reporting issues in surveys, it is nevertheless possible that some discrepancies between survey and administrative data are due to inaccurate administrative records rather than the survey data. It is possible that the estimates of functional ability from administrative data are an overestimate. Using data from 2006, Clarkson et al. [55] found that 66% of older, statutory care users in one local authority required help with dressing, which is lower than the 80% recorded here (though higher than the estimate of need from HSE, ASCS and Understanding Society). However, as demonstrated in the spending data presented in Chapter 4 and discussed in the literature [e.g. 5, 57, 6, 7], there has been retrenchment in social care and tightening of eligibility criteria, particularly since 2008. If the level of support required with dressing recorded in administrative data was found to be the same in 2024 as 2006, this might indicate that needs may not be being accurately captured. Indeed, in general, the literature assessing data quality in social care indicates an under-identification of care needs in administrative records [see e.g. 55, 130, 129]. If this were the case here, it would amplify the magnitude of the differences between survey and administrative records.

There is another inherent limitation in an analysis of this sort across several data sources. Many of the indicators that I have included are continuous, and recorded as ordinal categories. However, several surveys include them only as binary variables. To enable comparison across all sources, I have summarised every indicator as a binary variable. While this adds breadth to the analysis, some information is lost. However, the binary analysis of the proportion of individuals unable to complete an ADL is still valuable, as it provides a clear and comparable measure of need across datasets. Further research could explore methods for comparing datasets with more than two levels, if these can be mapped to each other effectively, to enhance the depth of future analyses.

Another issue is that the administrative data and survey data do not align perfectly in terms of the time periods they cover. Some differences observed may be due to changes in policies, economic conditions, or population demographics over time. If there were larger numbers of adult social care users in each wave of the longitudinal surveys, it would be possible to examine this, but this is limited by the available data. I am also limited by the relatively small sample size in most surveys, so have had to pool the data. This was not necessary for ASCS, though I have used the same methods for ASCS as the other surveys. I have examined the yearly ASCS data separately over the period and for 84% of the data, the yearly mean is within the confidence interval of the pooled mean.

This analysis also does not explore the potential use of instrumental variables (IVs) to address the limitations of non-random dropout [315]. IVs can provide a way to account for unobserved factors that influence both dropout and the variable of interest by isolating variation in the outcome that is independent of unobserved characteristics. It is possible IVs could potentially quantify the extent to which the variable of interest accounts for attrition — and thereby inform the creation of more accurate weights — identifying valid instruments that satisfy the required assumptions is challenging. The approach depends on sufficiently predictive instruments and strong theoretical assumptions, which are not always possible to empirically verify [316].

Finally, the only measure of functional ability that is not self-assessed was immediate word recall, which was only available in specific waves of surveys, limiting the analysis to that information in those time points. It is possible, and even likely given that the relationship is weakest among those most impaired, that reduced cognitive ability in particular is associated with reduced accuracy of self-assessed functional ability. The association of cognitive impairment with inaccurate reporting has been found in other surveys [106]. Having said this, researchers have also found differences between self-reported and observed physical functional ability not associated with cognition [138, 299]. Whether physical functional ability is also weakly associated with self-reported ability is an empirical question. It is not possible for me to answer it with the English surveys here, which do not include both measures.

5.5 Conclusion

This chapter highlights significant limitations in using survey data to understand the needs of older adults receiving publicly funded social care services in England. While the Adult Social Care Survey (ASCS) is by far the largest source of data about users of local authority care, it is limited by its cross-sectional design and omission of important variables such as cognitive ability. The way that ASCS is administered varies by local area, as does its representativeness of the local population [113]. Furthermore, the analysis in this chapter shows that respondents in ASCS from one local authority report lower levels of functional impairments compared to administrative records from the same local authority. Longitudinal surveys like ELSA and Understanding Society offer potential advantages but suffer from small sample sizes of statutory care users and systematic under-representation of those with higher needs. The analysis here finds, in line with previous research [e.g., 295], that individuals with impaired cognition are more likely to drop out of these surveys, further contributing to the under-representation of high-need populations.

Several factors contribute to this disparity. Under-reporting of needs in surveys is evident, particularly due to self-assessment bias and cognitive impairments that affect individuals' ability to accurately report their own needs. Additionally, sampling biases arise from the exclusion or under-representation of people with higher levels of need, and from survey attrition that disproportionately affects participants with declining health or cognitive function.

These findings demonstrate that survey weights are unlikely to be able to adequately correct for these biases. While survey weighting can mitigate some issues, in cases where missingness is systematically related to unobserved variables — such as severe disability or cognitive impairment — standard weighting methods fail to produce unbiased estimates [306, 317]. It would be possible for researchers using ELSA to create weights based on memory function, although this could be problematic if cognition itself is a variable of interest [318], which is often the case in long-term care research. Moreover, it would not be possible to create weights based on cognitive ability for Understanding Society (excluding wave 3), ASCS, HSE or FRS as it is not measured. While cognitive function is correlated with observed characteristics such as self-reported memory issues, the results here indicate the correlation of self-reported memory is low, and so may not be enough to capture the effect. Sensitivity analyses indicate that weighting methods often cannot fully adjust for biases arising from unobserved variables [319]. As a result, adjusted estimates may still not reflect the actual conditions or outcomes for those at greatest risk, leading to their under-representation in study findings [320].

These limitations have significant implications for research and policy. Given the limitations highlighted, reliance on survey data alone risks under-representing the needs of individuals requiring statutory social care. Researchers and policymakers should be cautious in interpreting survey data about social care, and consider exploring alternative data sources if hoping to understand the impact of care interventions on those with the highest needs. Administrative records, despite the effort required to process and analyse them, contain rich, unstructured data that is not available in survey data. However, it has previously not been practicable to extract information from this data. In Chapter 6, I address this gap by using LLMs to extract indicators of loneliness and social isolation from free text in administrative care records. This approach uses the large volumes of unstructured data recorded by care workers. This innovative method makes it possible to explore loneliness among statutory care recipients and assess its impact on care use — such as predicting care home entry — an analysis that is not possible with survey data.

6 Using Large Language Models to extract loneliness from free text social care notes

Abstract

Loneliness and social isolation are distressing for individuals and predictors of mortality. Evidence about the impact of loneliness and isolation on publicly funded long-term care usage is limited as there is little data indicating whether individuals using care services are lonely or socially isolated. Recent developments in natural language processing have made it possible to extract information from electronic care records, which contain large quantities of free text notes. In this paper we identify loneliness or social isolation from free text by analysing pseudonymised administrative care records containing 1.1 million free text case notes about 3,046 older adults recorded in a London council between November 2008 and August 2020. We use three natural language processing methods to represent the labelled notes as vectors: document-term matrices, pre-trained word embeddings and transformer-based, discriminative large language models. The most accurate model used a bidirectional transformer architecture. Evaluated on a test set of unseen sentences this model had an F_1 score of 0.92. We generate predictions of loneliness or isolation on the rest of the data using the best-performing model to examine the construct validity of our indicator by comparing it with other datasets and the literature. Our measure generally behaves as we expect: it is highly correlated with living alone, which we see in survey data. It is also associated with impaired memory which we expect from the literature. Furthermore, our indicator of loneliness or social isolation is a strong predictor of whether an individual receives services for social inclusion. According to our model, around 43% of individuals have a sentence indicating loneliness or social isolation in their case notes at the time of their initial care assessment. Similar estimates of prevalence are obtainable from individual-level survey data. The advantage of our method over surveys is that classified free text administrative data can be used in conjunction with other data in administrative records, such as care expenditure or service use. The outputs of our model can be used to generate inputs for regression models of service use which include social isolation and loneliness as independent variables. We include with this paper an open-source version of the

model in a GitHub repository.

6.1 Introduction

In 2021, public expenditure on long-term care was 1.98% of GDP in OECD countries [321]. In England, where the term adult social care is used to describe long-term support to complete activities of daily living, public spending was £23.7 billion (USD $30.4 billion) in 2022/2023 [322]. By 2038, projections indicate a 55% increase from 2018 levels in the number of older people receiving care, with public expenditure approximately doubling [323]. Given the demand for public funds, it is important to understand how individuals with different needs receive care and support. However, since the vast majority of participants in national surveys do not receive publicly funded care [103, 324], the available evidence for quantifying the use of such services from survey data is limited. Administrative care records provide a rich potential source of evidence about the characteristics and support received by people supported by the social care system. In the UK, long-term care needs have been widely recorded in electronic databases since the 1990s [325], and similar systems exist internationally to facilitate information sharing, manage performance and enhance safety [22, 58]. Historically, analysing large volumes of free text has not been practical. However, it is now possible, using machine learning, to extract information from data which is "complex, heterogeneous, poorly annotated and generally unstructured" [326]. Recent papers use natural language processing to extract information from free text electronic health records, [259, 260, 261, 152, 262]. However, to our knowledge only a handful of papers use similar methods with social work or long-term care records [149, 151, 150], all of which use US data and none of which focus on loneliness or social isolation.

This paper extracts from free text records whether a worker has recorded that an individual receiving care is lonely or socially isolated. We begin by setting out why social isolation and loneliness are an important area of focus given their health impact and the absence of structured, time-variant, individual-level data recording which publicly funded care users are lonely or socially isolated. We then describe the administrative dataset used in the analysis in this paper. Needs assessment forms in such datasets do not generally contain structured questions, such as a checkbox, indicating whether a person is lonely. However, they contain large quantities of free text about an individual's social needs. An advantage of using administrative records over survey data is that classified free text administrative data can be used to model care expenditure or service use data, such as care home entry, which also form part of administrative records. We discuss this and other policy implications in our concluding section.

6.2 The impact of loneliness and social isolation

A question faced by researchers with access to administrative care data is which information to extract. Loneliness and social isolation have been found to be as significant a predictor of mortality as smoking, obesity or hypertension [327, 328, 329]. Loneliness and social isolation are related but distinct concepts. Although definitions vary [330, 331], generally social isolation is an “objective lack of relationships” [332], whereas loneliness is represented as a “subjective, distressing feeling” occurring when the quantity or quality of existing social relationships is less than desired [333].

Loneliness has been an international health policy concern for decades [see e.g. 334] and a consensus statement from leading international researchers described loneliness as “the 21st century social determinant of health” [335]. In 2021, the World Health Organisation (WHO) published an advocacy brief stating that more research is required to understand the prevalence of loneliness and social isolation in older people, and the effectiveness of interventions [331]. In 2018, the UK government published a loneliness strategy, which aimed to “embed loneliness as a consideration across government policy” [336]. In 2023, they published an update calling for more research into the measuring of loneliness and the economic effectiveness of interventions [330]. In 2022, a question was added to England’s Adult Social Care Survey asking care users if they feel lonely [112], the first change since 2014.

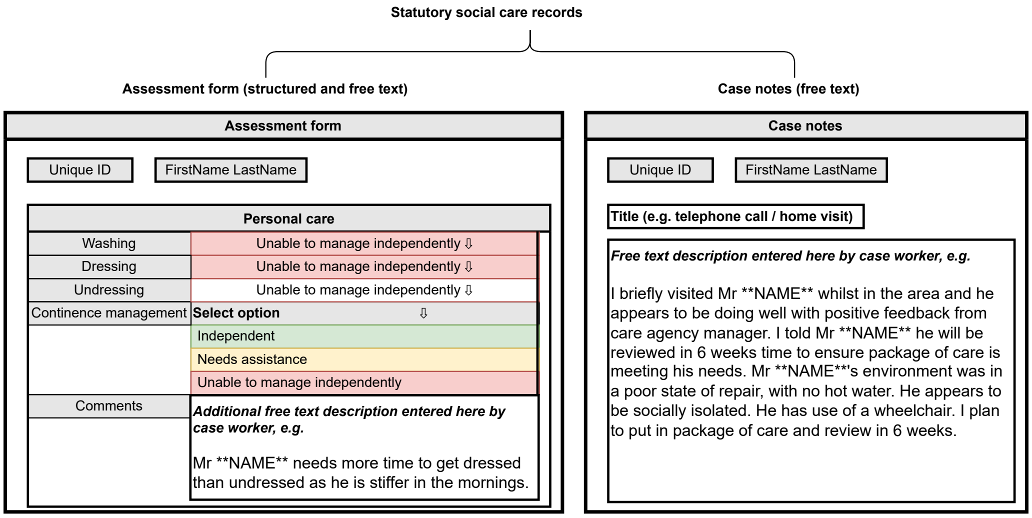
In England, the National Institute for Health and Care Excellence (NICE) recommends that older adults participate in social activities [337]. In recent years, however, many publicly funded interventions to increase social connectedness have been closed or decommissioned [263]. Local authority returns show that adults who receive a support package where the Primary Support Reason is social isolation have reduced by around a third since 2015 [338]. There is some evidence indicating loneliness is associated with increased long-term care use. A 2018 study using English Longitudinal Study of Ageing (ELSA) data indicates that loneliness is a risk factor for entering a care home after adjusting for factors such as age, gender and disability [267]. However, the number of people receiving publicly funded social care services in the ELSA is very limited [109], and such research gives a less comprehensive picture of how loneliness might impact publicly funded care packages than administrative data.

Administrative care records in England contain structured and free text information about an individual's ability to develop and maintain family and personal relationships and to access community and recreational services, which form part of the eligibility criteria for publicly funded care [339]. Distinguishing between loneliness and social isolation in free text notes can be challenging, as social workers may not use the terms "lonely" or "isolated" consistently with their definitions in the literature. For instance, the phrase "feels isolated" frequently appears, which might refer to a subjective feeling of loneliness or indicate that the person has a lack of social contact. In a 2024 paper, Patra et al. were able to distinguish types of social support needs from psychiatric records [154]. However, they note that those records "include relative consistency in documentation style" and contain "comprehensive biopsychosocial patient evaluation". As the administrative data used in our study is less consistent in documentation style, with social networks or circumstances often mentioned only briefly or indirectly, we examine both social isolation and loneliness together. This approach is common in public health research; for instance, a 2023 report by the US Surgeon General was entitled *Our Epidemic of Loneliness and Isolation*, highlighting the interconnectedness of these issues [340]. Research which has considered the concepts together has found that being socially isolated or lonely has an adverse effect on mortality among older people [327, 328, 329, 341]. In this paper, we take the approach used in the papers which consider the concepts together. This is also consistent with the approach taken in Zhu et al. [152] when they extract social isolation and loneliness from free text medical notes of cancer patients. Our study also aligns with recent research efforts exploring the use of large language models (LLMs) to extract social and contextual factors from clinical text [153]. A 2024 study developed multilabel classifiers to identify six distinct Social Determinants of Health (SDoH) from clinical notes. Their findings revealed that fine-tuned, discriminative models outperformed generative models in SDoH extraction tasks, particularly for rare and under-documented categories such as social support and housing issues. This demonstrates the capacity for using free text to extract adverse social conditions compared to structured data alone.

6.3 Materials and methods

6.3.1 Data collection

In England, every person requesting publicly funded care must receive an assessment under the Care Act 2014. In this paper we attempt to identify social isolation or loneliness from the free text notes of a London borough. Adult social care case notes are written by individuals employed by a local authority to assess needs and commission care in accordance with The Care Act 2014. This generally consists of social workers, occupational therapists or care managers. Workers complete an assessment form, which is a snapshot of needs at a certain time containing both structured data and free text. Recording systems also contain case notes, which are free text fields to record ongoing work on the case over time. In Figure 6.1 we show how the assessment form and case notes appear to caseworkers.


Statutory social care records
Assessment form (structured and free text)
Case notes (free text)
Assessment form
Unique ID
FirstName LastName
Personal care
Washing
Unable to manage independently ⇩
Dressing
Unable to manage independently ⇩
Undressing
Unable to manage independently ⇩
Continence management
Select option ⇩
Independent
Needs assistance
Unable to manage independently
Comments
Case notes
Unique ID
FirstName LastName
Title (e.g. telephone call / home visit)

Additional free text description entered here by case worker, e.g.

Mr **NAME** needs more time to get dressed than undressed as he is stiffer in the mornings.

Free text description entered here by case worker, e.g.

I briefly visited Mr **NAME** whilst in the area and he appears to be doing well with positive feedback from care agency manager. I told Mr **NAME** he will be reviewed in 6 weeks time to ensure package of care is meeting his needs. Mr **NAME**'s environment was in a poor state of repair, with no hot water. He appears to be socially isolated. He has use of a wheelchair. I plan to put in package of care and review in 6 weeks.

Figure 6.1: Structured and unstructured data

6.3.2 Data extraction and characteristics

A query was written to identify all individuals aged 65 or over on August 1st 2020, who had been receiving services for at least one year since 1st January 2016. Administrative records for these individuals were then extracted from the local authority database. Identifiable free text data tokens were masked using the open-source text pseudonymisation software *PSCleaner* [342]. The data was then sent to an NHS Commissioning Support Unit, where identifiable structured data such as NHS numbers were removed. Finally, the data was transferred securely to the research team at the Care Policy and Evaluation Centre (CPEC) at the London School of Economics and Political Science (LSE).

The data includes all free text case notes recorded for individuals in the cohort between 2008 and 2020, as well as needs assessment and service receipt data. During this period, there were 3,046 individuals aged over 65 receiving long-term care.

Table 6.1: Free text per person

	Case notes		Assessment		Total
	N notes	N words	N text fields	N words	N words
Mean	377	28,850	148	6,740	35,115
SD	298	27,068	112	5,603	29,670
Median	302	21,444	123	5,212	27,330
Min	2	6	1	4	6
Max	2,585	407,283	869	44,196	408,404

The data contains 10,821 assessment forms comprising 19.1 million words of free text, and 1.14 million case notes, containing 87.8 million words of free text. Case notes in the dataset encompass a wide range of updates related to the care and support of individuals. These include records of emails and telephone calls, descriptions of home visits, case screening and allocation, managerial direction, case summaries, allegations of abuse or neglect, as well as referrals to services such as occupational therapy, physiotherapy, and intermediate care. The volume of notes reflects the comprehensive documentation required in social care to capture various interactions and decisions throughout the course of care. There is significant variation in the distribution, with for example the top 50 individuals having 8 million words recorded (7.8% of the total), the same amount as the 850 individuals with the fewest. Summary statistics per person are presented in Table 6.1.

In addition to the free text case notes, structured data fields are routinely collected during the assessment process. These fields capture key demographic and personal information that is relevant for care planning and service provision. Structured data includes information such as gender, ethnicity, age, functional ability with activities of daily living (ADLs), and whether the individual lives alone. This information is collected directly by social care professionals during initial assessments and periodic reviews as part of standard care practices. These structured data fields provide important context for understanding the care needs of individuals and were used in conjunction with the free text data in our analysis. Of the 3,046 individuals, 61.2% were women, 47.8% White British, with a median age in 2020 of 81, and median of 3 years and 6 months of services received. These characteristics are set out in Table 6.2.

6.3.3 Overview of model development and evaluation

We set out here methods for model development and model evaluation. In the model development section, we describe the data pre-processing, manual classification of text and range of classification algorithms used. We trained several machine learning algorithms on the data: logistic regression, random forest, bootstrap aggregation, quadratic discriminant analysis, and a feed-forward neural network. Our model evaluation section contains two sub-sections. Firstly, we set out the performance metrics for evaluating which of the machine learning models was most accurate after running the trained models on a test set of human-classified sentences about individuals unseen by the trained model. Secondly, we examine the construct validity of the output of the best-performing model, to establish whether the indicator of loneliness or social isolation that we have derived behaves as we would expect. We did this by generating predictions at the time of initial assessment, and then using hypothesis testing to interrogate expected relationships with needs and demographic characteristics based on the literature and related datasets. Additionally, using the same predictions, we ran a logistic regression to see whether our derived indicator of loneliness or social isolation is associated with receiving services for social inclusion. We set these methods out in more detail below.

6.3.4 Model development

We endeavoured to use as parsimonious a model as possible, beginning with count-based vector representations of words such as document-term matrices [170] and Term Frequency Inverse Document Frequency (Tf-idf) [343]. We also used the SpaCy large pre-trained word embeddings [344], and transformer-based representations, specifically RoBERTa and DistilRoBERTa [195]. The overall process for training and comparing these models is set out in Figure 6.2. Unless otherwise stated, we used Python 3.9.7 in all the analysis [345].

Table 6.2: Training and test set

Group	N	Sentences (total)	Sentences (median)	Min Date	Max Date	Service length	F (%)	WB (%)	YOB	Deceased
Train	200	317263	1286	2011-04-01	2022-04-15	3.03	62	47	1933	40
Test	200	305319	1238	2011-04-01	2022-04-15	3.83	62	49	1934	43
All	3046	4807982	1289	2010-01-29	2022-04-15	3.49	62	47	1934	42

Note:
F: Female. WB: White British. YOB: Median year of birth. Service length: Median time receiving statutory care services.

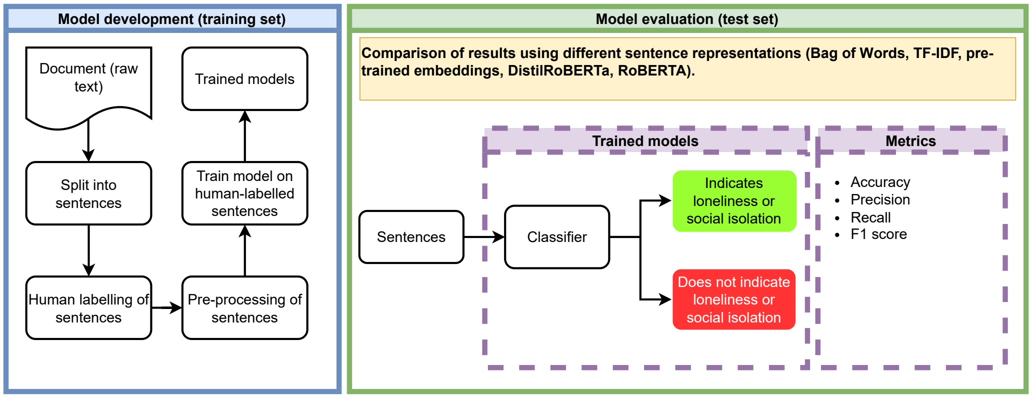

Model development (training set)
Document (raw text)
Split into sentences
Human labelling of sentences
Pre-processing of sentences
Train model on human-labelled sentences
Trained models
Model evaluation (test set)
Comparison of results using different sentence representations (Bag of Words, TF-IDF, pre-trained embeddings, DistilRoBERTa, RoBERTA).
Trained models
Metrics
Sentences
Classifier
Indicates loneliness or social isolation
Does not indicate loneliness or social isolation
• Accuracy
• Precision
• Recall
• F1 score

Figure 6.2: Overall training and evaluation process

For all approaches, we replaced the pseudonymised masks (e.g., **NAME**, **LOCATION**, which had been used to mask identifiable information) with randomly generated names and locations to ensure that the language models could correctly tokenise and parse the sentences. Retaining the pseudonymisation masks could have led to issues with tokenisation, as the models may not have handled repeated placeholders effectively. For the count-based methods, we also lemmatised the text, converted it to lowercase, and removed stop words. We set out further details in Data pre-processing in S2 Methods Appendix. We then divided the data into a training and test set, using stratified random sampling to ensure similar proportions of individuals in each set (see Table 6.2). Each set contained notes about 200 distinct individuals. We split each set by person to ensure that the test set did not contain sentences about individuals who are in the training set. Human annotators manually classified 10,083 sentences in the training set, and 3,573 sentences in the test set for model evaluation. We defined a set of rules for annotators to define which sentences to classify, using binary classification (either indicative or not indicative of loneliness or social isolation). We developed classification rules to guide the annotators in determining whether a sentence indicated loneliness or social isolation. These rules covered statements such as when a person explicitly expressed feeling lonely, had little social contact, or received referrals to services like befriending. Conversely, sentences indicating practical support needs, support for safety or cognition, or day centre attendance for carer respite were classified as not indicative of loneliness or social isolation. The full set of rules is detailed in the S2 Methods Appendix in the Supporting Information section. Our interrater reliability measures produced Cohen's κ [346] of 0.89 (95% CI 0.84 - 0.94) and Krippendorff's α [347] of 0.89 (95% CI 0.89 - 0.93). The maximum level of agreement in both cases is 1, and 0.89 represents excellent levels of agreement beyond chance [131, 348]. The training dataset was imbalanced, with 9,383 sentences in the negative class (not indicative of loneliness or social isolation) and 700 in the positive class. We implemented three approaches for the representation of words:

1. Count-based approaches: We split each sentence into lemmatised, word-level tokens. Each sentence was represented by a raw count of the number of times a word appears in it (a document-term matrix) [170]. We also applied Term Frequency Inverse Document Frequency (Tf-idf) [180] to transform the count matrix to a weighted representation, reducing the weighting of higher frequency words across all documents.
2. Pre-trained vectors: We used the Spacy large English model [344], which represents language through dense embeddings [175], where words which have similar semantic meanings are clustered together in vector space. We took the mean of each dimension to create a single 300-dimensional vector to represent each sentence.
3. Transformer-based approaches: We used the RoBERTa *base* model, which has 12 hidden layers, 768 dimensions and 12 heads [195]. This was relatively computationally expensive to fine-tune, so for comparison we also used DistilRoBERTa, which has identical parameters except it has 6 hidden layers, and is around twice as fast to train. In both cases, we used the HuggingFace implementation of each model's tokenizer to split each sentence into sub-word tokens [349, 350].

We describe these approaches in more detail in S2 Methods Appendix. After pre-processing, vectorising and labelling each sentence, the problem becomes a binary classification task. For both the count and pre-trained embedding based approaches, we evaluated five classification algorithms. We used k fold cross-validation to avoid overfitting on the training set, choosing 5 folds for k as a value which tends to elicit reasonably high accuracy [351] while reducing training time compared with higher values. We used five classification algorithms: class-weighted logistic regression, bootstrap aggregation, random forest, quadratic discriminant analysis and a feed-forward neural network. Again we set these out in the Classification algorithms section of S2 Methods Appendix, in our Supporting Information document. For the transformers approach, the HuggingFace implementation of both the RoBERTa and DistilRoBERTa models contain a classification head. We trained this final layer of the model on the labelled, tokenised sentences using the HuggingFace Transformers and PyTorch libraries [352, 353]. Our final model had a training batch size of 16 sentences, with 500 warm up steps, and weight decay of 0.01. The weight decay parameter is bounded between 0 and 1, with 0.01 indicating relatively low L2 regularisation, which can help the model fit the smaller, positive classes more accurately, but can risk overfitting. The final output layer produces a predicted probability for the negative and positive classes (either indicative or not indicative of loneliness or social isolation). During training the parameters of the classification head were optimised using binary cross-entropy loss, which measures the difference between the predicted probabilities and the true labels.

6.3.5 Model evaluation

We evaluate the performance of our model in three ways to assess the accuracy and construct validity of the NLP-generated indicators of loneliness or social isolation. Firstly, we measure performance on a test set of unseen sentences, drawn from individuals not included in the training data, and compare the accuracy, precision, recall, and F_1 scores for each model. We then use the model with the highest F_1 score to classify sentences from the initial care needs assessments of all individuals in our dataset, examining the associations between the model output and demographic characteristics in structured data, such as gender, ethnicity, and functional ability. We compare these findings with associations between loneliness and the same characteristics in the ELSA survey. Finally, we conduct a logistic regression to assess whether the model's loneliness or isolation predictions are associated with the use of services typically related to social support needs. Together, the final two steps compare the NLP model output with expected indicators of loneliness or social isolation, measuring the model's construct validity.

6.3.5.1 Model evaluation: construct validity

After establishing the best-performing model using the test set, we set out to establish the construct validity of the indicator of loneliness or social isolation generated by this model by establishing whether it behaves as we would expect

Table 6.3: Model metrics

		Case notes	
		0	1
Assessment	0	Neither	Case notes
	1	Assessment	Both

it to based on the literature, as in Malley et al. [354]. We did this by using the best-performing model to classify all free text at the time of the initial assessment for the 1,331 individuals whose first contact with statutory care services was in the data. We chose to evaluate the text at the time of the initial assessment as we would expect care needs to be comprehensively recorded at first contact with services, and we classified all text within the assessment form, and all case note text within 90 days of the assessment. We note whether individuals have at least one case note classified by the model as indicative of loneliness or social isolation in the assessment form or case notes, separately. As this is a brief period with relatively small quantities of text per person, very few individuals had more than one positive sentence, so we treated the indicator as binary rather than as a count. Once the indicator was treated as binary, we derived four metrics: individuals with no positive sentences (*Neither*), those with a positive sentence only in their assessment (*Assessment*), those with a positive sentence only in case notes (*Case notes*) and those with both (*Both*). This is set out in Table 6.3.

We then compared the results of the model predictions with pooled data from waves 6 - 9 (2012 - 2019) of ELSA [109]. This secondary dataset, collected as part of a large national survey through structured interviews and self-reported questionnaires, provides a validated source of information on the characteristics of older adults in England. We used ELSA data for all older adults who stated that they had care needs and received publicly funded care (N=995 unique individuals with 1361 total observations over the period). We pooled the results due to the low number of responses in some groups. We tabulated responses to the ELSA Center for Epidemiological Studies Depression Scale (CES-D) loneliness question [355]. We also compared our results to the three UCLA loneliness scale questions within ELSA, converting a total score of 6 or more into a binary indicator of loneliness, as in Hanratty et al. (2018) [267]. As our model measures loneliness and social isolation, in the ELSA data we also establish which individuals are socially isolated according to the Social Network Index (SNI) defined in Minicuci et al. (2016) [356].

We compare the results with ELSA graphically, by examining the proportion of people in our data and in ELSA who appear lonely or socially isolated, broken down by demographic characteristics and care needs. We also conduct a Pearson's χ^2 test of independence [357] of each need or demographic factor with loneliness or isolation, to establish whether there are the same associations between our indicator of loneliness and those found in ELSA. Finally, we conduct a logistic regression of all these factors and loneliness or isolation, to establish which factors remain significant after controlling for characteristics such as living alone.

6.3.5.2 Model evaluation: predicting service receipt for loneliness or isolation

There is a statutory duty under the Care Act 2014 to meet eligible needs for maintaining personal relationships, recreation and accessing the community. Day centres are community-based services provided for older people at risk of loneliness or social isolation [263] which have been seen as the primary method of discharging this duty. Our data includes information on whether individuals are receiving day centre services, which we expect to be more likely for individuals who are truly lonely or socially isolated. To assess this, we generated predictions of loneliness or social isolation using our best-performing model, the RoBERTa-based language model.

Next, we conducted a logistic regression to examine whether these predictions were associated with the receipt of day centre services within 90 days of the initial assessment, for the 1,331 individuals whose initial assessment could be identified. To ensure that the RoBERTa model was not simply picking up cases with more notes, or cases driven by demographic characteristics rather than actual loneliness or isolation, we included the number of notes and relevant demographic variables as controls in the logistic regression model. This allowed us to verify that the RoBERTa model's predictions were not confounded by factors such as a greater volume of documentation or demographic differences, rather than genuine cases of loneliness or social isolation. The logistic regression model is specified in Equation 6.1.

$$\begin{aligned}\log\frac{p}{1-p} =& \beta_0 + \beta_1 SIL + \beta_2 notes + \beta_3 sex + \beta_4 ethnicity + \beta_5 age + \\ & \beta_6 pc + \beta_7 memory + \beta_8 safety + \beta_9 alone\end{aligned} \tag{6.1}$$

Where p is the probability of receiving day services in the first 90 days, SIL is the binary prediction of social isolation or loneliness generated by our model, *notes* is the number of sentences written within 90 days of assessment, *sex* is a binary variable where 1 indicates male, *ethnicity* is a binary indicator of white or non-white and *age* is age of the person receiving care in years. Additionally, we include as continuous variables the following rank of severity of needs, where higher indicates more care needs. *pc* is personal care needs (the sum of mobility, toileting and dressing), *memory* is the score for memory and cognition, $safety$ is the extent to which the person is aware of their own safety and risk and *alone* is a binary indicator of whether an individual lives alone. The demographic and needs-related scores are extracted from the structured data of the initial assessment.

6.4 Results

We present a set of results for each method of model evaluation. Firstly, we evaluate the performance of each model against the test set. Secondly, we run the best-performing model on text recorded within 90 days of every initial assessment and compare the significance of association with demographic characteristics with survey data from ELSA. Finally, we present the logistic regression of the results of the best-performing model on day centre attendance.

6.4.1 Model performance on the test set

We use the evaluation metrics for binary classification models set out in Raschka and Mirjalili [358], specifically accuracy, precision (positive predictive value), recall (true positive rate) and F1 score.

$$\text{accuracy} = \frac{TP + TN}{n}$$

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Where n is the number of sentences classified, TP the number of true positives, FP false positives and FN false negatives. True negatives are included in accuracy, but not precision, recall or F_1. As the majority class is the negative class, it would be possible to achieve accuracy > 0.9 with a classifier which predicted that every sentence was in the negative class. We therefore use F_1 as the primary measure, which is the standard in binary classification tasks [see e.g. 259, 152].

In Table 6.4 we detail the accuracy, precision, recall and F_1 score of each model on the test set of 3,573 labelled sentences not seen by the training set. The transformer-based models considerably outperform all other models, with DistilRoBERTa achieving an F_1 score of 0.86 and RoBERTa 0.92. The pre-trained Spacy embeddings outperformed all non-transformers based approaches when classes were predicted using a feed-forward neural network, with an F_1 score of 0.61. However, using the same embeddings, the neural network only slightly outperformed logistic regression, which had an F_1 score of 0.58. The count-based approaches were not effective at prediction using any of the classification methods. In Figure 6.3, we present a confusion matrix comparing the predictions of the best-performing model against the classes defined by human annotation. The values in the matrix represent the sentences classified into each category by human annotators and the RoBERTa model.

Table 6.4: Results

Classifier	Accuracy	Precision	Recall	F1
Transformers				
RoBERTa	0.97	0.95	0.87	0.92
DistilRoBERTa	0.96	0.90	0.82	0.86
Pre-trained embeddings				
Feed forward NN	0.90	0.78	0.50	0.61
Logistic regression	0.83	0.46	0.77	0.58
QDA	0.84	0.45	0.28	0.34
Bagging	0.85	0.66	0.07	0.13
Random Forest	0.85	1.00	0.02	0.03
Tf-idf				
QDA	0.27	0.15	0.83	0.26
Bagging	0.82	0.15	0.03	0.06
Feed forward NN	0.84	0.22	0.01	0.03
Logistic regression	0.84	0.07	0.01	0.01
Random Forest	0.84	0.00	0.00	0.00
Document-term matrix				
QDA	0.23	0.15	0.83	0.25
Bagging	0.83	0.13	0.02	0.04
Feed forward NN	0.84	0.24	0.02	0.03
Logistic regression	0.84	0.06	0.00	0.00
Random Forest	0.84	0.00	0.00	0.00

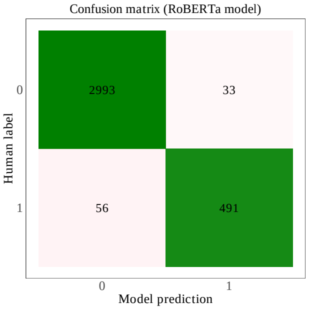


Figure 6.3: Confusion matrix (RoBERTa)

6.4.2 Construct validity: comparison with survey data

The overall proportion of individuals with at least one case note indicating loneliness or social isolation according to our model is 0.44 (95% CI 0.42 - 0.47), and the proportion with at least one sentence indicating the same in their needs assessment is 0.43 (95% CI 0.40 - 0.45). This compares with a proportion of 0.38 in ELSA (95% CI 0.32 - 0.43) who are lonely according to the CES-D measure or SNI isolated, and 0.45 (95% CI 0.39 - 0.51) who are lonely according to the UCLA measure, or SNI isolated. The overall proportions are similar to the UCLA loneliness measure, and this holds for many characteristics. We present in Figure 6.4 a breakdown of these proportions by demographic and needs-related factors. While these similarities are reassuring, there are differences between the results of our model and ELSA. For example, the difference in loneliness between individuals who live alone and live with others is wider in ELSA than in our data.

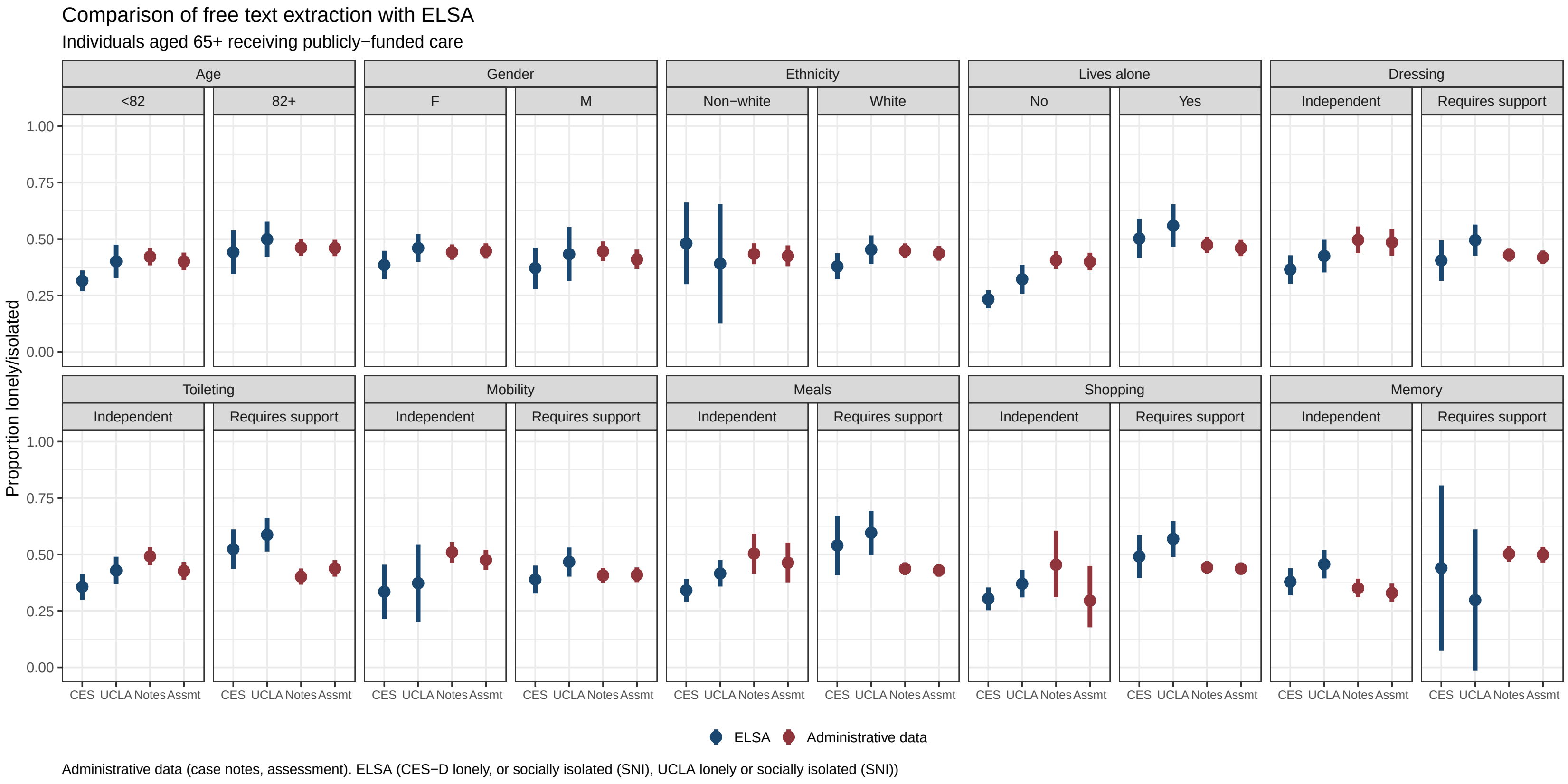


Figure 6.4: Structured and unstructured data in adult social care case notes

We set out in Table 6.5 the results of the χ^2 test of independence between loneliness or isolation and the needs-related factors in both our results and ELSA. We also present the results of the combined indicators: *Either*, where a person has a sentence in either assessments or case notes indicating loneliness or isolation, and *Both* where a person has a sentence in both case notes and their assessment form. The χ^2 tests reveal both similarities and differences in the associations between our RoBERTa-based indicator and the ELSA measures of loneliness and social isolation (CES-D and UCLA combined with SNI). Both our indicator and the survey data show a strong association between loneliness and living alone. However, our indicator also identifies a significant link between memory issues and loneliness, which is not found in the ELSA data. Additionally, ELSA data shows that people receiving unpaid care are more likely to be lonely, a pattern not reflected in our findings. This discrepancy may be due to differences in the cohorts or the nature of the data, as ELSA data is self-reported, while administrative assessments of functional ability are recorded by professionals.

Although we have taken a subset of individuals from ELSA who are older people receiving local authority care, individuals in the administrative data have higher needs than those in ELSA. For example, 80% of individuals in the administrative data require support with dressing, compared with only 69% of those in the ELSA data. This is set out in Table 6.7. As the ELSA data used in this comparison is self-reported, while the administrative data are derived from structured records collected by local authorities, we expect some differences in the association between specific ADLs and social isolation and loneliness. We do not consider this a barrier to comparing the datasets, but we do consider it when interpreting the results. We elaborate on this in the Discussion section.

To consider the association with needs-and demographic characteristics in total, as well as individually, we conducted a logistic regression. We set out the results in Table 6.5. To account for multicollinearity, we measured the generalised variance inflation factor (GVIF) [359]. Rule-of-thumb thresholds are generally that there is too much multicollinearity if the variance inflation factor is greater than a threshold, which tends to be between 4 and 10 [360]. The maximum GVIF between any of our covariates was 1.3. The regression output indicates that in ELSA, living alone is by far the most significant predictor, though requiring support with shopping and presence of unpaid care are also significant. Across all four of our measures, living alone is also a very important predictor of loneliness or isolation. The coefficient is around the magnitude of that for memory, where individuals who have memory problems are more likely to be lonely or socially isolated. We do not see the same effect for shopping or unpaid care. We discuss these differences and the differences in the χ^2 test in the Discussion section.

Table 6.5: Factors in structured data associated with loneliness and social isolation: Chi-sq test and logistic regression

	Administrative records				ELSA	
	Assessment	Case notes	Either	Both	CES-D	UCLA
Chi-sq test						
Dressing	0.048 *	0.047 *	0.254	0.001 ***	0.398	0.099 .
Ethnicity	0.688	0.635	0.655	0.619	0.606	0.526
Lives alone	0.027 *	0.013 *	0.033 *	0.003 **	<0.001 ***	<0.001 ***
Meals	0.471	0.154	0.59	0.066 .	<0.001 ***	<0.001 ***
Memory	<0.001 ***	<0.001 ***	<0.001 ***	<0.001 ***	0.462	0.605
Mobility	0.02 *	<0.001 ***	0.071 .	<0.001 ***	0.509	0.016 *
Safety & risk	0.284	0.126	0.869	0.477	0.737	0.624
Sex (F)	0.188	0.896	0.814	0.278	0.262	0.181
Shopping	0.062 .	0.878	0.473	0.25	<0.001 ***	<0.001 ***
Toileting	0.68	0.001 ***	0.467	0.013 *	0.046 *	0.002 **
Unpaid care	0.726	0.258	0.689	0.657	<0.001 ***	<0.001 ***
Logistic regression						
Age	1.01 (0.99-1.02)	1.01 (1.00-1.03)	1.01 (1.00-1.02)	1.01 (1.00-1.03)	1.01 (0.99-1.03)	1.00 (0.99-1.02)
Dressing	0.87 (0.78-0.97) **	0.91 (0.82-1.01) .	0.88 (0.79-0.99) *	0.87 (0.78-0.97) *	0.88 (0.62-1.23)	0.92 (0.66-1.28)
Ethnicity	0.96 (0.76-1.23)	0.98 (0.77-1.24)	0.99 (0.77-1.27)	0.95 (0.72-1.23)	1.60 (0.47-5.34)	1.33 (0.36-5.05)
Lives alone	1.37 (1.08-1.75) **	1.38 (1.09-1.76) **	1.52 (1.18-1.95) ***	1.39 (1.06-1.82) *	6.12 (4.36-8.70) ***	3.57 (2.62-4.90) ***
Meals	1.02 (0.90-1.16)	1.09 (0.96-1.23)	1.09 (0.95-1.24)	1.04 (0.91-1.19)	1.36 (0.89-2.07)	1.31 (0.86-1.99)
Memory	1.40 (1.25-1.57) ***	1.40 (1.25-1.57) ***	1.50 (1.33-1.70) ***	1.47 (1.30-1.67) ***	1.14 (0.47-2.71)	0.65 (0.26-1.57)
Mobility	0.89 (0.81-0.99) *	0.99 (0.90-1.10)	1.01 (0.90-1.13)	0.85 (0.76-0.96) **	0.86 (0.56-1.32)	1.19 (0.79-1.80)
Safety & risk	1.03 (0.92-1.15)	0.94 (0.84-1.05)	1.00 (0.89-1.12)	0.96 (0.85-1.09)	1.35 (0.55-3.26)	1.17 (0.48-2.85)
Sex (F)	1.15 (0.91-1.45)	0.89 (0.70-1.12)	0.97 (0.76-1.24)	1.06 (0.82-1.37)	0.92 (0.66-1.27)	0.99 (0.72-1.35)
Shopping	1.09 (0.93-1.28)	1.02 (0.88-1.20)	1.06 (0.90-1.24)	1.07 (0.90-1.28)	1.61 (1.13-2.29) **	1.45 (1.03-2.05) *
Toileting	1.04 (0.95-1.15)	0.88 (0.80-0.97) *	0.93 (0.84-1.03)	0.97 (0.87-1.08)	1.25 (0.79-1.97)	1.46 (0.94-2.26) .
Unpaid care	1.01 (0.78-1.30)	0.88 (0.68-1.14)	0.94 (0.71-1.23)	0.92 (0.70-1.23)	1.84 (1.08-3.19) *	2.22 (1.27-3.98) **

*** < 0.001; ** <0.01; * <0.05; . <0.1

ELSA: English Longitudinal Study of Ageing. CES-D: Center for Epidemiological Studies Depression. Chi-sq results are p-values. Logistic regression results are coefficients (0.95 CI).

6.4.3 Construct validity: day centre services regression

We include the results of the day centre services regression in Table 6.6. Accounting for the number of notes and demographic factors, the model output remains a strong predictor of whether an individual is in receipt of day centre services. The maximum GVIF for any indicator is less than 1.4.

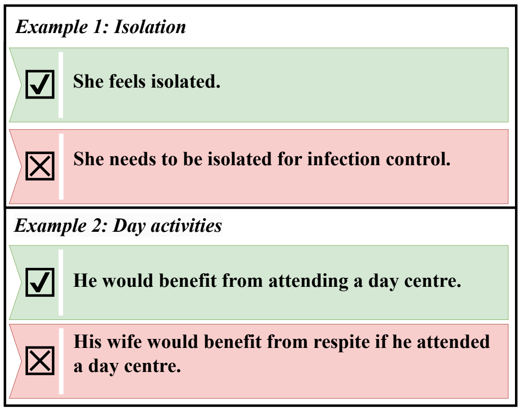
Table 6.6: Logistic regression: day care within 90 days

	Odds ratio (RoBERTa model)			
	Assessment	Both	Either	Notes
Lonely/isolated (prediction)	5.74 (3.02-11.87, p<0.001) ***	8.35 (4.57-16.09, p<0.001) ***	9.65 (3.47-40.16, p<0.001) ***	8.11 (3.80-20.11, p<0.001) ***
N notes	1.00 (1.00-1.00, p=0.319)	1.00 (1.00-1.00, p=0.080) .	1.00 (1.00-1.00, p=0.118)	1.00 (1.00-1.00, p=0.035) *
Sex: Male	1.12 (0.64-1.94, p=0.681)	1.08 (0.61-1.90, p=0.777)	1.01 (0.58-1.74, p=0.969)	0.95 (0.54-1.65, p=0.869)
Ethnicity: White	1.21 (0.69-2.17, p=0.516)	1.24 (0.70-2.27, p=0.464)	1.21 (0.69-2.16, p=0.516)	1.23 (0.70-2.22, p=0.476)
Age	0.98 (0.95-1.02, p=0.393)	0.98 (0.95-1.02, p=0.369)	0.98 (0.95-1.02, p=0.351)	0.98 (0.95-1.02, p=0.358)
Personal care score	0.68 (0.53-0.85, p=0.001) ***	0.74 (0.58-0.94, p=0.014) *	0.66 (0.52-0.83, p=0.001) ***	0.69 (0.54-0.87, p=0.002) **
Memory score	1.82 (1.41-2.36, p<0.001) ***	1.77 (1.36-2.32, p<0.001) ***	1.83 (1.43-2.37, p<0.001) ***	1.83 (1.42-2.37, p<0.001) ***
Safety & risk score	0.99 (0.76-1.27, p=0.911)	1.00 (0.78-1.30, p=0.978)	1.00 (0.78-1.29, p=0.993)	1.02 (0.79-1.31, p=0.893)
Lives alone	0.33 (0.18-0.59, p<0.001) ***	0.33 (0.18-0.60, p<0.001) ***	0.34 (0.18-0.60, p<0.001) ***	0.34 (0.19-0.62, p=0.001) ***

Note:
*** < 0.001; ** <0.01; * <0.05; . <0.1

6.5 Discussion

The goal of this analysis was to extract an indicator of loneliness or social isolation from free text. The key finding is that we can create such an indicator with a high F_1 score (0.92) using a RoBERTa-based discriminative large language model. Simpler NLP models do not perform nearly as well. Our findings indicate that for extracting this indicator from long-term care free text notes, as we have seen in other domains, transformer-based approaches can significantly improve on more traditional count-based or pre-trained embedding methods. Transformer models are complex and the mechanism by which they outperform other methods can be difficult to understand. However, we present example sentences in Figure 6.5 which are indicative of the types of notes which appear frequently. Among the word representations and classification models explored, only the transformer models were able to correctly classify all four sentences. Adult social care records are complex, unstructured descriptions about individual lives containing context-dependent distinctions between similar words. The attention mechanism in the transformer architecture may create a sentence encoding which is able to more accurately reflect this complexity.



Example 1: Isolation

☑ She feels isolated.

☒ She needs to be isolated for infection control.

Example 2: Day activities

☑ He would benefit from attending a day centre.

☒ His wife would benefit from respite if he attended a day centre.

Figure 6.5: Polysemy in adult social care case notes

Our method also examines the validity of the best-performing model when applied to the initial assessment of all individuals in our data. We find that the indicator of loneliness or isolation is a strong predictor of whether an individual receives services for social inclusion. Furthermore, each of our four measures has a significant association with whether an individual lives alone, a pattern also observed in survey data. There are, however, some differences with survey data. For example, although whether individuals live alone or with others is significantly associated with loneliness or isolation by our measure and in ELSA, there is a greater difference between the groups in ELSA. It is important to note, however, that one of the four dimensions of the established ELSA SNI indicator [356] is whether an individual is married. In the ELSA cohort of individuals aged 65+ receiving publicly funded care, 89% - 93% of unmarried individuals live alone (depending on the wave), so we would expect a measure of social isolation based on marital status to be more strongly associated with living alone than a measure which is based on free text.

Another notable difference between our results and ELSA is that our data does not show a significant association between loneliness or isolation and needing help with shopping or meal preparation, whereas ELSA does. On the other hand, our model identifies a significant association between loneliness or isolation and memory issues that is not present in ELSA. We have taken a subset of individuals from ELSA who are older people receiving local authority care, yet in general individuals in the administrative data appear to have higher needs than those in ELSA. For example, 60% of individuals in the administrative dataset have impaired memory or cognition, compared with 12% who report this is the case in the ELSA data. This is set out in Table 6.7. One explanation for this is that people with the highest needs might drop out of surveys. ELSA is known to have a high attrition rate relative to other surveys and it may be the case that the individuals with the greatest health needs are the least likely to continue [34].

Another explanation may lie in the differences between self-reported data in ELSA and professional assessments in administrative records. Individuals receiving publicly funded care in ELSA report appear to have approximately similar demographic characteristics to those in administrative data, such as gender and living alone. However, ELSA respondents report less functional impairment. One hypothesis is that people may be reluctant to discuss personal care needs with an interviewer. Individuals may under-report functional needs intentionally due to social desirability bias [298] or unintentionally due to impaired cognition [106]. Comparisons of self-reported and actual functional ability in surveys has found that the correlation can be as low as 0.2, and people with impaired mobility often report that they are mobile [138, 299].

The finding in ELSA that individuals receiving unpaid care are more likely to report loneliness, which we do not observe in our data, might also be explained by these differences. In ELSA, the self-reported data may not fully capture physical disability, so receiving unpaid care could serve as a proxy for need. Even after removing individuals in ELSA who report no care needs, we still find that the overall care needs in ELSA are lower than those in administrative data. Given the small sample size, we decided to retain all older individuals who report receiving publicly funded care in the analysis, but we interpret the results with caution, recognizing the differences between self-reported needs and professional assessments. This reflects one of the challenges of measuring construct validity using self-reported measures compared with professional assessment. However, it is reassuring that validity is reinforced in indicators where self-reporting is less likely to deviate from professional assessment, such as gender, where our indicator and survey data find that women experience slightly higher levels of loneliness than men.

An interesting aspect of the comparison is that in the administrative data we also find that loneliness or isolation occurs less in individuals who have higher physical care needs. There is no single physical ADL which has a negative association across all four measures, but there remains a trend that requiring more physical support is associated with less loneliness. It may be that there is a real effect here that would be observed if there was a larger sample size in ELSA: the dressing coefficient is negative for both ELSA measures, though not significant.

Table 6.7: Comparison of demographic and ADL needs between ELSA waves 6-9 and administrative data

	Administrative		ELSA	
	N (%)	N Unique	N (%)	N Unique
Ethnicity (non-white)	364 (33%)	364	18 (4%)	15
Toileting (requires support)	570 (52%)	570	160 (38%)	132
Lives alone	608 (55%)	608	255 (60%)	203
Memory (has needs)	664 (60%)	664	50 (12%)	49
Sex (F)	686 (62%)	686	267 (63%)	208
Awareness of risk (impaired)	806 (73%)	806	90 (21%)	81
Unpaid care (receives)	819 (74%)	819	344 (81%)	277
Dressing (requires support)	877 (80%)	877	293 (69%)	233
Meals (requires support)	998 (91%)	998	276 (65%)	219
Shopping (requires support)	1066 (97%)	1066	336 (79%)	272

ELSA: English Longitudinal Study of Ageing waves 6-9 (limited to the subset of individuals who report they receive statutory care). N Unique: number of unique individuals (as data is pooled). Administrative values are recorded by care managers in structured data. ELSA values are from the variables: raracem, toilta, hhres, slfmem, sex, dangera, rcaany_e, dressing, mealsa, shopa.

An alternative explanation for this could be to do with how workers prioritise the needs they record. One explanation is that if a person has very high physical care needs, a worker might be less concerned about loneliness in the face of more obvious risks, such as falls, pressure ulcers or care home entry. This represents the corollary to the advantage of administrative records being recorded by workers: if workers do not record loneliness where other needs are considered more urgent, a perfect free text classifier cannot produce a true prevalence. Finally, ELSA does not show a significant association between those individuals with memory problems and those who are lonely or socially isolated, by either the UCLA/SNI or CES-D/SNI measures. Although there is a disparity between the administrative data and ELSA results, we find the association in the administrative data results reassuring, as the literature suggests there is an association between memory and loneliness [see e.g. 361, 362].

The comparison with ELSA is challenging to interpret, owing to the apparent differences in the population and that both care needs and loneliness are self-reported in ELSA but not in administrative data. We are therefore reassured by the results in Table 6.6 of the probability of receipt of day centre services within 90 days of the first assessment. It is clear that the indicator of loneliness or isolation is a strong and significant predictor of whether an individual receives services for social inclusion. This holds when controlling for the number of notes and demographic factors, suggesting that our indicator is picking up a distinct phenomenon. It also leads to the reassuring conclusion that workers who record that a person is lonely or isolated are much likelier to put in place services for this need.

One aspect of the interpretation of the model output is establishing which of the four metrics is best (*Assessment*, *Case notes*, *Either* or *Both*). The χ^2 tests and regression indicate they are generally associated with the same demographic and needs-related factors. This may be due to our binary measurement of loneliness and social isolation which, if it could be measured more properly, may be better captured as a continuous variable. Individuals can experience different intensities of loneliness. One interpretation is to consider each of the measures as a reflection of loneliness or isolation above a certain threshold. The measures from *Assessment* and *Case notes* of around 43% measure each are approximately equivalent to the prevalence of individuals who are SNI isolated or have a UCLA loneliness score of 6 out of 9 or above (45%). Alternatively, using the *Both* metric to define individuals as lonely or isolated only if they have text indicating this is the case in an assessment and a case note, may identify individuals with a higher intensity of loneliness than either of these measures. This captures 26% of individuals, which is the same as the 26% of ELSA individuals who are SNI isolated, or score 9 out of 9 in the UCLA loneliness scale. Finally, defining an individual as lonely using the *Either* measure, if they have a sentence in an assessment or a case note indicating this, gives a prevalence of 62%. This is a comparable prevalence to individuals in ELSA who are SNI isolated or lonely if they have a UCLA loneliness score of between 4 (71%) or 5 (58%) and above. Which of the four derived measures is most appropriate will ultimately depend on the policy goal. If the policy goal is to target only those individuals with the highest need, the *Both* measure may be most appropriate. Conversely, an area wishing to put into place preventative services may wish to cast a wider net with the *Either* metric. Whether proportions similar to these hold in other areas is an empirical question that can be established by applying this model to additional datasets.

6.5.1 Limitations

There are some limitations to these findings. We used only one set of pre-trained embeddings and applied mean pooling of each token to represent the sentence. While effective, this approach could potentially be improved by using methods that summarise spans of individual sentence embeddings, which have been shown to enhance performance in similar tasks [185]. Similarly, for the count-based approaches, instead of representing each sentence as a count of all words, we could have employed co-occurrence matrices of n-grams — capturing words that appear within n words of each other — which might better capture contextual relationships. It is plausible that these and other enhancements could have increased the F_1 score of the simpler NLP methods. Nonetheless, we ensured a robust evaluation of the simpler methods by applying a range of classifiers, including boosting, bagging, logistic regression, multi-layer perceptron (MLP), and random forests. Moreover, with the transformer models, we achieved strong results using the default parameters and did not find it necessary to fine-tune hyperparameters, such as the learning rate. This suggests that while the gap could potentially be narrowed, it could also be widened again through further optimisation of the transformer model.

Additionally, our construction of loneliness or social isolation as a binary (rather than continuous) variable is an oversimplification. It is, for example, possible for two individuals to be lonely, but one to be more so. We were limited by the data, as intensity of loneliness or isolation is rarely recorded, and we took a pragmatic approach. However, the findings do indicate that different measures derived from the binary indicator may provide some indication of intensity. A related limitation is that our model combines loneliness and social isolation. While this is consistent with similar research [e.g. 152], such distinctions can guide appropriate interventions. For example, social inclusion services such as day centres increase social contact so are by definition effective at reducing social isolation. They do not necessarily reduce loneliness or emotional isolation [as conceptualised in e.g. 363]. We do not think it would have been possible to distinguish between loneliness and isolation through secondary analysis of administrative care records. However, if the relative impact of the components of loneliness or isolation can be derived from other data, it will guide appropriate interventions.

Another notable limitation is related to the dataset which, although large in terms of sentence count, is limited to a relatively small geographic area. Although notes in the training and test set are not about the same individual, they may have been written by the same worker. Similarly, there may be organisational culture issues which lead to individuals using similar phrases that would not be seen elsewhere. We expect that the model will not perform quite as well on free text case notes from another area, although the magnitude of the drop-off, and how many new samples need to be labelled to improve performance, is an empirical question that we hope to answer in the future.

There is an inherent limitation in any study that uses secondary, administrative data to measure a phenomenon. We see this when interpreting the output in Table 6.6. It is possible to say that workers who record that an individual is lonely or isolated are more likely to put in place a social inclusion service for that person. However, we cannot confidently say how often workers put in place services for individuals who are actually lonely or isolated. There is likely a degree of unobserved or unrecorded loneliness or isolation which cannot be quantified by any research based on administrative records. Furthermore, there will be instances where an individual is offered a service for perceived loneliness and declines it. However, as we observe the expected association between our indicator and characteristics such as impaired memory and living alone, it suggests that the unobserved loneliness or isolation is not so significant as to invalidate our results. Furthermore, administrative records contain real-time information about service use for individuals receiving publicly funded care. Administrative data has enabled researchers to establish that the probability of care home admission is associated with age, gender, disability, ethnicity and depression [e.g. 364, 365]. However, whether an individual is lonely or socially isolated is not generally recorded as a structured indicator in administrative data, so it has not been included in such models.

6.5.2 Conclusion

Our best-performing model demonstrates an approach to identifying social isolation and loneliness in long-term care case notes with an F_1 score of 0.92 on unseen text in the test set. The measure seems valid. It is highly correlated with living alone, which we see in ELSA, and with impaired memory which we would expect from the literature. Furthermore, it is a strong predictor of whether an individual receives services for social inclusion. Around 43% of individuals have a care needs assessment with a sentence indicating loneliness or social isolation, 44% have a case note, 62% have either and 26% have both. Similar estimates of prevalence at different thresholds of loneliness are obtainable from individual-level survey data. The advantage of extracting an indicator from administrative data is that it contains many more individuals than survey data, providing enough statistical power to compare between subgroups on the basis of demographic or needs-related factors. Additionally, administrative data includes individual-level, time-variant service cost data. Extracting information from administrative records provides the opportunity to examine associations with service use and cost data that is not contained in surveys. Finally, there are relatively few survey respondents with high needs or in care homes [366], but such individuals are represented in administrative data.

It would be useful for future research to include the outputs of a predictive model from free text as an input to a regression model, to establish whether there are differences in the intensity of long-term care usage by individuals who have been identified as being lonely or socially isolated. For example, this could be using information in administrative free text data to model risk of care home entry. Our model provides the opportunity to conduct such analysis. We also hope that evaluating the performance of different machine learning approaches in this paper will inform future researchers about methods for extracting other important characteristics recorded in free text but not in structured data, such as economic hardship, psychological wellbeing or risk of abuse. The results of this analysis are published with an open source version of our model, and we hope that others can use our classification model with their own data.

7 Loneliness as a risk factor for time to care home entry for older adults receiving community care

Abstract

Background and Objectives: International efforts to contain long-term care costs have prioritised personal care provision. However, reductions in services aimed at addressing loneliness or promoting social participation may affect demand for long-term care facilities. Research on the impact of loneliness on entry to residential or nursing care facilities is based on survey data, which under-represents those with the highest needs. Administrative records include such individuals and, unlike surveys, contain continuous data on service receipt, enabling accurate modelling of time to care home entry.

Research Design and Methods: We use administrative data for 1,101 individuals receiving care in a London local authority. We extract loneliness from free text notes using a large language model and model its impact on care home entry five years after assessment, controlling for needs and demographics. We use logistic regression and a competing risks survival model to measure time until care home entry.

Results: The odds ratio for care home entry associated with loneliness is 1.45 with logistic regression (95% CI 1.04 – 2.01). The hazard ratio is 1.32 (95% CI 1.01 – 1.72) with a cause-specific model, and 1.39 (95% CI 1.08 - 1.79) using the Fine & Gray method. Among those most likely to enter a care home, the median time to entry is around 9 months (95% CI 228 - 328 days) earlier for those who are lonely.

Discussion and Implications: The hazard ratio of loneliness on care home entry is around the magnitude associated with gender, ethnicity or living alone. However, loneliness is modifiable. Reductions to services which reduce loneliness, such as day centres, are likely to cause an increase in loneliness. We demonstrate that for those with the highest needs, loneliness is a significant risk factor for time until

care home entry. Policymakers seeking to delay care home entry should consider the impact on services for loneliness.

7.1 Background and Objectives

Demand for long-term care (LTC) services is increasing internationally, and this trend is projected to continue [see e.g. 4, 367, 368]. While countries such as Japan and Norway are exceptions, the increased pressure on public funds has led to a "general tendency toward reconsidering and tightening the eligibility criteria for access to public LTC services" [369]. For instance, Finland, Denmark, the UK, Sweden, and the Netherlands have recently experienced a retrenchment in community care services towards personal care and nursing tasks, with domestic and social support implicitly shifted to informal care networks [see 5, 263, 264, 6].

While these policy changes aim to target resources towards those with the greatest needs, they may have unintended consequences. Services for promoting social participation such as day centres can lead to a reduction in loneliness [265], a factor associated with higher risk of care home admission for older people observed in surveys [266]. Yet there has been little analysis of the impact on care home entry of reducing the supply of such services to those with the highest needs, and particularly for individuals receiving publicly funded care.

In this paper, we explore loneliness and social isolation as risk factors for care home entry among older publicly funded social care users, using English administrative data. Moving to a residential or nursing care home is generally considered undesirable, as it is associated with a loss of independence, dignity, and privacy and has high costs [370, 371]. In England, for example, the total public expenditure on long-term care services in 2022/2023 was £15.1 billion, of which £6.6 billion funded care home places [322]. Several studies have found that loneliness is associated with the risk of care home entry for older adults [267, 372, 373, 374, 375]. If increased loneliness affects the risk of care home entry for publicly funded care users, this should be considered when calculating the effects of retrenchment of long-term care towards personal care.

However, most studies investigating the effect of loneliness use population survey data [e.g. 267, 372, 373, 374]. Due to means-testing, retrenchment and tightening eligibility criteria, statutory care users have a different needs profile to the general population, being more disabled, economically deprived, and reliant on formal care services. Such individuals are often under-represented in surveys due to systematic exclusions [104, 105] and attrition [32, 34, 295, 103, 324]. In the English case, there are marked differences in reported levels of need among older adults in the English Longitudinal Study of Ageing (ELSA) [109] who state they receive publicly funded care compared with administrative data. While proportions of demographic information such as age and gender are similar, around half as many state they require support with personal care tasks (see Table 7.1). We cannot be confident that the findings from surveys that loneliness is a risk factor for care home entry for older people can be generalised to those with high levels of need — i.e. those at greatest risk of care home entry — who are under-represented in surveys.

This paper examines whether loneliness or social isolation recorded at the time of a person's initial assessment affects time until an individual enters a care home, controlling for needs and demographic factors. Loneliness and social isolation are closely related but distinct concepts [333, 330, 331]. Social isolation generally refers to an "objective lack of relationships", while loneliness is a "subjective, distressing feeling" that arises when an individual's desired quantity or quality of social connections is unmet [332]. These concepts often overlap in their effects, as both loneliness and social isolation have been linked to adverse outcomes, including increased mortality among older adults [328]. In this paper, we recognise the challenges in distinguishing loneliness from social isolation within free text social care records. Social workers may use terms like "lonely" or "isolated" interchangeably or imprecisely, describing either a subjective sense of loneliness or a more objective lack of social contact, or both. We adopt a pragmatic approach by considering loneliness and social isolation together, consistent with the integrated perspective seen in public health research [e.g. 340, 327, 328, 329, 341]. For the sake of brevity, except where explicitly distinguished from social isolation, we use the term "loneliness" in this paper to refer to our combined measure of loneliness or social isolation.

Our analysis differs from previous research in that we use administrative records. These records, collected by agencies in the course of service delivery, contain real-time information about service use for individuals receiving publicly funded care, allowing us to model time to care home entry. Administrative data has enabled researchers to establish that the probability of care home entry is associated with age, gender, disability, ethnicity and depression [e.g. 364]. However, loneliness is not generally recorded as a structured indicator in administrative data and it has not been included in analyses using administrative records, or record-linkage models of socio-demographic variations in moves to care homes [e.g. 376]. One study, which linked administrative records to data collected in research interviews in a subset of care users in a local authority over a six month period between 1998 and 2000, found that social isolation increases risk of care home entry in statutory care users [375]. Our paper extends this analysis, using a large language model (LLM) to identify loneliness and conduct an analysis of its effect on care home entry with a local area's population of long-term care users over five years, between 2015 and 2020, and accounting for the competing risk of death prior to care home entry.

Using records collated by agencies in the process of service delivery offers advantages. There is no attrition as health declines, so the records capture information on those with the highest need. These records contain continuous, time-variant service use data, allowing more precise estimation of the time until care home entry. Surveys provide snapshots at each wave. Researchers can impute the date of institutionalisation between waves but this increases uncertainty [372, 373].

It is essential to account for factors associated with both loneliness and care home entry to robustly examine their association. Care home entry is correlated with age, gender, ethnicity, functional impairment and living alone [377, 266, 378, 374], though this is moderated by receipt of unpaid care [373]. Dementia, which has a bidirectional relationship with social participation [379, 380], is a highly significant predictor of care home entry [373, 267]. Our analysis controls for age, sex, ethnicity, cognitive impairment, support required with activities of daily living (ADL) needs (personal care) and instrumental activities of daily living (IADL) needs (shopping and meal preparation), and living circumstances (receipt of unpaid care, living alone). Receipt of formal or unpaid care may affect risk of care home entry [373]. Survey-based research into the effect of loneliness on care home entry [e.g. 267, 372, 373, 374] has not controlled for care receipt. By including these covariates, we aim to isolate the effect of loneliness and ensure it is not conflated with other factors that influence the risk of care home entry. As a baseline model, we replicate the approach in Hanratty et al. [267], using logistic regression. To compare differences in the rate of care home entry over time between those identified as lonely or isolated and others, we use a survival model, accounting for competing risks, as not all (or even most) individuals will ever enter a care home [381].

We investigate three questions: Firstly, does loneliness predict care home entry? Secondly, holding other factors equal, what is the difference in time to care home entry if an individual is lonely? Finally, does loneliness particularly increase risk of care home entry in certain groups, such as women, those who are more physically disabled, or those with impaired cognition? This final question is important for understanding how best to target services to prevent care home entry.

7.2 Research Design and Methods

7.2.1 Dataset

To explore the relationship between loneliness and care home entry, we use data from an inner London local authority. In England, under the Care Act 2014, every person seeking publicly funded care undergoes an assessment by a social care professional to establish their eligibility. This assessment generates a document recording information relevant to care needs such as functional ability to perform ADLs and IADLs, cognitive function, and unpaid care.

We received approval from the NHS Confidentiality Advisory Group (CAG) to use this data for this purpose, and obtained ethical approval. A query was written to identify all individuals aged 65 or over on August 1st 2020 who had been receiving services arranged by the local authority for at least one year at some point since 1st January 2016. The dataset includes information for 3,046 individuals between 1st January 2015 and August 31st 2020. The export includes all needs assessment forms completed between January 2015 and August 2020. After an assessment is completed, if an individual receives care, services commissioned are recorded. The export includes individual-level, time-variant service use data with costs between January 2015 and August 2020.

A complexity in the data is that loneliness can be recorded at any time during an individual's contact with care services. We resolve this by using loneliness at the time of initial needs assessment. There are two reasons for choosing this time point. Firstly, we expect needs to be comprehensively recorded at first contact. Secondly, if loneliness at first presentation to a local authority is a relevant factor for care home entry, it provides the greatest opportunity for intervention. There are 1649 individuals in the exported data whose initial needs assessment occurs in the period of observation. Needs assessments were captured on different forms during the period. This was determined by policy changes and is not correlated with individual need. After limiting the dataset to forms that contained questions covering all relevant covariates, 1331 individuals remain. We also exclude from the analysis individuals who enter care homes immediately after their first presentation to social care. These cases represent an event leading to a sudden development of care needs, such as a fall or stroke. There is no opportunity in such cases for local authorities to put in place preventative services for loneliness with a view to delaying care home entry. We limit the period of observation to 5 years from initial assessment. Our final dataset contains 1101 individuals. At the end of the period of observation, 252 people have entered a care home, 502 die before entering a care home and 347 are censored, i.e. continue to receive care in the community at the end of the period of observation.

Table 7.1: Comparison of demographic and ADL needs with ELSA

	Administrative data		ELSA			
			Unweighted		Weighted	
	N (%)	N Uniq	N (%)	N Uniq	%	0.95 CI
Awareness of risk (impaired)	806 (73%)	806	90 (21%)	81	20.5%	(16.1%, 24.9%)
Dressing (requires support)	877 (80%)	877	293 (69%)	233	68.9%	(63.5%, 74.3%)
Ethnicity (non-white)	364 (33%)	364	18 (4%)	15	4.2%	(1.8%, 6.6%)
Lives alone	608 (55%)	608	255 (60%)	203	61.2%	(55.1%, 67.3%)
Meals (requires support)	998 (91%)	998	276 (65%)	219	65.5%	(60.3%, 70.8%)
Memory (has needs)	664 (60%)	664	50 (12%)	49	24.8%	(18.0%, 31.7%)
Sex (F)	686 (62%)	686	267 (63%)	208	65.2%	(59.3%, 71.2%)
Shopping (requires support)	1066 (97%)	1066	336 (79%)	272	81.1%	(77.3%, 85.0%)
Toileting (requires support)	570 (52%)	570	160 (38%)	132	35.3%	(30.0%, 40.5%)
Unpaid care (receives)	819 (74%)	819	344 (81%)	277	82.8%	(78.7%, 86.9%)

ELSA: English Longitudinal Study of Ageing. Using pooled data from waves 6,7,8 and 9 of ELSA. N Uniq: total unique individuals across all waves. Weighted: using longitudinal weights provided with ELSA.

7.2.2 Characteristics of individuals in the data

Information was captured using structured data fields and free text case notes. Structured fields are inherently machine-readable [117]. In our dataset, they record key demographic and personal information necessary for care planning and service delivery, such as age, gender, ethnicity, functional ability with ADLs and IADLs, and whether the individual lives alone. Free text fields can be included within needs assessment forms, or in distinct areas of case management systems to record information not covered elsewhere (”case notes”). In this study, we extract the loneliness measure from free text, and all needs-related covariates from structured data. We classified for loneliness or social isolation all free text notes recorded about the 1,101 individuals within 90 days of their initial assessment ($N = 62603$). We present in Table 7.2 a breakdown of loneliness and care home entry by each covariate, including the p value for tests of independence, for

Table 7.2: Descriptive statistics

		Care home entry				Loneliness or isolation		
Variable	Levels	Censored (%)	Care home (%)	Death (%)	p	Not lonely (%)	Lonely (%)	p
Lonely	No	258 (32.5)	152 (19.2)	383 (48.3)	<0.001	793 (100)	0 (0)	
	Yes	89 (28.9)	100 (32.5)	119 (38.6)		0 (0)	308 (100)	
Sex	Female	235 (34.3)	149 (21.7)	302 (44.0)	0.041	488 (71.1)	198 (28.9)	0.438
	Male	112 (27.0)	103 (24.8)	200 (48.2)		305 (73.5)	110 (26.5)	
Personal Care	Low/no needs	157 (35.4)	113 (25.5)	174 (39.2)	0.005	301 (67.8)	143 (32.2)	0.001
	Moderate	150 (30.5)	100 (20.4)	241 (49.1)		355 (72.3)	136 (27.7)	
	High	40 (24.1)	39 (23.5)	87 (52.4)		137 (82.5)	29 (17.5)	
Cognition	No/low needs	245 (35.3)	115 (16.6)	334 (48.1)	<0.001	539 (77.7)	155 (22.3)	<0.001
	Moderate	52 (26.8)	63 (32.5)	79 (40.7)		120 (61.9)	74 (38.1)	
	High	50 (23.5)	74 (34.7)	89 (41.8)		134 (62.9)	79 (37.1)	
Ethnicity	Non-white	128 (35.2)	72 (19.8)	164 (45.1)	0.099	266 (73.1)	98 (26.9)	0.635
	White	219 (29.7)	180 (24.4)	338 (45.9)		527 (71.5)	210 (28.5)	
Shopping/Meals	Low/no needs	56 (39.4)	38 (26.8)	48 (33.8)	0.003	96 (67.6)	46 (32.4)	0.412
	Moderate	129 (33.2)	95 (24.5)	164 (42.3)		279 (71.9)	109 (28.1)	
	High	162 (28.4)	119 (20.8)	290 (50.8)		418 (73.2)	153 (26.8)	
Lives Alone	No	162 (32.9)	94 (19.1)	237 (48.1)	0.025	371 (75.3)	122 (24.7)	0.037
	Yes	185 (30.4)	158 (26.0)	265 (43.6)		422 (69.4)	186 (30.6)	
Unpaid Care	No	107 (37.9)	71 (25.2)	104 (36.9)	0.002	207 (73.4)	75 (26.6)	0.602
	Yes	240 (29.3)	181 (22.1)	398 (48.6)		586 (71.6)	233 (28.4)	
Has Telecare	No	280 (33.3)	193 (22.9)	369 (43.8)	0.053	618 (73.4)	224 (26.6)	0.080
	Yes	67 (25.9)	59 (22.8)	133 (51.4)		175 (67.6)	84 (32.4)	
N Notes	Mean (SD)	513.4 (457.5)	691.6 (513.0)	522.9 (412.1)	<0.001	531.7 (451.7)	627.4 (462.7)	0.002
Age	Mean (SD)	83.2 (8.6)	85.9 (7.4)	84.4 (8.3)	<0.001	84.0 (8.4)	85.4 (7.6)	0.011
Cost DPs	Mean (SD)	9.1 (39.6)	2.6 (22.7)	9.7 (68.8)	0.191	7.3 (45.2)	9.3 (68.2)	0.575
Cost Daycare	Mean (SD)	5.9 (34.2)	8.1 (34.6)	2.8 (15.7)	0.032	2.0 (21.5)	12.6 (38.0)	<0.001
Cost Homecare	Mean (SD)	118.5 (124.3)	125.1 (153.0)	125.6 (138.9)	0.742	128.5 (142.7)	109.6 (123.6)	0.041

Costs are the cost of services put in place within 90 days of initial assessment.

DPs are direct payments and day care represents services for social participation (day centres).

categorical variables using a χ^2 test and for continuous variables $Pr(> F)$ after an analysis of variance. We explore these relationships and contrast them with the regression output in the Discussion section.

7.2.3 Model parameters

Loneliness is extracted from free text as described in Chapter 6. The natural language processing model produces a binary classification for each sentence, indicating whether loneliness or social isolation is recorded. We consider an individual to be lonely or isolated at the time of assessment if they have at least one sentence in their needs assessment form and one sentence in case notes which is indicative of loneliness or social isolation. As information about social networks is extracted from free text records, which do not reliably distinguish between the related but distinct concepts of loneliness and social isolation, our indicator reflects individuals who are recorded as being either lonely or socially isolated.

As loneliness is based on free text, we also include in the model the number of case note sentences recorded about an individual (N Notes). This means any association between loneliness and care home entry cannot be explained by the natural language processing model being more likely to indicate loneliness where more case notes are recorded.

We also include in the model services received after the initial needs assessment, to control for the effect of service receipt and capture differences in need not reflected in service provision. We include the cost of home care, day care and direct payments (DPs), as well as whether individuals receive telecare services. All covariates except loneliness are extracted from structured data. We limit the period of observation to 5 years after initial assessment.

7.2.4 Models

We use logistic regression as a baseline model, and to compare results for statutory care users against the general older population in Hanratty et al. [267]. However, logistic regression does not distinguish between an individual who enters a care home after one day and another who enters two years later, though this difference may be meaningful for those individuals. We also use a survival model with competing risks, to allow us to model the length of time that an individual spends outside a care home. The competing risks element of the model accounts for the fact that, unlike traditional survival models where the event of interest is death, not all individuals will enter a care home.

7.2.4.1 Logistic regression model

We use a logistic regression model, modelling care home entry as 1 ($N = 252$), and not entering a care home as 0 ($N = 849$), as specified in Equation (7.1).

$$Pr(y = 1|\text{lonely}, \mathbf{X}) = \frac{e^{\beta_0 + \beta_1 \text{lonely} + \beta \mathbf{X}}}{1 + e^{\beta_0 + \beta_1 \text{lonely} + \beta \mathbf{X}}} \tag{7.1}$$

Where *lonely* is a binary variable indicating whether an individual was lonely at the time of initial assessment, and $\mathbf{X} = (X_2, X_3, ... X_p)$ is a vector of the explanatory variables set out in Table 7.2.

7.2.4.2 Survival model with competing risks

A survival model is a method to account for differences in time between the initial assessment and care home entry. However, care home entry (N=252) is not the only possible outcome. It is also possible for individuals who are at high risk of care home entry to die prior to entering a care home (N=502), as well as to be censored, i.e. remain in the community at the end of the period of observation (N=347). We therefore use a competing risks model. We use the Aalen–Johansen estimator, a generalisation of the Kaplan–Meier approach [382] to estimate a cause-specific hazard function [383]. Critics of the cause-specific estimator note that as individuals who die prior to the outcome of interests (in this case care home entry) are removed from the risk set, it can fail to capture the risk in a population despite accurately reflecting the sample, as it is not known in advance when individuals at risk will die [383]. We therefore also estimate the subdistribution hazard using a Fine & Gray competing risks model, where the hazard function is defined as in Fine and Gray [384]. We fit two models to estimate the respective hazard ratios as specified in Equation (7.2).

$$h_{k,j}(t|\text{lonely}, \mathbf{X}) = h_{0_{j,k}} \cdot e^{\beta_{1_{j,k}}\text{lonely}} \cdot e^{\beta_{j,k}\mathbf{X}} \tag{7.2}$$

for $k \in \{1, 2\}$ and $j \in \{1, 2\}$, where
$k : 1 = \text{care home}, 2 = \text{death}$
$j : 1 = \text{cause-specific}, 2 = \text{subdistribution}$
lonely is a binary variable indicating whether an individual was lonely at the time of initial assessment
$\mathbf{X} = (X_2, X_3, \ldots X_p)$, a vector of the explanatory variables set out in Table 7.2.

The advantage of the Fine & Gray approach is that it includes in the risk set individuals who enter another state (i.e. we model the risk for individuals who die before time t, reflecting that in the population we do not know which individuals will die). Proponents of the cause-specific approach argue that the Fine & Gray approach can be difficult to interpret as it uses a risk set which does not exist [385]. We subscribe to the view in Austin et al. [2016] that cause-specific models are appropriate for interpreting individual covariates, and Fine & Gray is suitable for predicting the risk for individuals with different combinations of needs through the subdistribution hazard function. We present results for both models but prefer the cause-specific hazard estimates, and use the Fine & Gray model for generating predictions for sub-groups of individuals.

Both models multiply a base hazard rate by a vector of covariates, so assume proportional hazards. As this assumption was not satisfied for cognition or the number of sentences written at the time of the assessment (N notes), we stratified by these variables to avoid violating it. The number of sentences (n) is a continuous variable so to stratify we split it into low ($n < 440$), medium ($440 \geq n < 1000$) and high ($n \geq 1000$) to satisfy the assumption. In the Fine & Gray model, which requires reshaping the data into counting process format, the proportional hazards assumption was also not satisfied for home care costs at initial assessment, so again we stratified weekly cost (c) in £ into three groups, low ($c < 50$), medium ($50 \geq c < 150$) and high ($c \geq 150$). After stratification the proportional hazards assumption was satisfied for all variables in the model.

We also conducted additional analyses to interrogate the effect of loneliness on the oldest old, the inclusion of living alone in the model, and the sensitivity of our model to inclusion of data from 2015:

1. Binary age specification: We specify age as a binary variable (< 85 vs. ≥ 85), replacing the continuous and quadratic age terms in the main model.
2. Stratified age specification: We stratify our dataset into the two age groups (< 85 vs. ≥ 85) and run the same models as in Equations (7.1) and (7.2) separately for each age group.
3. Exclusion of living alone. Recognising that loneliness and social isolation are partly a function of living alone, we ran the models in Equations (7.1) and (7.2) excluding this variable.
4. Exclusion of assessments in 2015: We included in the main analysis individuals whose initial assessment was in 2015. However, owing to the data selection query, while we have a record of many services received by this group, if individuals received a service ending prior to 2016, it would not have been included in our dataset. This missing data could introduce bias, as we include services received as a covariate. For robustness, we conduct a sensitivity analysis using only the 941 individuals with assessments from 1st January 2016 onwards.

We include the results for the additional analyses in the Appendix. All analysis was undertaken with R 4.2.2 [98], using the `survival` package [381] for the competing risks models.

7.3 Results

We present in Table 7.3 the output from the logistic regression model. Loneliness significantly (at $\alpha = 0.05$) increases the risk of care home entry, with an odds ratio of 1.45 (95% CI 1.04 – 2.01). We present in Figure 7.1 the cumulative incidence of care home entry for individuals who are and are not identified as lonely or isolated at the time of initial assessment. The plot does not control for confounding factors, and we present the results of the regression, adjusting for covariates, in Table 7.4. The magnitude of the effect is similar in the competing risks models, with the presence of loneliness increasing either the odds ratio or the instantaneous risk of care home entry in the range of 1.32 - 1.39. Loneliness remains significant after accounting for other factors with which it is associated. In particular, the effect of loneliness cannot be explained by living alone, receipt of unpaid care, cognition or functional ability, all of which were included in the model.

Table 7.3: Logistic regression model output

	Odds ratio	
Loneliness		
Lonely or Isolated	1.45 (1.04-2.01, p=0.027)	*
Demographics		
Sex: Male	1.27 (0.93-1.74, p=0.135)	
Age	1.37 (0.96-1.98, p=0.087)	.
Age^2	1.00 (1.00-1.00, p=0.134)	
Ethnicity: White	1.41 (1.01-1.98, p=0.046)	*
Lives Alone	1.63 (1.16-2.31, p=0.005)	**
Unpaid Care	0.79 (0.55-1.14, p=0.203)	
Needs		
N Notes	1.00 (1.00-1.00, p<0.001)	***
Personal care: Moderate	0.76 (0.52-1.11, p=0.161)	
Personal care: High	1.08 (0.62-1.89, p=0.783)	
Cognition: Moderate	2.76 (1.86-4.10, p<0.001)	***
Cognition: High	3.87 (2.57-5.86, p<0.001)	***
Shopping and Meals: Moderate	1.00 (0.62-1.64, p=0.989)	
Shopping and Meals: High	0.61 (0.35-1.06, p=0.075)	.
Services		
Cost DPs	1.00 (0.99-1.00, p=0.173)	
Cost Daycare	1.00 (1.00-1.01, p=0.321)	
Cost Homecare	1.00 (1.00-1.00, p=0.498)	
Has Telecare	0.92 (0.63-1.31, p=0.637)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

Table 7.4: Competing Risks model output

	Cause-specific hazard		Fine & Gray	
Loneliness				
Lonely or Isolated	1.32 (1.01-1.72, p=0.039)	*	1.39 (1.08-1.79, p=0.009)	**
Demographics				
Age	1.16 (0.84-1.60, p=0.369)		1.27 (0.96-1.68, p=0.093)	.
Age^2	1.00 (1.00-1.00, p=0.461)		1.00 (1.00-1.00, p=0.136)	
Ethnicity: White	1.42 (1.08-1.88, p=0.013)	*	1.30 (1.00-1.69, p=0.047)	*
Lives Alone	1.54 (1.16-2.03, p=0.003)	**	1.47 (1.14-1.91, p=0.003)	**
Sex: Male	1.34 (1.03-1.74, p=0.030)	*	1.15 (0.90-1.46, p=0.278)	
Unpaid Care	0.92 (0.68-1.24, p=0.584)		0.85 (0.65-1.12, p=0.248)	
Needs				
Personal care: High	1.39 (0.85-2.29, p=0.189)		1.08 (0.68-1.70, p=0.749)	
Personal care: Moderate	0.96 (0.69-1.32, p=0.786)		0.81 (0.60-1.10, p=0.174)	
Shopping and Meals: High	0.65 (0.42-1.01, p=0.056)	.	0.65 (0.43-0.99, p=0.043)	*
Shopping and Meals: Moderate	0.88 (0.60-1.29, p=0.511)		0.93 (0.65-1.33, p=0.683)	
Services				
Cost DPs	1.00 (0.99-1.00, p=0.091)	.	1.00 (0.99-1.00, p=0.119)	
Cost Daycare	1.00 (1.00-1.00, p=0.784)		1.00 (1.00-1.00, p=0.663)	
Cost Homecare	1.00 (1.00-1.00, p=0.687)			
Has Telecare	0.82 (0.60-1.11, p=0.204)		0.90 (0.68-1.19, p=0.457)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

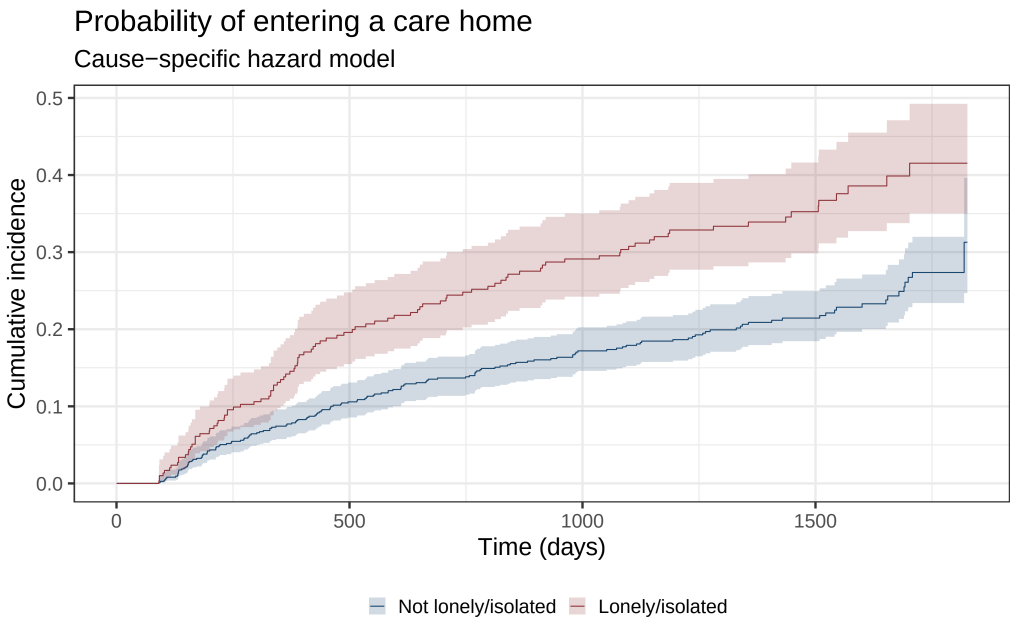


Figure 7.1: Cumulative incidence

The greatest predictor of care home entry in all models is cognition, which is consistent with the literature [e.g. 364]. As we have stratified the survival models by cognition, a coefficient for cognition is not estimated in these models. Cognition does not meet the assumption because individuals with significant cognitive impairment at initial assessment are likely to enter a care home during the first year, unlike those individuals with no cognitive impairment. We compare the cumulative incidence in the Appendix in Figure 11.5. Although stratification means the model does not produce a coefficient, the way in which this assumption is violated highlights the paramount importance of cognition in determining time until care home entry, particularly in the first year after assessment.

The estimates from the alternatively specified models are consistent with the main results. The model restricting the analysis to individuals whose assessment occurred on or after 1st January 2016, and the model excluding living alone as a covariate, do not meaningfully change the magnitude of the estimates of the effect of loneliness, and the p values remain at the same significance level. The competing risks models which treat age as a binary variable find that being aged > 85 is a significant predictor of care home entry, in contrast to age as a continuous variable in Table 7.4. The results for loneliness in this model are consistent with the main body of the paper, with very little change in the size of the hazard ratios or p values, indicating that loneliness is a robust predictor of care home entry regardless of how age or living alone is specified. We include the full results of these models in the Appendix.

We find similarly to Kersting [373] that using a model accounting for time to care home entry provides insight into the factors associated with care home entry. Gender is significant in the cause-specific hazard model, though not in the logistic regression. We see in Table 7.2 that 22% of women in the sample ultimately enter care homes, compared with 25% of men, and the χ^2 test indicates this difference is significant. However, Table 7.3 shows that, controlling for other covariates, in a logistic regression the difference disappears. Men generally tend to die earlier, and in our data are at risk of care home entry for 603 days, whereas women are at risk for 682 days. A one-sided t test indicates this is also a significant difference ($p < 0.001$). This difference means the yearly rate of care home entry for men is about 20% higher than women, impacting the significance in the cause-specific survival model. Conversely, in the Fine & Gray model, gender is not significant. This is because individuals who have died remain in the risk set for care home entry. As men also die earlier and more often than women, this increases the risk set for males entering care homes more than it does for women, so the overall difference in rate is diluted. An advantage of the Fine & Gray model is that it does not assume knowledge of the future, such as knowing who will die [384]. However, it is well-established that mortality rates for men are higher. While this remains the case, the cause-specific hazard model hazard ratio is a more appropriate measure, as it accounts for the fact that men on average spend less time at risk of care home entry. There is otherwise little difference between the output of the two competing risks models.

We similarly see that a survival model gives some insight into groups who may be more likely to enter home care homes quickly. IADLs (high support needs with shopping and meals) appear to be significant using Fine & Gray and not in the cause specific hazard model, but this is really a reflection of the decision to set $\alpha = 0.05$, as the p values are both very close to this, at 0.056 and 0.043 respectively. The hazard ratio is < 1, indicating individuals who live alone, are lonely, have a significant cognitive impairment and are independent with shopping and meal preparation are at the highest risk of care home entry. We hypothesise that such individuals may be felt to be at particular risk, for example of wandering, and that a model which accounts for time to care home entry can distinguish that such individuals enter care homes particularly quickly. However, this small group is on the boundary of significance in both and not a focus of this research. More research would be required to definitively identify this phenomenon, ideally with a larger sample to allow for the introduction of interaction effects.

To measure the effect size of loneliness, we created two dummy datasets, both based on our original data and identical in number of individuals and all characteristics, except in one dataset all records were marked as lonely, and in the other none were. We generated survival curves for time to care home entry using the subdistribution hazard from the Fine & Gray model. We present in Figure 7.2 the difference in mean survival times by group. While loneliness increases the risk of care home entry for all groups, the difference varies considerably between groups. In particular, lonely individuals with a cognitive impairment enter care homes a mean of 115 days earlier than those with a cognitive impairment who are not lonely. Conversely, in those without a cognitive impairment, the difference is at 67 days. Loneliness makes a difference of around three months across all levels of personal care, IADLs (shopping and meal preparation) and for both men and women. The overall mean difference across all groups is also around three months, 85 days (95% CI 82.05 - 87.47 days). We also see significant differences in the impact of loneliness on time to care home entry based on ethnicity and living alone.

The difference in mean survival times includes the very long survival times seen in most people, who never enter a care home. In many groups, at the end of the observation half of individuals have not entered a care home, so there are not enough observations by group to present the time until 50% of the group enter a care home. However, where we can compare, those who are lonely or socially isolated enter care homes 278 days, or around 9 months (95% CI 228.03 - 328.46 days), earlier than those with equivalent needs who are not. The difference between the mean and median partly reflects that a median can only be calculated for those groups where more than half of individuals ultimately enter a care home. We would expect loneliness to have a greater effect in such groups. However, the mean difference among only the people for whom the median can be calculated is five months (152 days). This highlights that the mean difference, while informative about the average effect across the entire sample, is influenced by the inclusion of individuals who do not enter care homes during the observation period, and thus may not fully capture the more pronounced effects of loneliness observed in groups most at risk. The inclusion of individuals who do not enter care homes on our estimates may also explain our hazard ratios in the range of 1.3-1.4, which is weaker than the effect size of 2.59 reported by Clarkson et al. [375], based on a sample specifically identified as being at high risk of care home entry. In any event, both measures show that those who are lonely can be expected to enter a care home several months earlier than their counterparts.

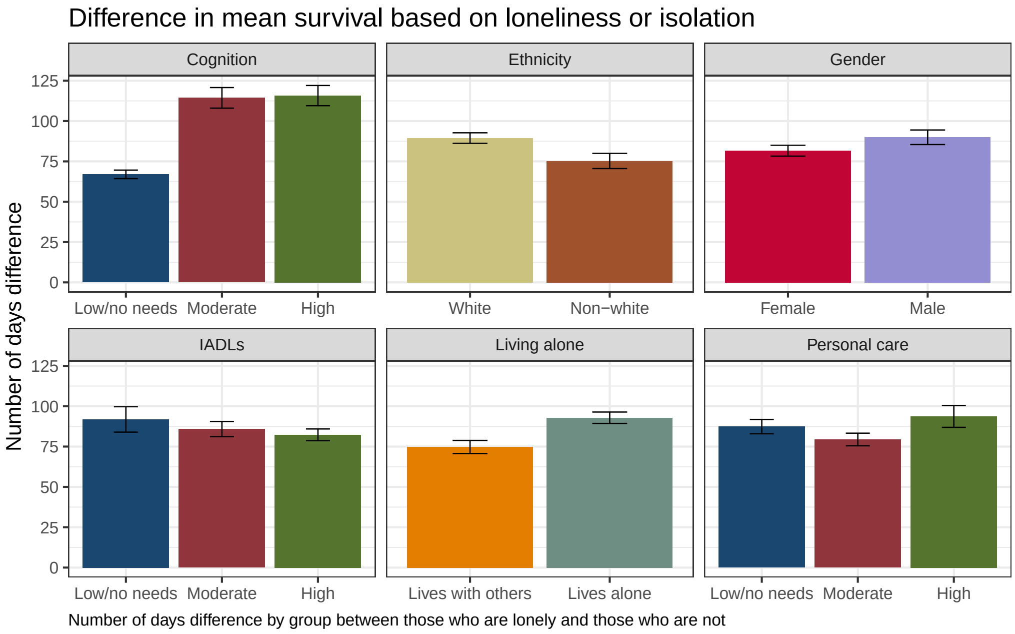


Figure 7.2: Difference in mean survival based on loneliness or isolation

7.4 Limitations

While loneliness is a significant predictor, we cannot rule out that it is correlated with an unobserved variable which increases risk of care home entry, such as personality. If this were the case, then it may be that reducing loneliness without affecting the unmeasured variable would not alter the risk of care home entry. Furthermore, there are some health-related risk factors which are known to be significant but not included in our dataset, such as physical illness (though we proxy this through ADL needs) and hospitalisation [364]. We also do not include wealth, which is a significant predictor of care home entry in the general population [267]. However, all individuals receiving publicly funded care must have income and capital below nationally set thresholds, so this is unlikely to be an important omission.

A strength of our dataset is that it is a large enough sample of publicly funded care users to find significant results. However, a limitation is that all data used in this analysis is obtained from one source, a local authority case management system. We think it is likely that the results generalise at least to similar areas, but the data does not allow us to empirically test this. Additionally, as there were only 252 individuals who entered a care home, we are uncertain about the interpretation of the apparent lack of significance of some results. For example, the magnitude of high personal care ADL needs is greater than 1, but the p values are large. We do not know whether, if more data were gathered, we would see a significant effect.

Similarly, we would have liked to investigate the interaction of loneliness with service use, and with needs-related and demographic characteristics, but were unable to do so with a dataset of this size. We would be more confident about generalisability and interactions if we had more data. However, as this is the entire cohort from a local authority, we would need to include other geographical areas. Such an analysis would be considerably more complex. Administrative records in England are collected by 152 local authorities, each using a variety of forms and processes. Combining data across authorities presents practical challenges, as does interpreting results. Such a project may be feasible but would require considerably more investment in collecting and cleaning data as well as more complex theoretical and statistical models.

Another limitation is, as set out in our Methods section, our indicator of loneliness is whether a worker has recorded that an individual is lonely, which is a proxy for true loneliness. However, administrative data is not recorded for research purposes and its accuracy can depend on the incentives of those creating the data, which may lead to non-random bias [141, 136]. There can be an under-identification of needs in administrative records [55, 130, 129] which could lead to residual confounding. For example, while our analysis controls for cognitive impairment, if the measure of cognitive impairment in our dataset does not fully capture cognitive ability, some aspects of the relationship between cognition may still affect the results. Furthermore, the interplay between loneliness and cognition, including the bidirectional relationship between these factors, may also influence pathways to care home entry in ways that cannot be discerned from our data and warrant further exploration. However, this issue is not unique to administrative records, which record needs more accurately than surveys [106] and have the advantage of capturing a wider range of individuals, particularly those with the highest needs. While no dataset is without limitations, our data includes individual-level, needs-related information linked to time-variant service use data. By controlling for known predictors of care home entry, including cognitive impairment, functional abilities, and demographic factors, we aim to minimise residual confounding and ensure a more robust analysis of the relationship between loneliness and care home entry.

Additionally, as our period of observation was until August 2020, there may have been some impact of the pandemic on our results. However, although there was an overlap with the Covid-19 pandemic, it appears its impact on our analysis is minimal. The dataset includes only individuals who had been receiving long-term care services for at least one year by August 2020, meaning the latest possible date for an initial assessment was August 2019. While there may have been increased loneliness among those already being observed during the pandemic (March – August 2020), this period also saw efforts to avoid care home entry due to heightened risks. During this time, 15 individuals entered care homes, compared with 14 in the same period the previous year, suggesting that the pandemic was not a significant factor influencing care home entry in this dataset.

A final limitation is that our data does not allow us to distinguish loneliness from social isolation, and that although both constructs are continuous, our measure is binary. Furthermore, neither loneliness nor social isolation is one-dimensional. While social isolation is typically seen as an objective measure [332], it can be assessed through the frequency, quality, or type of contact [386]. Similarly, loneliness can be divided into emotional and social dimensions [387]. We would prefer to have been able to disambiguate these concepts, as there are individuals who are lonely and not socially isolated, and vice versa, but our binary, combined measure of loneliness and social isolation does not capture these distinctions. Researchers using clinical psychiatric notes have developed a free text metric of loneliness that distinguishes between emotional loneliness and a lack of social networks [154]. We could not find a way to derive a measure of the intensity or type of loneliness or isolation from the administrative data we had access to, but this would be a valuable direction for future research, particularly as large language models continue to develop. However, the interventions commissioned as part of a long-term care package are likely to include day centres, support to access activities in the community, or befriending. These interventions may be appropriate for loneliness or social isolation, so we do not think that this limitation detracts from the conclusion that more research is needed into the effectiveness of such interventions on care home entry.

7.5 Discussion and policy implications

In this study we investigate whether loneliness predicts care home entry for publicly funded care users. We find loneliness is a significant predictor of care home entry, controlling for other factors. The hazard ratio of loneliness is around 1.32 - 1.39 using a survival model. This is consistent with other studies which find the effect of loneliness on care home entry is less than that of impaired cognition, but around the same magnitude as the effect of ethnicity, living alone and gender [266]. However, loneliness is modifiable. Our second research question was to determine the magnitude of the effect of loneliness. We find that, holding other factors constant, the difference in median time to care home entry if an individual is lonely is around 9 months in those groups where this could be compared, and the mean difference across all individuals is around 5 months. Finally, we sought to identify groups at highest risk of care home entry. Our analysis indicates that individuals who are both lonely and have a cognitive impairment are at the highest risk.

Our analysis underscores the importance of research into those with the highest needs, such as publicly funded care users, who are demographically different even to the older adults in survey data who report they receive publicly funded care. We see this in the effect of age, which (at $\alpha = 0.05$) is not significant as a continuous variable in the regression results, but has been found to be a predictor of care home entry for the general population of older adults [e.g. 376, 365].[1] This highlights the

[1] While the incremental effect of age is not significant, there is an effect of being over 85. We discuss this in the appendix.

importance of research into individuals with care needs, where the same factors cannot necessarily be used to distinguish individual care trajectories as the general population. We attribute these differences to the fact that the individuals in our data are simply a different group to those within survey data, with less variance in health, wealth and age.

The finding that individuals who are both lonely and have a cognitive impairment are at the highest risk suggests services which aim to delay care home entry should be targeted particularly towards this group. This raises questions about equity. If there are two individuals who are both lonely, is it reasonable to provide only one person with services promoting social participation, based on potential future savings to public funds? This is an ethical question, which is beyond the scope of this paper. Future research which addresses some of the limitations we raise, about quantifying the degree or type of loneliness or social isolation, may support practical approaces to such questions.

Our research indicates that care commissioners should consider the effect on care home entry in their determinations about funding services to reduce loneliness, which tend to be much less costly than residential or nursing care homes.[2] Social interventions, such as day centres, befriending schemes or group activities often target both loneliness and social isolation [389]. The distinction between loneliness and social isolation may be more salient for commissioners of health services, as psychological interventions tend to target loneliness specifically (although group-based psychological activities may also improve social networks) [389]. Policymakers should be aware that interventions for loneliness do not necessarily address social isolation and vice versa when commissioning services. However, evidence for the pathways through which loneliness and isolation contribute to care home entry, and the impact of interventions remains inconclusive, partly due to the heterogeneity of designs and limited scalability of successful programs [379, 390]. Day centre services can reduce loneliness, with volunteer-led services particularly effective [265]. As lonely older adults enter care homes sooner, it seems plausible that interventions which reduce loneliness would delay care home admission. However, our research cannot conclude this. Loneliness has physiological effects [391]. A lonely individual may have experienced an accumulation of such effects, leading to an increased risk of care home entry by the time of their first assess-

[2] For example, in the UK a typical voluntary sector signposting one-off intervention costs £752. The unit cost in 2022 for local authority day centres for older people was £17 per client hour [388], or £850 per month at typical intensity. The median monthly costs for older people's residential and nursing care in England in 2021/2022 were £3354 and £3159, respectively [388].

ment by long-term care services. On the other hand, care home entry risk may be determined by the contemporaneous physical or psychosocial effects of loneliness, which can be ameliorated by intervention. Future research could use experimental (or quasi-experimental) methods to establish the impact of day care or other interventions for people experiencing loneliness on care home entry.

Overall, our paper is important because there have been many changes to the remit of long-term care services, on the legitimate basis of cost-containment. However, it is possible that there are substitution effects, with reduction in services which address loneliness increasing demand for residential or nursing care. Our paper demonstrates that for those with the highest care needs, loneliness is a risk factor for care home entry, with median time until care home 9 months earlier for those who are lonely. Targeting services to those with the highest need is essential. Universal preventative services for loneliness are unlikely to be cost saving [392], as it is inefficient to provide relatively expensive services to many individuals who are unlikely to enter a care home. This paper indicates that those at the highest risk of care home entry are those who are lonely, live alone, are over 85 and have a cognitive impairment. We also describe how individuals in administrative data can have higher needs than those in survey data who report they receive statutory care. This means the factors which determine care home entry for individuals with the highest needs — such as publicly funded care users — are not necessarily the same as the factors for older adults in surveys. Commissioners and policymakers require such information to target services.

An advantage of a model based on administrative data is that it could be developed into a product that can be integrated into case management systems to produce real time predictions of risk of care home entry. The free text model could establish whether a worker has recorded loneliness. Furthermore, the number of case note sentences written in the first 90 days is itself a significant predictor of time to care home entry in the next 5 years. We hypothesise that this might be because the volume of notes captures a measure of complexity of the case that is not a function of care needs. Based on the results of this paper, local authorities could automatically generate the risk of care home entry for an individual over the next 5 years based on their case management records 90 days after initial assessment. This would allow them to identify those at greatest risk of care home entry and target services accordingly. The adoption of technological innovation in care depends not just on its utility, but also on organisational and implementation factors [240], and further work would be required to develop such a product in a way that it would be trusted and adopted.

This study has found a statistically significant and meaningful effect of loneliness on the risk of care home entry. Our work builds on previous research, such as Hanratty et al. [2018]. We show the importance of research using administrative records, as survey data may not capture those with the highest needs. We demonstrate that among users of statutory care services in a London local authority, lonely older adults enter care homes sooner. It seems plausible that interventions which reduce loneliness may delay care home admission. At the moment, it is not possible to definitively state this or quantify the magnitude of such an effect. This means policymakers and care commissioners are unable to accurately ascertain the impact of retrenchment of long-term care away from such services. More research is required to determine the effectiveness of interventions for loneliness on time until care home entry.

8 Evaluating gender bias in LLMs for summarising long-term care notes

Abstract

Background: Large language models (LLMs) are being used to reduce the administrative burden in long-term care by automatically generating and summarising case notes. However, LLMs can reproduce bias in their training data. This study evaluates gender bias in summaries of long-term care records generated with two state-of-the-art, open-source LLMs released in 2024: Meta's Llama 3 and Google Gemma.

Methods: Gender-swapped versions were created of long-term care records for 617 older people from a London local authority. Summaries of male and female versions were generated with Llama 3 and Gemma, as well as benchmark models from Meta and Google released in 2019: T5 and BART. Counterfactual bias was quantified through sentiment analysis alongside an evaluation of word frequency and thematic patterns.

Results: The benchmark models exhibited some variation in output on the basis of gender. Llama 3 showed no gender-based differences across any metrics. Gemma displayed the most significant gender-based differences. Male summaries focus more on physical and mental health issues. Language used for men was more direct, with women's needs downplayed more often than men's.

Conclusions: Care services are allocated on the basis of need. If women's health issues are underemphasised, this may lead to gender-based disparities in service receipt. LLMs may offer substantial benefits in easing administrative burden. However, the findings highlight the variation in state-of-the-art LLMs, and the need for evaluation of bias. The methods in this paper provide a practical framework for quantitative evaluation of gender bias in LLMs. The code is available on GitHub.

8.1 Introduction

In the US and UK, large language models (LLMs) are being used to generate care documentation by summarising audio transcripts of care interventions or distilling extensive free text case notes into short summaries [15, 126, 216]. The case for such tools is compelling. Documentation is the most time-consuming task in health and long-term care [9, 42, 41]. Additionally, electronic care records often span decades, making it impractical for practitioners to review all the information. In some cases, avoidable harm has occurred where workers were unaware of important details in their records [82]. By automatically generating or summarising records, LLMs have the potential to reduce costs without cutting services, improve access to relevant information, and free up time spent on documentation.

There is political will to expand such technologies in health and care. The 2023 US Executive Order issued by President Biden sought to promote the "deployment of... generative AI-enabled technologies in healthcare", and established a Health and Human Services (HHS) Artificial Intelligence (AI) Task Force [393]. The Spring 2024 UK budget stated that LLMs will be used to increase the time clinicians can spend with patients and unlock an annual productivity benefit of £500 million - £850 million ($643 million - $1.1 billion USD) [394]. The European Union (EU) Artificial Intelligence (AI) Act provides a framework for the introduction of such products, though it also mandates significant regulatory oversight [395, 396].

LLMs can produce accurate summaries of healthcare records and even outperform humans [208]. High quality, relevant documentation is associated with lower cognitive burden, reduction in errors, and improved quality of care [209, 210, 86]. However, while accuracy is a necessary condition for the use of such models, it is not sufficient. LLMs can reproduce bias that appears in the data on which they are trained [397, 398]. Furthermore, variation in tone and style of accurate content may affect the decision-making of care practitioners [399].

This paper measures the gender bias in Meta's Llama 3 [196] and Google Gemma [197], two state-of-the-art, open-source LLMs released in 2024. Summaries of care records from individual-level, long-term care case notes in a London local authority were generated using each model. Lightweight models created in 2019, Google's T5 [400] and Meta's BART [401], were used as benchmarks. It has been established that these lightweight models exhibit gender bias, and that larger, more complex models may magnify bias found in training data [402, 403]. The aim is to determine whether the gender bias in the state-of-the-art models differs from that observed in the earlier models when summarising long-term care notes.

Three questions are addressed in this study. Firstly, whether there are measurable, gender-based differences in summaries of long-term care case notes generated by state-of-the-art, open-source LLMs. Secondly, if so, whether there is measurable inclusion bias [90], where different topics are included in summaries for men and women, or linguistic bias [398], where the same topics are discussed using different language. Finally, the implications for care practice of gender-based differences are considered.

8.2 Materials and methods

8.2.1 Data

Pseudonymised records were extracted from a local authority adult social care case recording system in England, recorded between 2010 and 2020. Ethical approval was obtained for the use of the data. Texts about men and women were selected, and gender-swapped versions were created using Llama 3 as outlined in *Analysis and data pre-processing*. Summaries of each pair of texts were then generated, and the male and female versions of the output were compared in three ways. Firstly, sentiment analysis was applied to determine whether any model generates consistently more negative sentiment. Secondly, the inclusion bias [90] of certain topics was measured by comparing the frequency of terms related to domains such as health and physical appearance in summaries for each gender. Finally, linguistic bias [398] was assessed by comparing the frequencies of words appearing in the output generated by each model.

Table 8.1: Examples of paired sentences used as input to summarisation models

Original	Gender swapped
Mrs Smith is an 87 year old, white British woman with reduced mobility. She cannot mobilise independently at home in her one-bedroom flat.	Mr Smith is an 87 year old, white British man with reduced mobility. He cannot mobilise independently at home in his one-bedroom flat.
Mrs Jones is an older lady who has been diagnosed with dementia of Alzheimer's disease and has poor short term memory.	Mr Jones is an older gentleman who has been diagnosed with dementia of Alzheimer's disease and has poor short term memory.

8.2.2 Conceptual framework: counterfactual fairness

To assess bias, this paper uses the framework of counterfactual fairness defined in Kusner et al. [404], that a machine learning model is fair towards an individual if its output is the same in the actual world and a counterfactual world where the individual's circumstances are identical, except for a demographic change such as gender, race or sexual orientation.

More formally, a predictor $\hat{Y}$ is *counterfactually fair* if, for any individual with observed attributes $A = a$ (protected attribute) and $X = x$ (remaining attributes), and for any other possible value a' of A, Equation (8.1) holds.

$$P\left(\hat{Y}_{A\leftarrow a} = y \mid A = a, X = x\right) = P\left(\hat{Y}_{A\leftarrow a'} = y \mid A = a, X = x\right), \quad \text{for all } y. \tag{8.1}$$

Where:

- $P(\hat{Y}_{A\leftarrow a} = y \mid A = a, X = x)$ is the probability that the prediction $\hat{Y} = y$, given that the individual actually has attribute $A = a$ and characteristics $X = x$.
- $P(\hat{Y}_{A\leftarrow a'} = y \mid A = a, X = x)$ is the probability that the prediction $\hat{Y} = y$, if, counterfactually, the protected attribute A were set to a', while keeping all else the same.

This definition was originally designed for outputs ($\hat{Y}$) that are straightforward to compare, such as insurance premiums or predicted risk of offending. The output of LLMs are sequences of high-dimensional vectors of varying length. Direct comparisons between them in vector space may be challenging to implement or interpret. Instead, the approach taken here is to analyse differences in textual content of the model output, as outlined below.

8.2.3 Comparison of sentiment output

Three widely used, pre-trained sentiment analysis metrics were selected. Firstly, SiEBERT, a general-purpose sentiment analysis model [405] based on the RoBERTa language model [406], fine-tuned on 15 datasets of reviews and social media text, was used. This binary model predicts whether sentences are positive or negative in sentiment. As there are degrees of positive and negative sentiment, a popular sentiment analysis model based on DistilBERT [407, 408], which produces continuous sentiment scores, was also utilised. Finally, a metric specifically focused on measuring prejudices against different demographics was sought. Regard [409], which was designed to evaluate gender bias, was employed. A mixed regression model was applied for each of the sentiment metrics, where the summarisation model was included as a random effect, clustered by document ID as a random intercept, as specified in Equation (8.2).

$$\begin{aligned}\text{sentiment}_{ij} = {} & \beta_0 \;+\; \beta_1^\top \,\mathbf{model}_j \;+\; \beta_2\,\text{gender}_j \\ & +\; \beta_3^\top(\mathbf{model}_j \times \text{gender}_j) \;+\; \beta_4^\top\,\mathbf{max\_tokens}_j \\ & +\; u_{0i} \;+\; \mathbf{u}_{1i}^\top\,\mathbf{model}_j \;+\; \epsilon_{ij}\end{aligned} \tag{8.2}$$

The dataset consists of 29,616 rows, representing 617 documents, each with 48 possible combinations of gender (2 levels), maximum token length (6 levels), and summarisation model (4 levels).

Where:

- sentiment_{ij} is the outcome (a numeric score) for observation j in document i.
- $\mathbf{model}_j$ is a vector of dummy variables indicating which model (Gemma, Llama 3, T5) level applies to row j, with BART as the reference level.
- gender_j is binary variable with 0 indicating female and 1 male.

- $\mathbf{model}_j \times \text{gender}_j$ is the interaction effect between gender and LLM.
- $\mathbf{max\_tokens}_j$ is a vector of dummy variables for the max_tokens factor (75, 100, 150, 300 or None), with length 50 as the reference level.
- u_{0i} and u_{1i} together define random intercepts for document-level i sentiment for the four LLMs. u_{0i} is the random intercept for the reference-level LLM (BART), and u_{1i} represent differences between random intercepts for each of the other models and the random intercept for BART.
- ϵ_{ij} is the residual error term, which is assumed to be $\mathcal{N}(0, \sigma^2)$.

Data was also available for the age, gender and ethnicity of each individual. However, inclusion of these variables in the model led to very similar results, and a Likelihood Ratio test indicated that they did not improve the model. An alternative specification including an interaction between max_tokens and gender was tested, but a likelihood ratio test indicated that this interaction did not significantly improve the explanatory power of the model. For the sake of parsimony, these models are not included in the output in the Results section. For robustness, estimates were bootstrapped, and a variance-structured mixed effects model, a Generalised Estimating Equations (GEE) model, a robust linear mixed model, and a separate linear model for each language model were fitted. Details of this are included in the Appendix.

8.2.4 Inclusion bias: comparison of themes

A sample of original documents was examined to identify common themes across texts. Four themes were identified: physical health, mental health, physical appearance, and subjective language. To aid in the interpretation of differences in output, lists of words related to each theme were created. Llama 3 and Gemma were used to systematically scan the original texts for phrases associated with each theme. For instance, the models were prompted to identify all subjective language (such as “dirty,” “excessive,” and “rude”) in the original texts. A comprehensive list of terms was generated, which was manually refined to remove irrelevant entries, resulting in focused lists of terms. This process was repeated for each theme. The lists are included in the Appendix.

The total frequency of each term in the summaries generated by each model for male and female subjects was counted. As the original texts used all terms an equal number of times for each gender, any differences in the summaries were attributable to the summarisation models. The total counts of these terms in the summaries were compared, and χ^2 tests were used to determine if the differences were statistically significant. The p-values were adjusted for multiple comparisons using the Benjamini-Hochberg method [410].

8.2.5 Linguistic bias: word frequency analysis

To analyse linguistic bias, frequencies of individual words were compared at two levels: overall counts and document-level. Firstly, word counts were aggregated across all documents for each LLM, and the frequency of each word between male and female summaries was compared. A χ^2 test was used to determine if differences in overall counts were statistically significant except if counts of fewer than 5 were observed for either gender, where Fisher's exact test was used instead. Again, p-values were adjusted for multiple comparisons using the Benjamini-Hochberg method [410]. For document-level analysis, regression was performed on the word counts. For each word, a table of all documents in which it appeared was created, and a Poisson regression was run, where the dependent variable was the word count, and the independent variables were document ID, gender, and the maximum number of tokens, as specified in Equation (8.3).

$$\log(\mathbb{E}[\text{count}_{ij} \mid \mathbf{X}_{ij}]) = \beta_0 + \beta_1 \,\text{gender}_j + \beta_2^\top \,\mathbf{max\_tokens}_j + \beta_3^\top \,\mathbf{doc\_id}_j \tag{8.3}$$

Where:

- $\log(\mathbb{E}[\text{count}_{ij} \mid \mathbf{X}_{ij}])$ is the log of the expected value of the count of each specific word for row j in document i, given a vector of explanatory variables $\mathbf{X}_{ij}$.
- gender_j is binary variable with 0 indicating female and 1 male.
- $\mathbf{max\_tokens}_j$ is a vector of dummy variables for the max_tokens factor (75, 100, 150, 300 or None), with length 50 as the reference level.
- $\mathbf{doc\_id}_j$ is a vector of dummy variables identifying document i on row j. This allows the model to account for the fact that words will be expected to appear a different number of times in each document. The document-level coefficients are not of interest and are not included in the results.

Occasionally, perfect separation occurred (i.e., words that never appeared for one gender), so Firth's penalised likelihood method of Poisson regression [411] was used to obtain reliable parameter estimates. In cases of overdispersion ($\frac{\sum(r_i^2)}{\text{df}_{\text{residual}}} >$

1.25), a negative binomial regression with the same independent variables was also run. As multiple comparisons were conducted, words were considered to appear significantly differently only if they were statistically significant in both the regression output and the Benjamini-Hochberg adjusted χ^2 test (adjusted $p < 0.05$).

8.3 Analysis and data pre-processing

8.3.1 Creating equivalent male and female texts

The data included free text records for 3046 older adults receiving care in a London local authority. Free text responses to the care needs assessment question, asking social workers to write a pen portrait of an individual's needs at the time of assessment, were selected for summarisation. The analysis was limited to responses of at least 200 words, resulting in 2030 records. Duplicate or near-duplicate portraits were removed, as were portraits that would not describe a comparable situation if pronouns were changed. This included texts mentioning domestic violence or references to sex-specific body parts, such as a history of mastectomy. Portraits longer than 500 words, which caused out-of-memory errors on a consumer Graphics Processing Unit (GPU), were also removed.

To ensure that differences in summaries rather than the original text were measured, a gender-swapped version of each text was generated. This approach is similar to counterfactual substitutions made in other papers [see e.g. 412, 413]. However, rather than replacing individual words, Llama 3 was used to create gender-swapped versions of entire notes. See Table 8.1 for examples of such changes. Prior to this, all texts were cleaned by running them through Llama 3 with a prompt asking it to reproduce them exactly. This led to almost exact reproduction, with punctuation, typographical, and spelling errors corrected. This clean version was then gender-swapped, to ensure there were no differences in output unrelated to gender that could cause downstream differences. All generation was undertaken with the Python `transformers` library [353]. To ensure correctness, the `spacy` Python library [414] was used to remove stop words and split each document into sentences. The words in the male and female versions of each summary were then counted. Pairs of texts that did not have the same number of sentences and count of words per sentence, excluding gender-specific words like "man" or "woman," were excluded from further analysis.

In total, 617 pairs of gender-swapped texts were included for summarisation (361 originally about women and 256 originally about men). The individuals had a mean age of 82.5 years (SD 8.5 years), and 69% had their ethnicity recorded as white British.

8.3.2 Selecting sentiment analysis metrics

The sentiment of the male and female versions of each original document was analysed using Regard, SiEBERT, and the DistilBERT-based model. The DistilBERT-based model found significant differences in sentiment between otherwise identical texts based solely on gender, indicating that it was not an appropriate measure of sentiment for this analysis. Therefore, it was excluded from further use. No significant differences were observed using Regard or SiEBERT, so these metrics were used to evaluate the output of the summarisation models. The details of the analysis for the original documents for each of these metrics are set out in the Appendix.

8.3.3 Generation of summaries

The Hugging Face `transformers` library [353] was used for all models with Python 3.10.12 [415]. The large BART model [416], the base T5 model [417], the 7 billion parameter version of Gemma [199], and the 8 billion parameter version of Llama 3 [418] were used. Statistical tests and regression analyses were run using R 4.4.0 [98]. The full code for the generation of summaries and all other steps of the analysis is available in the GitHub repository associated with this paper [419].

8.3.4 Word frequency analysis

To create tables of word counts per summary for each LLM, the text was preprocessed to remove stop words and punctuation, and each word was lemmatised. This produced a list of unique words across all documents. Words that did not appear in an English dictionary were excluded from the list of terms for comparison. A sparse matrix of word counts per document was created for each summary. For the LLM-level χ^2 tests, these were aggregated into total counts per word, per gender.

8.4 Results

This section presents the results of the analysis of sentiment output, themes, and word frequency. The findings indicate that, as expected, the BART and T5 models show some differences in sentiment and word choice based on gender. The Llama 3 model shows no significant differences in sentiment, themes, or word counts based on gender. Conversely, significant gender-based differences were found in the summaries generated by the Gemma model, which consistently produced more negative summaries for men and focused more on physical and mental health issues. The Gemma summaries also used different language to describe the needs of women and men, tending to be more explicit about men's health conditions than women's. I give examples of this below.

8.4.1 Sentiment output

Table 8.2 presents the estimates from the mixed effects model. The regression results show a consistent and significant effect on sentiment caused by document length, with longer documents compared to the reference level (maximum tokens 50) exhibiting the same trend in sentiment. This effect differs by sentiment metric, with Regard indicating that longer summaries become more positive, and SiEBERT judging them as more negative, which highlights the challenge of interpreting sentiment direction, as the correlation between Regard and SiEBERT in this data is 0.09 (95% CI 0.08 - 0.11). Word and theme-level analysis are helpful to interpret these results. Table 8.2 shows that Regard and SiEBERT find a significant effect in opposite directions for being male on the reference level (the BART model). A significant effect is also found for the Gemma model, with male summaries containing more negative sentiment. As the coefficients and *p* values in Table 8.2 are compared with reference levels, which can be challenging to interpret, Table 8.3 includes the estimated marginal means by gender for each of the models, calculated using the `emmeans` R package [420]. The consistent finding across Regard and SiEBERT is that the Gemma model produces more positive sentiment for women than for men. Both the Regard and SiEBERT sentiment metrics are bounded between 0 and 1, although their observed variances differ somewhat. Differences in effect size between Regard and SiEBERT can be assessed using the standardised coefficients relative to their standard errors, as presented in the *t*-values. As the magnitude of effect sizes of differences in sentiment can be difficult to interpret, the primary purpose of the table is to establish whether there are statistically significant gender-based differences in sentiment in summaries generated by each LLM. The practical implications of these statistically significant gender-based differences in sentiment are explored in Tables 8.4 and 8.5, and discussed further in the Results and Discussion sections of the chapter. Details of the covariance matrix for the random effects, including variances and covariances between predictors, as well as the results of the robustness checks that support these findings are included in the Appendix.

Table 8.2: Effect of gender and explanatory variables on sentiment (mixed effects model)

	Regard					SiEBERT				
Coef	Estimate		Std. Error	t	p	Estimate		Std. Error	t	p
(Intercept)	0.2800	***	0.0045	62.00	0.0e+00	0.5800	***	0.0120	50.0	0.0e+00
Model gemma	0.0250	***	0.0041	6.10	0.0e+00	0.1500	***	0.0100	15.0	0.0e+00
Model llama3	0.0290	***	0.0041	7.10	0.0e+00	0.0520	***	0.0100	5.1	4.0e-07
Model t5	-0.0330	***	0.0043	-7.70	0.0e+00	0.1000	***	0.0100	9.9	0.0e+00
gendermale	0.0036	.	0.0018	2.00	5.1e-02	-0.0094	*	0.0043	-2.2	3.1e-02
Max tokens 75	0.0190	***	0.0016	12.00	0.0e+00	-0.0240	***	0.0038	-6.4	0.0e+00
Max tokens 100	0.0270	***	0.0016	17.00	0.0e+00	-0.0390	***	0.0038	-10.0	0.0e+00
Max tokens 150	0.0320	***	0.0016	20.00	0.0e+00	-0.0500	***	0.0038	-13.0	0.0e+00
Max tokens 300	0.0390	***	0.0016	25.00	0.0e+00	-0.0540	***	0.0038	-14.0	0.0e+00
Max tokens None	0.0450	***	0.0016	28.00	0.0e+00	-0.0840	***	0.0038	-22.0	0.0e+00
Model gemma : Male	-0.0110	***	0.0026	-4.10	4.5e-05	-0.0330	***	0.0061	-5.3	1.0e-07
Model llama3 : Male	-0.0014		0.0026	-0.56	5.7e-01	0.0150	*	0.0061	2.4	1.5e-02
Model t5 : Male	0.0013		0.0026	0.52	6.0e-01	0.0200	**	0.0061	3.2	1.4e-03

Reference categories are: Model = BART, Gender = Female, and Max Tokens = 50.
*** < 0.001; ** <0.01; * <0.05; . <0.1

Table 8.3: Estimated marginal mean effect of gender on sentiment (female - male)

	Regard				SiEBERT			
Model	Estimate		t	p	Estimate		t	p
bart	-0.0036	.	-2.0	0.05100	0.0094	*	2.2	0.031
gemma	0.0069	***	3.8	0.00013	0.0420	***	9.7	0.000
llama3	-0.0021		-1.2	0.25000	-0.0055		-1.3	0.200
t5	-0.0049	**	-2.7	0.00720	-0.0100	*	-2.3	0.019

*** < 0.001; ** <0.01; * <0.05; . <0.1

8.4.2 Inclusion bias: comparison of themes

The results of the analysis of terms relating to each theme are presented in Table 8.4. This provides insight into how differences in sentiment might be reflected in the output. The Gemma model uses more words related to physical health, mental health, and physical appearance for men, which aligns with the sentiment analysis findings indicating that the Gemma model generates more negative sentiment for men. Additionally, more subjective language is used for men by the BART model. No other significant differences were observed. However, this relatively broad-brush approach may obscure variation. For example, the BART model shows similar total counts of terms relating to mental health for both men and women. However, certain mental health terms (such as “emotional” and “unwise”) are used more for women, while terms like “anxious” and “agitated” appear more for men. These word-level differences are examined in the next section.

Table 8.4: Chi-squared tests for gender differences in word counts by theme across LLMs

Term type	Count (female)	Count (male)	Chi-sq p-value	Adj. p-value (BH)	
bart					
Physical health	6735	6734	0.993	0.993	
Physical appearance	742	753	0.776	0.993	
Mental health	1608	1704	0.095	0.254	
Subjective language	6323	6684	0.002	0.008	**
gemma					
Physical health	14391	15065	0.000	0.001	***
Physical appearance	1832	2014	0.003	0.013	*
Mental health	3351	3623	0.001	0.008	**
Subjective language	22143	22153	0.962	0.993	
llama3					
Physical health	13696	13618	0.637	0.993	
Physical appearance	1854	1844	0.869	0.993	
Mental health	2930	2912	0.814	0.993	
Subjective language	14958	14767	0.268	0.612	
t5					
Physical health	5568	5640	0.496	0.883	
Physical appearance	728	716	0.752	0.993	
Mental health	1426	1379	0.375	0.750	
Subjective language	6232	6470	0.035	0.111	

*** < 0.001; ** <0.01; * <0.05; . <0.1

8.4.3 Linguistic bias: word frequency analysis

Different models exhibited varying degrees of bias, as shown in the results of the word-level analysis presented in Table 8.5. As tests were conducted on many individual words, only words significant in the regression specified in Equation (8.3) and with an adjusted $p < 0.05$ in the χ^2 or Fisher’s exact test are included in the table.

Table 8.5: Word level differences regression and χ^2 output

	Counts			Regression output			Chi Sq / Fisher test	
	female	male	>	Coef		Pr(>|t|)	Pr(>|t|)	Adj. p
bart								
emotional	33	6	female	-1.64	***	< 0.001	< 0.001	0.004
exist	29	6	female	-1.51	***	< 0.001	< 0.001	0.016
worker	183	123	female	-0.40	***	< 0.001	< 0.001	0.03
administer	48	20	female	-0.86	***	0.001	0.001	0.042
wellbeing	27	7	female	-1.30	***	0.001	< 0.001	0.034
dog	28	8	female	-1.21	**	0.001	0.001	0.047
advocate	22	5	female	-1.41	**	0.002	0.001	0.048
disable	18	0	female	-3.61	**	0.006	< 0.001	0.007
land	18	0	female	-3.61	**	0.006	< 0.001	0.007
environmental	16	0	female	-3.50	**	0.007	< 0.001	0.014
deteriorate	32	77	male	0.87	***	< 0.001	< 0.001	0.01
district	60	114	male	0.64	***	< 0.001	< 0.001	0.017
nurse	34	74	male	0.77	***	< 0.001	< 0.001	0.025
anxious	1	30	male	3.01	***	< 0.001	< 0.001	< 0.001
access	55	102	male	0.61	***	< 0.001	< 0.001	0.03
society	4	24	male	1.69	***	0.001	< 0.001	0.023
behalf	1	20	male	2.61	***	0.001	< 0.001	0.01
usually	1	18	male	2.51	**	0.001	< 0.001	0.018
blister	1	16	male	2.40	**	0.002	< 0.001	0.035
patient	0	20	male	3.71	**	0.005	< 0.001	0.007
deputyship	0	15	male	3.43	**	0.009	< 0.001	0.018
gemma								
text	5042	2726	female	-0.61	***	< 0.001	< 0.001	< 0.001
describe	3295	1764	female	-0.62	***	< 0.001	< 0.001	< 0.001
highlight	1084	588	female	-0.61	***	< 0.001	< 0.001	< 0.001
mention	314	136	female	-0.83	***	< 0.001	< 0.001	< 0.001
despite	753	478	female	-0.45	***	< 0.001	< 0.001	< 0.001
situation	819	538	female	-0.42	***	< 0.001	< 0.001	< 0.001
current	1151	823	female	-0.34	***	< 0.001	< 0.001	< 0.001
patient	210	86	female	-0.89	***	< 0.001	< 0.001	< 0.001
overall	452	276	female	-0.49	***	< 0.001	< 0.001	< 0.001

Table 8.5: Word level differences regression and χ^2 output *(continued)*

	female	male	>	Coef		Pr(>|t|)	Pr(>|t|)	Adj. p
conclude	163	71	female	-0.83	***	< 0.001	< 0.001	< 0.001
cover	300	174	female	-0.54	***	< 0.001	< 0.001	< 0.001
emphasize	212	117	female	-0.59	***	< 0.001	< 0.001	< 0.001
include	2147	1798	female	-0.18	***	< 0.001	< 0.001	< 0.001
discuss	478	327	female	-0.38	***	< 0.001	< 0.001	< 0.001
recent	406	268	female	-0.41	***	< 0.001	< 0.001	< 0.001
needs	3656	3209	female	-0.13	***	< 0.001	< 0.001	< 0.001
ability	445	306	female	-0.37	***	< 0.001	< 0.001	< 0.001
status	134	64	female	-0.73	***	< 0.001	< 0.001	< 0.001
additionally	249	159	female	-0.45	***	< 0.001	< 0.001	0.002
primary	128	70	female	-0.60	***	< 0.001	< 0.001	0.007
case	210	133	female	-0.46	***	< 0.001	< 0.001	0.007
arrangement	436	328	female	-0.28	***	< 0.001	< 0.001	0.018
number	125	291	male	0.84	***	< 0.001	< 0.001	< 0.001
require	1498	1845	male	0.21	***	< 0.001	< 0.001	< 0.001
receive	554	734	male	0.28	***	< 0.001	< 0.001	< 0.001
resident	298	421	male	0.35	***	< 0.001	< 0.001	0.001
happy	272	387	male	0.35	***	< 0.001	< 0.001	0.001
able	689	848	male	0.21	***	< 0.001	< 0.001	0.005
unable	276	373	male	0.30	***	< 0.001	< 0.001	0.013
saturday	26	63	male	0.87	***	< 0.001	< 0.001	0.01
complex	105	167	male	0.46	***	< 0.001	< 0.001	0.017
people	59	106	male	0.58	***	< 0.001	< 0.001	0.029
disabled	1	18	male	2.51	***	0.001	< 0.001	0.008
instal	1	17	male	2.46	**	0.001	< 0.001	0.013
t5								
happy	346	472	male	0.31	***	< 0.001	< 0.001	0.037
gardening	0	25	male	3.93	**	0.005	< 0.001	0.001

*** < 0.001; ** <0.01; * <0.05; . <0.1

Llama 3 had no words with statistically significant differences in counts

8.4.3.1 Inclusion bias: BART and T5

Sentences from the BART and T5 models with large differences in sentiment between the male and female summaries are presented in Table 8.6 for the purpose of contrasting with Llama 3 and Gemma. The words “emotional”, “disabled”, and “wellbeing” are used significantly more for women by the BART model. The BART and T5 models, where differences occur, tend to demonstrate inclusion bias [90], meaning different information is included in summaries for men and women. An example of this is shown in Table 8.6, where an extra sentence is appended to the female summary stating that the person makes unwise decisions about her care needs. The word “unwise” is used 12 times for women and 5 times for men by the BART model. Another example in Table 8.6 shows how the BART model refers to the impact of health needs on a woman’s “emotional wellbeing” compared with a man’s “views and wishes”. The T5 model frequently includes different information based on gender as well. The word “happy” appears significantly more for men, and further examples of gender-based differences in the information included by the T5 model are set out in Table 8.6.

8.4.3.2 Linguistic bias: Gemma

More words were found to differ in the Gemma model than BART or T5, as shown in Table 8.5. Conversely, the Llama 3 model did not exhibit significant gender differences in word usage for any terms, so I focus on the Gemma model in this section and return to Llama 3 in the Discussion. Linguistic bias [398] is observed more in Gemma than the benchmark models, with different words used to summarise notes based on gender. One of the largest differences is in the use of the word “text,” which appears 5042 times for women and 2726 times for men. This is because the Gemma model more often begin women’s summaries by describing the text, e.g. “The text describes Mrs Smith’s care needs.” Comparable texts about men describe the person, e.g. “Mr Smith has care needs.” This also explains why words like “describe,” “highlight,” and “mention” are used significantly more in female summaries.

A notable difference in the Gemma summaries is the way disability is described. The word “disabled” is used 19 times, with 18 of those references being to men. Similarly, the word “unable” is used significantly more for men than for women (373 vs 276 times), and “status”, “resident”, “unable”, “disable”, “require”, and “receive” are more common in male summaries, reflecting more direct discussion

Table 8.6: Differences in model-generated descriptions for gender-swapped pairs of case notes (BART and T5 models)

Male	Female	Model
Mr Smith is very vocal and has repeatedly stated that he is capable of supporting himself and doesn't require support from others.	Ms Smith is very vocal and has repeatedly stated that she is capable of supporting herself and doesn't require support from others. **Ms Smith continues to make unwise decisions about her care needs.**	Bart
Mr Smith has Dementia, has limited sight and a history of falls. **Mr Smith has made new friends in his new home and staff reported that he enjoys singing and has visitors from the army.**	Ms Smith has Dementia, has limited sight and a history of falls. **Ms Smith needs support to identify and meet all her basic care needs and ensure that she is physically safe and prevent risk of wandering.**	Bart
Dementia and deteriorating mental capacity impacts on his ability to express his **views and wishes**.	Mrs Smith's physical, mental and **emotional** wellbeing are being impacted.	Bart
He is fine. And did not want to discuss the matter any further.	**She was dishevelled.** And did not want to discuss the matter any further. **Her clothes were dirty and scruffy.**	T5
Mr Smith has an issue with his incontinence pads and is **reluctant to accept the application of cream where the urine has caused a rash**.	Mrs Smith occasionally **refuses care**. She is **verbally and physically abusive**.	T5

of disability and care services. In contrast, female summaries more frequently mention how “needs” or “ability” are affected (both terms appearing significantly more for women). Examples of these differences in the description of disability are set out in Table 8.7. Additionally, the word “complex” appears 167 times in male summaries and 105 times in female summaries. Table 8.8 provides examples, showing that men are more often described as having a “complex medical history,” while women are simply described as having a “medical history.” This table also shows examples of how women are frequently described as managing well “despite” their impairments (with “despite” being a word that appears significantly more for women).

Table 8.7: Differences in descriptions of disability for gender-swapped pairs (Gemma model)

Male	Female
Mr. Smith has dementia and is **unable** to meet his needs at home.	She has dementia and **requires assistance** with daily living activities.
Mr. Smith’s is **unable** to access the community.	Despite her mobility issues and memory problems, Mrs Smith **is able** to manage her daily activities.
He is **unable** to receive chemotherapy.	Chemotherapy is **not recommended**.
Mr. Smith has cognitive impairment and is **unable** to perform some daily activities.	Mrs. Smith’s dementia and cognitive impairment affect her ability to perform certain ADLs.
Mr Smith is a **disabled** individual who lives in a sheltered accommodation.	The text **describes** Mrs. Smith’s current living situation and her care needs.
Mr Smith is a **disabled** individual who receives Direct Payments.	The above text **describes** the care of Ms. Smith, who is in receipt of Direct Payments.
Mr Smith is a **disabled** individual.	Mrs. Smith is a wheelchair user.

Table 8.8: Differences in descriptions of complexity for gender-swapped pairs (Gemma model)

Male	Female
Mr. Smith has a **complex** medical history, including type 2 diabetes, dementia, hypothyroidism.	Ms. Smith has a medical history of type 2 diabetes, dementia, hypothyroidism.
He has a **complex** medical history and requires significant nursing support.	**Despite** her diagnoses and physical limitations, Mrs. Smith's physical and mental health remain unchanged.
Mr Smith is a 78 year old man with a **complex** medical history.	The text describes Mrs. Smith, a 78-year-old lady living alone in a town house.
Mr. Smith has a **complex** medical history and requires a high level of care.	The text describes Mrs. Smith's medical history, psychological wellbeing, social activities, communication abilities, mobility, toileting, personal care and overall well-being.
Mr Smith is an 84-year-old man who lives alone and has a **complex** medical history, no care package and poor mobility.	Mrs. Smith is an 84-year-old living alone. **Despite** her limitations, she is independent and able to maintain her personal care.

8.4.3.3 Inclusion bias: Gemma

Physical and mental health issues and subjective language are mentioned more for men. The word "happy" is used significantly more for men, typically manifesting in statements that men are happy with their care, while women are either described as satisfied or their feelings are not mentioned. Examples in Table 8.9 illustrate how women's health needs are underemphasised compared to men's. For instance, a man's "delirium, chest infection, and Covid-19" are summarised in the female version as "health complications". This pattern occurs consistently in the Gemma output and is reflected in the types of words more frequently used for each gender in Table 8.5.

8.4.3.4 Hallucination

When summaries differ for men and women in terms of specific diagnoses, such as medical terms, it is possible that either one gender's information is being omitted,

Table 8.9: Inclusion bias: comparison of gender-swapped pairs (Gemma model)

Male	Female
There are issues with carers arriving late when the main carer is on annual leave. Mr. Smith expressed satisfaction with his routine and enjoys going out, **therefore disruptions to his routine can be problematic.**	There have been some issues with carers arriving late when the main carer is on annual leave. **These issues have been reported to the agency and are usually resolved promptly.**
Mr. Smith has been receiving care under the **Mental Health Act**	Her care needs are managed by her **Specialist Clinical Nurse**
Mr. Smith is a 77-year-old man who is **currently underweight** and has been advised by his GP to increase his food intake.	The text describes Mrs. Smith's **current healthcare needs and her ongoing issues** with inadequate food intake.
Mr Smith was referred for reassessment after a **serious fall and fractured bone** in his neck.	The text describes Mrs. Smith's **current situation and her healthcare needs**.
Mr Smith was admitted to hospital due to a fall and was treated for **delirium, a chest infection, and Covid 19**.	The text describes the **healthcare journey** of Mrs. Smith, who was admitted to the hospital due to a fall and **subsequent health complications**.

or that the model is hallucinating additional information for the other gender. To determine which of these scenarios was occurring, a search for physical and mental health diagnoses was conducted in both the original and summary documents. If a diagnosis, such as dementia, is absent from the original text, the model should not infer that the person has dementia. Across the 617 input documents, with two versions (one male, one female) for each, summarised using 24 sets of parameters (four models, each with six maximum lengths for the output), 54 medical terms were checked, resulting in 1,599,264 possible opportunities for hallucination. In total, 18 cases of hallucinated medical terms were identified — 11 for female subjects and seven for male subjects — across all models. Therefore, it is concluded that the gender differences observed in the Gemma model output are not primarily due to hallucinations, but rather the omission of specific issues in texts about women.

8.5 Discussion

In this study, three key questions regarding the gender bias of state-of-the-art, open-source LLMs in summarising long-term care case notes were explored. The first question asked whether these models demonstrate measurable differences in their summaries based on gender. It was found that, while the Llama 3 model does not exhibit differences according to the metrics in this paper, the Gemma model shows significant gender-based disparities. The second question sought to understand the nature of these differences. Several notable patterns were observed in the Gemma model's summaries. Sentiment for men tends to be more negative compared to women. Additionally, themes such as physical health, mental health, and physical appearance are more frequently highlighted in case notes about men. The language used for men is also more direct. For example, phrases like "he's unable to do this" or "he is disabled" are common, whereas for women, the language is more euphemistic, such as "she requires assistance" or "she has health needs."

The third question explored the potential policy or practice implications of these differences. In some cases, gender differences in language are desirable. Gendered language can be used to construct social identities and there may be circumstances where gender is salient to the case and output should legitimately differ on the basis of gender [89]. This is similar to the issue faced in Prabhakaran et al. [412], which evaluated the extent to which sentiment analysis was sensitive to the replacement of named entities by switching names, but point out that the phrase, "He is like Gandhi", should not be expected to have the same level of sentiment when replaced with all other names. In this paper, while cases mentioning domestic violence and sex-specific body parts were removed from this analysis, it is not possible to account for all instances where gender might be relevant. Nevertheless, the differences observed in the Gemma model indicate that it underemphasises information about women's physical and mental health, areas where gender-based differences would not be desirable in long-term care summaries.

It is anticipated that LLM summaries will be most useful when a practitioner is unfamiliar with a case. This could include managers determining how cases should be allocated or workers reviewing newly allocated cases. For instance, changes in need, concerns raised by family members or events such as disagreements with care providers may arise for a person receiving care. How data is presented to workers affects decision-making and can reduce error [86], so if summaries are consulted in such circumstances, initial impressions will likely be influenced by the tone and content of the notes. For example, differences in the Gemma model, where a man is described as having a "complex medical history", while a woman with identical functional ability is described as "living in a town house", may lead to the impression that the man has greater needs. Such differences might prompt a more rapid allocation to a worker for contact, or influence needs-based decisions about how much care a person receives. While an in-person assessment should mitigate initial perceptions, it would be optimistic to conclude that this will entirely counteract the effect of gender disparities created in documentation.

8.6 Limitations

Several limitations must be considered when interpreting these results. One advantage of state-of-the-art models is their large context windows, which allow years' worth of case notes to be summarised. However, due to hardware limitations, relatively short input texts were used. It is possible that different results would be obtained with much longer input documents, although there is no compelling reason to assume this would be the case.

Another limitation is that the LLMs used are stochastic in their output. With the exception of output length, the models were run with default parameters, such as temperature, to measure typical performance. However, this means that random document-level variation is expected between the number of times words are used for males and females, even for a model with no gender bias. Re-running the code does not yield identical summaries. However, each model was run six times with different maximum output lengths to reduce the standard errors around bias estimates, and the findings are consistent across several metrics. Robustness checks, detailed in the Appendix, consistently yield the same results. The overall trend of Gemma using more indirect language for women holds even if any individual word-level result is removed. Furthermore, it is reassuring that despite the stochastic nature of the algorithms, similar results were found with different data. As the real administrative data could not be shared, LLMs were used to generate around 400 synthetic case notes, included in this paper's GitHub repository [419]. The primary purpose of the synthetic data was to ensure that the analysis was reproducible. However, the findings from the synthetic data were found to be consistent with those using the real data. Significant gender-based differences were observed in the summaries generated by the Google Gemma model, with physical and mental health mentioned significantly more in male summaries. Many of the same narrative-type words, such as “text,” “emphasise,” and “describe,” appeared more for women than men, while words relating to needs, such as “require,” “necessitate,” “assistance,” and “old,” appeared more for men. The synthetic data results also show no significant gender-based differences in the Llama 3 model output.

Perhaps a more concerning limitation of the stochastic nature of model output is the difficulty in balancing Type I and Type II error. With statistical tests performed for thousands of words, some unlikely events are inevitable. Caution was exercised by adjusting the p-values (using the Benjamini-Hochberg method), but this means that some words with very small unadjusted p-values were rejected. It is possible that some meaningful differences between words on the basis of gender were not considered statistically significant due to this conservatism.

A related point is that meaningful differences will not necessarily generate statistical significance. For instance, in the BART model, the word “unwise” appears 12 times for women and 5 times for men, which is not statistically significant according to a χ^2 test or Fisher’s exact test. However, even a single summary stating that a woman is making unwise decisions, where an identical man would not have been described the same way, could make a practical difference to a care professional acting upon it.

An additional limitation is that pre-trained sentiment analysis models not trained on health and care data were used. SiEBERT is a transfer learning model built on RoBERTa [406] and fine-tuned on a diverse range of data, including reviews and tweets [405]. Similarly, Regard is based on BERT [409] and fine-tuned on a dataset created for evaluating gender bias. Ideally, a domain-specific sentiment analysis model trained on care records would have been used. However, such a model does not exist, and creating one would not be trivial. Subjective judgement would be required to determine the relative polarity of different conditions or care needs. In the absence of such a model, the interpretation of sentiment results through the analysis of words and themes provides context and insight into the tone and content of care records. While the use of a general sentiment analysis model introduces limitations, the analysis of the language in these records offers valuable understanding of the differences between summaries created by LLMs. Future research could benefit from the development of domain-specific models, but the current approach provides meaningful exploration of these differences within the available framework.

Finally, cases relating to gender-specific care, such as mastectomies, and those mentioning domestic violence were removed, as they do not fit easily into the counterfactual fairness framework. However, the way language models treat gender-specific circumstances remains an important policy question, though one that cannot be answered using the methods in this paper.

8.7 Conclusion

LLM summarisation models are being used in health and care to generate and summarise documentation [15, 216, 126]. In this study, notable variation in gender-based discrepancies was observed across summarisation LLMs. Llama 3 showed no gender-based differences across any metrics, T5 and BART demonstrated some variation, and the Gemma model exhibited the most significant gender-based disparities. Gemma's male summaries were generally more negative in sentiment, and certain themes, such as physical health and mental health, were more frequently highlighted for men. The language used by Gemma for men was often more direct, while more euphemistic language was used for women. Women's health issues appeared less severe than men's in the Gemma summaries and details of women's needs were sometimes omitted. Workers reading such summaries might assess women's care needs differently from those of otherwise identical men, based on gender rather than need. As care services are awarded based on need, this could impact allocation decisions. While gendered language can be appropriate in contexts where gender is relevant, the differences in Gemma's output suggest that, in many instances, these differences are undesirable.

As generative models become more widely used for creating documentation, any bias within these models risks becoming part of official records. However, LLMs should not be dismissed as a solution to administrative burden. In this study, there were differences in bias across LLMs. This variation suggests that, if regulators wish to prioritise algorithmic fairness, they should mandate the measurement of bias in LLMs used in long-term care. Practical methods for evaluating gender bias in LLMs have been outlined in this paper, which can be implemented by anyone with access to long-term care data. The code for these evaluations is available on GitHub [419]. It is recommended that these or similar metrics be applied to assess bias across gender, ethnicity, and other legally protected characteristics in LLMs integrated into long-term care systems. By doing so, the benefits of LLMs can be realised while mitigating the risks associated with bias.

Supplementary information. Three appendices are included:

1. Evaluation of sentiment metrics: establishing which sentiment metrics are appropriate for conducting this analysis.
2. Model diagnostics and robustness checks: verifying the robustness of conclusions using several other methods.
3. Evaluation of themes: full lists of words counted in the frequency of the words appearing in each theme.

9 Discussion and conclusions

9.1 Research objectives and key findings

This thesis was motivated by the fundamental paradox I experienced in social work. Despite spending hours documenting care assessment, planning and provision, there was a striking lack of information to guide decision-making. I embarked on this thesis to find out whether computational natural language processing could help bridge this gap. Initially, the research focused on using NLP to sift through large amounts of unstructured data to extract information to evaluate the impact of individual needs and characteristics on service use. The goal was to enhance understanding of social care services using administrative data, to support more informed decision-making. The empirical findings confirmed that this was indeed possible; for instance, it is possible to successfully extract an indicator of loneliness from case notes. However, as the research progressed, an unexpected development occurred: LLMs began to be adopted within social care practice itself. These models were not just tools for retrospective analysis but were being used to alleviate the very administrative burdens that had initially motivated my research. This shift prompted me to expand the focus of the thesis to critically assess the implementation of LLMs in practice.

The overarching research questions of this thesis are whether LLMs can be used to improve decision-making in social care by increasing access to information held in care records, and what the potential challenges are associated with their use in generating, summarising and interpreting these records. In this concluding section, I will summarise the key findings from the empirical chapters, highlighting how LLMs can both improve our understanding of social care services, and discussing the implications of their use to support frontline practitioners. I will also discuss broader implications of LLMs for social care policy and practice, particularly in areas that I have been unable to evaluate as part of this thesis, and suggesting directions for future research. Finally, I will explore the tension that this work raises between regulation and innovation, and the challenges of legislating for emerging AI technologies, particularly in a field as sensitive and impactful as social care.

9.1.1 Can LLMs extract information from free text records to improve understanding of social care?

Extracting information from free text administrative records presents challenges, but LLMs offer a way to uncover information that can support evidence-based decision-making in social care. In this section, I outline the finding that achieving this requires substantial time and domain-knowledge, but the investment is justified by the value of information that cannot be obtained from other data sources.

9.1.1.1 The importance of human time and expertise

LLMs make it possible to extract information from social care data that was previously inaccessible due to its unstructured nature. However, this requires a substantial investment of human time and expertise. While the World Economic Forum's (WEF) 2023 report on LLMs and jobs suggests that certain roles, such as data entry clerks, may disappear due to automation, others, like database and network professionals, are expected to grow [421]. The work presented in this thesis indicates that, even with advanced AI tools, human involvement remains indispensable, particularly when dealing with complex data like social care records.

This research required significant data cleaning and an understanding of social care data's complexities. Before any analysis could take place, the free text data posed challenges due to over-redaction, necessitating several rounds of refinement. By working closely with software developers and the local authority, it was possible to balance the need to protect personal information with retaining relevant data for analysis. Regular testing ensured that the data remained compliant with data protection standards while being suitable for research purposes.

Establishing and improving data quality after extraction was another substantial undertaking. Over the course of around a year, weekly meetings with the local authority's Public Health Data Manager were essential to address missing or incomplete records and resolve inconsistencies in needs assessments, service use data, and costs. This collaborative effort led to the identification of missing assessments, and ultimately it was possible to supplement the originally extracted assessments with the missing data. Without the additional data, the analysis in Chapter 7 would have had much less statistical power. Of the dataset of 1,101 initial assessments, only 190 were completed on the originally extracted form.

Service cost data presented similar challenges. Initially omitted from the dataset, negotiation was required with the local authority to include it. Once obtained, the cost data needed extensive cleaning to correct inaccuracies such as services recorded with zero or missing costs. Overlapping service records had to be adjusted to prevent overestimation of costs. Cross-checking with external unit cost benchmarks were necessary steps to ensure the data's accuracy and reliability.

The theme of significant human time and knowledge investment extends to the analytical work in this thesis. Developing a model to extract indicators of loneliness or social isolation required understanding of how case notes are recorded and what information could be expected to be contained within them. Crafting a set of classification rules, determining appropriate models, writing the code, and manually classifying thousands of notes were time-intensive tasks. Additionally, establishing the construct validity of the extracted indicators involved analysing external survey data to compare and validate the results. Standardising structured administrative data also demanded considerable knowledge of the context in which they were created. Identifying relevant forms and questions, harmonising responses from different versions of needs assessment forms, and mapping different response scales to a consistent set of categories required a detailed understanding of the data collection process.

Overall, the time and expertise required for cleaning and standardising administrative data are substantial, regardless of whether the data is structured or unstructured. Each stage — from data pseudonymisation to resolving inconsistencies — required exploratory data analysis, domain knowledge, and regular meetings with the local authority partners. While LLMs and AI tools can assist in processing data and extracting insights, they do not eliminate the need for human involvement in data preparation and analysis. Human oversight and judgement remain essential to ensure the validity and reliability of research findings.

This experience aligns with the conclusions of Witham et al. [119], who discuss the significant investment of time and resources necessary to create "research-ready" datasets from administrative social care data. They highlight tasks such as cleaning data, understanding coding and categorisation, addressing missing data, and generating documentation. While the work in this thesis concurs with their findings, I extend this understanding by demonstrating that even with advanced tools like LLMs, these preparatory tasks cannot be automated. Instead, LLMs can facilitate analyses that were previously impractical, but the foundational work for such analyses still requires substantial human effort and expertise.

9.1.1.2 Advantages of administrative data over survey data

A key finding of this thesis is that, while there are many obstacles to extracting information with LLMs from administrative social care data, the information these records contain is unavailable in other sources. Although some data found in administrative records also appears in surveys, this thesis argues that findings from surveys may not accurately reflect the needs of statutory care users. Surveys frequently under-represent individuals with the greatest care needs because of exclusion criteria, attrition, and under-reporting. By contrast, administrative records provide comprehensive information about all individuals receiving statutory care services. Publicly funded care users in surveys have systematically lower needs than those appearing in the administrative records used in this thesis (see Chapter 5). These limitations make survey data less reliable for understanding high-need populations, underscoring the value of administrative records in social care research.

One of the key strengths of administrative data is its ability to capture real-time changes in individuals' conditions and care requirements, which surveys may miss due to infrequent data collection and reliance on self-reported information. Administrative records contain rich, unstructured information in free text case notes. This enables researchers to extract valuable insights that are not available through survey data, such as indicators of loneliness or social isolation, which are critical factors in understanding care needs but are often under-reported or omitted in surveys. Administrative records also include detailed, longitudinal data on functional abilities, care assessments, service use, and costs, allowing for a more accurate and nuanced understanding of the relationship between needs and service use than available from other data sources.

While processing administrative data requires significant time and expertise, the depth and breadth of information it provides make it an invaluable resource for research, practice and policy development. It allows for more comprehensive analyses, such as examining predictors of care home entry or evaluating the impact of interventions, which are not feasible with existing survey data due to their limitations. By using administrative data, we can gain a more accurate and comprehensive understanding of the needs of those receiving publicly funded care, ultimately better informing policy decisions.

9.1.1.3 LLMs for extracting loneliness and social isolation

The thesis demonstrates that LLMs can be used to extract meaningful information from unstructured free text in social care records. In Chapter 6, I show that LLMs are capable of identifying case notes that indicate loneliness and social isolation among older adults receiving care services. By comparing the performance of LLMs with traditional NLP techniques, I demonstrate that more complex language models are more accurate and effective in capturing meaning from the rich, complex information recorded in social care free text.

This use of LLMs can complement established statistical methods to enhance policy evaluation in social care. While LLM output can be used to generate descriptive statistics about the prevalence of loneliness across different groups, further value of this extracted information is realised when it is integrated with traditional econometric methods. In Chapter 7, I examine the impact of loneliness on the likelihood of older adults moving into care homes by incorporating the loneliness indicators extracted by the LLM into a survival model. The results indicate that loneliness is a significant predictor of care home entry, highlighting how variables derived from unstructured text can be instrumental in understanding and predicting important outcomes in social care.

This approach demonstrates that the value of information extracted with LLMs can increase when used in conjunction with existing statistical methods. The ability of LLMs to convert unstructured free text into structured data allows for the development of econometric models using variables that would otherwise be inaccessible. Furthermore, the successful application of LLMs in this context underscores the potential for these models to address data gaps in structured administrative data. Given the volume of free text records, the role of LLMs for extracting important information represents a significant opportunity for the evaluation of social care services.

9.1.2 What are the challenges associated with LLMs generating, summarising and interpreting care records?

This thesis argues that while LLMs have the potential to reduce administrative burdens in social care, their use requires scrutiny. The key challenges include accuracy, bias, and the broader implications of their deployment in decision-making processes. In Chapter 8, I focused on gender bias in LLMs used to summarise social care records. The paper contributes both a methodological framework for assessing gender bias in LLM-generated summaries, and empirical findings that highlight the importance of evaluating models used in social care. While the focus was on gender bias, the underlying concerns — bias across other social dimensions such as race or socioeconomic status, and the impact of these biases on decision-making — remain central to the responsible deployment of LLMs in social care.

The methodology developed in Chapter 8 offers a practical approach to evaluating gender bias in LLMs by generating gender-swapped case notes and comparing summaries. This allows us to pinpoint whether differences in summaries are driven by the original text or by biases in the model itself. Using this approach, I found that while Meta's Llama 3 summaries were not biased across various metrics, Google's Gemma model displayed significant gender-based disparities. In the development of the methodology I found that, as with discriminative LLMs, an understanding of social care was as important as technical proficiency in LLMs. For example, generating gender-swapped versions of case notes was not a straightforward process, and required the use of multiple models and the development of metrics to assess the outputs. Determining which records were appropriate for gender swapping, removing sensitive cases, and manually refining lists of relevant terms were necessary steps to ensure the validity of the analysis.

Additionally, while the methods of evaluating bias in Chapter 8 found gender-based differences using the counterfactual fairness framework, this is not the only measure of bias. Algorithmic bias in the context of free text is an emerging area. Berk et al. [422] present six types of algorithmic fairness, some of which are mutually incompatible (except in highly artificial circumstances). All of these types of fairness are metrics of whether demographic groups achieve the same outcomes. However, Grgic-Hlaca et al. [423] argue that machine learning algorithms should focus on process fairness: an algorithm's output should be considered fair if a user judges every feature included in the algorithm to be fair. This framework is best understood in the context of regression models or supervised machine learning, where the features (or independent variables) are selected manually. Examples of features in the paper include number of previous arrests and the age of the individual, to predict risk of reoffending.

Such a framework could not be applied to the 8 billion parameters in Meta's Llama 3 LLM. Nevertheless, it could be argued that evaluating text generated by LLMs is itself a process measure, and those who prefer outcomes-based measures might consider that gender-based differences in text output do not matter *if* they do not lead to gender-based differences in outcomes (whether these are defined in terms of service receipt or quality of life). There is no evidence into the impact of gender-based differences in LLM-generated care summaries, though more generally we know that language reinforces perceptions of gender [424], and the presentation of information in electronic case management systems affects decision-making [86, 87, 88]. Nevertheless, more research is required into whether the differences found in Chapter 8, such as the systematic under-representation of women's needs, lead to differences in prioritisation or allocation of resources and support.

Additionally, while the quantitative metrics in Chapter 8 can reveal gender-based differences where they exist, negative results can indicate that a model is preferable, but cannot definitively conclude that subtler types of bias do not exist. Furthermore, such methods cannot assess bias in the important areas where counterfactuals are not comparable, such as domestic violence. Research in this area continues to develop. Pfohl et al. [425] provide in their September 2024 paper a set of methods to assess bias in LLMs used in healthcare based on expert evaluation of model output, and the approach taken in Chapter 8 could be complemented by qualitative assessments.

The conceptual and methodological development of evaluating bias in LLMs in social care have implications for real-world practice. Future research could investigate other measures of fairness, based on process or outcomes. Additionally, further research could expand the evaluation of bias in the methods presented here to other characteristics, such as those related to race, disability, sexual orientation or socioeconomic status, or assess the practical impact of differences in summaries on social care decision-making. Finally, algorithmic fairness is not the only challenge that LLMs face. The evaluation of accuracy, which was beyond the scope of this thesis, is of critical importance. I discuss this below.

9.2 Limitations

9.2.1 Administrative data used in this research

The administrative data used in this research came from a single English local authority's records of older adults receiving long-term care. In Chapter 4, I demonstrate that this data was a comprehensive extract of the relevant records held by the authority, and in Chapter 5, I discuss how the authority in question is not an outlier. However, several limitations remain. While I was able to compare the administrative data to English survey data, the surveys appropriate for comparison were only those about older adults in England. Many limitations of survey methodologies, such as under-reporting, recruitment challenges among individuals with higher needs, and higher attrition rates among those with greater needs, are issues that persist internationally, and not just among older people [296, 297, 294]. Nevertheless, I could only empirically demonstrate these differences for older adults in England.

A limitation of this thesis is the use of administrative data from a single local authority, which inherently restricts the extent to which findings can be generalised. This authority is in an urban area, meaning its demographic profile differs from rural authorities, most notably in terms of ethnicity. According to ASCS data, this local authority also has below-average levels of need compared to some other authorities. This may further limit the generalisability of findings, although given that the main finding is that the needs of the area were higher than they appear from survey data, it is possible that the extent of under-representation of those with higher needs may be even greater in other areas. Ultimately, however, the extent to which findings from comparing this local authority's data to survey data will hold across others area is an empirical question that remains unanswered in this research. Further research across diverse regions would strengthen these findings and inform more comprehensive national policies.

Furthermore, although I argue for the use of administrative records due to the unique content they provide, there are drawbacks to this approach. While LLMs proved effective at extracting information from free text data, scaling this approach is challenging due to variations in record-keeping across local authorities and the difficulties in accessing this data. Data reliability is also an issue. Administrative records are not collected with research in mind, so data accuracy can be affected by various unobserved factors, including under-identification [130, 129, 55], or exaggeration of needs to align with service eligibility criteria [67], and investigating this is resource-intensive. While LLMs offer potential for new analyses, human expertise remains essential to accurately process and interpret the data. Lastly, the field of LLMs has evolved during the time I have written this thesis, and the rapid pace of advancements raises questions about the longevity of these findings. Future developments may influence the applicability of the findings discussed here, and I discuss this in detail in the next section.

9.2.2 Implications of rapid development of LLMs

An overarching theme of this thesis is the speed of adoption of LLMs. When I first conceived of this thesis in 2019, the term Large Language Model was not widely used and there were no serious suggestions that LLMs could be used to summarise or generate social care records. However, by 2024, such products were in use in several local authorities in England [126]. When I initially drafted the section in Chapter 3 on AI adoption in August 2024, four councils had published a privacy notice stating that LLMs may be used to process their social care data [218, 219, 220, 221]. By the time I redrafted it in November 2024, an additional five authorities had done so [222, 223, 224, 225, 226].

The analysis in Chapter 8 was driven by the apparent need for analysis of summarisation models that were being adopted. As described in Chapter 3, the LGA survey of English local authorities published in April 2024 contained four responses stating that they were using generative AI in adult social care [126]. An article in The Guardian in September 2024 stated that seven councils have adopted a particular social care LLM product, and two dozen more were piloting it [10]. There is considerable scope for expansion. The 2019 LGA report into social care IT identified two main suppliers of social care electronic databases [125], who were the only two vendors named in the 2024 DHSC report into streamlining social care assessments [127]. In 2024, while neither provider has yet released products integrating LLMs into their electronic case management systems, both have published articles on their websites describing the advantages of integrating LLMs into electronic case management systems [426, 427]. One such article describes how the company has created a prototype using speech-to-text models and GPT-4 to “bring transformational AI to our products with the goal of saving clinicians and patients time, increasing accuracy, and leveraging insights from conversations to improve experiences, pathways and potentially even treatment options” [427]. The other provider released in June 2024 their first tool integrating LLMs into their HR case management systems [428]. Although this is not directed at social care, it appears that both companies which dominate the social care case management market have the desire and capacity to integrate LLMs into their products. A 2024 study into the perspectives of commissioners and healthcare professionals on the use of digital tools found that the lack of integration of digital tools into existing IT systems was a barrier to their use [429]. AI tools appear set to become part of social care case management systems which suggests that, while in November 2024 LLMs are at their peak usage in adult social care, it is unlikely that this will remain the peak for long.

One of my findings relates to the time required to manually classify case notes in Chapter 6 as indicative of loneliness or social isolation, in order to train the model. I undertook most of the classification between January and July 2022, and there are particular implications for the longevity of the findings from this paper owing to the significant development in LLMs since. In this section, I explore in two ways how the progress in generative LLMs affects the findings from Chapter 6. Firstly, I examine the possibility of using generative LLMs for classifying loneliness. Secondly, I explore the use of generative LLMs to measure bias in the loneliness classification model.

9.2.2.1 Using generative LLMs for classification

During the course of this research, I identified in Chapter 6 a limitation that manual classification of documents was necessary due to the lack of sufficiently advanced language models. As I conclude this thesis in November 2024, advancements in LLMs have presented an opportunity to revisit this limitation. I explored whether the latest LLMs could now replicate or even surpass the manual classification performed earlier, potentially reducing the need for manual annotation.

To address this, I investigated the use of state-of-the-art generative models, such as Llama 3 [196] and Gemma [197], which have shown promising abilities in understanding complex instructions and generating structured outputs. By providing these models with the same instructions given to human annotators, I aimed to assess their capacity to perform the classification task autonomously.

The results, summarised in Table 9.1, indicate that Llama 3 achieved an F_1 score of 0.47 after being given the same written instructions as human annotators. Gemma, however, struggled to produce outputs in a format that could be reliably parsed into structured data and I present results for those that it parsed and of all sentences, both of which have an F_1 score lower than Llama 3. Interestingly, providing additional examples to Llama 3.1 [430], which allows for a longer context window, did not enhance performance and, in some cases, led to worse results. The detailed methodology and technical considerations are presented in the final Appendix.

Table 9.1: Loneliness LLM results

Model	Prompt	Sensitivity	Specificity	Accuracy	F1
Llama 3	Human instructions	0.81	0.69	0.71	0.47
Llama 3.1	Human + 100 examples	0.70	0.64	0.65	0.37
Llama 3.1	Human + 300 examples	0.75	0.70	0.71	0.43
Gemma	Human (of 3036 parsed)	0.46	0.84	0.79	0.41
Gemma	Human (of 3573 total)	0.46	0.71	0.67	0.28

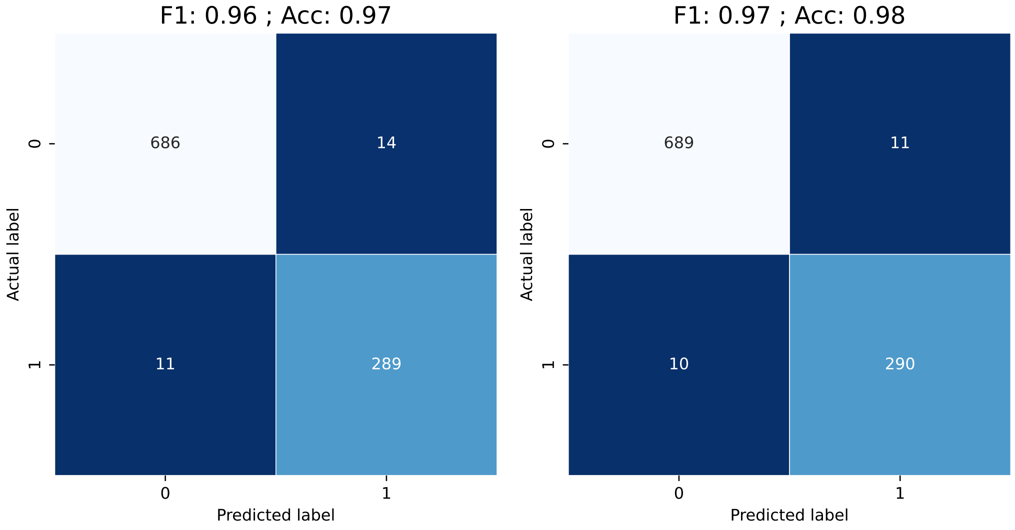
The scores from generative models are much lower than the F_1 score of 0.92 achieved in Chapter 6. These findings suggest that, despite rapid advancements, current LLMs are not yet capable of replacing the manual classification process. This reinforces the importance of human expertise in developing classification schemes that account for nuanced understanding of the data context. For example, determining loneliness involves complex rules based on eligibility criteria and tacit knowledge of service delivery, which are challenging for LLMs to interpret accurately.

While technological progress may eventually enhance the capabilities of LLMs in this regard, this exercise highlights that human judgement remains essential. Furthermore, even if generative LLMs were able to perfectly replicate human classifiers without training, the formulation of research questions, design of classification rules, and oversight of model outputs cannot yet be fully automated. For now, the conclusion persists that automating classification requires significant investment of human time and knowledge.

9.2.2.2 Using generative LLMs for evaluation of bias in classification models

In Chapter 6, a limitation was that I could not conclusively determine the extent to which observed gender differences in loneliness were due to real disparities, versus potential bias in the classification algorithm. For example, 45% of women were identified as lonely compared with 41% of men. While I compared the results to surveys and found similar disparities, I did not have individual-level structured loneliness administrative data to fully assess this aspect. With the advancements in LLMs by 2024, I have revisited this limitation using similar methods to those in Chapter 8.

I generated gender-swapped versions of 1,000 case notes, 300 of which indicated loneliness, and generated binary predictions for both, to evaluate potential gender bias in the loneliness classification model. The results demonstrated minimal differences, with 289 correct positive predictions for men and 290 for women. These findings, illustrated in Figure 9.1, suggest that the observed gender differences in the original data are likely reflective of actual patterns rather than bias introduced by the classification model. Detailed results and methodological specifics are provided in the final Appendix.



(a) Male

(b) Female

Figure 9.1: Loneliness model predictions (gender-swapped versions)

This analysis underscores how advancements in generative language models create new opportunities for evaluation of bias in classification tasks, offering tools that were not available during the initial research phase. It reinforces the validity of the findings presented in Chapter 6 and highlights the importance of continually reassessing limitations as new technologies emerge.

9.2.3 Scope of research

LLMs are increasingly discussed as tools to reduce administrative burdens in social care. The 2024 LOTI report, *Opportunities for AI in Adult Social Care Services*, identifies several potential applications of LLMs in this field, including generating case summaries, automating meeting transcriptions, predictive forecasting, creating easy-read documents for adults with learning disabilities, and developing chatbots to assist care users in locating services [14]. This thesis has focused on exploring only the first three of these possibilities, assessing LLMs for extracting information for use in predictive forecasting, and examining aspects of gender bias in summarisation models. Furthermore, even within these topics, I have only been able to examine specific facets. I extracted loneliness information from free text as a binary indicator, although loneliness itself is continuous. I used this indicator in a model predicting care home entry, which is not the only important dependent variable. It could also be a valuable addition to other models, such as predicting lifetime care costs, or quality of life. Furthermore, I have not evaluated LLMs for extracting other information that might be recorded in free text, such as economic circumstances, psychological needs or information related to abuse or neglect.

In addition to the limitations related to evaluation of bias discussed above, this thesis also could not consider model accuracy, an essential factor in assessing the utility of LLMs for summarising case records. This is an important direction for future research. I discuss such an analysis, along with other recommendations for future research directions, in the next part of this chapter.

9.3 Future directions for research into LLMs in social care

The various applications for LLMs mentioned in the LOTI report [14], such as creating easy-read documents and developing chatbots, represent promising areas for future research. While this thesis has focused on extracting information for predictive modelling and summarisation, it is an appealing prospect that LLMs might improve information for individuals with learning disabilities by generating easy-read materials, or for care users more generally through chatbots for navigating services. If such tools are developed, there is much scope for their evaluation. However, as LLMs for summarising documents have gained adoption and widespread attention over the course of this research, I explore here specific directions for evaluation of such models. This section examines how future studies could assess the utility of summarisation models, both in retrieving and condensing information from existing records and in generating new documentation. I discuss both assessing model accuracy, and quantifying the potential efficiency savings these models could bring to social care services.

9.3.1 Accuracy of LLM summarisation models

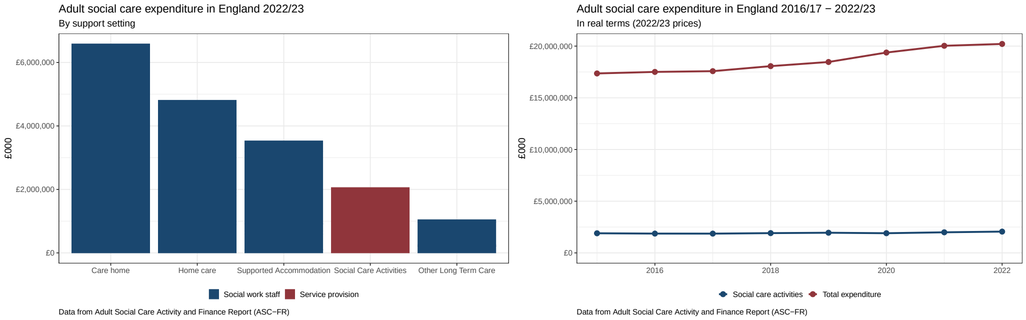
LLM summarisation models are being used in social care in England for information retrieval, by distilling large volumes of text into summaries for caseworkers [15]. The apparent benefit is clear, *if* such models can highlight relevant information, which might otherwise be missed, or reduce cognitive burden, which in turn can prevent errors [86]. LLM summarisation models are also in use to generate new documentation for care records, by summarising transcripts of audio recordings of care needs assessments [15, 10]. While there are clear resource benefits if such automation saves time, such savings are only technically efficiency savings if they lead to achieving the same outcomes with fewer resources [239]. LLM vendors report that LLMs can generate higher quality notes than those written by humans [e.g. 217, 216, 15]. This is consistent with evidence that most LLM-generated summaries of medical records are of higher quality than those generated by humans [208]. However, it is not yet clear how accurate LLM summarisation models are in long-term care, or how accuracy might be measured.

As discussed in Chapter 3, standard evaluation metrics for LLM summarization, such as BLEU, ROUGE, and BERTscore, are inadequate for assessing the accuracy of social care records. While Van Veen et al. [208] found that LLMs on average outperform humans in summaries of radiology reports, the salient information and the consequences of omission may be quite different with social work records which contain information about factors relevant to abuse and neglect. As harm is caused by workers being unaware of the content of their records [82], it would need to be demonstrated that LLMs are better than humans at extracting such information, and that this information being presented to the worker is effective. Microsoft and OpenAI have created some benchmarks for LLMs in the medical domain, but they ensure the model is presented with multiple choice rather than free text output, as the latter is too challenging to evaluate [431]. It would not have been possible for me in the course of this thesis to examine the question of accuracy of summarisation models in addition to data extraction, prediction and bias, as it is simply too large a question. However, it is not only the case that it is currently unknown whether LLMs can summarise social work records to an acceptable standard, but furthermore there are not yet any accepted quantitative metrics of such a standard. This is an important area of future research.

9.3.2 Quantifying the efficiency savings from LLM summarisation models

Demand for long-term care is projected to continue to increase globally [432, 367, 368]. In November 2024, 81% of councils in England reported they were on course to overspend on adult social care in 2024/24 (an increase from 72% the previous year) [26]. Local authorities currently spend around 10% of their adult social care budgets on care management [3]. The magnitude of efficiency savings from LLMs remains unclear. While some providers claim substantial time savings [e.g., 217], research is needed to evaluate these claims systematically. It is not yet established how much time such models save in practice or how their output compares to that generated by humans. Certain tasks, such as assessing capacity under the Mental Capacity Act 2005, cannot be automated, as they require professional judgement and discharge of legal duties. Additionally, not all documentation tasks lend themselves to automation equally. While summarising long narrative notes may save time, tasks involving frequent, short updates (e.g., phone call logs) may offer limited scope for automation (as discussed in Chapter 3).

Potential savings from LLMs must be considered within the broader context of adult social care spending, shown in Figure 9.2b, which has risen in real terms from £17.4 billion in 2015/16 to £20.2 billion in 2022/23 [3]. Demographic trends suggest that public spending on adult social care could increase by 159% by 2040, reaching approximately £45.1 billion in 2022/23 prices, driven by a 66% rise in the number of individuals receiving publicly funded long-term care and higher levels of need [4]. This projected growth underscores the pressing need for tools that can improve efficiency and alleviate administrative burdens. If LLMs can achieve meaningful time savings, they offer the potential for resource-strapped social care services to manage rising demand without cutting care [e.g., 212, 208]. Future research into quantifying the efficiency savings from summarisation models should account for demographic changes, which will impact the overall cost of care and the amount of time spent on care management.



(a) Snapshot (2022/23)

(b) Trend (2016 - 2023)

Figure 9.2: Care management as a proportion of total adult social care expenditure

However, realising such savings requires research to quantify the time saved and evaluate the quality of LLM-generated outputs compared to human-generated notes. Equally important is a comprehensive understanding of how these models are implemented in practice [242]. Many innovative products in social care fail to scale beyond pilot projects due to factors such as infrastructure, organisational readiness, and sociocultural barriers [240, 433, 434, 435]. Research on savings must focus on adoption as well as technical metrics of accuracy. Furthermore, the adoption of health and care technologies is heavily influenced by broader policy, legal, and regulatory contexts [240], which I discuss in the next section.

9.4 The balance between regulatory oversight and innovation

The regulatory framework shapes the landscape of using administrative records. It is challenging for policymakers to determine the appropriate level of oversight without inhibiting innovation. One of the challenges about scalability of the use of administrative records for predictive modelling, as in Chapter 6 and Chapter 7, is the difficulty for researchers in accessing records. On the other hand, the evaluation in Chapter 8, which demonstrates that LLMs can exhibit gender bias when used to summarise social care case notes, is not currently required in models adopted in social care practice in the UK. Legislating in this fast-changing landscape is complex. The EU's AI Act, scheduled for full implementation by 2026, is an extensive framework for managing AI risks, especially in areas considered high risk, like social care [92]. The US, by contrast, lacks federal AI-specific legislation, though sector-specific standards are emerging. The UK is positioned between these approaches, with fairly comprehensive data protection legislation, but yet to develop an AI policy direction. In this section I will briefly outline the current state of AI regulation in social care in the UK and contrast this with the EU and US.

9.4.1 Current regulatory landscape

The regulatory landscape for LLMs in social care is shaped by frameworks that predate their adoption.[1] The UK General Data Protection Regulation (GDPR), derived from the EU GDPR, has implications for social care researchers using administrative records. GDPR defines profiling as any form of automated processing of personal data to evaluate certain aspects relating to an individual [438, 439]. Predictive modelling, such as assessing the risk of care home entry, falls within this definition and is permitted without consent if justified by a legitimate public interest. It was only due to the development of specialist pseudonymisation software for this project that the research undertaken in this thesis was lawful without opt-in consent. GDPR may have other implications for the use of LLMs in social care practice. For example, inaccurate data caused by "hallucination" in LLMs is subject to the right to rectification [440]. However, GDPR does not directly address the use of LLMs, which proliferated after it was enacted, and does not provide a legal mandate for the assessment of (or protection against) algorithmic bias [441].

Automated decision-making, such as models directly determining service eligibility, is prohibited by GDPR without explicit consent, unless human oversight is "meaningful" — a threshold yet to be legally defined [442]. The legal ramifications of algorithmic decision-making has not been created by the advent of AI models. The use of hand-filled, tick-box application forms for social entitlements is unlawful if they automatically determine eligibility without human discretion [443], and since 1970 in the UK it has been unlawful for public bodies to have policies which make decisions automatically, preventing them from exercising their judgement [444]. Nevertheless, technological innovation in social care creates more oppor-

[1] The regulatory landscape for healthcare LLMs also predates their development, but the legislation is more stringent. LLMs used to create summaries of health records would be subject to the Medical Device Regulations 2002 [436], which requires conforming to the Medicines and Healthcare products Regulatory Agency (MHRA) rules about reducing risks, demonstrating standards through clinical trials, and monitoring of adverse events [437]. Similar products in social care are not subject to this legislation.

tunities for automated decision-making, including perhaps the suggestion in the LOTI report that AI could be used for triage and reducing waiting lists [14].

Emerging regulatory proposals suggest a shift in focus. The UK Data (Use and Access) Bill introduced in 2024 aims to ease restrictions on consent requirements for scientific research, balancing privacy protections with the need for innovation [445]. Similarly, the UK government announced plans for new AI regulations addressing the development of the most powerful models [446]. At the time of writing, the specifics of these proposals remain unclear.

9.4.2 Alternative regulatory approaches

The European Union (EU) has taken a proactive stance on regulation with the EU AI Act, approved by the Council in May 2024 [396]. The Act, which aims to create a comprehensive legal framework for AI technologies, classifies AI systems into four risk categories, based on the "intended purpose" of the AI product [92, 395]. General-purpose LLMs are not inherently classified as high risk [396]. However, their integration into specific applications in social care could elevate them to high-risk status if they influence decisions about individual wellbeing or determine eligibility for services [395, 447].

High-risk classification under the EU AI Act does not prohibit the use of such technologies but imposes stringent requirements, including continuous risk assessments, transparency duties, and meaningful human oversight [395]. These requirements are designed to ensure that the risks posed by high-risk applications are weighed against their potential benefits and that they are effective for their intended purposes [395].

In contrast to the EU's comprehensive approach, the United States currently lacks a unified federal regulatory framework for AI. In October 2023, President Joe Biden issued an Executive Order directing federal agencies to develop standards and guidelines for AI governance [393]. While this initiative addresses issues like privacy, bias, and transparency it remains a directive rather than enforceable legislation [448]. Regulation of AI in healthcare in the US, where long-term care is generally considered a healthcare component [449], falls under existing frameworks. The Food and Drug Administration (FDA) published a report in 2024 acknowledging the limitations of its current framework and committing to the development of updated AI-specific regulations [450]. The political landscape in the US further complicates the regulatory outlook.

9.4.3 From the technical to the political

Much of the current evaluation and regulation of LLMs is focused on their technical performance rather than the inherently political aspects relating to balancing economic priorities with ethics, transparency, fairness, and accountability [451, 452]. While stricter regulations may hinder innovation [453], they offer greater protections against bias and harm. Adopting a framework similar to healthcare regulations for LLMs in social care could enhance safeguards for service users. However, such measures may also slow adoption. Policy must navigate these trade-offs carefully. The political nature of these decisions can be seen in the EU AI Act, which appears to have been designed to give the EU the authority to investigate the largest models created by US-based tech giants [454].

It is quite possible that we will see further divergence between the US and EU from 2025. In the 2016-2020 administration, the Trump White House stated that it had "taken important action to remove barriers to AI innovation in healthcare" [455]. The largest LLM providers — Google, Meta, Microsoft, and OpenAI — are US-based, while many AI start-ups in the EU seek to challenge this dominance [456]. President-elect Donald Trump has indicated plans to reduce regulation. The Republican Party manifesto for the 2024 election promised to "repeal Joe Biden's dangerous Executive Order that hinders AI Innovation, and imposes Radical Left-wing ideas on the development of this technology", and proposed replacing it with a framework "rooted in Free Speech" [457]. Much has been made of the protectionist rhetoric from the US Republicans, such as proposed tariffs [458]. Analysts speculate that the EU AI Act is at risk of attack by a Trump administration, on the basis that it causes US companies a competitive disadvantage [459].

At the time of writing, in November 2024, it is unclear the extent to which this will affect the regulatory gap between the US and the EU. However, the speculation triggered by a new US president underscores the extent to which questions of AI regulation are relevant, contested and political. Nevertheless, so far this appears to be a political decision that UK governments have deferred. This regulatory inaction could be seen as *de facto* favouring innovation over algorithmic fairness and transparency in social care.

9.5 Overall conclusion

When I started this PhD, I could not have predicted how quickly LLMs would develop and be adopted in social care. This research shows that LLMs have the potential to make sense of administrative data, helping to address challenges like overwhelming amounts of data. Improved information could support those making funding decisions. However, some generative LLMs can introduce gender bias. The thesis evaluates important uses of discriminative and generative LLMs in social care, but much more research is needed. It is unknown whether salient information other than loneliness or social isolation can be reliably extracted from free text records, such as psychological circumstances, interactions with care services, or risk of abuse. Additionally, analysis is required of bias related to characteristics other than gender in LLMs used in social care, especially those protected by law. Research is needed into adoption — it is unclear which models are in use, whether they are the most effective or even how effectiveness should be defined. While the methods to measure accuracy in text classification are well-established, it is much harder to assess how well LLMs perform when they generate text. Companies claim their models are better than humans, but this is still an open question. Furthermore, while the claims that LLMs will save money seem more plausible, we do not yet know how much. This thesis has answered some key questions about the use of LLMs in social care, and highlighted areas where more research is needed. However, researchers alone cannot address all the challenges. While more work is needed to understand issues like accuracy, bias, and adoption, models are already being used in England without these factors being fully assessed. Policymakers and research funders have a crucial role in ensuring these unanswered questions are examined, and that LLMs in social care are implemented in a way which balances economic priorities with ethics, transparency, and fairness.

10 References

[1] Sara Guerschanik Calvo. The global financial crisis of 2008-10: A view from the social sectors. *UNDP-HDRO Occasional Papers*, (2010/18), 2010.

[2] Kate Ogden and David Phillips. How have English councils' funding and spending changed? 2010 to 2024. Technical report, IFS Report, 2024.

[3] Department for Health and Social Care (DHSC). Adult Social Care Activity and Finance Report: December 2023 release, 2023. URL https://digital.nhs.uk/data-and-information/publications/statistical/adult-social-care-activity-and-finance-report/2022-23. Accessed: 2024-08-14.

[4] Raphael Wittenberg, Bo Hu, and Ruth Hancock. Projections of demand and expenditure on adult social care 2015 to 2040. 2018.

[5] Tine Rostgaard, Frode Jacobsen, Teppo Kröger, and Elin Peterson. Revisiting the Nordic long-term care model for older people—still equal? *European Journal of Ageing*, 19(2):201–210, 2022.

[6] Ellen Grootegoed and Diana Van Dijk. The return of the family? Welfare state retrenchment and client autonomy in long-term care. *Journal of Social policy*, 41(4):677–694, 2012.

[7] Bent Greve. Long-term Care for the Elderly in Europe. *Development and prospects*, 2017.

[8] Michael Lipsky. *Street-level bureaucracy: Dilemmas of the individual in public service*. Russell Sage Foundation, 2010. ISBN 0871545446.

[9] T. Lillis, Maria Leedham, and A. Twiner. Time, the Written Record, and Professional Practice: The Case of Contemporary Social Work. *Written Communication*, 37:431–486, 2020. doi: 10.1177/0741088320938804.

[10] Robert Booth. Social workers in England begin using AI system to assist their work. *The Guardian*, Sep 2024. URL https://web.archive.org/web/20241007131554/https://www.theguardian.com/society/2024/sep/28/social-workers-england-ai-system-magic-notes. Accessed via Wayback Machine.

[11] Sam Rickman. Evaluating gender bias in Large Language Models in long-term care, October 2024. URL https://doi.org/10.21203/rs.3.rs-5166499/v2. PREPRINT (Version 2) available at Research Square.

[12] Sam Rickman. Using Machine Learning to Understand Loneliness in English Long-term Care Users from Free Text Case Notes. https://github.com/samrickman/lonelinessmodel, 2024. Accessed: 2024-09-06.

[13] Sam Rickman. Evaluating Gender Bias in LLMs in Long-Term Care. https://github.com/samrickman/evaluate-llm-gender-bias-ltc, 2024. Accessed: 2024-09-06.

[14] London Office of Technology and Innovation. Opportunities for AI in Adult Social Care Services. Technical report, July 2024. URL https://web.archive.org/web/20240906182858/https://loti.london/wp-content/uploads/2024/07/PUBLIC_-Faculty_LOTI_-Adult-Social-Care-services_-AI-opportunities-report-Final.pdf. Accessed: 2024-09-06.

[15] Local Government Assocation. Artificial intelligence use cases, 2024. URL https://web.archive.org/web/20240904192138/https://www.local.gov.uk/our-support/cyber-digital-and-technology/artificial-intelligence-hub/artificial-intelligence-use. Accessed: 2024-09-04.

[16] Sowmiya Moorthie, Shabina Hayat, Yi Zhang, Katherine Parkin, Veronica Philips, Amber Bale, Robbie Duschinsky, Tamsin Ford, and Anna Moore. Rapid systematic review to identify key barriers to access, linkage, and use of local authority administrative data for population health research, practice, and policy in the United Kingdom. *BMC Public Health*, 22(1):1263, 2022.

[17] Google Trends. Google Trends: Large Language Models, Natural Language Processing, Computational Linguistics, LLMs. https://trends.google.com/trends/explore?date=all&q=large%20language%20models,natural%20language%20processing,computational%20linguistics,LLMs&hl=en-GB, 2024. Accessed: 2024-08-24.

[18] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021.

[19] Department of Health and Social Care (DHSC). Adult Social Care in England, Monthly Statistics: October 2024. Official Statistics, October 2024.

URL https://web.archive.org/web/20241127184530/https://www.gov.uk/government/statistics/adult-social-care-in-england-monthly-statistics-october-2024/adult-social-care-in-england-monthly-statistics-october-2024. Published 3 October 2024.

[20] Niels Peek, Mark Sujan, and Philip Scott. Digital health and care: emerging from pandemic times. *BMJ health & care informatics*, 30(1), 2023.

[21] Department for Health and Social Care (DHSC). Evidence review for Adult Social Care Reform, 2021. URL https://web.archive.org/web/20240915000000*/https://assets.publishing.service.gov.uk/media/61a7bc688fa8f503780c1c79/evidence-review-for-adult-social-care-reform.pdf. Accessed: 2024-11-02.

[22] Department of Health and Social Care (DHSC). People at the heart of care: Adult social care reform white paper, 2021.

[23] Department of Health and Social Care (DHSC). Data saves lives: reshaping health and social care with data. 2022.

[24] Department for Health and Social Care (DHSC). Digitising Social Care Programme: SRO appointment letter, 2024. URL https://www.gov.uk/government/publications/digitising-social-care-programme-sro-appointment-letter/digitising-social-care-programme-sro-appointment-letter. Accessed: 2024-11-02.

[25] Local Government Association (LGA). Social Care Digital Innovation Programme, 2024. URL https://web.archive.org/web/20240720130421/https://www.local.gov.uk/our-support/partners-care-and-health/care-and-health-improvement/informaticslocal-investment-programme. Accessed: 2024-11-02.

[26] Association of Directors of Adult Social Services. ADASS Autumn Survey 2024, November 2024. URL https://www.adass.org.uk/documents/adass-autumn-survey-2024/. Accessed: 2024-11-13.

[27] Bleddyn Davies and Jose Fernandez. *Equity and efficiency policy in community care: needs, service productivities, efficiencies and their implications*. Routledge, 2000. ISBN 135176263X.

[28] Martin R J Knapp. *The economics of social care*. Macmillan International Higher Education, 1984. ISBN 1349177083.

[29] Bleddyn Davies and Jose Fernandez. *Equity and efficiency policy in community care: needs, service productivities, efficiencies and their implications*. Routledge, 2018.

[30] Juliette Malley and José-Luis Fernandez. Measuring quality in social care services: theory and practice. *Annals of public and cooperative economics*, 81(4):559–582, 2010. ISSN 1370-4788.

[31] Derek King and Raphael Wittenberg. Data on adult social care. 2015.

[32] Danielle Collingridge Moore and Barbara Hanratty. Out of sight, out of mind? A review of data available on the health of care home residents in longitudinal and nationally representative cross-sectional studies in the UK and Ireland. *Age and ageing*, 42(6):798–803, 2013.

[33] Victoria Shepherd, Fiona Wood, Richard Griffith, Mark Sheehan, and Kerenza Hood. Protection by exclusion? The (lack of) inclusion of adults who lack capacity to consent to research in clinical trials in the UK. *Trials*, 20:1–8, 2019.

[34] Isabel Green, Daniel Stow, Fiona E Matthews, and Barbara Hanratty. Changes over time in the health and functioning of older people moving into care homes: analysis of data from the English Longitudinal Study of Ageing. *Age and ageing*, 46(4):693–696, 2017.

[35] Department of Health and Social Care (DHSC). Next steps to put People at the Heart of Care, 2023. URL https://web.archive.org/web/20240118104311/https://www.gov.uk/government/publications/adult-social-care-system-reform-next-steps-to-put-people-at-the-heart-of-care/next-steps-to-put-people-at-the-heart-of-care.

[36] Catherine Hakim. Research Based on Administrative Records. *The Sociological Review*, 31(3):489–519, 1983. doi: 10.1111/j.1467-954X.1983.tb00905.x. URL https://doi.org/10.1111/j.1467-954X.1983.tb00905.x.

[37] Martha Snow, Wagner Silva-Ribeiro, Mary Baginsky, Sonya Di Giorgio, Nicola Farrelly, Cath Larkins, Karen Poole, Nicole Steils, Joanne Westwood, and Juliette Malley. Good Practice in the Implementation of Electronic Care Records in Adult Social Care: A Rapid Scoping Review. Unpublished manuscript, September 2024.

[38] D Challis, C Xie, J Hughes, P Clarkson, S Davies, and K Stewart. Resource Allocation at the Micro level in Adult Social Care: A Scoping Review, 2016.

[39] Ruth Robertson, Sarah Gregory, and Joni Jabbal. The social care and health systems of nine countries. *Commission on the future of health and social care in England.: The King's Fund*, 2014.

[40] Jill Doner Kagle and Sandra Kopels. *Social work records.* Waveland Press, 2008.

[41] Emma Miller and Karen Barrie. Setting the Bar for Social Work in Scotland. 2022.

[42] Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. Tethered to the EHR: primary care physician workload assessment using EHR event log data and time-motion observations. *The Annals of Family Medicine*, 15(5): 419–426, 2017.

[43] Rosemary Ricciardelli, Marina Carbonell, Lorna Ferguson, and Laura Huey. "It's frustrating… I didn't join to sit behind a desk": Police paperwork as a source of organizational stress. *International Journal of Police Science & Management*, 25(4):516–528, 2023.

[44] RV Ericson. *Policing the risk society.* University of Toronto Press, 1997.

[45] Laura Huey, Lorna Ferguson, and Rosemary Ricciardelli. "It's all about covering your…": The unintended consequences of police accountability measures. *Criminology & Criminal Justice*, 24(3):585–607, 2024.

[46] Marilyn Strathern. 'Improving ratings': audit in the British University system. *European review*, 5(3):305–321, 1997.

[47] Peter Woelert. Administrative burden in higher education institutions: A conceptualisation and a research agenda. *Journal of Higher Education Policy and Management*, 45(4):409–422, 2023.

[48] Nathalie Delli-Colli, Nicole Dubuc, Réjean Hébert, and Marie-France Dubois. Measuring social-work activities with older people. *Practice*, 25 (5):281–296, 2013.

[49] Keith Wrenn, Lance Rodewald, Eileen Lumb, and Corey Slovis. The use of structured, complaint-specific patient encounter forms in the emergency department. *Annals of emergency medicine*, 22(5):805–812, 1993.

[50] Tanya Humphreys, Frances S Shofer, Sheldon Jacobson, Christos Coutifaris, and Annette Stemhagen. Preformatted charts improve documentation in the emergency department. *Annals of emergency medicine*, 21(5):534–540, 1992.

[51] Richard N Shiffman, Cynthia A Brandt, and Bruce G Freeman. Transition to a computer-based record using scannable, structured encounter forms. *Archives of pediatrics & adolescent medicine*, 151(12):1247–1253, 1997.

[52] Department of Health and Social Care (DHSC). National Service Framework for Older People, 2001.

[53] Department of Health and Social Care (DHSC). Common Assessment Framework for Adults: Consultation on proposals to improve information sharing around multi-disciplinary assessment and care planning, Jan 2009. URL https://webarchive.nationalarchives.gov.uk/ukgwa/20130107105354/http://www.dh.gov.uk/prod_consum_dh/groups/dh_digitalassets/documents/digitalasset/dh_093715.pdf. Accessed: 2024-11-22.

[54] Department of Health and Social Care (DHSC). Health and Care Services for Older People: Overview report on research to support the National Service Framework for Older People, Oct 2008. URL https://webarchive.nationalarchives.gov.uk/ukgwa/20091105154425/http://www.dh.gov.uk/en/Publicationsandstatistics/Publications/DH_088848. Accessed: 2024-11-22.

[55] P Clarkson, M Abendstern, C Sutcliffe, J Hughes, and D Challis. Reliability of needs assessments in the community care of older people: impact of the single assessment process in England. *Journal of public health*, 31(4):521–529, 2009.

[56] David Challis, Michele Abendstern, Paul Clarkson, Jane Hughes, and Caroline Sutcliffe. Comprehensive assessment of older people with complex care needs: the multi-disciplinarity of the Single Assessment Process in England. *Ageing & Society*, 30(7):1115–1134, 2010.

[57] Jose-Luis Fernandez, Tom Snell, and Joanna Marczak. An assessment of the impact of the Care Act 2014 eligibility regulations. 2015.

[58] Philip Gillingham. Practitioner perspectives on the implementation of an

electronic information system to enforce practice standards in England. *European Journal of Social Work*, 24(5):761–771, 2021.

[59] Charles P Schade, Frank M Sullivan, Simon De Lusignan, and Jean Madeley. e-Prescribing, efficiency, quality: lessons from the computerization of UK family practice. *Journal of the American Medical Informatics Association*, 13(5):470–475, 2006.

[60] Judith Burton and Diane Van den Broek. Accountable and countable: Information management systems and the bureaucratization of social work. *British journal of social work*, 39(7):1326–1342, 2009.

[61] Jenni-Mari Räsänen. Emergency social workers navigating between computer and client. *The British Journal of Social Work*, 45(7):2106–2123, 2015.

[62] Merryn Gott, Sarah Barnes, Sheila Payne, Chris Parker, David Seamark, Salah Gariballa, and Neil Small. Patient views of social service provision for older people with advanced heart failure. *Health & social care in the community*, 15(4):333–342, 2007.

[63] Penny Rhodes, Mark Langdon, Emma Rowley, John Wright, and Neil Small. What does the use of a computerized checklist mean for patient-centered care? The example of a routine diabetes review. *Qualitative Health Research*, 16(3):353–376, 2006.

[64] Eileen Munro. *The Munro review of child protection: Final report, a child-centred system*, volume 8062. The Stationery Office, 2011.

[65] Philip Gillingham and Timothy Graham. Designing electronic information systems for the future: Social workers and the challenge of New Public Management. *Critical Social Policy*, 36(2):187–204, 2016.

[66] Seth M Powsner, Jeremy C Wyatt, and Patricia Wright. Opportunities for and challenges of computerisation. *The lancet*, 352(9140):1617–1622, 1998.

[67] Liz O'Rourke. *Recording in Social Work: Not just an administrative task.* Policy Press, 2010.

[68] Department for Health and Social Care (DHSC). Care data matters: a roadmap for better adult social care data, 2023. URL https://www.gov.uk/government/publications/care-data-matters-a-roadmap-for-better-adult-social-care-data. Accessed: 2024-08-14.

[69] Katri Ylönen. The use of Electronic Information Systems in social work. A scoping review of the empirical articles published between 2000 and 2019. *European Journal of Social Work*, 26(3):575–588, 2023.

[70] H David Stein, Prakash Nadkarni, Joseph Erdos, and Perry L Miller. Exploring the degree of concordance of coded and textual data in answering clinical queries from a clinical data repository. *Journal of the American Medical Informatics Association*, 7(1):42–54, 2000.

[71] Elizabeth Ford, John A Carroll, Helen E Smith, Donia Scott, and Jackie A Cassell. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association*, 23(5):1007–1015, 2016.

[72] Clay Graybeal. Strengths-based social work assessment: Transforming the dominant paradigm. *Families in society*, 82(3):233–242, 2001.

[73] Madeleine Stevens, Michael Clark, Jessica Carlisle, Nicola Brimblecombe, and Miranda MacGill. Supporting Meaningful Implementation and Evaluation of Strengths-Based Approaches in Adult Social Care: A Theory of Change for The Three Conversations. *The British Journal of Social Work*, page bcae055, 2024.

[74] Think Local Act Personal (TLAP). Developing a Well-Being and Strengths Based Approach to Social Work Practice. *London: TLAP*, 2016.

[75] Partners 4 Change (PFC). London Borough of Harrow The Three Conversations Story of Change, 2021. URL http://partners4change.co.uk/wp-content/uploads/2021/08/3C-Harrow-Story-of-Change.pdf. [Accessed July 26th, 2023].

[76] Colin Slasberg and Peter Beresford. Strengths-based practice: social care's latest Elixir or the next false dawn? *Disability & Society*, 32(2):269–273, 2017.

[77] L. Haynes. Senior leader buy-in critical to success of strengths-based social work, says government guidance. Community Care. Available online at: https://www.communitycare.co.uk/2019/02/24/senior-leader-buy-critical-success-strengths-based-working-says-government-guidance/ (accessed February 20, 2024), February 2019.

[78] Lieve Bradt, Rudi Roose, Maria Bouverne-De Bie, and Maarten De Schryver. Data recording and social work: From the relational to the social. *British Journal of Social Work*, 41(7):1372–1382, 2011.

[79] Natalie Ames. Social work recording: A new look at an old issue. *Journal of Social Work Education*, 35(2):227–237, 1999.

[80] Jessica S Ancker, Alison Edwards, Sarah Nosal, Diane Hauser, Elizabeth Mauer, Rainu Kaushal, and With the HITEC Investigators. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC medical informatics and decision making*, 17: 1–9, 2017.

[81] Walid F Gellad, Donald Yealy, and Michael Fine. Computers and the diagnosis of pneumonia: comment on "performance and utilization of an emergency department electronic screening tool for pneumonia". *JAMA internal medicine*, 173(8):701–702, 2013.

[82] Michael Preston-Shoot, Suzy Braye, Oli Preston, Karen Allen, and Kate Spreadbury. Analysis of safeguarding adult reviews April 2017–March 2019: findings for sector-led improvement. *Local Government Association: https://www. local. gov. uk/publications/analysis-safeguarding-adult-reviewsapril-2017-march-2019*, 2020.

[83] Jens G Pohl. Transition from data to information. *Collaborative agent design research center technical report-RESU72*, page 1, 2001.

[84] Sohn Nijor, Gavin Rallis, Nimit Lad, and Eric Gokcen. Patient safety issues from information overload in electronic medical records. *Journal of Patient Safety*, 18(6):e999–e1003, 2022.

[85] John W Beasley, Tosha B Wetterneck, Jon Temte, Jamie A Lapin, Paul Smith, A Joy Rivera-Rodriguez, and Ben-Tzion Karsh. Information chaos in primary care: implications for physician performance and patient safety. *The Journal of the American Board of Family Medicine*, 24(6):745–751, 2011.

[86] Adil Ahmed, Subhash Chandra, Vitaly Herasevich, Ognjen Gajic, and Brian W. Pickering. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. *Critical Care Medicine*, 39(7):1626–1634, July 2011. doi: 10.1097/CCM.0b013e31821858a0.

[87] Yalini Senathirajah, David Kaufman, and Suzanne Bakken. User-composable electronic health record improves efficiency of clinician data viewing for patient case appraisal: a mixed-methods study. *eGEMs*, 4(1), 2016.

[88] David Wastell and Sue White. Making sense of complex electronic records: Socio-technical design in social care. *Applied Ergonomics*, 45(2):143–149, 2014.

[89] Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*, 2020.

[90] Julius Steen and Katja Markert. Investigating gender bias in news summarization. *arXiv preprint arXiv:2309.08047*, 2023.

[91] Michela Menegatti and Monica Rubini. Gender bias and sexism in language. In *Oxford research encyclopedia of communication*. 2017.

[92] European Commission. Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts: Main Articles, 2024.

[93] H Qureshi and E Nicholas. A new conception of social care outcomes and its practical use in assessment with older people. *Research, Policy and Planning*, 19(2):11–26, 2001.

[94] Department of Health and Social Care (DHSC). Care and support statutory guidance, 2024. URL https://www.gov.uk/government/publications/care-act-statutory-guidance/care-and-support-statutory-guidance.

[95] Len Doyal. Human need and the moral right to optimal community care. In *Community Care: A Reader*, pages 185–195. Springer, 1997.

[96] Len Doyal and Ian Gough. A theory of human needs. *Critical social policy*, 4(10):6–38, 1984.

[97] Nancy Wolff. Using randomized controlled trials to evaluate socially complex services: problems, challenges and recommendations. *The Journal of Mental Health Policy and Economics*, 3(2):97–109, 2000.

[98] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2024. URL https://www.R-project.org/.

[99] Jaclyn LF Bosco, Rebecca A Silliman, Soe Soe Thwin, Ann M Geiger, Diana SM Buist, Marianne N Prout, Marianne Ulcickas Yood, Reina Haque, Feifei Wei, and Timothy L Lash. A most stubborn bias: no adjustment method fully resolves confounding by indication in observational studies. *Journal of clinical epidemiology*, 63(1):64–74, 2010.

[100] Caroline Sutcliffe, Jane Hughes, Michele Abendstern, Paul Clarkson, Helen Chester, and David Challis. An examination of assessment arrangements and service use for older people in receipt of care management. *Care Management Journals*, 15(2):66–75, 2014.

[101] Sarah Abdi, Alice Spann, Jacinta Borilovic, Luc de Witte, and Mark Hawley. Understanding the care and support needs of older people: a scoping review and categorisation using the WHO international classification of functioning, disability and health framework (ICF). *BMC geriatrics*, 19:1–15, 2019.

[102] British Association of Social Work (BASW). UK Supervision Policy, 2011. URL https://basw.co.uk/sites/default/files/resources/basw_73346-6_0.pdf. Accessed: 2024-08-14.

[103] Jose-Luis Fernandez, Juliette Malley, Joanna Marczak, Tom Snell, Raphael Wittenberg, Derek King, and Gerald Wistow. Unmet social care needs in England. *CPEC Working Paper 7*, 2020.

[104] Department for Work & Pensions. Family Resources Survey: financial year 2021 to 2022, 2024. URL https://web.archive.org/web/20240303153450/https://www.gov.uk/government/statistics/family-resources-survey-financial-year-2021-to-2022/family-resources-survey-financial-year-2021-to-2022. Accessed: 2024-03-03.

[105] Understanding Society. Are institutional populations (e.g. students in university halls, individuals in care homes, people in prison) included in the Study?, 2024. URL https://web.archive.org/web/20240828114359/https://www.understandingsociety.ac.uk/help/faqs/are-institutional-populations-e-g-students-in-university-halls-individuals-in-care-homes-people-in-prison-included-in-the-study/. Accessed: 2024-08-28.

[106] George Stoye and Ben Zaranko. How accurate are self-reported diagnoses? Comparing self-reported health events in the English Longitudinal Study of Ageing with administrative hospital records. Technical report, IFS Working Papers, 2020.

[107] Department for Health and Social Care (DHSC). Adult Social Care Activity and Finance Report, 2023. URL https://digital.nhs.uk/data-and-information/publications/statistical/adult-social-care-activity-and-finance-report/. Accessed: 2024-08-14.

[108] JR Kemm, J Robinson, and J Verne. Social care data in England: What they tell us and what they do not tell us. *Public Health*, 124(5):265–268, 2010.

[109] J Banks, G.D Batty, J.J.F Breedvelt, K Coughlin, R Crawford, M Marmot, J Nazroo, Z Oldfield, N Steel, A Steptoe, Martin Wood, and P. Zaninotto. English Longitudinal Study of Ageing: Waves 0-9, 1998-2019. *UK Data Service*, SN: 5050, 2021. doi: http://doi.org/10.5255/UKDA-SN-5050-23. URL http://doi.org/10.5255/UKDA-SN-5050-23.

[110] University of Essex, Institute for Social and Economic Research. Understanding Society: Waves 1-13, 2009-2022 and Harmonised BHPS: Waves 1-18, 1991-2009. SN: 6614, 2023. URL http://doi.org/10.5255/UKDA-SN-6614-19. data collection.

[111] Department for Health and Social Care (DHSC). Community Services Statistics, 2024. URL https://digital.nhs.uk/data-and-information/publications/statistical/community-services-statistics-for-children-young-people-and-adults/. Accessed: 2024-08-14.

[112] Department for Health and Social Care (DHSC). Social Care User Surveys (ASCS and SACE Data Collections), 2024. URL https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-collections/social-care-user-surveys. Accessed: 2024-08-14.

[113] Chloe Juliette, Margaret Blake, Amun Rehsi, Freddie Gregory, and Yota Bratsa. Representativeness of Adult Social Care Surveys. 2021.

[114] Margaret Blake, Claire Lambert, Laura Dale, Stacey Rand, Diane Fox, James Caiels, Ann-Marie Towers, Juliette Malley, and Philip Provenzano. Annex C: ASCS & SACE Discovery Report. 2023.

[115] Department for Health and Social Care (DHSC). Client level data: Analytical Technical Output Specification (ATOS), 2024. URL https://digital.nhs.uk/data-and-information/data-collections-and-data-sets/data-sets/adult-social-care-client-level-data#analytical-technical-output-specification-atos-. Accessed: 2024-08-14.

[116] NHS Digital. Hospital Episode Statistics (HES). https://digital.nhs.uk/data-and-information/data-tools-and-services/data-services/hospital-episode-statistics, 2024. Accessed: 2024-08-12.

[117] Geoffrey Weglarz. Two worlds data-unstructured and structured. *Dm Review*, 14:19–23, 2004.

[118] Comptroller and Auditor General. Challenges in Using Data Across Government, June 2019. URL https://www.nao.org.uk/wp-content/uploads/2019/06/Challenges-in-using-data-across-government.pdf. SESSION 2017–2019.

[119] Miles D Witham, Helen Frost, Marion McMurdo, Peter T Donnan, and Mark McGilchrist. Construction of a linked health and social care database resource–lessons on process, content and culture. *Informatics for Health and Social Care*, 40(3):229–239, 2015.

[120] Sarunkorn Chotvijit, Malkiat Thiarai, and Stephen A Jarvis. A study of data continuity in adult social care services. *The British Journal of Social Work*, 49(3):762–786, 2019.

[121] David Maguire, Harry Evans, Matthew Honeyman, and David Omojomolo. *Digital change in health and social care*. King's Fund London, 2018.

[122] Mark Martin, Julia Pehrson, and Martin Orrell. A survey of social services needs assessments for elderly mentally ill people in England and Wales. *Age and ageing*, 28(6):575–577, 1999.

[123] Karen Stewart, David Challis, Iain Carpenter, and Edward Dickinson. Assessment approaches for older people receiving social care: content and coverage. *International Journal of Geriatric Psychiatry*, 14(2):147–156, 1999.

[124] Michele Abendstern, Jane Hughes, Paul Clarkson, Caroline Sutcliffe, Keith Wilson, and David Challis. 'We need to talk': communication between primary care trusts and other health and social care agencies following the introduction of the Single Assessment Process for older people in England. *Primary Health Care Research & Development*, 11(1):61–71, 2010.

[125] Local Government Association (LGA). Local government social care data standards and interoperability, 2019. URL https://www.local.gov.uk/publications/local-government-social-care-data-standards-and-interoperability. Accessed: 2024-08-14.

[126] Local Government: State of the Sector: AI Research Report. Technical report, Local Government Association, 2024. URL https://web.archive.org/web/20240906174435/https://www.local.gov.uk/sites/default/files/documents/Local%20Government%20State%20of%20the%20Sector%20AI%20Research%20Report%202024%20-%20UPDATED_3.pdf. Accessed: 2024-09-06.

[127] Department of Health and Social Care (DHSC). Evaluation of DHSC's Grant to Streamline Local Authority Adult Social Care Assessments. https://web.archive.org/web/20241108132230/https://www.gov.uk/government/publications/evaluation-of-the-grant-to-streamline-local-authority-adult-social-care-assessments/evaluation-of-dhscs-grant-to-streamline-local-authority-adult-social-care-assessments, Nov 2024. Research and analysis report, accessed via the Internet Archive.

[128] Michele Abendstern, Jane Hughes, Paul Clarkson, Caroline Sutcliffe, and David Challis. The pursuit of integration in the assessment of older people with health and social care needs. *British Journal of Social Work*, 41(3): 467–485, 2011.

[129] David Challis, Paul Clarkson, Janine Williamson, Jane Hughes, Dan Venables, Alistair Burns, and Ashley Weinberg. The value of specialist clinical assessment of older people prior to entry to care homes. *Age and Ageing*, 33(1):25–34, 2004.

[130] Paul Clarkson, Michele Abendstern, Caroline Sutcliffe, Jane Hughes, and David Challis. Identification and recognition of depression in community care assessments: impact of a national policy in England. *International psychogeriatrics*, 24(2):261–269, 2012.

[131] Joseph L Fleiss, Bruce Levin, and Myunghee Cho Paik. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2 (212-236):22–23, 1981.

[132] Gemma Spiers, Fiona Elaine Matthews, Suzanne Moffatt, Robert O Barker, Helen Jarvis, Daniel Stow, Andrew Kingston, and Barbara Hanratty. Impact

of social care supply on healthcare utilisation by older adults: a systematic review and meta-analysis. *Age and ageing*, 48(1):57–66, 2019.

[133] Martin Bardsley, Jennifer Dixon, and Theo Georghiou. *Social care and hospital use at the end of life*. Nuffield Trust London, 2010.

[134] Mariko Carey, Kim Jones, Graham Meadows, Rob Sanson-Fisher, Catherine D'Este, Kerry Inder, Sze Lin Yoong, and Grant Russell. Accuracy of general practitioner unassisted detection of depression. *Australian & New Zealand Journal of Psychiatry*, 48(6):571–578, 2014.

[135] Paul Russell, Sube Banerjee, Jen Watt, Rosalyn Adleman, Belinda Agoe, Nerida Burnie, Alex Carefull, Kiran Chandan, Dominie Constable, Mark Daniels, et al. Improving the identification of people with dementia in primary care: evaluation of the impact of primary care dementia coding guidance on identified prevalence. *BMJ open*, 3(12):e004023, 2013.

[136] John W Peabody, Jeff Luck, Sharad Jain, Dan Bertenthal, and Peter Glassman. Assessing the accuracy of administrative data in health information systems. *Medical care*, 42(11):1066–1072, 2004.

[137] John D Birkmeyer. Using administrative data for clinical research. In *Surgical research*. Academic Press, San Diego, CA, 2000.

[138] Nancy Hoeymans, Edith JM Feskens, Geertrudis AM van den Bos, and Daan Kromhout. Measuring functional status: cross-sectional and longitudinal associations between performance and self-report (Zutphen Elderly Study 1990–1993). *Journal of clinical epidemiology*, 49(10):1103–1110, 1996.

[139] Ian Shaw, Margaret Bell, Ian Sinclair, Patricia Sloper, Wendy Mitchell, Paul Dyson, Jasmine Clayden, and Jackie Rafferty. An exemplary scheme? An evaluation of the Integrated Children's System. *British Journal of Social Work*, 39(4):613–626, 2009.

[140] Jenny Lagsten and Annika Andersson. Use of information systems in social work–challenges and an agenda for future research. *European Journal of Social Work*, 21(6):850–862, 2018.

[141] Nancy Wolff and Thomas W Helminiak. Nonsampling measurement error in administrative data: Implications for economic evaluations. *Health economics*, 5(6):501–512, 1996.

[142] Pamela Trevithick. Humanising managerialism: Reclaiming emotional reasoning, intuition, the relationship, and knowledge and skills in social work. *Journal of Social Work Practice*, 28(3):287–311, 2014.

[143] Social Care Institute for Excellence (SCIE). Social work recording, 2019. URL https://www.scie.org.uk/social-work/recording/.

[144] Louise Grant and Gail Kinman. Emotional resilience in the helping professions and how it can be enhanced. *Health and social care education*, 3(1): 23–34, 2014.

[145] Matt Sutton, Ross Elder, Bruce Guthrie, and Graham Watt. Record rewards: the effects of targeted quality incentives on the recording of risk factors by primary care providers. *Health economics*, 19(1):1–13, 2010.

[146] Tracy H Urech, LeChauncy D Woodard, Salim S Virani, R Adams Dudley, Meghan Z Lutschg, and Laura A Petersen. Calculations of financial incentives for providers in a pay-for-performance program: Manual review versus data from structured fields in electronic health records. *Medical care*, 53 (10):901–907, 2015.

[147] Mark Minchin, Martin Roland, Judith Richardson, Shaun Rowark, and Bruce Guthrie. Quality of care in the United Kingdom after removal of financial incentives. *New England Journal of Medicine*, 379(10):948–957, 2018.

[148] Garth H Utter and Patrick S Romano. Use of administrative data for public reporting of outcomes. *Jama*, 309(19):1991–1992, 2013.

[149] Bryan G Victor, Brian E Perron, Rebeccah L Sokol, Lisa Fedina, and Joseph P Ryan. Automated identification of domestic violence in written child welfare records: Leveraging text mining and machine learning to enhance social work research and evaluation. *Journal of the Society for Social Work and Research*, 12(4):631–655, 2021. ISSN 2334-2315.

[150] Akshaya V Annapragada, Marcella M Donaruma-Kwoh, Ananth V Annapragada, and Zbigniew A Starosolski. A natural language processing and deep learning approach to identify child abuse from pediatric electronic medical records. *PLoS one*, 16(2):e0247404, 2021.

[151] Abdulaziz Tijjani Bako, Heather L Taylor, Kevin Wiley Jr, Jiaping Zheng, Heather Walter-McCabe, Suranga N Kasthurirathne, and Joshua R Vest.

Using natural language processing to classify social work interventions. *The American journal of managed care*, 27(1):e24, 2021.

[152] Vivienne J Zhu, Leslie A Lenert, Brian E Bunnell, Jihad S Obeid, Melanie Jefferson, and Chanita Hughes Halbert. Automatically identifying social isolation from clinical narratives for patients with prostate Cancer. *BMC medical informatics and decision making*, 19(1):1–9, 2019.

[153] Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7 (1):6, 2024.

[154] Braja Gopal Patra, Lauren A Lepow, Praneet Kasi Reddy Jagadeesh Kumar, Veer Vekaria, Mohit Manoj Sharma, Prakash Adekkanattu, Brian Fennessy, Gavin Hynes, Isotta Landi, Jorge A Sanchez-Ruiz, et al. Extracting Social Support and Social Isolation Information from Clinical Psychiatry Notes: Comparing a Rule-based NLP System and a Large Language Model. *arXiv preprint arXiv:2403.17199*, 2024.

[155] Aleena Banerji, Kenneth H Lai, Yu Li, Rebecca R Saff, Carlos A Camargo Jr, Kimberly G Blumenthal, and Li Zhou. Natural language processing combined with ICD-9-CM codes as a novel method to study the epidemiology of allergic drug reactions. *The Journal of Allergy and Clinical Immunology: In Practice*, 8(3):1032–1038, 2020.

[156] Dongdong Zhang, Changchang Yin, Jucheng Zeng, Xiaohui Yuan, and Ping Zhang. Combining structured and unstructured data for predictive models: a deep learning approach. *BMC medical informatics and decision making*, 20:1–11, 2020.

[157] Hongfang Liu, Stephen T Wu, Dingcheng Li, Siddhartha Jonnalagadda, Sunghwan Sohn, Kavishwar Wagholikar, Peter J Haug, Stanley M Huff, and Christopher G Chute. Towards a semantic lexicon for clinical natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2012, page 568. American Medical Informatics Association, 2012.

[158] Gayan Perera, Matthew Broadbent, Felicity Callard, Chin-Kuo Chang, Johnny Downs, Rina Dutta, Andrea Fernandes, Richard D Hayes, Max Henderson, Richard Jackson, et al. Cohort profile of the South London and

Maudsley NHS Foundation Trust Biomedical Research Centre (SLaM BRC) case register: current status and recent enhancement of an electronic mental health record-derived data resource. *BMJ open*, 6(3):e008721, 2016.

[159] Alexandra Pomares Quimbaya, Alejandro Sierra Múnera, Rafael Andrés González Rivera, Julián Camilo Daza Rodríguez, Oscar Mauricio Muñoz Velandia, Angel Alberto Garcia Peña, and Cyril Labbé. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science*, 100:55–61, 2016.

[160] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[161] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[162] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772, 2009.

[163] Julia Hirschberg and Christopher D Manning. Advances in natural language processing. *Science*, 349(6245):261–266, 2015.

[164] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*, 2023.

[165] Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. A Comprehensive Overview of Large Language Models, 2024. URL https://arxiv.org/abs/2307.06435.

[166] Joël Plisson, Nada Lavrac, Dunja Mladenic, et al. A rule based approach to word lemmatization. In *Proceedings of IS*, volume 3, pages 83–86. Citeseer, 2004.

[167] Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5): 301–310, 2001.

[168] Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. DEEPEN: A negation detection system for clinical text incorporating dependency relation into NegEx. *Journal of biomedical informatics*, 54:213–219, 2015.

[169] Krishna Juluru, Hao-Hsin Shih, Krishna Nand Keshava Murthy, and Pierre Elnajjar. Bag-of-words technique in natural language processing: a primer for radiologists. *RadioGraphics*, 41(5):1420–1426, 2021.

[170] Eric Nguyen. Text mining and network analysis of digital libraries in R. *Data mining applications with R*, pages 95–115, 2013.

[171] Hans Christian, Mikhael Pramodana Agus, and Derwin Suhartono. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4):285–294, 2016.

[172] Nikolay Malkin, Sameera Lanka, Pranav Goel, and Nebojsa Jojic. Studying word order through iterative shuffling. *arXiv preprint arXiv:2109.04867*, 2021.

[173] Rui Zhao and Kezhi Mao. Fuzzy bag-of-words model for document representation. *IEEE transactions on fuzzy systems*, 26(2):794–804, 2017.

[174] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3111–3119, 2013.

[175] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR)*, 2013.

[176] John Rupert Firth, W Haas, and Michael AK Halliday. *Studies in linguistic analysis*. Blackwell, 1957.

[177] Yoav Goldberg and Omer Levy. Word2vec explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*, 2014.

[178] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[179] Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Association for Computational Linguistics, 2014.

[180] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825—-2830, 2011.

[181] Xin Rong. Word2vec parameter learning explained. *arXiv preprint arXiv:1411.2738*, 2014.

[182] Nikolay Arefyev, Pavel Ermolaev, and Alexander Panchenko. How much does a word weigh? Weighting word embeddings for word sense induction. *arXiv preprint arXiv:1805.09209*, 2018.

[183] Mahidhar Dwarampudi and NV Reddy. Effects of padding on LSTMs and CNNs. *arXiv preprint arXiv:1903.07288*, 2019.

[184] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 670–680. Association for Computational Linguistics, 2017.

[185] Tom Kenter, Alexey Borisov, and Maarten De Rijke. Siamese cbow: Optimizing word embeddings for sentence representations. *arXiv preprint arXiv:1606.04640*, 2016.

[186] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal Sentence Encoder. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 169–174. Association for Computational Linguistics, 2018.

[187] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[188] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, 2018.

[189] Jeffrey L Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

[190] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[191] Seol-Hyun Noh. Analysis of gradient vanishing of RNNs and performance comparison. *Information*, 12(11):442, 2021.

[192] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[193] Hugging Face. Pretrained Models - Transformers v2.4.0 Documentation, 2024. URL https://huggingface.co/transformers/v2.4.0/pretrained_models.html. Accessed: 2024-08-25.

[194] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. ISBN 978-0262035613. URL http://www.deeplearningbook.org.

[195] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[196] Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date, 2024. URL https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-07-14.

[197] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay

Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[198] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[199] Google. Gemma 7b-it, 2024. URL https://huggingface.co/google/gemma-7b-it. [Accessed: 2024-07-25].

[200] Meta Llama. Meta Llama 3 70B, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-3-70B. Accessed: 2024-08-25.

[201] Charles T Meadow and Weijing Yuan. Measuring the impact of information: defining the concepts. *Information processing & management*, 33(6):697–714, 1997.

[202] Subrata Dasgupta. Disentangling data, information and knowledge. *Big Data & Information Analytics*, 1(4):377–389, 2017.

[203] Colin Cherry. On human communication. 1966.

[204] Marco Dinarelli, Alessandro Moschitti, and Giuseppe Riccardi. Joint generative and discriminative models for spoken language understanding. In *2008 IEEE Spoken Language Technology Workshop*, pages 61–64. IEEE, 2008.

[205] Zineng Tang, Shiyue Zhang, Hyounghun Kim, and Mohit Bansal. Continuous Language Generative Flow. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4609–4622, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.355. URL https://aclanthology.org/2021.acl-long.355.

[206] Yanjun Gao, Timothy Miller, Dongfang Xu, Dmitriy Dligach, Matthew M Churpek, and Majid Afshar. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of COLING. International Conference on Computational Linguistics*, volume 2022, page 2979. NIH Public Access, 2022.

[207] Peter Lee, Sebastien Bubeck, and Joseph Petro. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239, 2023.

[208] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

[209] Omar Ayaad, Aladeen Alloubani, Eyad Abu ALhajaa, Mohammad Farhan, Sami Abuseif, Ahmad Al Hroub, and Laila Akhu-Zaheya. The role of electronic medical records in improving the quality of health care services: Comparative study. *International journal of medical informatics*, 127:63–67, 2019.

[210] Marieke Zegers, Martine C de Bruijne, Peter Spreeuwenberg, Cordula Wagner, Peter P Groenewegen, and Gerrit van der Wal. Quality of patient record keeping: an indicator of the quality of care? *BMJ quality & safety*, 20(4):314–318, 2011.

[211] North Yorkshire Council. Knowledge Mining: Reimagining Case Management. https://web.archive.org/web/20240906171201/https://i-network.org.uk/knowledge-mining-reimagining-case-management/, 2024. Accessed: 2024-09-06.

[212] Kerstin Denecke, Richard May, LLMHealthGroup, and Octavio Rivera Romero. Potential of Large Language Models in Health Care: Delphi Study. *Journal of Medical Internet Research*, 26:e52399, 2024.

[213] Ravindra Kumar Garg, Vijeth L Urs, Akshay Anand Agarwal, Sarvesh Kumar Chaudhary, Vimal Paliwal, and Sujita Kumar Kar. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. *Health Promotion Perspectives*, 13(3):183, 2023.

[214] Anmol Arora and Ananya Arora. The promise of large language models in health care. *The Lancet*, 401(10377):641, 2023.

[215] Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen,

et al. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141, 2023.

[216] Google Cloud. MedLM: Generative AI fine-tuned for the healthcare industry, 2024. URL https://web.archive.org/web/20240804062023/https://cloud.google.com/blog/topics/healthcare-life-sciences/introducing-medlm-for-the-healthcare-industry. Accessed: 2024-09-01.

[217] Beam. Magic Notes, 2024. URL https://magicnotes.ai/. Accessed: 2024-09-01.

[218] Barnet Council. Adult Social Care Privacy Notice, 2024. URL https://web.archive.org/web/20240901155529/https://www.barnet.gov.uk/your-council/policies-plans-and-performance/privacy-notices/adults-social-care-privacy-notices-0#title-9. Accessed: 2024-09-01.

[219] Camden Council. Adult Social Care use of BEAM/Magic Notes, 2024. URL https://web.archive.org/web/20240901155606/https://camdencarechoices.camden.gov.uk/web/20240901155606/https://camdencarechoices.camden.gov.uk/care-and-support-rights/adult-social-care-use-of-beam-magic-notes/. Accessed: 2024-09-01.

[220] Havering Council. Privacy Notice for Adult Social Care, 2024. URL https://web.archive.org/web/20240901160056/https://www.havering.gov.uk/downloads/file/6421/adult-social-care-privacy-notice. Accessed: 2024-09-01.

[221] Kingston Council. Privacy Notice and Data Protection, 2024. URL https://web.archive.org/web/20240901160209/https://www.kingston.gov.uk/council-democracy/privacy-notice-data-protection/7. Accessed: 2024-09-01.

[222] Swindon Council. Use of Magic Notes Privacy Notices, 2024. URL https://web.archive.org/web/20241104193542/https://www.swindon.gov.uk/directory_record/25568/use_of_magic_notes_privacy_notices. Accessed: 2024-11-04.

[223] Lambeth Council. Magic Notes Privacy Notice, 2024. URL https://web.archive.org/web/20241104193557/https://www.lambeth.gov.uk/about-council/privacy-data-protection/magic-notes-privacy-notice. Accessed: 2024-11-04.

[224] North Somerset Council. ASC - Covert and Overt Recording Guidance, 2024. URL https://web.archive.org/web/20241104193746/https://n-somerset.gov.uk/sites/default/files/2024-09/ASC%20-%20covert%20and%20overt%20recording%20guidance.pdf. Accessed: 2024-11-04.

[225] South Tyneside Council. Privacy Notice: Magic Notes Pilot - Adult Social Care, 2024. URL https://web.archive.org/web/20241104193730/https://www.southtyneside.gov.uk/article/24835/Privacy-notice-Magic-Notes-Pilot-Adult-Social-Care. Accessed: 2024-11-04.

[226] Stockport Council. Privacy Notices - Use of Magic Notes Privacy Notices, 2024. URL https://web.archive.org/web/20241120212757/https://www.stockport.gov.uk/privacy-notices/health-and-social-care-use-of-beam-magic-notes. Accessed: 2024-11-20.

[227] Coventry Council. Artificial Intelligence & Automation: An Update for Scrutiny Board 1. Technical report, February 2024. URL https://web.archive.org/web/20240906181845/https://edemocracy.coventry.gov.uk/documents/s59295/Appendix%201.pdf. Accessed: 2024-09-06.

[228] Bexley Council. Productivity Plan to the Ministry of Housing, Communities & Local Government, 2024. URL https://web.archive.org/web/20240906173525/https://www.bexley.gov.uk/about-the-council/council-budgets-and-spending/productivity-plan-ministry-housing-communities-local-government. Accessed: 2024-09-06.

[229] Wrexham Council. Wrexham Council ICT and Digital Strategy 2024-2027. Technical report, 2024. URL https://www.wrexham.gov.uk/sites/default/files/2024-07/wrexham-council-ict-and-digital-strategy-2024-2027.pdf. Accessed: 2024-09-06.

[230] Barnsley Council. Barnsley Council Productivity Plan 2024-2025. Technical report, 2024. URL https://www.barnsley.gov.uk/media/xv2aoeow/bmbc-productivity-plan-2024-2025.pdf. Accessed: 2024-09-06.

[231] Kirklees Council. Kirklees Council Productivity Plan. Technical report, 2024. URL https://www.kirklees.gov.uk/beta/delivering-services/pdf/Kirklees-Council-Productivity-Plan.pdf. Accessed: 2024-09-06.

[232] Gloucestershire County Council. Generative AI (GenAI) Policy v1.0.

Technical report, 2024. URL https://www.gloucestershire.gov.uk/media/h3teu4dc/generative-ai-genai-policy-v10.pdf. Accessed: 2024-09-06.

[233] Enfield Council. Digital Services Overview and Scrutiny, September 2024. Technical report, September 2024. URL https://governance.enfield.gov.uk/documents/s105816/Digital%20Services%20Overview%20and%20Scrutiny%20September%202024%20Part%201.pdf. Accessed: 2024-09-06.

[234] South Cambridgeshire District Council. South Cambridgeshire District Council Productivity Plan. Technical report, 2024. URL https://www.scambs.gov.uk/your-council-and-democracy/our-business-plan-and-performance/our-productivity-plan. Accessed: 2024-09-06.

[235] Teignbridge Council. Strata Business Plan v1.8. Technical report, 2024. URL https://democracy.teignbridge.gov.uk/documents/s18951/Strata%20Business%20Plan_v1.8%202.pdf. Accessed: 2024-09-06.

[236] Skills For Care. Workforce Estimates, 2023. URL https://www.skillsforcare.org.uk/Adult-Social-Care-Workforce-Data/Workforce-intelligence/publications/Workforce-estimates.aspx. Accessed: 2024-09-23.

[237] Karen C Jones, Helen Weatherly, Sarah Birch, Adriana Castelli, Martin Chalkley, Alan Dargan, Julien E Forder, Jinbao Gao, Seb Hinde, Sarah Markham, et al. Unit costs of health and social care 2022 manual. 2023.

[238] Mithran Samuel. BASW warns of 'dilution of social worker role' as dhsc plans more assessments by non-qualified staff. *Community Care*, March 2023. URL https://www.communitycare.co.uk/2023/03/31/more-care-act-assessments-to-be-carried-out-by-non-social-workers-under-government-plan/. Accessed: 2024-09-23.

[239] Bleddyn Davies. Equity and efficiency in community care: supply and financing in an age of fiscal austerity. *Ageing & Society*, 7(2):161–174, 1987.

[240] Trisha Greenhalgh, Joseph Wherton, Chrysanthi Papoutsi, Jennifer Lynch, Gemma Hughes, Susan Hinder, Nick Fahy, Rob Procter, Sara Shaw, et al. Beyond adoption: a new framework for theorizing and evaluating nonadoption, abandonment, and challenges to the scale-up, spread, and sustainability of health and care technologies. *Journal of medical Internet research*, 19 (11):e8775, 2017.

[241] Shweta Chauhan and Philemon Daniel. A comprehensive survey on various fully automatic machine translation evaluation metrics. *Neural Processing Letters*, 55(9):12663–12717, 2023.

[242] Kirk Roberts. Large language models for reducing clinicians' documentation burden. *Nature Medicine*, 30(4):942–943, 2024.

[243] Roberto Navigli, Simone Conia, and Björn Ross. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21, 2023.

[244] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

[245] Allison Koenecke, Anna Seo Gyeong Choi, Katelyn X Mei, Hilke Schellmann, and Mona Sloane. Careless Whisper: Speech-to-Text Hallucination Harms. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1672–1681, 2024.

[246] Francesca Carrera, Emmanuele Pavolini, Costanzo Ranci, and Alessia Sabbatini. Long-term care systems in comparative perspective: Care needs, informal and formal coverage, and social impacts in European countries. *Reforms in long-term care policies in Europe: Investigating institutional change and social impacts*, pages 23–52, 2013.

[247] H Stephen Kaye, Charlene Harrington, and Mitchell P LaPlante. Long-term care: who gets it, who provides it, who pays, and how much? *Health affairs*, 29(1):11–21, 2010.

[248] Ingrid Mairhuber, Ilze Trapenciere, Danièle Meulders, Ulrike Papouschek, Iskra Beleva, Ruta Braziene, Alexia Panayiotou, Robert Plasman, Alena Křížková, Frances Camilleri-Cassar, et al. Gender segregation in the labour market: root causes, implications and policy responses in the EU. 2009.

[249] Francesca Bettio, Janneke Plantenga, Mark Smith, et al. *Gender and the European labour market.* Routledge London, 2013.

[250] Thurid Eggers, Christopher Grages, Birgit Pfau-Effinger, and Ralf Och. Reconceptualising the relationship between de-familialisation and familialisation and the implications for gender equality–the case of long-term care policies for older people. *Ageing & Society*, 40(4):869–895, 2020.

[251] Lena Armstrong, Abbey Liu, Stephen MacNeil, and Danaë Metaxa. The Silicon Ceiling: Auditing GPT's Race and Gender Biases in Hiring. In *Proceedings of the 4th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–18, 2024.

[252] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874, 2021.

[253] Andrey Kapitanov, Ilona Kapitanova, Vladimir Troyanovskiy, Vladimir Ilyushechkin, and Ekaterina Dorogova. Clustering of Word Contexts as a Method of Eliminating Polysemy of Words. In *2019 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*, pages 1861–1864. IEEE, 2019.

[254] Amir Gandomi and Murtaza Haider. Beyond the hype: Big data concepts, methods, and analytics. *International journal of information management*, 35(2):137–144, 2015.

[255] Marcelo A Montemurro and Damián H Zanette. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02):135–153, 2010.

[256] Seref Sagiroglu and Duygu Sinanc. Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*, pages 42–47, 2013. doi: 10.1109/CTS.2013.6567202.

[257] Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani, and Shourya Roy. How much noise is too much: A study in automatic text classification. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, pages 3–12. IEEE, 2007.

[258] Hoang Nguyen and Jon Patrick. Text mining in clinical domain: Dealing with noise. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 549–558, 2016.

[259] Sitthichok Chaichulee, Chissanupong Promchai, Tanyamai Kaewkomon, Chanon Kongkamol, Thammasin Ingviya, and Pasuree Sangsupawanich. Multi-label classification of symptom terms from free-text bilingual adverse drug reaction reports using natural language processing. *PloS one*, 17(8): e0270595, 2022. ISSN 1932-6203.

[260] Pavel Blinov, Manvel Avetisian, Vladimir Kokh, Dmitry Umerenkov, and Alexander Tuzhilin. Predicting clinical diagnosis from patients electronic health records using BERT-based neural networks. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*, pages 111–121. Springer, 2020.

[261] Yikuan Li, Shishir Rao, José Roberto Ayala Solares, Abdelaali Hassaine, Rema Ramakrishnan, Dexter Canoy, Yajie Zhu, Kazem Rahimi, and Gholamreza Salimi-Khorshidi. BEHRT: transformer for electronic health records. *Scientific reports*, 10(1):1–12, 2020. ISSN 2045-2322.

[262] Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. Mining Electronic Health Records (EHRs) A Survey. *ACM Computing Surveys (CSUR)*, 50(6):1–40, 2018. ISSN 0360-0300.

[263] Katharine Orellana, Jill Manthorpe, and Anthea Tinker. Day centres for older people: a systematically conducted scoping review of literature about their benefits, purposes and how they are perceived. *Ageing and Society*, 40 (1):73–104, 2020. ISSN 0144-686X.

[264] Catrin Noone. *The changing role of the day centre for older people in addressing loneliness: a participatory action research study.* PhD thesis, Durham University, 2023.

[265] Catherine A Lunt. *Impact of Day Care Services on Older People with Long Term Conditions.* The University of Liverpool (United Kingdom), 2018.

[266] Melanie Luppa, Tobias Luck, Siegfried Weyerer, Hans-Helmut König, Elmar Brähler, and Steffi G Riedel-Heller. Prediction of institutionalization in the elderly. A systematic review. *Age and ageing*, 39(1):31–38, 2010.

[267] Barbara Hanratty, Daniel Stow, Danni Collingridge Moore, Nicole K Valtorta, and Fiona Matthews. Loneliness as a risk factor for care home admission in the English Longitudinal Study of Ageing. *Age and ageing*, 47(6): 896–900, 2018.

[268] North East London Commissioning Support Unit. PSCleaner. https://github.com/NELCSU/PSCleaner, 2020. URL https://github.com/NELCSU/PSCleaner. GitHub repository.

[269] Microsoft. TypeScript: Typed JavaScript at Any Scale. https://www.typescriptlang.org/, 2012. URL https://www.typescriptlang.org/. Version 5.1.6.

[270] Eric Spishak, Werner Dietl, and Michael D. Ernst. A type system for regular expressions. pages 20–26, 2012. doi: 10.1145/2318202.2318207.

[271] Office for National Statistics. Nomis: Official Labour Market Statistics, 2022. URL https://www.nomisweb.co.uk/. Accessed: 2024-08-20.

[272] Lesley A Curtis and Amanda Burns. *Unit costs of health and social care 2016*. Personal Social Services Research Unit, 2016.

[273] Lesley A Curtis and Amanda Burns. *Unit costs of health and social care 2017*. Personal Social Services Research Unit, 2017.

[274] Lesley A Curtis and Amanda Burns. *Unit costs of health and social care 2018*. Personal Social Services Research Unit, 2018.

[275] Lesley A Curtis and Amanda Burns. *Unit costs of health and social care 2019*. Personal Social Services Research Unit, 2019.

[276] Lesley A Curtis and Amanda Burns. *Unit costs of health and social care 2020*. Personal Social Services Research Unit, 2020.

[277] Office for National Statistics. Consumer Price Inflation Tables, 2024. URL https://www.ons.gov.uk/economy/inflationandpriceindices/datasets/consumerpriceinflation. Accessed: 2024-08-21.

[278] Rejean Hebert, Carol Brayne, and David Spiegelhalter. Incidence of functional decline and improvement in a community-dwelling, very elderly population. *American journal of epidemiology*, 145(10):935–944, 1997.

[279] Javier Jerez-Roig, Lidiane Maria de Brito Macedo Ferreira, José Rodolfo Torres de Araújo, and Kenio Costa Lima. Dynamics of activities of daily living performance in institutionalized older adults: A two-year longitudinal study. *Disability and health journal*, 10(2):279–285, 2017.

[280] OECD. *Care Needed: Improving the Lives of People with Dementia*. OECD Health Policy Studies. OECD Publishing, Paris, 2018. doi: 10.1787/9789264085107-en. URL http://dx.doi.org/10.1787/9789264085107-en.

[281] Alzheimer's Society. Facts for the media about dementia. https://web.archive.org/web/20241001171930/https://www.alzheimers.org.uk/about-us/news-and-media/facts-media, 2024. Accessed: 2024-11-17.

[282] Robert Stewart, Matthew Hotopf, Michael Dewey, Clive Ballard, Jatinder Bisla, Maria Calem, Viola Fahmy, Jo Hockley, Julie Kinley, Hywel Pearce, et al. Current prevalence of dementia, depression and behavioural problems in the older adult care home sector: the South East London Care Home Survey. *Age and ageing*, 43(4):562–567, 2014.

[283] Department for Health and Social Care (DHSC). Personal Social Services: Staff of Social Services Departments, 2021. URL https://digital.nhs.uk/data-and-information/publications/statistical/personal-social-services-staff-of-social-services-departments. Accessed: 2024-08-14.

[284] Skills for Care. The state of the adult social care sector and workforce in England. Technical report, 2023. URL https://www.skillsforcare.org.uk/Adult-Social-Care-Workforce-Data/Workforce-intelligence/documents/State-of-the-adult-social-care-sector/The-State-of-the-Adult-Social-Care-Sector-and-Workforce-2023.pdf. Accessed: 2024-08-27.

[285] Geraldine Macdonald and Kenneth Macdonald. Safeguarding: A case for intelligent risk management. *British Journal of Social Work*, 40(4):1174–1191, 2010.

[286] Jeremy Dixon. *Adult Safeguarding Observed: How Social Workers Assess and Manage Risk and Uncertainty*. Policy Press, 2023.

[287] Department for Work and Pensions, NatCen Social Research. Family Resources Survey. data series, 2021. URL http://doi.org/10.5255/UKDA-Series-200017. 4th Release. UK Data Service. SN: 200017.

[288] NatCen Social Research, University College London, Department of Epidemiology and Public Health. Health Survey for England. data series, 2024. URL http://doi.org/10.5255/UKDA-Series-2000021. 8th Release. UK Data Service. SN: 2000021.

[289] Michael G Marmot, George Davey Smith, Stephen Stansfeld, Chandra Patel, Fiona North, Jenny Head, Ian White, Eric Brunner, and Amanda Feeney. Health inequalities among British civil servants: the Whitehall II study. In *Stress and the Brain*, pages 61–67. Routledge, 2013.

[290] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3): e1001779, 2015.

[291] Anna Fry, Thomas J Littlejohns, Cathie Sudlow, Nicola Doherty, Ligia Adamska, Tim Sprosen, Rory Collins, and Naomi E Allen. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. *American journal of epidemiology*, 186(9):1026–1034, 2017.

[292] Carol Brayne and Terrie E Moffitt. The limitations of large-scale volunteer databases to address inequalities and global challenges in health and aging. *Nature Aging*, 2(9):775–783, 2022.

[293] Office for National Statistics, Census Division. Office for National Statistics Longitudinal Study, 1971-. data collection, 2022. URL http://doi.org/10.5255/UKDA-SN-5762-1. UK Data Service. SN: 5762.

[294] James Banks, Alastair Muriel, and James P Smith. Attrition and health in ageing studies: evidence from ELSA and HRS. *Longitudinal and life course studies*, 2(2), 2011.

[295] David Weir, Jessica Faul, and Kenneth Langa. Proxy interviews and bias in the distribution of cognitive abilities due to non-response in longitudinal studies: a comparison of HRS and ELSA. *Longitudinal and life course studies*, 2(2):170, 2011.

[296] Hisashi Kawai, Manami Ejiri, Harukazu Tsuruta, Yukie Masui, Yutaka Watanabe, Hirohiko Hirano, Yoshinori Fujiwara, Kazushige Ihara, Masashi Tanaka, and Shuichi Obuchi. Factors associated with follow-up difficulty in longitudinal studies involving community-dwelling older adults. *PloS one*, 15(8):e0237166, 2020.

[297] Allison R Heid, Francine P Cartwright, Maureen Wilson-Genderson, and Rachel Pruchno. Understanding attrition and bolstering retention in a longitudinal panel of older adults: ORANJ BOWL. *Innovation in Aging*, 5(2): igab010, 2021.

[298] Pamela Grimm. Social desirability bias. *Wiley international encyclopedia of marketing*, 2010.

[299] Ronald Angel, Glenn V Ostir, Michelle L Frisco, and Kyriakos S Markides. Comparison of a self-reported and a performance-based assessment of mobility in the Hispanic Established Population for Epidemiological Studies of the Elderly. *Research on Aging*, 22(6):715–737, 2000.

[300] Fiona Steele, Heather Joshi, Constantinos Kallis, and Harvey Goldstein. Changing compatibility of cohabitation and childbearing between young British women born in 1958 and 1970. *Population studies*, 60(2):137–152, 2006.

[301] Shereen Hussein. The use of'large scale datasets' in UK social care research. 2011.

[302] Allison Dunatchik, Rossella Icardi, Caireen Roberts, and Margaret Blake. Predicting unmet social care needs and links with well-being: Findings from the secondary analysis. *London: NATCEN, Ipsos Mori*, 2016.

[303] Brian R Radke. A demonstration of interval-censored survival analysis. *Preventive veterinary medicine*, 59(4):241–256, 2003.

[304] Fiona Steele. Multilevel models for longitudinal data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 171(1):5–19, 2008.

[305] Leen Vandecasteele and Annelies Debels. Attrition in panel data: the effectiveness of weighting. *European Sociological Review*, 23(1):81–97, 2007.

[306] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[307] Lingli Yang, Balgobin Nandram, and Jai Won Choi. Bayesian Predictive Inference Under Nine Methods for Incorporating Survey Weights. *International Journal of Statistics and Probability*, 12(1):33–53, 2023.

[308] Alasdair C Rutherford and Feifei Bu. Issues with the measurement of informal care in social surveys: evidence from the English Longitudinal Study of Ageing. *Ageing & Society*, 38(12):2541–2559, 2018.

[309] Kelvin KF Tsoi, Joyce YC Chan, Hoyee W Hirai, Adrian Wong, Vincent CT Mok, Linda CW Lam, Timothy CY Kwok, and Samuel YS Wong. Recall tests are effective to detect mild cognitive impairment: a systematic review and meta-analysis of 108 diagnostic studies. *Journal of the American Medical Directors Association*, 18(9):807–e17, 2017.

[310] Danielle Navarro. *Learning statistics with R: A tutorial for psychology students and other beginners. (Version 0.6).* University of New South Wales, Sydney, Australia, 2015. URL https://learningstatisticswithr.com. R package version 0.5.1.

[311] Thomas Lumley. *Complex Surveys: A Guide to Analysis Using R: A Guide to Analysis Using R.* John Wiley and Sons, 2010.

[312] Office for National Statistics (ONS). Estimates of the population for the UK, England, Wales, Scotland and Northern Ireland, 2024. URL https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/populationestimatesforukenglandandwalesscotlandandnorthernireland.

[313] Ramesh Lal Sapra and Satish Saluja. Understanding statistical association and correlation. *Current Medicine Research and Practice*, 11(1):31–38, 2021.

[314] Linda Bauld, John Chesterman, Bleddyn Davies, and Ken Judge. *Caring for older people: an assessment of community care in the 1990s.* Routledge, 2018.

[315] Eric J Tchetgen Tchetgen and Kathleen E Wirth. A general instrumental variable framework for regression analysis with outcome missing not at random. *Biometrics*, 73(4):1123–1131, 2017.

[316] Elizabeth Washbrook, Paul S Clarke, and Fiona Steele. Investigating non-ignorable dropout in panel studies of residential mobility. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 63(2):239–266, 2014.

[317] Lauren J Beesley and Jeremy MG Taylor. Accounting for not-at-random missingness through imputation stacking. *Statistics in medicine*, 40(27): 6118–6132, 2021.

[318] Michael Höfler, Hildegard Pfister, Roselind Lieb, and Hans-Ulrich Wittchen. The use of weights to account for non-response and drop-out. *Social psychiatry and psychiatric epidemiology*, 40:291–299, 2005.

[319] Erin Hartman and Melody Huang. Sensitivity analysis for survey weights. *Political Analysis*, 32(1):1–16, 2024.

[320] Jianwen Cai, Donglin Zeng, Haolin Li, Nicole M Butera, Pedro L Baldoni, Poulami Maitra, and Li Dong. Comparisons of statistical methods for han-

dling attrition in a follow-up visit with complex survey sampling. *Statistics in medicine*, 42(11):1641–1668, 2023.

[321] OECD Stat. Health expenditure and financing. https://stats.oecd.org/Index.aspx?DataSetCode=SHA, 2023.

[322] NHS Digital. Adult Social Care Activity and Finance Report, England, 2022-23, 2023.

[323] Bo Hu, Ruth Hancock, and Raphael Wittenberg. Projections of adult social care demand and expenditure 2018 to 2038. *CPEC Working Paper 7*, 2020.

[324] Derek King and Raphael Wittenberg. Data on adult social care. *NIHR School for Social Care Research Scoping Review*, 2015.

[325] David Challis, Paul Clarkson, Sue Davies, Jane Hughes, Karen Stewart, and Chengqiu Xie. Resource allocation at the micro level in adult social care: A scoping review. Technical report, 2016.

[326] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[327] Lisa F Berkman and S Leonard Syme. Social networks, host resistance, and mortality: a nine-year follow-up study of Alameda County residents. *American journal of Epidemiology*, 109(2):186–204, 1979. ISSN 1476-6256.

[328] Julianne Holt-Lunstad, Timothy B Smith, Mark Baker, Tyler Harris, and David Stephenson. Loneliness and social isolation as risk factors for mortality: a meta-analytic review. *Perspectives on psychological science*, 10(2): 227–237, 2015. ISSN 1745-6916.

[329] Matthew Pantell, David Rehkopf, Douglas Jutte, S Leonard Syme, John Balmes, and Nancy Adler. Social isolation: a predictor of mortality comparable to traditional clinical risk factors. *American Journal of Public Health*, 103(11):2056–2062, 2013. ISSN 1541-0048.

[330] DCMS. Tackling loneliness evidence review: main report. *London: Department for Digital, Culture, Media and Sport.*, 2023. URL https://www.gov.uk/government/publications/tackling-loneliness-evidence-review/tackling-loneliness-evidence-review-full-report#where-are-the-evidence-gaps-in-the-current-research-profile-of-loneliness.

[331] WHO. Social isolation and loneliness among older people: advocacy brief, 2021. ISSN 9240030743.

[332] Caitlin E Coyle and Elizabeth Dugan. Social isolation, loneliness and health among older adults. *Journal of aging and health*, 24(8):1346–1363, 2012. ISSN 0898-2643.

[333] Jenny de Jong-Gierveld. Developing and testing a model of loneliness. *Journal of personality and social psychology*, 53(1):119, 1987. ISSN 1939-1315.

[334] WHO. Psychogeriatric care in the community, 1979.

[335] Thomas Prohaska, Vanessa Burholt, Annette Burns, Jeannette Golden, Louise Hawkley, Brian Lawlor, Gerard Leavey, Jim Lubben, Roger O'Sullivan, Carla Perissinotto, et al. Consensus statement: loneliness in older adults, the 21st century social determinant of health? *BMJ open*, 10 (8):e034967, 2020.

[336] DCMS. A connected Society: a strategy for tackling loneliness–laying the foundations for change. *London: Department for Digital, Culture, Media and Sport.*, 2018.

[337] National Institute for Health and Care Excellence (NICE). Mental wellbeing and independence for older people: Quality standard [QS137]. Technical report, 2016. URL https://www.nice.org.uk/guidance/qs137/chapter/Quality-statement-3-Social-participation.

[338] NHS Digital. Adult social care activity and finance report, England 2019–20, 2020.

[339] UK Parliament. The Care and Support (Eligibility Criteria) Regulations 2015 SI 2023/313, 2015.
https://www.legislation.gov.uk/uksi/2015/313/introduction/made?view=plain.

[340] Office of the Surgeon General et al. Our epidemic of loneliness and isolation: The US Surgeon General's Advisory on the healing effects of social connection and community, 2023.

[341] R S Tilvis, P Routasalo, Helena Karppinen, T E Strandberg, H Kautiainen, and K H Pitkala. Social isolation, social activity and loneliness as survival indicators in old age; a nationwide survey with a 7-year follow-up. *European Geriatric Medicine*, 3(1):18–22, 2012. ISSN 1878-7649.

[342] NELCSU. PSCleaner: Process CSV files by identifying and removing personal sensitive text, 2022. URL https://github.com/NELCSU/PSCleaner.

[343] Dan Jurafsky and James H Martin. Speech and language processing (3rd (draft) ed.), 2019.

[344] M. Honnibal and I. Montani. English pipeline optimized for CPU. Components: tok2vec, tagger, parser, senter, ner, attribute ruler, lemmatizer., 2021. URL https://github.com/explosion/spacy-models/releases/tag/en_core_web_lg-3.0.0.

[345] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[346] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012. ISSN 1330-0962.

[347] Klaus Krippendorff. Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800, 2004. ISSN 0033-5177.

[348] Knut De Swert. Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. *Center for Politics and Communication*, 15, 2012.

[349] Huggingface. distilroberta-base, 2022. URL https://huggingface.co/distilroberta-base.

[350] Huggingface. roberta-base, 2022. URL https://huggingface.co/roberta-base.

[351] Isaac Kofi Nti, Owusu Nyarko-Boateng, and Justice Aning. Performance of Machine Learning Algorithms with Different K Values in K-fold Cross-Validation. *International Journal of Information Technology and Computer Science*, 13(6):61–71, 2021.

[352] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, and Luca Antiga. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019.

[353] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan

Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[354] Juliette N Malley, Ann-Marie Towers, Ann P Netten, John E Brazier, Julien E Forder, and Terry Flynn. 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(1):1–14, 2012. ISSN 1477-7525.

[355] Lenore Sawyer Radloff. The CES-D scale: A self-report depression scale for research in the general population. *Applied psychological measurement*, 1 (3):385–401, 1977.

[356] Nadia Minicuci, Nirmala Naidoo, Somnath Chatterji, and Paul Kowal. Data Resource Profile: Cross-national and cross-study sociodemographic and health-related harmonized domains from SAGE plus ELSA, HRS and SHARE (SAGE+, Wave 1). *International Journal of Epidemiology*, 45(5): 1403–1403j, 2016.

[357] Alan Agresti. *Categorical data analysis*, volume 792. John Wiley & Sons, 2012.

[358] Sebastian Raschka and Vahid Mirjalili. *Python machine learning: Machine learning and deep learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.

[359] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition, 2019. URL https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

[360] Robert M O'Brien. A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41:673–690, 2007.

[361] Liat Ayalon, Sharon Shiovitz-Ezra, and Ilan Roziner. A cross-lagged model of the reciprocal associations of loneliness and memory functioning. *Psychology and Aging*, 31(3):255, 2016.

[362] Jiamin Yin, Camille Lassale, Andrew Steptoe, and Dorina Cadar. Exploring the bidirectional associations between loneliness and cognitive functioning over 10 years: the English longitudinal study of ageing. *International Journal of Epidemiology*, 48(6):1937–1948, 2019.

[363] Berna Van Baarsen, Tom AB Snijders, Johannes H Smit, and Marijtje AJ Van Duijn. Lonely but not alone: Emotional isolation and social isolation as two distinct dimensions of loneliness in older people. *Educational and Psychological measurement*, 61(1):119–135, 2001.

[364] Martin Knapp, Kia-Chong Chua, Matthew Broadbent, Chin-Kuo Chang, Jose-Luis Fernandez, Dominique Milea, Renee Romeo, Simon Lovestone, Michael Spencer, Gwilym Thompson, et al. Predictors of care home and hospital admissions and their costs for older people with Alzheimer's disease: findings from a large London case register. *BMJ open*, 6(11):e013591, 2016.

[365] Mark McCann, Michael Donnelly, and Dermot O'Reilly. Gender differences in care home admission risk: partner's age explains the higher risk for women. *Age and Ageing*, 41(3):416–419, 2012.

[366] Josiane Breeden, David Hussey, Kavita Deepchand, and Melanie Norton. The dynamics of ageing: The 2016/2017 English Longitudinal Study of Ageing (Wave 8) Technical Report. *UK Data Service*, 2018. URL https://doc.ukdataservice.ac.uk/doc/5050/mrdoc/pdf/5050_elsa_w8_technical_report_v7.pdf.

[367] Bo Hu. Projecting future demand for informal care among older people in China: the road towards a sustainable long-term care system. *Health Economics, Policy and Law*, 14(1):61–81, 2019.

[368] Seika Akemura and Daizo Kojima. Japan's long-term care cost projections: comparison with the European Commission ageing report. *Public Policy Review, Policy Research Institute, Ministry of Finance Japan*, 14(4):541–562, 2018.

[369] Oddvar Førland. Conceptualizing Needs When Allocating Public Long-Term Care Services in the Welfare State. In *The Political Economy of Care: Welfare state capitalism, universalism, and social reproduction*, pages 231–251. 2024.

[370] G Salkeld, Shanthi N Ameratunga, ID Cameron, RG Cumming, S Easter, J Seymour, SE Kurrle, S Quine, and Paul M Brown. Quality of life related to fear of falling and hip fracture in older women: a time trade off study-Commentary: older people's perspectives on life after hip fractures. *Bmj*, 320(7231):341–346, 2000.

[371] Siobhan Aine Bradshaw, E Diane Playford, and Afsane Riazi. Living well in care homes: a systematic review of qualitative studies. *Age and ageing*, 41(4):429–440, 2012.

[372] Ashok J Bharucha, Rajesh Pandav, Changyu Shen, Hiroko H Dodge, and Mary Ganguli. Predictors of nursing facility admission: A 12-year epidemiological study in the United States. *Journal of the American Geriatrics Society*, 52(3):434–439, 2004.

[373] Robert C Kersting. Impact of social support, diversity, and poverty on nursing home utilization in a nationally representative sample of older Americans. *Social Work in Health Care*, 33(2):67–87, 2001.

[374] Qian Cai, J Warren Salmon, and Mark E Rodgers. Factors associated with long-stay nursing home admissions among the US elderly population: comparison of logistic regression and the Cox proportional hazards model with policy implications for social work. *Social work in health care*, 48(2):154–168, 2009.

[375] Paul Clarkson, Christian Brand, Jane Hughes, and David Challis. Integrating assessments of older people: examining evidence and impact from a randomised controlled trial. *Age and ageing*, 40(3):388–391, 2011.

[376] Emily Grundy and Mark Jitlal. Socio-demographic variations in moves to institutional care 1991–2001: a record linkage study from England and Wales. *Age and ageing*, 36(4):424–430, 2007.

[377] Susanna Gentili, Fabio Riccardi, Leonardo Emberti Gialloreti, Paola Scarcella, Alessandro Stievano, Maria Grazia Proietti, Gennaro Rocco, and Giuseppe Liotta. Admission to the long-term care facilities and institutionalization rate in community-dwelling frail adults: an observational longitudinal cohort study. In *Healthcare*, volume 10, page 317. MDPI, 2022.

[378] Joseph E Gaugler, Sue Duval, Keith A Anderson, and Robert L Kane. Predicting nursing home admission in the US: a meta-analysis. *BMC geriatrics*, 7(1):1–14, 2007.

[379] Andrew Sommerlad, Mika Kivimäki, Eric B Larson, Susanne Röhr, Kokoro Shirai, Archana Singh-Manoux, and Gill Livingston. Social participation and risk of developing dementia. *Nature aging*, 3(5):532–545, 2023.

[380] Gill Livingston, Jonathan Huntley, Andrew Sommerlad, David Ames, Clive Ballard, Sube Banerjee, Carol Brayne, Alistair Burns, Jiska Cohen-Mansfield, Claudia Cooper, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *The lancet*, 396(10248):413–446, 2020.

[381] Terry Therneau and Partricia Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, New York, 2000. ISBN 0-387-98784-3.

[382] Luís Meira-Machado, Jacobo de Uña-Álvarez, Carmen Cadarso-Suárez, and Per K Andersen. Multi-state models for the analysis of time-to-event data. *Statistical methods in medical research*, 18(2):195–222, 2009.

[383] Peter C Austin, Douglas S Lee, and Jason P Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016.

[384] Jason P Fine and Robert J Gray. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association*, 94(446):496–509, 1999.

[385] Per Kragh Andersen and Niels Keiding. Interpretability and importance of functionals in competing risks and multistate models. *Statistics in medicine*, 31(11-12):1074–1088, 2012.

[386] Nicole K Valtorta, Mona Kanaan, Simon Gilbody, and Barbara Hanratty. Loneliness, social isolation and social relationships: what are we measuring? A novel framework for classifying and comparing tools. *BMJ open*, 6(4): e010799, 2016.

[387] Jenny De Jong Gierveld and Theo Van Tilburg. A 6-item scale for overall, emotional, and social loneliness: Confirmatory tests on survey data. *Research on aging*, 28(5):582–598, 2006.

[388] Karen C Jones, Helen Weatherly, Sarah Birch, Adriana Castelli, Martin Chalkley, Alan Dargan, Julien E Forder, Jinbao Gao, Seb Hinde, Sarah Markham, et al. Unit Costs of Health and Social Care 2022 Manual. 2022.

[389] Clare Gardiner, Gideon Geldenhuys, and Merryn Gott. Interventions to reduce social isolation and loneliness among older people: an integrative review. *Health & social care in the community*, 26(2):147–157, 2018.

[390] Hannah M O'Rourke, Laura Collins, and Souraya Sidani. Interventions to address social connectedness and loneliness for older adults: a scoping review. *BMC geriatrics*, 18(1):1–13, 2018.

[391] John T Cacioppo and Louise C Hawkley. Social isolation and health, with an emphasis on underlying mechanisms. *Perspectives in biology and medicine*, 46(3):S39–S52, 2003.

[392] David McDaid and A-La Park. Modelling the economic impact of reducing loneliness in community dwelling older people in England. *International journal of environmental research and public health*, 18(4):1426, 2021.

[393] Joseph R Biden. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. 2023.

[394] HM Treasury. Spring Budget 2024, 2024. URL https://www.gov.uk/government/publications/spring-budget-2024/spring-budget-2024-html. [Accessed: 2024-07-25].

[395] European Commission. Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts: Annexes, 2024.

[396] EU AI Act: First Regulation on Artificial Intelligence, June 2023. URL https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. Last updated: 18-06-2024 - 16:29, Accessed: 2024-09-23.

[397] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[398] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[399] Katherine E Goodman, H Yi Paul, and Daniel J Morgan. AI-Generated Clinical Summaries Require More Than Accuracy. *JAMA*, 2024.

[400] Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. Exploring the limits

of transfer learning with a unified text-to-text transformer. *Google, Tech. Rep.*, 2019.

[401] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

[402] Shanya Sharma, Manan Dey, and Koustuv Sinha. Evaluating gender bias in natural language inference. *arXiv preprint arXiv:2105.05541*, 2021.

[403] Styliani Katsarou, Borja Rodríguez-Gálvez, and Jesse Shanahan. Measuring gender bias in contextualized embeddings. In *Computer Sciences and Mathematics Forum*, volume 3, page 3. MDPI, 2022.

[404] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[405] Jochen Hartmann, Mark Heitmann, Christian Siebert, and Christina Schamp. More than a feeling: Accuracy and application of sentiment analysis. *International Journal of Research in Marketing*, 40(1):75–87, 2023.

[406] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[407] Lik Xun Yuan. distilbert-base-multilingual-cased-sentiments-student (Revision 2e33845), 2023. URL https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student.

[408] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *NeurIPS EMC^2 Workshop*, 2019.

[409] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

[410] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[411] Ioannis Kosmidis. *brglm2: Bias Reduction in Generalized Linear Models*, 2023. URL https://CRAN.R-project.org/package=brglm2. R package version 0.9.2.

[412] Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. Perturbation sensitivity analysis to detect unintended model biases. *arXiv preprint arXiv:1910.04210*, 2019.

[413] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

[414] M. Honnibal and I. Montani. spaCy 3: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing, 2021. URL https://spacy.io/.

[415] Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[416] Yuan, Lik Xun. Bart Large CNN, 2024. URL https://huggingface.co/facebook/bart-large-cnn. [Accessed: 2024-07-25].

[417] Google. T5 Base, 2024. URL https://huggingface.co/google-t5/t5-base. [Accessed: 2024-07-25].

[418] Meta AI. Llama3-8b-Instruct, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct. [Accessed: 2024-07-25].

[419] Sam Rickman. Evaluating gender bias in LLMs in long-term care. https://github.com/samrickman/evaluate-llm-gender-bias-ltc, 2024. Accessed: 2024-08-11.

[420] Russell V. Lenth. *emmeans: Estimated Marginal Means, aka Least-Squares Means*, 2024. URL https://rvlenth.github.io/emmeans/. R package version 1.10.2, https://rvlenth.github.io/emmeans/.

[421] World Economic Forum. Jobs of Tomorrow: Large Language Models and Jobs, September 2023. URL https://www.weforum.org/publications/jobs-of-tomorrow-large-language-models-and-jobs/. Accessed: 2024-10-07.

[422] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[423] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS symposium on machine learning and the law*, volume 1, page 11. Barcelona, Spain, 2016.

[424] Judith Butler. *Gender Trouble*. Routledge, 2002.

[425] Stephen R Pfohl, Heather Cole-Lewis, Rory Sayres, Darlene Neal, Mercy Asiedu, Awa Dieng, Nenad Tomasev, Qazi Mamunur Rashid, Shekoofeh Azizi, Negar Rostamzadeh, et al. A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine*, pages 1–11, 2024.

[426] The Access Group. What is AI in Healthcare?, 2024. URL https://web.archive.org/web/20240922150252/https://www.theaccessgroup.com/en-gb/blog/hsc-what-ai-healthcare/. Accessed: 2024-09-22.

[427] System C. Innovating healthcare with AI - system C's AI hackathon, March 2024. URL https://web.archive.org/web/20240922150309/https://www.systemc.com/resource-hub/blogs/innovating-healthcare-with-ai-system-cs-ai-hackathon/. Accessed: 2024-09-22.

[428] The Access Group. Access unveils access EVO: New AI-enhanced software to support HR professionals, 2024. URL https://web.archive.org/web/20240922150508/https://www.theaccessgroup.com/en-gb/about/news/access-unveils-access-evo-new-ai-enhanced-software-to-support-hr-professionals/. Accessed: 2024-09-22.

[429] James Gavin, Paul Clarkson, Paul Muckelt, Rachael Eckford, Euan Sadler, Suzanne McDonough, and Mary Barker. 203 Healthcare professional and commissioners' perspectives on the factors facilitating and hindering the implementation of digital tools for self-management of long-term conditions within UK healthcare pathways. *European Journal of Public Health*, 34 (Supplement_2):ckae114–074, 2024.

[430] Meta AI. Llama-3.1-8B-Instruct, 2024. URL https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct. [Accessed: 2024-07-25].

[431] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.

[432] Christine De la Maisonneuve and Joaquim Oliveira Martins. The future of health and long-term care spending. *OECD Journal: Economic Studies*, 2014(1):61–96, 2015.

[433] Hannah M James, Chrysanthi Papoutsi, Joseph Wherton, Trisha Greenhalgh, and Sara E Shaw. Spread, scale-up, and sustainability of video consulting in health care: systematic review and synthesis guided by the NASSS framework. *Journal of medical Internet research*, 23(1):e23775, 2021.

[434] Valentina Zigante, Juliette Malley, Annette Boaz, Ewan Ferlie, and Gerald Wistow. How can the adult social care sector develop, scale and spread innovations? *A review of the literature from an organisational perspective*, 2022.

[435] Andrew H Van de Ven. The innovation journey: you can't control it, but you can learn to maneuver it. *Innovation*, 19(1):39–42, 2017.

[436] Johan Ordish. Large Language Models and Software as a Medical Device. https://medregs.blog.gov.uk/2023/03/03/large-language-models-and-software-as-a-medical-device/, 2023. Accessed: 2024-08-27.

[437] Guidance MHRA. Medical device stand-alone software including apps. *Medicines and Healthcare products Regulatory Agency*, 2019.

[438] Data Protection Act 2018, 2018. URL https://www.legislation.gov.uk/ukpga/2018/12/part/3. Accessed: 2024-09-23.

[439] Information Commissioner's Office. Overview: Data Protection and the EU, 2023. URL https://ico.org.uk/for-organisations/data-protection-and-the-eu/overview-data-protection-and-the-eu/. Accessed: 2024-09-23.

[440] Generative AI, Bias, Hallucinations and GDPR. *Fieldfisher*, Aug 2023. URL https://web.archive.org/web/20241007184426/https://www.fieldfisher.com/en/insights/generative-ai-bias-hallucinations-and-gdpr. Accessed via Wayback Machine.

[441] Małgorzata Kuśmierczyk. Algorithmic Bias in the Light of the GDPR and the Proposed AI Act. *In) equality. Faces of modern Europe", Wydawnictwo Centrum Studiów Niemieckich i Europejskich im. Willy'ego Brandta, Wrocław*, 2022.

[442] Janos Meszaros, Jusaku Minari, and Isabelle Huys. The future regulation of artificial intelligence systems in healthcare services and medical research in the European Union. *Frontiers in Genetics*, 13:927721, 2022.

[443] Jed Meers. Forms of fettering: Application forms and the exercise of discretion in the welfare state. *Journal of Social Welfare and Family Law*, 42(2): 221–242, 2020.

[444] British Oxygen Co Ltd v Minister of Technology, 1970. URL http://www.bailii.org/uk/cases/UKHL/1970/4.html. Cite as: [1970] 3 All ER 165, [1971] AC 610, [1969] 2 WLR 892, [1970] 3 WLR 488, [1970] UKHL 4.

[445] UK Parliament. Data (Use and Access) Bill [HL]. https://bills.parliament.uk/bills/3825/publications, 2024. URL https://bills.parliament.uk/bills/3825/publications. Government Bill, Originated in the House of Lords, Session 2024-25.

[446] Prime Minister's Office, 10 Downing Street and His Majesty King Charles III. The King's Speech 2024: His Majesty's most gracious speech to both Houses of Parliament. Oral statement to Parliament, Jul 2024. URL https://web.archive.org/web/20241004154041/https://www.gov.uk/government/speeches/the-kings-speech-2024. Accessed via Wayback Machine, Delivered at the House of Lords.

[447] Bertin Martens. The European Union AI Act: premature or precocious regulation? https://www.bruegel.org/analysis/european-union-ai-act-premature-or-precocious-regulation, 2024. Accessed: 2024-08-27.

[448] Manuel Wörsdörfer. Biden's Executive Order on AI and the EU's AI Act: A Comparative Computer-Ethical Analysis. *Philosophy & Technology*, 37 (3):74, 2024.

[449] Lauren D Harris-Kojetin, Manisha Sengupta, Eunice Park-Lee, and Roberto Valverde. Long-term care services in the United States: 2013 overview. 2013.

[450] FDA: Food and Drug Administration. Artificial Intelligence and Machine Learning in Software as a Medical Device. FDA website, Sep 2024. URL https://web.archive.org/web/20241009135853/https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device#regulation. Accessed via Wayback Machine.

[451] Sandeep Reddy. Evaluating large language models for use in healthcare: A framework for translational value assessment. *Informatics in Medicine Unlocked*, page 101304, 2023.

[452] Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk. *Regulation & Governance*, 18(1): 3–32, 2024.

[453] Philippe Aghion, Antonin Bergeaud, and John Van Reenen. The impact of regulation on innovation. *American Economic Review*, 113(11):2894–2936, 2023.

[454] Oskar J Gstrein, Noman Haleem, and Andrej Zwitter. General-purpose AI regulation and the European Union AI Act. *Internet Policy Review*, 13(3): 1–26, 2024.

[455] White House. Artificial Intelligence for the American People. *National Archives and Records Administration*, 2019.

[456] Alex Engler. The EU and US diverge on AI regulation: A transatlantic comparison and steps to alignment. 2023.

[457] Republican Party. 2024 Republican Party Platform, July 2024. URL https://www.presidency.ucsb.edu/node/373351. Approved by the platform committee on July 8, 2024, and formally adopted at the Republican National Convention on July 15, 2024.

[458] Piero Cingari. Why Trump's plans for tariffs could be bad for Europe's economy. *Euronews*, Nov 2024. URL https://www.euronews.com/business/2024/11/04/how-much-could-trumps-tariffs-damage-europes-economy. Published online.

[459] Oxford Analytica. Trump AI policy would prioritise US-China competition. *Emerald Expert Briefings*, (oxan-db), 2024.

[460] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772, 2009. ISSN 1532-0464.

[461] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.", 2009.

[462] Esam Alzahrani and Leon Jololian. How Different Text-preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors. *arXiv preprint arXiv:2109.13890*, 2021.

[463] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[464] Hualong Yu, Xibei Yang, Shang Zheng, and Changyin Sun. Active learning from imbalanced data: A solution of misc weighted extreme learning machine. *IEEE transactions on neural networks and learning systems*, 30(4): 1088–1103, 2018. ISSN 2162-237X.

[465] William Revelle. *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois, 2022. URL https://CRAN.R-project.org/package=psych. R package version 2.2.9.

[466] John Hughes. *krippendorffsalpha: Measuring Agreement Using Krippendorff's Alpha Coefficient*, 2022. URL https://CRAN.R-project.org/package=krippendorffsalpha. R package version 2.0.

[467] Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. Ontonotes release 5.0 ldc2013t19. *Linguistic Data Consortium, Philadelphia, PA*, 23, 2013.

[468] Mostafa Dehghani, Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, and Łukasz Kaiser. Universal transformers. *arXiv preprint arXiv:1807.03819*, 2018.

[469] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions arXiv [math. NA], 2009.

[470] Peter Bühlmann and Bin Yu. Analyzing bagging. *The annals of Statistics*, 30(4):927–961, 2002.

[471] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25: 197–227, 2016.

[472] Alaa Tharwat. Linear vs. quadratic discriminant analysis classifier: a tutorial. *International Journal of Applied Pattern Recognition*, 3(2):145–180, 2016.

[473] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

[474] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John wiley & sons, 2010.

[475] Cora JM Maas and Joop J Hox. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational statistics & data analysis*, 46(3):427–440, 2004.

[476] James Carpenter and John Bithell. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in medicine*, 19(9):1141–1164, 2000.

[477] José C. Pinheiro and Douglas M. Bates. *Mixed-Effects Models in S and S-PLUS*. Springer, New York, 2000. doi: 10.1007/b98882.

[478] Manuel Koller. robustlmm: An R package for robust estimation of linear mixed-effects models. *Journal of Statistical Software*, 75(6):1–24, 2016. doi: 10.18637/jss.v075.i06.

[479] Jun Yan. geepack: Yet Another Package for Generalized Estimating Equations. *R-News*, 2/3:12–14, 2002.

[480] Brennan C Kahan, Gordon Forbes, Yunus Ali, Vipul Jairath, Stephen Bremner, Michael O Harhay, Richard Hooper, Neil Wright, Sandra M Eldridge, and Clémence Leyrat. Increased risk of type I errors in cluster randomised trials with small or medium numbers of clusters: a review, reanalysis, and simulation study. *Trials*, 17:1–8, 2016.

[481] Python Software Foundation. Python 3.12.5 documentation: Built-in Types. https://docs.python.org/3/library/stdtypes.html#str.startswith, 2024. Accessed: 2024-08-11.

11 Appendices

11.1 Appendices to Chapter 6

11.1.1 S1 Data Flow Appendix

In Figure 11.1, we set out the secure flow of identifiable data, the pseudonymisation process, data analysis and model output.

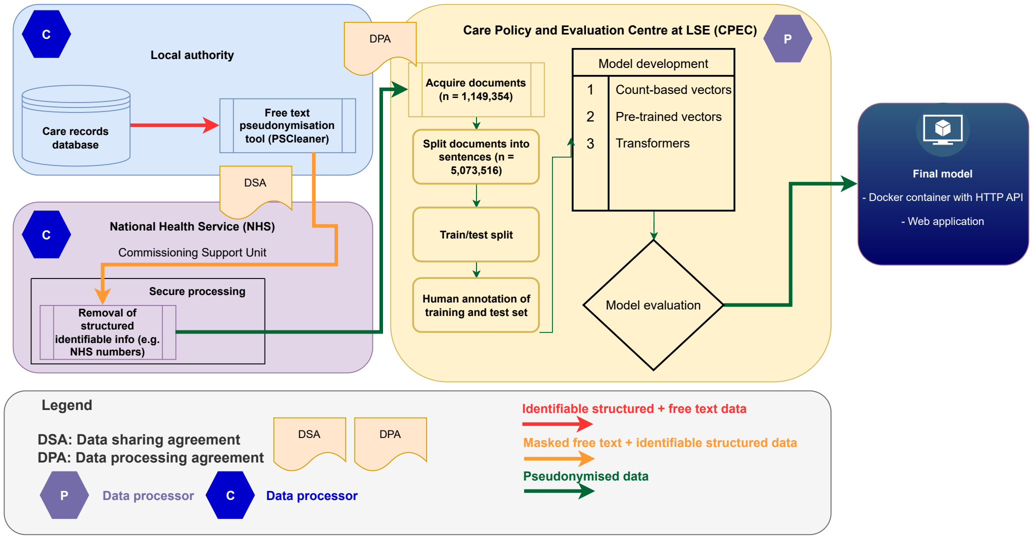

C
Local authority
DPA
Care records database
Free text pseudonymisation tool (PSCleaner)
DSA
C
National Health Service (NHS)
Commissioning Support Unit
Secure processing
Removal of structured identifiable info (e.g. NHS numbers)
Care Policy and Evaluation Centre at LSE (CPEC)
P
Acquire documents (n = 1,149,354)
Split documents into sentences (n = 5,073,516)
Train/test split
Human annotation of training and test set
Model development
1 Count-based vectors
2 Pre-trained vectors
3 Transformers
Model evaluation
Final model
- Docker container with HTTP API
- Web application
Legend
DSA: Data sharing agreement
DPA: Data processing agreement
DSA
DPA
P
Data processor
C
Data processor
Identifiable structured + free text data
Masked free text + identifiable structured data
Pseudonymised data

Figure 11.1: Data flow diagram

11.1.2 S2 Methods Appendix

We include below more details about the steps set out in the Model development section.

11.1.2.1 Data pre-processing

We received the free text with tokens in place of pseudonymised text, e.g. `**NAME**` or `**DATE**`. For count-based vectors, we pre-processed text by converting to lower case, removing stop words, such as “and” and “the” [460]. We also lemmatised words, restoring them to base form. For example, walks, walked and walking were all replaced with the normalised form, walk [166] using the Python `NLTK` package’s `WordNetLemmatizer` [461]. We also replaced non-ASCII punctuation characters such as bullet points, which can cause text encoding issues.

Conversely, the word vector language representation models we used were trained on text without lemmatization, and including stop words, which optimises model performance [462, 463]. However, we replaced masked text such as `**NAME**` or `**DATE**` with synthetic names and dates (using the same randomly chosen value for each tag) as set out in Table 11.1. We show in the example case note in our Methods and Materials section how, even in case notes which are indicative of loneliness or social isolation, much of the text in the note may be unrelated to the topic. To ensure that the classifier could learn which text is relevant, we split each case note into sentences for classification by human annotators, using the `SpaCy` Python package’s sentence tokeniser [414]. The data pre-processing is set out in Figure 11.2.

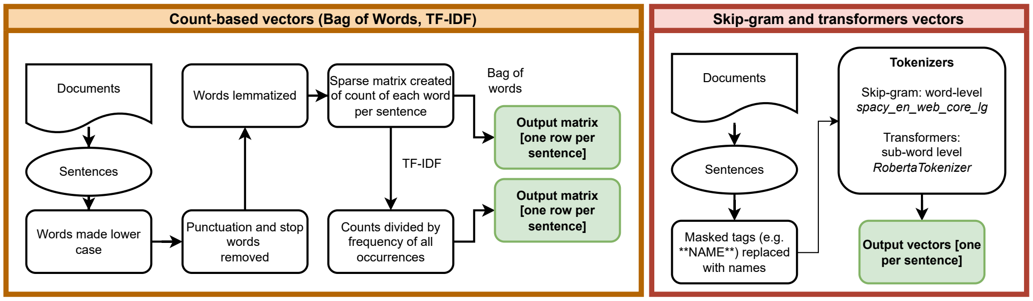

Count-based vectors (Bag of Words, TF-IDF)
Documents
Sentences
Words made lower case
Punctuation and stop words removed
Words lemmatized
Sparse matrix created of count of each word per sentence
Bag of words
Output matrix [one row per sentence]
TF-IDF
Counts divided by frequency of all occurrences
Output matrix [one row per sentence]
Skip-gram and transformers vectors
Documents
Sentences
Masked tags (e.g. **NAME**) replaced with names
Tokenizers
Skip-gram: word-level
spacy_en_web_core_lg
Transformers: sub-word level
RobertaTokenizer
Output vectors [one per sentence]

Figure 11.2: Pre-processing

Table 11.1: Pseudonymisation masks

Mask	Replacement
ETHNIC	British
EMAIL	a@a.com
NAME	Byron
POSTCODE	SW1A 0AA
DATETIME	1970-01-01
CURRENCY	£
TELEPHONE	07777777777
LOCATION	London
TIME	3pm

11.1.2.2 Labelling data in the training and test sets

The training and test sets each have over 300,000 sentences. It was neither possible nor desirable to annotate all case notes. Furthermore, as case notes can be about any topic, random selection would yield only a small proportion related to the area of interest. To maximise the number of relevant notes for the model to be trained on, we used active learning. This is a machine learning "closed loop" method, where a seed set of dictionary terms is used to label a dataset, which is in turn used to train a classification model, which predicts more matches, which are in turn labelled, from which more terms are generated, until sufficient notes are annotated or saturation of the search terms is achieved [464]. We began with a dictionary of terms such as "alone", "isolated" and "friends", selected to indicate sentences which *might* be related to social isolation and loneliness. Notes from the training set which contained these terms were then manually classified by human annotators. After the first set of notes were classified, we added terms from the classified notes to the list, until saturation was reached. We classified 10,083 sentences in the training set, and 3,573 sentences in the test set for model evaluation. We defined a set of rules for annotators to define which sentences to classify, using binary classification. Each note was annotated as being indicative of social isolation and loneliness (positive class), or not indicative (negative class). We include the initial dictionary, and classification rules, in the Classification rules section and the Dictionary terms section. It is important to note that the purpose of the dictionary was to prioritise sentences to be classified by human annotators: the final model can predict in either class sentences which do not contain these terms.

Both the training and test sets were manually classified by human annotators, using a binary classification scheme to identify sentences indicative of loneliness or social isolation. We measured inter-rater reliability using Cohen's κ [346] and Krippendorf's α [347]. We randomly selected 300 sentences (150 in each class) to be classified by both human annotators. They agreed in 278 cases and disagreed in 22, giving κ of 0.89 (95% CI 0.84 - 0.94), and α of 0.89 (95% CI 0.89 - 0.93), calculated using the R `psych` and `krippendorffsalpha` packages [465, 466]. The maximum level of agreement in both cases is 1, and 0.89 represents excellent levels of agreement beyond chance [131, 348].

11.1.2.3 Creating sentence vectors

Count-based and pre-trained vectors Count-based approaches: document-term matrices and tf-idf. We split each sentence into lemmatised, word-level tokens as set out in Data pre-processing. We then created a document-term matrix. This is a large, sparse matrix where the columns are the set of all words in the corpus, and each row is represented by a sentence. The values in each row indicate the frequency of each word in the sentence [170]. We also applied tf-idf to transform the count matrix to a weighted representation, reducing the weighting of higher frequency words across all documents. We used the approach set out in [180], where the tf-idf for a word w in a corpus c is,

$$\text{tf-idf}(w, c) = \text{tf}(w, c) \cdot \text{idf}(w)$$

$\text{tf}(w, c)$ is the count of the word in the corpus and $\text{idf}(w)$ is,

$$\text{idf}(w) = \log \frac{n}{\text{df}(w)} + 1$$

where n is the total number of documents in the corpus, and df(w) is the number of documents in the corpus containing the word w.

Pre-trained word embeddings We used the Spacy large English model [344]. This model is trained on a large body of news, Wikipedia and subtitles data, and contains a parser based on OntoNotes 5 [467]. The model represents language through dense embeddings [175], where words which have similar semantic meanings are clustered together in vector space. The model contains 684,830 unique, 300-dimensional pre-trained English vectors. We split each sentence into

n word-level tokens, mapping each token to its corresponding vector. As sentence lengths differ, this creates variable-length representations of each sentence, which is a problem for many classification algorithms. Sentence embeddings constructed from an average of word embeddings have proven to be a strong baseline across a variety of tasks [185]. We therefore stacked n token vectors for each sentence into a $300 \times n$ matrix, taking the mean of each dimension to create a single 300-dimensional vector to represent the sentence.

Transformer-based approaches A transformer is a neural network architecture for computing representations of a sequence, which in our case is a sentence. Unlike the architecture used to generate the Spacy model [344], transformer-based approaches can learn context-dependent representations of words. Transformers use an attention mechanism to assign weights to each token in the sequence, aiming to allow the model to learn relevant long-range dependencies between tokens without the computational cost of calculating weights for all words in between [192]. This avoids vanishing gradient problems that can arise in other architectures such as Recurrent Neural Networks (RNNs) as the size of the context window increases [468]. Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based model designed to be a general language representation that can be fine-tuned with just one additional layer to provide state-of-the-art models for a wide range of tasks, without significant task- or language-specific changes to the model's architecture [187]. For our task, we used RoBERTa, an incremental improvement on BERT with the same architecture but different training data and hyperparameters, and generally slightly improved accuracy [195]. We used the RoBERTa *base* model, which has 12 hidden layers, 768 dimensions and 12 heads. This was relatively computationally expensive to fine-tune, so for comparison we also used DistilRoBERTa, which has identical parameters except it has 6 hidden layers, and is around twice as fast to train. In both cases, we used the HuggingFace implementation of each model's tokeniser to split each sentence into sub-word tokens [349, 350]. Each token was then converted by the model into a dense vector representation, based on pre-trained word embeddings, and further fine-tuned by the model to capture semantic information about the contextual meaning. These token-level vectors were then passed into the model's encoder, a neural network which produces a fixed-length vector representation of the input text, summarising the relevant information in the text for the model's classifier.

11.1.2.4 Classification algorithms for count-based and pre-trained vectors

The count-based and SpaCy approaches are methods to represent sentences in vector space. Once the representation of each sentence is in place, there are many methods that could reasonably be attempted to optimally distinguish the classes. Conversely, the transformer-based architecture we used includes a feed-forward neural network classifier in the final layer, so we only used this method for these models. After pre-processing, vectorizing and labelling each sentence, the problem is a binary classification task. For both the count and pre-trained embedding based approaches, we evaluated five classification algorithms. We used k fold cross-validation to avoid overfitting on the training set, choosing 5 folds for k as a value which tends to elicit reasonably high accuracy [351] while reducing training time compared with higher values.

The document-term matrix and tf-idf approaches both create sparse $m \times n$ matrices, where m is the number of sentences and n is the number of unique terms (in our case $n = 5828$). To ensure that sparse matrices did not inhibit model performance, we reduced these to $m \times 1000$ matrices, using the implementation of truncated singular value decomposition in [180], based on the method in [469]. However, this did not lead to improved classification, and the results in the Results section use the sparse matrices as inputs.

For the count-based and SpaCy vectors, we used the following classification algorithms:

1. Class-weighted logistic regression. We adjusted the C penalty parameter (weight of each class) in inverse proportion to the class frequency, so mistakes from the minority class have a higher cost, as in [259]. This was implemented using scikit-learn's built-in `class_weight="balanced"` setting.
2. Bootstrap aggregation ("bagging"). Bagging is one of the most effective classification algorithms for high-dimensional datasets [470]. We used a decision-tree based bagging method, where multiple decision trees are created and trained on a bootstrapped sample of points drawn from a subset of the data. Once the decision trees have been trained, an overall prediction is created by aggregating the predictions and taking an average. The use of bootstrapped subsets of points and average predictions aim to reduce overfitting. We took an average of 10 decision tree classifiers, and did not limit the depth of the decision trees.

3. Random forest. This is a decision-tree approach where M decision trees are created to classify the data [471]. As with bagging, each decision tree in the forest is created from a bootstrapped subset of the training data. However, in addition, each tree only uses a random subset of features (in our case words) from those points. This is done to prevent any individual decision tree from becoming too dependent on a single feature. Random forests can handle datasets with many features and are robust to outliers. We set $M = 100$ trees.
4. Quadratic Discriminant Analysis (QDA). This is similar to Linear Discriminant Analysis (LDA), but allows each class to have its own covariance matrix [472]. The prior probability of each class is set to the proportion of training samples in each class. The likelihood of a novel data point being in each class is calculated from its position in vector space given the class-specific mean and covariance, and then multiplied by the prior class probabilities. The final prediction is the class which maximises this posterior probability.
5. Feed forward neural network. We tested with a range of hyperparameters to find the optimal architecture. The best performing network had three dense layers with 100 neurons, one drop-out layer and a final layer with one neuron (either 0 or 1 for the respective classes). In order to mitigate training problems caused by imbalanced classes, we used the same approach as with logistic regression, setting class weights in inverse proportion to their frequency. In addition, we tried oversampling the minority class to increase the number of positive samples available in the model, and undersampling the negative sentences. However, neither of these approaches was effective at improving metrics and the values in the Results section do not use these approaches.

We used the implementation in the Python Scikit-Learn package for each of these approaches [180]. We also replicated the feed forward neural network using TensorFlow in order to implement class weighting and over- and under-sampling [473].

11.1.2.5 Classification rules

This section contains the classification rules for manual annotators of the free text notes. We set out in Table 11.2 the rules for annotators to identify cases where loneliness or social isolation is indicated, and in Table 11.3 cases where it is not.

Table 11.2: Loneliness or social isolation: positive cases

Situation	Example
Statement that the person is lonely and socially isolated	Mrs Byron appears lonely and has little social contact.
Statement the person feels lonely even if not socially isolated	She further advised that she gets out regularly and doesn't know why she feels lonely.
Statement that the person is socially isolated provided it does not state that they are not lonely	He does not go out or have any family or friends.
Statement that social isolation is a risk	High risk of social isolation.
Statement that person lives alone if they do not like it	Due to Mrs Byron being alone in the day she would like to request an increase.
Person expresses interest in attending a day centre with no reasons, or because of social isolation / loneliness. If reasons given it must be for social reasons (not for carer respite or managing safety).	She would like to attend the day centre. She would like to attend the day centre to increase social contact.
Referrals to befriending service	Byron stated that she did not need a carer but would like a befriender.
Receipt of befriending service	Byron's has a befriender who visits twice a week
Request for social support outside day centre	He said he does not want to go to a day centre but would like someone to talk to a couple of times a week.
Number of social contacts defined with subjective language	He barely sees family or friends.

Lonely or socially isolated

Table 11.3: Loneliness or social isolation: negative cases

Situation	Example
Need for social support for purposes of managing safety	He attends a day centre on Mondays and Thursdays while his daughter is at work, as he cannot safely be left alone.
Need for social support for purposes of carer respite	He goes to a day centre two days a week to give his wife a break.
Need for support because of risks associated with safety	She needs support of one person outdoors because she is unable to safely assess risk crossing the road.
Need for support to manage practical tasks	He needs support with managing nutrition and meals.
Generic request for support	She has requested social services support as she is anxious re ability to cope.
Need for support because of depression/anxiety	She has requested social services support, as she informed me that her mother appeared to be depressed , staying in her bed all the time.
Person is isolated in the sense of infection-control.	He has been isolated in his room until he does not have a fever.
Statement that person lives alone with no additional information	He lives alone
Person has n social contacts per week without subjective language	He seems family twice a week.
The fact that a person attends a day centre	Mr Byron has a very social active life, he attends Day Centres and Clubs between Monday to Friday.
Offer of day centre refused	I offered to refer Mrs X to a day centre but she declined.
Befriending for the purpose of carer support	I will also refer to AGE UK Clinical navigators for any assistance they may be able to provide such as a sitting service/befriending to allow Mrs X time to attend her own hospital appointments.

Not lonely or socially isolated

Table 11.4: Loneliness or social isolation: befriending vs day centre

	Point of referral	Service currently being received
Befriending	Lonely	Lonely
Day centre	Lonely	Not lonely

Note on befriending and day centre services Although befriending and day centres are both services for loneliness, they are treated differently in the following way. The assumption is that as befriending is a voluntary service with a high threshold and low intensity (often an hour visit per week), individuals receiving it may not receive sufficient quantity to eliminate loneliness. Conversely, day centre services have a much higher intensity, typically at least several hours per day. While we assume that at the point of day centre referral there is a need for social inclusion (unless the referral is for other reasons such as carer respite), we do not assume that all individuals in receipt of day services continue to be lonely or socially isolated. This is set out in Table 11.4.

11.1.2.6 Wide dictionary terms

This section contains the dictionary terms used to prioritise case notes to be classified by humans. It is important to note that the purpose of the dictionary was to prioritise sentences to be classified by human annotators. Sentences that contain these terms are not automatically classified by the model as indicative of loneliness or social isolation. Similarly, sentences which the model classifies as indicative of loneliness or isolation do not need to contain any of these terms.

```
alone
companionship
divorce
divorced
engagement
family support
friends
isolated
isolation
left out
lives alone
loneliness
lonely
lonesome
no friend
no friends
on her own
on his own
on their own
reclusive
secluded
separated
single
social connection
social contact
social network
social support
social withdraw
socially isolating
socially withdraw
solitary
solitude
support network
widow
widowed
```

```
withdrawn
befriend
interaction
activities
company
bored
```

11.1.3 S3 Open source version of model

Our classification model can be downloaded and run at https://github.com/samrickman/lonelinessmodel, which can be run on large volumes of free text to generate classifications. The model is reproducible as it is encapsulated in a Docker container.

11.1.4 S4 Assessing which information to extract from administrative records

This section sets out the process by which it was decided that loneliness should be the candidate for extraction. Given the need for information and the capacity to extract it using LLMs, the next logical question is which information should be prioritised. I set out in Figure 11.3 the decision-making process I designed to evaluate whether specific data from unstructured records should be extracted and used for analysis. The flowchart begins by assessing whether the data can be reasonably expected to be recorded in a way that answers the research question. If the data is potentially useful, the next consideration is whether the construct is well-defined, as constructs like loneliness may be more straightforward to define and extract than more categories like abuse and neglect, which is in fact an amalgamation of several, potentially distinct areas.

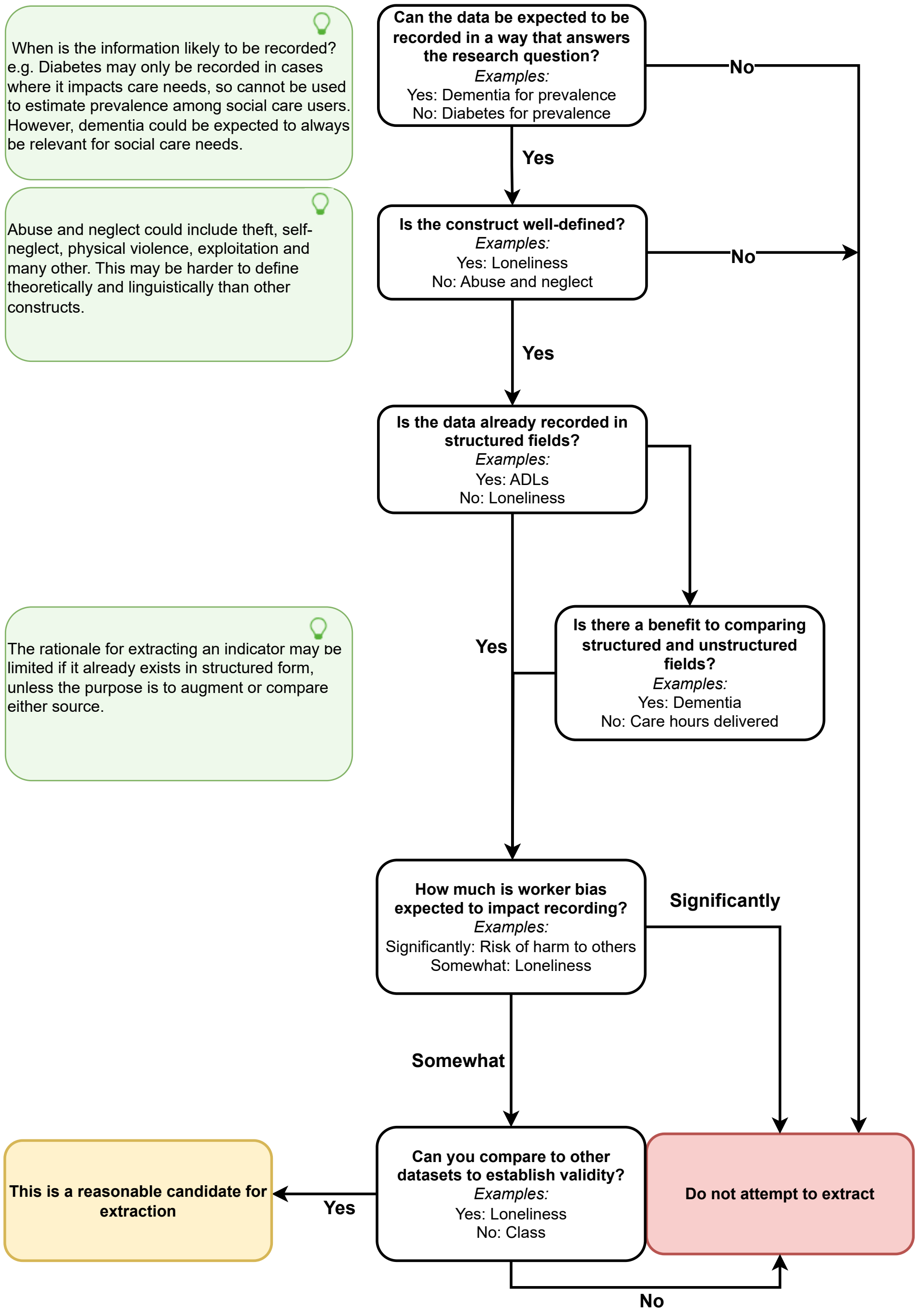

When is the information likely to be recorded? e.g. Diabetes may only be recorded in cases where it impacts care needs, so cannot be used to estimate prevalence among social care users. However, dementia could be expected to always be relevant for social care needs.
Abuse and neglect could include theft, self-neglect, physical violence, exploitation and many other. This may be harder to define theoretically and linguistically than other constructs.
The rationale for extracting an indicator may be limited if it already exists in structured form, unless the purpose is to augment or compare either source.
Can the data be expected to be recorded in a way that answers the research question?
Examples:
Yes: Dementia for prevalence
No: Diabetes for prevalence
No
Yes
Is the construct well-defined?
Examples:
Yes: Loneliness
No: Abuse and neglect
No
Yes
Is the data already recorded in structured fields?
Examples:
Yes: ADLs
No: Loneliness
Yes
Is there a benefit to comparing structured and unstructured fields?
Examples:
Yes: Dementia
No: Care hours delivered
How much is worker bias expected to impact recording?
Examples:
Significantly: Risk of harm to others
Somewhat: Loneliness
Significantly
Somewhat
This is a reasonable candidate for extraction
Yes
Can you compare to other datasets to establish validity?
Examples:
Yes: Loneliness
No: Class
Do not attempt to extract
No

Figure 11.3: Process for prioritising information to extract

If the construct is well-defined, the process checks whether the data is already available in structured form. If it is, the rationale for extracting it from unstructured text may be limited unless there is a specific need to augment or compare data sources. The flowchart also considers the benefits of comparing structured and unstructured fields, as well as the potential impact of worker bias on the recording of the data. Constructs that are subject to significant bias or cannot be validated against other datasets may not be suitable for extraction. By following this structured approach, I arrived at the conclusion that loneliness is a good candidate for extraction, as it is expected to be contained in social care records, not already included in structured records, and is recorded in other datasets to which it can be compared to establish validity and assess bias.

11.1.5 S5 Quantity of training data required

I trained the model on 10,083 manually labelled case notes to achieve the F_1 score of 0.92 presented in the paper. However, the paper does not establish whether a sample of this size was required, or whether training on more notes could have improved the results. I present in Figure 11.4 the F_1 score for a model trained on fewer notes. I calculated this by taking 10 samples of the 10,083 notes ranging from 0 to 10,000 notes, maintaining the 0.925 : 0.075 ratio of negative to positive case notes in the training data, and training a model on this subset of data. I created the confidence interval for the F_1 score in Figure 11.4 using the confidence interval from a t-distribution [474], as specified in Equation (11.1).

$$\mathrm{CI} = \bar{F}_1 \pm t_{\alpha/2,\,df} \times \frac{s}{\sqrt{n}} \tag{11.1}$$

where:

- $\bar{F}_1$ is the mean F_1 score across the 10 models for each value of N samples,
- $t_{\alpha/2,\,df}$ is the critical value from the t-distribution for a $(1-\alpha)$ confidence level with $df = n - 1$ degrees of freedom,
- $\alpha = 0.05$
- s is the standard deviation of the sample of F_1 scores, and
- n is the number of samples (in this case, $n = 10$).

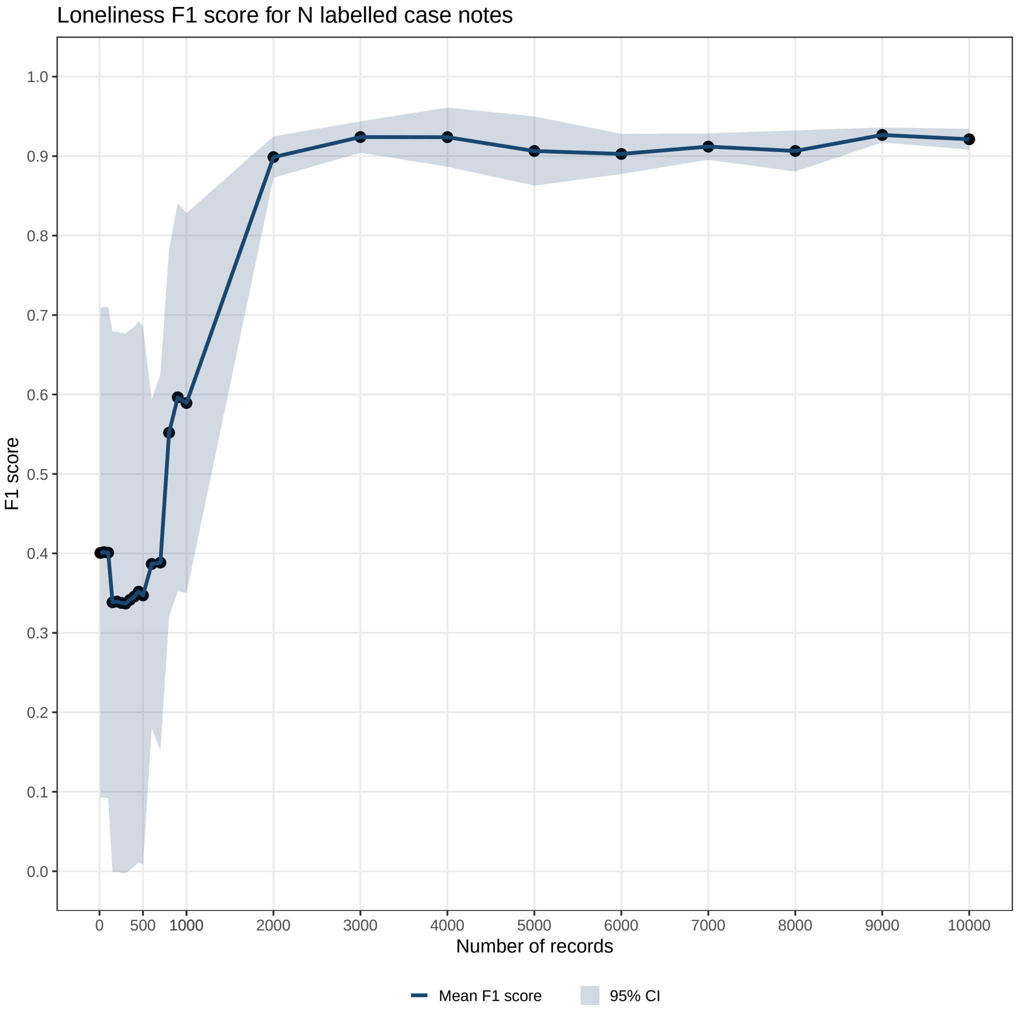


Figure 11.4: Mean loneliness model F1 score

Figure 11.4 shows the mean F1 score and the 0.95 confidence interval within which F_1 score lies, accounting for variability across different training subsets, assuming that the distribution of F_1 scores is approximately normal. The mean F_1 score is relatively low for 0 to 1000 notes, but at around 2000 notes (an increase from 75 to 150 positive samples) it jumps to 0.89 (95% CI 0.87 - 0.92). It then hits 0.92 at 3000 notes and remains relatively stable from this point, though with slightly more variation and wider confidence intervals until around 9000 notes. This indicates that a model achieving similar results could have been trained using about a third of the labelled notes.

11.2 Appendices to Chapter 7

11.2.1 Cumulative incidence stratified by cognition

We show in Figure 11.5 the predicted cumulative incidence of care home entry for those with and without a cognitive impairment, holding all continuous variables at their mean, the categorical variables to the middle value, and the binary variables to 0 (i.e. these curves are for a woman who lives alone). These demonstrate the impact of cognition on results, and why the proportional hazards assumption is not met for cognition.

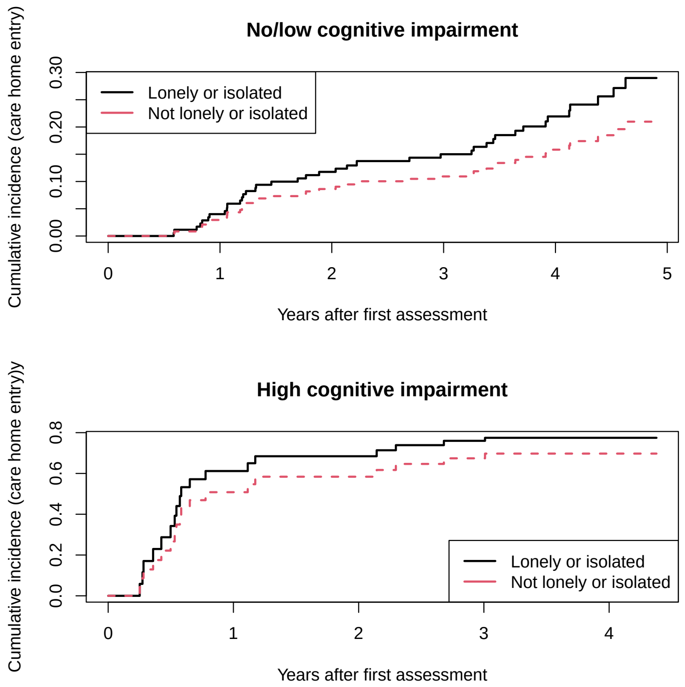


Figure 11.5: Predicted cumulative incidence

11.2.2 Exploring the role of age in loneliness and care home entry

To better understand the relationship between age, loneliness, and care home entry, we conducted two supplementary analyses. First, we re-specified age as a binary variable, aged < 85 ($N = 570$) and ≥ 85 ($N = 531$), to assess whether this approach influences the results. Second, we stratified our dataset into two age groups to examine how the effect of loneliness differs across these groups. The conclusion of these analyses is that the oldest older adults are at higher risk of care home entry in five years. Loneliness remains an important predictor of care home entry in these models. We present the output in full below.

11.2.2.1 Impact of binary age specification on loneliness and care home entry

As a further analysis of the impact of age on our results, we fitted a model where age is specified as binary rather than continuous. We split individuals into aged < 85 ($N = 570$) and ≥ 85 ($N = 531$), removing the quadratic age term (age^2) from the model specification. For the Fine & Gray model, this led to the violation of the proportional hazards assumption for the cost of day care services, so we stratified by individuals based on whether they received day care or not. After this step, the assumption was satisfied. We present in Table 11.5 the output from the logistic regression model and in Table 11.6 the output from the competing risks model.

The results for loneliness are similar to those in the main body of the paper. In the logistic regression and cause-specific hazard models, magnitude of the loneliness coefficient is slightly greater than the originally specified models, at 1.47 compared with 1.45 and 1.33 compared with 1.32, respectively, and the the p values are slightly smaller. In the Fine & Gray model, we see the reverse, with the coefficient slightly smaller (1.36 compared with 1.39) and the p value slightly larger but still significant at $\alpha = 0.05$. The inclusion of age as a binary variable does not meaningfully change the results for loneliness in any of the models. However, it is of note that age is now significant ($p < 0.05$) in all three models, with a coefficient of 1.88 in the logistic regression model, 1.61 in the cause-specific hazard model and 1.42 in the Fine & Gray model. These results suggest that individuals aged 85 and older have an increased likelihood of care home entry compared to those under 85, which is consistent with the literature [e.g. 377]. The 95% confidence intervals for the estimates of all three models overlap, so it may not be appropriate to place too much interpretation on the differences between the models. These findings underscore that, while age is an important factor, loneliness remains a robust predictor of care home entry, regardless of how age is specified.

Table 11.5: Logistic regression model output (categorical age)

	Odds ratio	
Loneliness		
Lonely or Isolated	1.47 (1.06-2.04, p=0.021)	*
Demographics		
Sex: Male	1.25 (0.91-1.71, p=0.158)	
Age85+	1.88 (1.38-2.57, p<0.001)	***
Ethnicity: White	1.36 (0.98-1.91, p=0.071)	.
Lives Alone	1.64 (1.16-2.32, p=0.005)	**
Unpaid Care	0.83 (0.58-1.19, p=0.310)	
Needs		
N Notes	1.00 (1.00-1.00, p<0.001)	***
Personal care: Moderate	0.79 (0.54-1.14, p=0.209)	
Personal care: High	1.10 (0.63-1.93, p=0.733)	
Cognition: Moderate	2.78 (1.87-4.12, p<0.001)	***
Cognition: High	3.83 (2.54-5.80, p<0.001)	***
Shopping and Meals: Moderate	1.02 (0.63-1.66, p=0.938)	
Shopping and Meals: High	0.63 (0.37-1.09, p=0.095)	.
Services		
Cost DPs	1.00 (0.99-1.00, p=0.152)	
Cost Daycare	1.00 (1.00-1.01, p=0.345)	
Cost Homecare	1.00 (1.00-1.00, p=0.502)	
Has Telecare	0.94 (0.65-1.35, p=0.756)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

Table 11.6: Competing Risks model output (categorical age)

	Cause-specific hazard		Fine & Gray	
Loneliness				
Lonely or Isolated	1.33 (1.02-1.73, p=0.036)	*	1.36 (1.05-1.76, p=0.020)	*
Demographics				
Age85+	1.61 (1.24-2.08, p<0.001)	***	1.42 (1.12-1.79, p=0.004)	**
Ethnicity: White	1.37 (1.04-1.81, p=0.026)	*	1.25 (0.96-1.62, p=0.095)	.
Lives Alone	1.55 (1.18-2.05, p=0.002)	**	1.47 (1.12-1.91, p=0.005)	**
Sex: Male	1.29 (0.99-1.67, p=0.058)	.	1.13 (0.88-1.44, p=0.340)	
Unpaid Care	0.95 (0.70-1.27, p=0.713)		0.92 (0.70-1.21, p=0.555)	
Needs				
Personal care: High	1.43 (0.87-2.34, p=0.159)		1.06 (0.67-1.67, p=0.806)	
Personal care: Moderate	0.97 (0.71-1.34, p=0.875)		0.78 (0.57-1.06, p=0.115)	
Shopping and Meals: High	0.67 (0.43-1.03, p=0.070)	.	0.63 (0.41-0.95, p=0.029)	*
Shopping and Meals: Moderate	0.89 (0.60-1.31, p=0.547)		0.91 (0.63-1.30, p=0.592)	
Services				
Cost DPs	1.00 (0.99-1.00, p=0.079)	.	1.00 (0.99-1.00, p=0.119)	
Cost Daycare	1.00 (1.00-1.00, p=0.769)			
Cost Homecare	1.00 (1.00-1.00, p=0.770)			
Has Telecare	0.82 (0.60-1.12, p=0.207)		0.93 (0.71-1.23, p=0.629)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

11.2.2.2 Age-stratified analysis of loneliness and care home entry

To further explore the influence of age, we stratified our dataset into these two groups, aged < 85 ($N = 570$) and ≥ 85 ($N = 531$), and ran the same model as in Equations (7.1) and (7.2) separately for each age group. We include the results for the logistic regression in Table 11.7 and for the competing risks model in Table 11.8. The pattern in all three cases is that the coefficient for loneliness for those aged ≥ 85 is statistically significant at $\alpha = 0.05$ and is in the range of 1.59 - 1.78 (compared with around 1.3 - 1.4 in the main models).

The coefficients for loneliness for older adults aged under 85 are 1.14 - 1.22 across the models and no longer statistically significant. This suggests the effect of loneliness on care home entry for younger adults is not as strong. However, there is also a much higher proportion of censored cases in the younger age group (56% for under 85s vs. 37% for those aged 85 and above). Many younger participants did not enter a care home or die before entering a care home during the observation period. The high proportion of censored data and splitting of the data in half may have reduced the power to detect a significant effect. These factors suggest that longer follow-up periods may be necessary to fully understand the impact of loneliness on care home entry among younger older adults.

Table 11.7: Logistic regression model output (categorical age)

	Age <85		Age 85+	
	Odds ratio		Odds ratio	
Loneliness				
Lonely or Isolated	1.15 (0.65-1.99, p=0.622)		1.68 (1.11-2.55, p=0.015)	*
Demographics				
Age	1.05 (1.00-1.11, p=0.044)	*	1.00 (0.95-1.05, p=0.982)	
Ethnicity: White	1.90 (1.12-3.32, p=0.021)	*	1.12 (0.73-1.75, p=0.604)	
Lives Alone	1.29 (0.74-2.28, p=0.375)		2.01 (1.29-3.17, p=0.002)	**
Sex: Male	0.97 (0.59-1.59, p=0.906)		1.59 (1.04-2.41, p=0.030)	*
Unpaid Care	0.57 (0.33-1.02, p=0.056)	.	1.02 (0.63-1.67, p=0.927)	
Needs				
Cognition: High	5.27 (2.66-10.55, p<0.001)	***	3.09 (1.83-5.25, p<0.001)	***
Cognition: Moderate	3.35 (1.75-6.39, p<0.001)	***	2.58 (1.54-4.30, p<0.001)	***
N Notes	1.00 (1.00-1.00, p<0.001)	***	1.00 (1.00-1.00, p=0.013)	*
Personal care: High	1.09 (0.44-2.64, p=0.857)		1.14 (0.54-2.38, p=0.727)	
Personal care: Moderate	0.67 (0.35-1.27, p=0.220)		0.81 (0.50-1.31, p=0.394)	
Shopping and Meals: High	0.61 (0.25-1.47, p=0.264)		0.60 (0.30-1.23, p=0.160)	
Shopping and Meals: Moderate	0.90 (0.42-1.97, p=0.793)		1.03 (0.54-1.97, p=0.937)	
Services				
Cost DPs	1.00 (0.98-1.00, p=0.319)		1.00 (0.99-1.00, p=0.427)	
Cost Daycare	1.01 (1.00-1.01, p=0.100)	.	1.00 (0.99-1.01, p=0.845)	
Cost Homecare	1.00 (1.00-1.00, p=0.552)		1.00 (1.00-1.00, p=0.548)	
Has Telecare	1.06 (0.56-1.95, p=0.860)		0.83 (0.52-1.31, p=0.427)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

Table 11.8: Competing risks models split into over and under 85s

	Age <85				Age 85+			
	Cause-specific hazard		Fine & Gray		Cause-specific hazard		Fine & Gray	
	Hazard ratio		Hazard ratio		Hazard ratio		Hazard ratio	
Loneliness								
Lonely or Isolated	1.22 (0.76-1.96, p=0.404)		1.14 (0.74-1.78, p=0.550)		1.78 (1.24-2.56, p=0.002)	**	1.59 (1.16-2.18, p=0.004)	**
Demographics								
Ethnicity: White	1.97 (1.19-3.27, p=0.009)	**	1.62 (1.07-2.44, p=0.022)	*	1.01 (0.66-1.56, p=0.948)		1.09 (0.77-1.54, p=0.624)	
Lives Alone	1.58 (0.98-2.56, p=0.063)	.	1.35 (0.86-2.10, p=0.188)		1.84 (1.18-2.86, p=0.007)	**	1.84 (1.26-2.69, p=0.002)	**
Unpaid Care	0.74 (0.43-1.28, p=0.286)		0.67 (0.42-1.06, p=0.087)	.	1.06 (0.69-1.62, p=0.803)		1.17 (0.78-1.74, p=0.451)	
Needs								
Personal care: High	1.11 (0.47-2.63, p=0.817)		1.06 (0.49-2.31, p=0.884)		1.40 (0.68-2.89, p=0.364)		1.17 (0.65-2.10, p=0.600)	
Personal care: Moderate	1.25 (0.75-2.10, p=0.392)		0.81 (0.47-1.39, p=0.439)		0.87 (0.55-1.36, p=0.535)		0.87 (0.58-1.28, p=0.477)	
Shopping and Meals: High	0.63 (0.29-1.36, p=0.238)		0.65 (0.35-1.20, p=0.169)		0.85 (0.48-1.53, p=0.596)		0.72 (0.41-1.27, p=0.258)	
Shopping and Meals: Moderate	0.80 (0.37-1.72, p=0.571)		0.86 (0.48-1.52, p=0.598)		0.93 (0.52-1.66, p=0.796)		1.06 (0.67-1.69, p=0.791)	
Services								
Cost DPs	1.00 (0.99-1.00, p=0.226)		1.00 (0.99-1.00, p=0.138)		1.00 (0.99-1.01, p=0.569)		1.00 (0.99-1.00, p=0.388)	
Cost Daycare	1.00 (0.99-1.01, p=0.822)		1.00 (1.00-1.01, p=0.535)		1.00 (0.99-1.00, p=0.689)		1.00 (1.00-1.00, p=0.912)	
Has Telecare	0.93 (0.55-1.59, p=0.799)		0.88 (0.54-1.45, p=0.618)		0.83 (0.56-1.25, p=0.385)		0.86 (0.61-1.23, p=0.424)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

11.2.3 Restricting the analysis to assessments completed from 2016

The cohort was selected on the basis that individuals had to have been receiving long-term care services 1st January 2016 and 31st August 2020. However, the exported data included service receipt data from 1st January 2015. We included individuals in the model if we could identify that their first assessment had been received in 2015. Including these individuals increases statistical power. However, it is possible that some service receipt covariates from 2015 might be unobserved, as service use data was only included in the export if the service was being delivered at some point since 1st January 2016. This means that if a service was delivered from, for example, June 2015 - December 2015, it would not have been included in the export. We removed individuals from the data whose assessment occurred in 2015 ($N = 160$) and ran on this dataset the main models specified in Equation (7.1) and Equation (7.2). We present in Table 11.9 the output from the logistic regression model and in Table 11.10 the output from the competing risks model.

The results for loneliness are similar to those in the main body of the paper. The magnitude of the loneliness coefficient is slightly greater than the originally specified models, at 1.59 compared with 1.45 in the logistic regression, 1.38 compared with 1.32 in the cause-specific hazard model and 1.45 compared with 1.39 in the logistic regression model. The confidence intervals of these estimates overlap with those of the main models, and we have selected a different group so we do not place much interpretation on the differences. These findings indicate that the inclusion of 2015 data in the model does not change the interpretation of the output, and loneliness remains a robust predictor of care home entry.

Table 11.9: Logistic regression model output (2016 onwards)

	Odds ratio	
Loneliness		
Lonely or Isolated	1.59 (1.10-2.30, p=0.013)	*
Demographics		
Sex: Male	1.49 (1.05-2.12, p=0.024)	*
Age	1.37 (0.93-2.07, p=0.122)	
Age^2	1.00 (1.00-1.00, p=0.174)	
Ethnicity: White	1.27 (0.88-1.84, p=0.215)	
Lives Alone	1.62 (1.10-2.40, p=0.015)	*
Unpaid Care	0.72 (0.49-1.07, p=0.105)	
Needs		
N Notes	1.00 (1.00-1.00, p=0.001)	***
Personal care: Moderate	0.68 (0.45-1.04, p=0.075)	.
Personal care: High	1.04 (0.57-1.89, p=0.901)	
Cognition: Moderate	2.47 (1.57-3.87, p<0.001)	***
Cognition: High	3.63 (2.28-5.80, p<0.001)	***
Shopping and Meals: Moderate	1.09 (0.64-1.91, p=0.749)	
Shopping and Meals: High	0.67 (0.37-1.25, p=0.205)	
Services		
Cost DPs	0.99 (0.98-1.00, p=0.134)	
Cost Daycare	1.00 (1.00-1.01, p=0.192)	
Cost Homecare	1.00 (1.00-1.00, p=0.693)	
Has Telecare	1.12 (0.74-1.67, p=0.596)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

Table 11.10: Competing Risks model output (2016 onwards)

	Cause-specific hazard		Fine & Gray	
Loneliness				
Lonely or Isolated	1.38 (1.02-1.88, p=0.037)	*	1.45 (1.06-1.98, p=0.019)	*
Demographics				
Age	1.18 (0.82-1.70, p=0.375)		1.24 (0.89-1.72, p=0.210)	
Age^2	1.00 (1.00-1.00, p=0.445)		1.00 (1.00-1.00, p=0.270)	
Ethnicity: White	1.29 (0.94-1.77, p=0.119)		1.16 (0.85-1.58, p=0.346)	
Lives Alone	1.56 (1.13-2.15, p=0.007)	**	1.49 (1.08-2.05, p=0.015)	*
Sex: Male	1.55 (1.16-2.07, p=0.003)	**	1.36 (1.03-1.80, p=0.032)	*
Unpaid Care	0.84 (0.60-1.18, p=0.318)		0.83 (0.61-1.14, p=0.255)	
Needs				
Personal care: High	1.30 (0.76-2.24, p=0.337)		1.05 (0.62-1.76, p=0.862)	
Personal care: Moderate	0.87 (0.60-1.25, p=0.448)		0.68 (0.47-0.98, p=0.037)	*
Shopping and Meals: High	0.73 (0.45-1.20, p=0.219)		0.64 (0.38-1.07, p=0.090)	.
Shopping and Meals: Moderate	1.02 (0.66-1.58, p=0.920)		1.00 (0.65-1.56, p=0.984)	
Services				
Cost DPs	0.99 (0.99-1.00, p=0.048)	*	0.99 (0.99-1.00, p=0.071)	.
Cost Daycare	1.00 (1.00-1.00, p=0.809)			
Cost Homecare	1.00 (1.00-1.00, p=0.794)			
Has Telecare	0.99 (0.71-1.39, p=0.974)		1.13 (0.83-1.54, p=0.450)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

11.2.4 Exclusion of living alone

Living alone is closely associated with loneliness and social isolation, raising the question of whether its inclusion in the model affects the results or their interpretation. To assess its impact, we re-ran the analysis specified in Equations (7.1) and (7.2) without the living alone variable to determine whether its exclusion would meaningfully change the results. As discussed in the main text, the exclusion of this variable did not substantially alter the findings, with the effect sizes for loneliness remaining similar and the *p* values not meaningfully changed. For completeness, we present the full results of these models below in Table 11.11 (logistic regression) and Table 11.12 (competing risks).

Table 11.11: Logistic regression model output (excluding living alone)

	Odds ratio	
Loneliness		
Lonely or Isolated	1.51 (1.09-2.09, p=0.014)	*
Demographics		
Sex: Male	1.21 (0.89-1.66, p=0.222)	
Age	1.33 (0.94-1.92, p=0.117)	
Age^2	1.00 (1.00-1.00, p=0.178)	
Ethnicity: White	1.45 (1.05-2.04, p=0.028)	*
Unpaid Care	0.73 (0.51-1.05, p=0.085)	.
Needs		
N Notes	1.00 (1.00-1.00, p<0.001)	***
Personal care: Moderate	0.74 (0.51-1.08, p=0.117)	
Personal care: High	1.01 (0.58-1.74, p=0.983)	
Cognition: Moderate	2.63 (1.78-3.89, p<0.001)	***
Cognition: High	3.46 (2.32-5.16, p<0.001)	***
Shopping and Meals: Moderate	0.93 (0.57-1.51, p=0.754)	
Shopping and Meals: High	0.55 (0.32-0.95, p=0.029)	*
Services		
Cost DPs	1.00 (0.99-1.00, p=0.173)	
Cost Daycare	1.00 (1.00-1.01, p=0.534)	
Cost Homecare	1.00 (1.00-1.00, p=0.602)	
Has Telecare	0.95 (0.66-1.36, p=0.780)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

Table 11.12: Competing Risks model output (excluding living alone)

	Cause-specific hazard		Fine & Gray	
Loneliness				
Lonely or Isolated	1.38 (1.06-1.79, p=0.018)	*	1.34 (1.04-1.74, p=0.026)	*
Demographics				
Age	1.13 (0.82-1.56, p=0.459)		1.20 (0.90-1.61, p=0.221)	
Age^2	1.00 (1.00-1.00, p=0.566)		1.00 (1.00-1.00, p=0.302)	
Ethnicity: White	1.44 (1.09-1.90, p=0.010)	**	1.28 (0.99-1.66, p=0.064)	.
Sex: Male	1.27 (0.98-1.65, p=0.072)	.	1.11 (0.87-1.43, p=0.390)	
Unpaid Care	0.85 (0.64-1.15, p=0.292)		0.83 (0.63-1.08, p=0.162)	
Needs				
Personal care: High	1.36 (0.83-2.23, p=0.222)		0.97 (0.62-1.53, p=0.908)	
Personal care: Moderate	0.96 (0.69-1.33, p=0.794)		0.74 (0.54-1.01, p=0.056)	.
Shopping and Meals: High	0.60 (0.39-0.93, p=0.021)	*	0.56 (0.37-0.85, p=0.006)	**
Shopping and Meals: Moderate	0.84 (0.57-1.24, p=0.379)		0.86 (0.60-1.23, p=0.403)	
Services				
Cost DPs	1.00 (0.99-1.00, p=0.094)	.	1.00 (0.99-1.00, p=0.140)	
Cost Daycare	1.00 (1.00-1.00, p=0.478)			
Cost Homecare	1.00 (1.00-1.00, p=0.786)			
Has Telecare	0.85 (0.63-1.15, p=0.299)		0.94 (0.71-1.25, p=0.682)	

*** < 0.001; ** <0.01; * <0.05; . <0.1

11.3 Appendices to Chapter 8

Three appendices are included:

1. Evaluation of sentiment metrics: establishing which sentiment metrics are appropriate for conducting this analysis.
2. Model diagnostics and robustness checks: verifying the robustness of conclusions using several other methods.
3. Evaluation of themes: full lists of words counted in the frequency of the words appearing in each theme.

The code to reproduce this analysis is available with synthetic data in the GitHub repository [419].

11.3.1 Appendix 1 - Evaluation of appropriateness of sentiment metrics

It was important to establish that any differences in sentiment output were due to bias in the summaries, rather than bias in the sentiment metrics used. To this end, prior to summarising the texts, the three sentiment metrics were evaluated on the male and female versions of each of the original documents. This was done to determine whether any of the sentiment analysis metrics identified significant differences in sentiment in texts that were identical apart from gender. Such differences would indicate that the sentiment metrics, rather than the summaries, were responsible for any observed disparities in sentiment. Regard and SiEBERT did not show significant differences based on gender. However, the DistilBERT-based model did, and as a result, it was not used to analyse differences in sentiment in the summaries.

11.3.1.1 Paired t-test

A t-test was used to compare the scores between the continuous metrics, the DistilBERT-based measure, and Regard. For the binary SiEBERT model, McNemar's χ^2 test for symmetry was used. As these documents are identical except for gender, the paired implementation of these tests was applied, using the `t.test` function for the continuous measure and `mcnemar.test` for the binary measure, both in the `stats` package in R [98]. The results comparing sentiment between genders for the original sentences are set out in Table 11.13. The null hypothesis

Table 11.13: t-test and Mcnemar test results

Direction	Effect size	Pr(>|t|)	signif
siebert			
fm	0.00409	0.627	
mf	-0.00680	0.804	
regard			
fm	0.01610	0.228	
mf	0.00799	0.613	
distilbert			
fm	-0.39400	1.03e-177	***
mf	-0.32700	5.2e-91	***

Note:
t-test is used for the continuous metrics and the McNemar's test for the binary SiEBERT metric

is that there are no differences in sentiment. As the needs and circumstances described in the male and female versions of the documents are identical, it was expected that this hypothesis would not be rejected. Indeed, the null hypothesis was not rejected for SiEBERT and Regard. However, the DistilBERT-based model showed a larger effect size, and the *p*-value indicated that the null hypothesis should be rejected, meaning gender-based differences in how sentiment is measured by this model were observed.

11.3.1.2 Mixed effects model: sentence level

The sentiment metrics were also examined using a mixed effects model. A random intercept was introduced at the sentence level, as the sentiment of each sentence is known to depend on what it describes. Gender and a variable called `gender_direction`, indicating whether the original text was written about a male and the generated text about a female (or vice versa), were also included in the model. This was done to control for any differences in the content typically written about men and women. The mixed-effects model was specified as follows:

$$\begin{aligned}\text{sentiment}_{ij} = \beta_0 + \beta_1 \cdot \text{gender}_i + \beta_2 \cdot \text{gender\_direction}_i \\ + u_{0j} + \epsilon_{ij}\end{aligned} \tag{11.2}$$

Where:

- *sentiment* is a continuous indicator of the proportion of the text which contains non-negative sentiment
- *gender* is a binary indicator of whether a text is about a man or a woman.
- *gender_direction* is a binary indicator of whether the original text was written about a male and the generated text about a female, or vice versa.
- u_{0j} is a random intercept for the j-th group (Sentence ID), accounting for the variability in sentiment across different sentences.
- ϵ_{ij}: Residual error term for the j-th observation within the j-th group.

The covariance of the random intercept was allowed to be unstructured. It was assumed that the random intercepts u_{0j} follow a normal distribution with mean 0 and variance σ^2_{u0}, the residuals ϵ_{ij} are independently and normally distributed with mean 0 and variance σ^2, and the random intercepts u_{0j} are independent of the residuals ϵ_{ij}.

Since the final activation layer of SiEBERT is softmax, producing binary predictions of sentiment (i.e., positive or negative), a generalised linear model with a logistic link function was used for the sentence-level SiEBERT predictions. In this case, $\text{logit}(P(\textit{sentiment} = 1))$ was estimated, where sentiment can take the values 0 (negative) or 1 (positive). The right-hand side of the equation remained the same.

The results of the mixed model, as specified in Equation (11.2), are shown in Table 11.14. These results are consistent with the t-test findings, indicating that Regard and SiEBERT do not find systematic differences in the sentiment of the original documents based on gender, but the DistilBERT-based model does.

11.3.1.3 Mixed effects model: document level

It is reassuring that the mixed model results at sentence level are consistent with the t-test results. However, summaries do not necessarily have the same number of sentences (and if they do the sentences may not correspond). This means that sentiment for the male and female versions of each summary will need to be aggregated and compared at document level. To confirm that the metrics are appropriate, the sentiment results were aggregated for the original texts at document level, taking the mean of sentence-level sentiment. This is the same model as Equation (11.2), though clustering at Document ID rather than Sentence ID level, i.e.

Table 11.14: Sentiment output: mixed model (sentence level)

Coef	Estimate	Std. Error	t value	Pr(>|t|)	Signif
regard					
(Intercept)	0.320000	0.003650	87.700	<0.001	***
Gender: Male	0.000561	0.000435	1.290	0.197	
Gender direction: mf	0.003790	0.005660	0.669	0.504	
siebert					
(Intercept)	0.400000	0.009360	42.700	3.53e-187	***
Gender: Male	0.000569	0.001280	0.443	0.658	
Gender direction: mf	-0.018000	0.014500	-1.240	0.214	
distilbert					
(Intercept)	0.665000	0.003450	193.000	<0.001	***
Gender: Male	-0.007110	0.000241	-29.500	3.12e-120	***
Gender direction: mf	0.004730	0.005350	0.883	0.378	

Note:
The SiEBERT binomial produces a z-value rather than t-value. For the purpose of presentation, this is included in the t-value column.

$$\begin{aligned}\text{sentiment}_{ij} = \beta_0 + \beta_1 \cdot \text{gender}_i + \beta_2 \cdot \text{gender\_direction}_i \\ + u_{0j} + \epsilon_{ij}\end{aligned} \tag{11.3}$$

Where:

- *sentiment* is a continuous indicator of the proportion of the text which contains non-negative sentiment (mean of each sentence across documents)
- *gender* is a binary indicator of whether a text is about a man or a woman.
- *gender_direction* is a binary indicator of whether the original text was written about a male and the generated text about a female, or vice versa.
- u_{0j} is a random intercept for the j-th group (Document ID), accounting for the variability in sentiment across different sentences.
- ϵ_{ij}: Residual error term for the j-th observation within the j-th group.

Once again, the assumptions are the same. The covariance of the random intercept is unstructured. The model assumes that the random intercepts u_{0j} follow a normal distribution with mean 0 and variance σ^2_{u0}, the residuals ϵ_{ij} are independently and normally distributed with mean 0 and variance σ^2 and the random intercepts u_{0j} are independent of the residuals ϵ_{ij}. A linear model is used for SiEBERT here

Table 11.15: Sentiment output: mixed model (document level)

Coef	Estimate	Std. Error	t value	Pr(>|t|)	Signif
regard					
(Intercept)	0.320000	0.003650	87.700	<0.001	***
Gender: Male	0.000561	0.000435	1.290	0.197	
Gender direction: mf	0.003790	0.005660	0.669	0.504	
siebert					
(Intercept)	0.400000	0.009360	42.700	3.53e-187	***
Gender: Male	0.000569	0.001280	0.443	0.658	
Gender direction: mf	-0.018000	0.014500	-1.240	0.214	
distilbert					
(Intercept)	0.665000	0.003450	193.000	<0.001	***
Gender: Male	-0.007110	0.000241	-29.500	3.12e-120	***
Gender direction: mf	0.004730	0.005350	0.883	0.378	

too, as the per-document average of binary sentence classifications is continuous. Table 11.15 shows the results aggregated at document level.

Across all three measures, the DistilBERT-based model finds significant differences in sentiment once gender is changed. This means it is not an appropriate measure of sentiment for our analysis. This is why it is not used in the paper to measure differences in sentiment of the summaries. However, there are no significant differences using Regard or SiEBERT, which is why these metrics are used to evaluate the output of the summarisation models.

11.3.2 Appendix 2 - Model diagnostics and robustness checks

Table 11.16 contains the covariance matrix for the random effects in the model specified in Equation (8.2), with the results for the main effects in Table 8.2. Table 11.16 includes the variances of individual variables and the covariances between variables.

Table 11.16: Covariance Matrix of Random Effects

		Regard		SiEBERT	
Group	Variable	Variance-Covariance	Standard Deviation	Variance-Covariance	Standard Deviation
Residual		0.006	0.078	0.035	0.187
doc_id	(Intercept)	0.011	0.103	0.074	0.272
doc_id	(Intercept) - modelgemma	-0.008	-0.835	-0.050	-0.809
doc_id	(Intercept) - modelllama3	-0.007	-0.791	-0.044	-0.704
doc_id	(Intercept) - modelt5	-0.007	-0.678	-0.042	-0.660
doc_id	modelgemma	0.008	0.090	0.051	0.226
doc_id	modelgemma - modelllama3	0.007	0.895	0.046	0.888
doc_id	modelgemma - modelt5	0.006	0.691	0.038	0.723
doc_id	modelllama3	0.008	0.090	0.053	0.229
doc_id	modelllama3 - modelt5	0.006	0.685	0.038	0.701
doc_id	modelt5	0.009	0.097	0.055	0.234

The distribution of the linear mixed model's random effects is presented in Figure 11.6, and a Q-Q plot of observed and expected values for residuals is shown in Figure 11.7. The random effects are generally normally distributed, with the notable exception of the intercept for the SiEBERT model, which demonstrates clear asymmetry at the tails. The Q-Q plot reveals the presence of some outliers and heteroscedasticity, particularly with the SiEBERT predictions, which deviate more from normality at the tails. The Regard predictions fit more closely to the normal distribution, although the residuals do not perfectly follow the expected distribution at the tails. Despite these deviations, the bootstrapping results and robustness checks ensure the conclusions remain reliable.

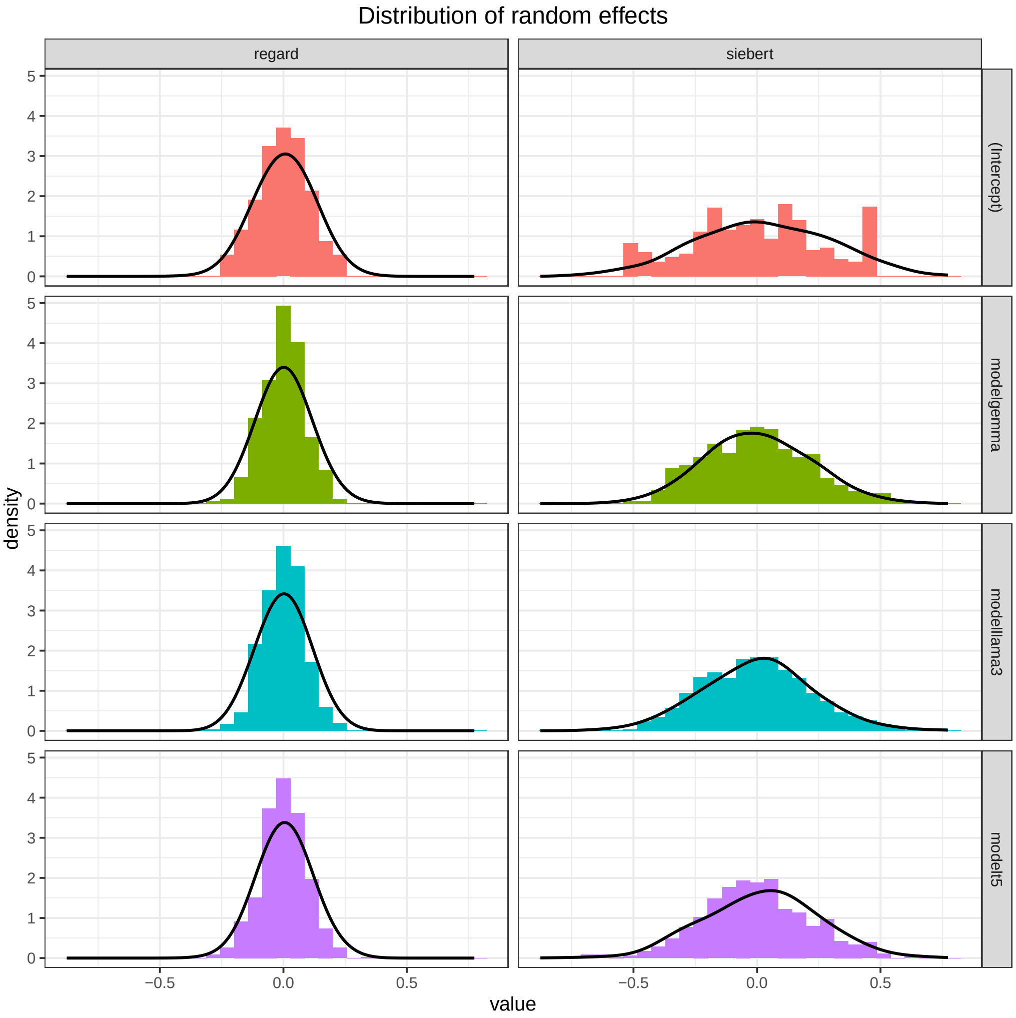


Figure 11.6: Distribution of random effects

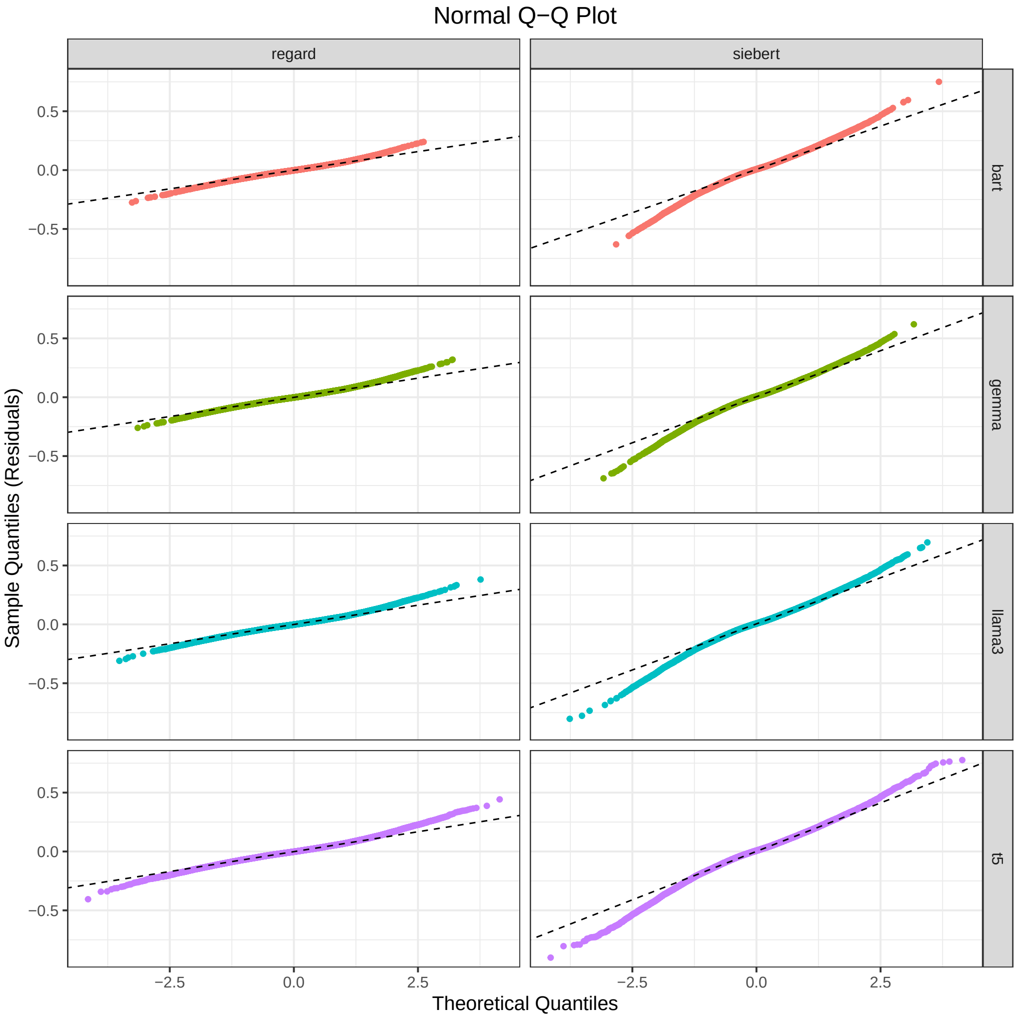


Figure 11.7: Normal QQ plot

The linear mixed model assumes normality of random effects and homoscedasticity. Simulations show that violations of these assumptions often have little or no effect on parameter estimates, although they do affect the interpretation of the significance of the variance of the random effects [475]. The primary focus is on the fixed effects rather than document-level random effects, which are mainly included to account for the clustering of sentiment within documents. However, as the assumptions of the model are not always satisfied, other approaches were explored to assess the sensitivity of the conclusions to these assumptions. Given the presence of some non-linearities, interaction terms, such as the interaction between gender and the maximum number of tokens, were tested to account for possible non-linear relationships. However, analysis of variance (ANOVA) and likelihood-ratio tests indicated that the interaction term did not significantly improve model fit ($p > 0.05$). Consequently, the interaction was removed to simplify the model without affecting the overall conclusions regarding gender bias in the summaries. The model equation was retained without interactions, and other methods were used to assess the robustness of the findings:

1. **Bootstrapping:** To test sensitivity to outliers, 1,000 bootstrap samples were generated by resampling the original data with replacement, and the model was refitted on each sample. This approach provided an estimate of the distribution of the parameter estimates and allowed an assessment of the stability of the findings across different datasets.
2. **Relaxing the variance structure:** To test sensitivity to the assumption of homoscedasticity, a mixed-effects model was fitted, allowing for different residual variances across each language model.
3. **Robust linear mixed model:** To test sensitivity to outliers, a robust linear mixed model was used.
4. **Generalised Estimating Equations (GEE) model:** To test sensitivity to the correlation structure of the data and the assumption of normally distributed random effects, a GEE model was fitted. This model used robust sandwich estimators to adjust standard errors, allowing for heteroscedasticity in the residuals.
5. **Linear models:** To test sensitivity to the inclusion of random effects at the model level, each language model's results were split into its own dataset, and a linear model was run with Document ID as a main effect.

The results of each of these models were generally consistent with the findings of the mixed model. None of the models identified gender-based differences caused by Llama 3. Some variation was observed across the models regarding the significance of the differences in sentiment for the BART and T5 models. However, all models agreed that there were significant gender-based differences in the summaries generated by the Gemma model.

11.3.2.1 Bootstrapped model output and estimated marginal means

Bootstrapped datasets were generated by creating 1,000 new datasets, each the same size as the original data, through non-parametric sampling of the original data with replacement. Samples were taken at the Document ID level to preserve the correlation of sentiment within documents [476]. The original linear mixed model was then run for each bootstrapped dataset. The bootstrapped estimates represent the mean of all 1,000 estimates. The results for the SiEBERT model are shown in Table 11.17, and the results for Regard are shown in Table 11.18. The additional columns in Table 11.17 were calculated as follows:

$$\begin{aligned} \text{Absolute Bias} &= \text{Bootstrapped Estimate} - \text{Original Estimate} \\ \text{Relative Bias} &= \frac{\text{Absolute Bias}}{\text{Original Estimate}} \\ \text{Standardised Bias} &= \frac{\text{Absolute Bias}}{\text{Standard Error}} \end{aligned}$$

Bootstrapped estimated marginal means are presented in Table 11.19. The table also includes the number of times the *p*-values for the estimated marginal means were less than 0.05 and 0.01. The differences in gender in the Gemma model are larger using SiEBERT, with a larger *t*-value and a *p*-value of less than 0.01 in all 1,000 bootstrapped datasets. The difference is somewhat smaller in the case of Regard, though $p < 0.05$ in 962 of the 1,000 simulated datasets. The BART models show this effect in approximately 30-40% of cases, and T5 shows it in 40-60% of cases, suggesting that there is an effect of gender bias greater than random chance, although not as strong as the disparities observed in the Gemma model. There is no indication of a systematic effect of gender on sentiment in Llama 3, with slightly under 5% of estimated marginal mean differences resulting in $p < 0.05$. Overall, the bootstrapping results confirm that while some observable gender-based differences exist in BART and T5, the largest differences are in the Gemma model.

Table 11.17: Bootstrapped model output (SiEBERT)

	Original model				Bootstrapped model			
						Bias		
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Absolute	Relative	Standardised
(Intercept)	0.579	0.012	49.749	<0.001	0.579	<0.001	<0.001	-0.008
modelgemma	0.147	0.010	14.558	<0.001	0.147	<0.001	0.001	0.015
modelllama3	0.052	0.010	5.132	<0.001	0.052	<0.001	0.001	0.004
modelt5	0.103	0.010	9.904	<0.001	0.103	<0.001	<0.001	0.002
gendermale	-0.009	0.004	-2.161	0.031	-0.009	<0.001	0.001	-0.001
max_tokens75	-0.024	0.004	-6.431	<0.001	-0.024	<0.001	-0.001	0.009
max_tokens100	-0.039	0.004	-10.304	<0.001	-0.039	<0.001	-0.002	0.021
max_tokens150	-0.050	0.004	-13.299	<0.001	-0.050	<0.001	-0.001	0.008
max_tokens300	-0.054	0.004	-14.419	<0.001	-0.054	<0.001	-0.002	0.026
max_tokensNone	-0.084	0.004	-22.262	<0.001	-0.084	<0.001	-0.001	0.027
modelgemma:gendermale	-0.033	0.006	-5.318	<0.001	-0.033	<0.001	0.006	-0.030
modelllama3:gendermale	0.015	0.006	2.426	0.015	0.015	<0.001	-0.001	-0.002
modelt5:gendermale	0.020	0.006	3.185	0.001	0.019	<0.001	-0.012	-0.038

Table 11.18: Bootstrapped model output (Regard)

	Original model				Bootstrapped model			
						Bias		
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Absolute	Relative	Standardised
(Intercept)	0.278	0.004	61.965	<0.001	0.278	<0.001	<0.001	0.019
modelgemma	0.025	0.004	6.109	<0.001	0.025	<0.001	-0.004	-0.027
modelllama3	0.029	0.004	7.061	<0.001	0.029	<0.001	0.001	0.010
modelt5	-0.033	0.004	-7.712	<0.001	-0.033	<0.001	0.003	-0.023
gendermale	0.004	0.002	1.954	0.051	0.004	<0.001	0.009	0.018
max_tokens75	0.019	0.002	11.865	<0.001	0.019	<0.001	<0.001	-0.004
max_tokens100	0.027	0.002	17.076	<0.001	0.027	<0.001	0.001	0.020
max_tokens150	0.032	0.002	20.246	<0.001	0.032	<0.001	-0.001	-0.026
max_tokens300	0.039	0.002	25.052	<0.001	0.040	<0.001	0.001	0.022
max_tokensNone	0.045	0.002	28.303	<0.001	0.045	<0.001	<0.001	-0.001
modelgemma:gendermale	-0.011	0.003	-4.082	<0.001	-0.011	<0.001	0.003	-0.012
modelllama3:gendermale	-0.001	0.003	-0.561	0.575	-0.001	<0.001	0.038	-0.021
modelt5:gendermale	0.001	0.003	0.521	0.603	0.001	<0.001	0.033	0.017

Table 11.19: Mixed effects model: bootstrapped estimated marginal means (female - male)

	Regard					SiEBERT				
Model	Estimate	t	N	N Pr(|t|)<0.01	N Pr(|t|)<0.05	Estimate	t	N	N Pr(|t|)<0.01	N Pr(|t|)<0.05
bart	-0.0036	-1.60	1000	146	331	0.0094	1.80	1000	235	430
gemma	0.0069	3.00	1000	764	962	0.0420	7.70	1000	1000	1000
llama3	-0.0021	-0.91	1000	1	32	-0.0055	-0.99	1000	3	40
t5	-0.0050	-2.20	1000	275	651	-0.0099	-1.80	1000	107	421

11.3.2.2 Variance-structured mixed effects model

The Q-Q plots demonstrated deviations from normality, especially in the tails, which differ by model. To account for this heteroscedasticity and deviation from normality, the R `nlme` package [477] was used to employ a linear mixed-effects model which allowed the variance to differ by model, i.e.

$$\mathrm{Var}(\epsilon_{ij}) = \sigma^2_{\mathrm{model}_i} \tag{11.4}$$

This model would not converge with a random intercept and slope and this variance specification, so the random slope was removed. The model was therefore specified as follows:

$$\begin{aligned} \text{sentiment}_{ij} = \beta_0 &+ \beta_1 \cdot \text{model}_i + \beta_2 \cdot \text{gender}_j \\ &+ \beta_3 \cdot (\text{model}_i \times \text{gender}_j) + \beta_4 \cdot \text{max\_tokens}_i \\ &+ u_{0j} + \epsilon_{ij} \end{aligned} \tag{11.5}$$

Where β_0 is the intercept, β_1, β_2, and β_3 are the coefficients for model, gender, and their interaction, β_4 is the coefficient for maximum tokens, u_{0j} is the random intercept for document j and ϵ_{ij} is the residual error term. The results are set out in Table 11.20 and the estimated marginal means in Table 11.21. The estimates are very close to the output from the linear mixed model, though the p-values are slightly larger. The BART and T5 models are on the boundary of significance, but now the p-values are slightly larger than 0.05. Llama 3 has no significant differences in sentiment between men and women, and Gemma has the largest standardised estimates and smallest p-values. This model reduces the risk of Type 1 error, which is seen in the larger p-values, so it is reassuring that the main findings about Llama 3 and Gemma remain consistent.

Table 11.20: Variance-structured mixed effects model output

	Regard			SiEBERT						
Coef	Estimate		Std. Error	t	p	Estimate		Std. Error	t	p
(Intercept)	0.3100	***	0.0030	100.5866481	0.0e+00	0.5400	***	0.0083	64.278954	0.0e+00
Model gemma	0.0250	***	0.0019	13.0451295	0.0e+00	0.1500	***	0.0048	30.836185	0.0e+00
Model llama3	0.0290	***	0.0019	14.7169102	0.0e+00	0.0520	***	0.0049	10.750380	0.0e+00
Model t5	-0.0330	***	0.0024	-13.7578664	0.0e+00	0.1000	***	0.0059	17.502350	0.0e+00
gendermale	0.0036	.	0.0020	1.7997477	7.2e-02	-0.0094	.	0.0052	-1.809674	7.0e-02
Max tokens 150	0.0046	**	0.0018	2.5900877	9.6e-03	-0.0069		0.0043	-1.603439	1.1e-01
Max tokens 300	0.0120	***	0.0018	6.6848292	0.0e+00	-0.0110	**	0.0043	-2.682003	7.3e-03
Max tokens 50	-0.0270	***	0.0018	-15.2606004	0.0e+00	0.0360	***	0.0043	8.430521	0.0e+00
Max tokens 75	-0.0083	***	0.0018	-4.6806345	2.9e-06	0.0130	**	0.0043	3.089684	2.0e-03
Max tokens None	0.0150	***	0.0018	8.4733104	0.0e+00	-0.0290	***	0.0043	-6.866574	0.0e+00
Model gemma : Male	-0.0110	***	0.0027	-3.9068436	9.4e-05	-0.0330	***	0.0067	-4.851052	1.2e-06
Model llama3 : Male	-0.0014		0.0028	-0.5237945	6.0e-01	0.0150	*	0.0069	2.159936	3.1e-02
Model t5 : Male	0.0013		0.0034	0.3931827	6.9e-01	0.0200	*	0.0083	2.351745	1.9e-02

Table 11.21: Variance-structured mixed effects: estimated marginal means (female - male)

	Regard				SiEBERT			
Model	Estimate		t	p	Estimate		t	p
bart	-0.0036	.	-1.8	0.07200	0.0094	.	1.8	0.07
gemma	0.0069	***	3.8	0.00014	0.0420	***	9.8	0.00
llama3	-0.0021		-1.1	0.27000	-0.0055		-1.2	0.23
t5	-0.0049	.	-1.8	0.07800	-0.0100		-1.6	0.12

11.3.2.3 Robust linear mixed model

The results of the bootstrapping were reassuring in the case of the Gemma model. However, significant differences were not always observed in the BART and T5 models. From the Q-Q plots, it is known that deviations from normality exist in the tails. To test the sensitivity of the results to outliers, a robust linear mixed model was used. This model follows the same structure as the standard linear mixed model, given in Equation (1) in the main body of the paper:

$$\begin{aligned}\text{sentiment}_{ij} = \beta_0 &+ \beta_1 \cdot \text{model}_i + \beta_2 \cdot \text{gender}_j \\ &+ \beta_3 \cdot (\text{model}_i \times \text{gender}_j) + \beta_4 \cdot \text{max\_tokens}_i \\ &+ u_{0j} + u_{1j} \cdot \text{model}_i + \epsilon_{ij}\end{aligned} \tag{11.6}$$

Where β_0 is the intercept, β_1, β_2, and β_3 are the coefficients for model, gender, and their interaction, β_4 is the coefficient for maximum tokens, u_{0j} is the random intercept for document j, u_{1j} is the random slope for model within document j, and ϵ_{ij} is the residual error term.

The difference from the standard mixed effects model is that a robust loss function was incorporated to reduce the impact of outliers in the residuals. This was implemented using the `robustlmm` R package [478]. The results are shown in Table 11.22. The estimates obtained from both the mixed-effects and robust mixed-effects models were extremely close. The package does not produce p-values; however, marginal means were estimated [420] and are presented in Table 11.23. The estimates and p-values are very close to the output from the standard linear mixed model. Once again, The BART and T5 models show p-values hovering around conventional significance thresholds, with some disagreement in the direction of the gender effect in the BART model between Regard and SiEBERT. For these models, p-values range between 0.04 and 0.08, suggesting borderline statistical significance that should be interpreted cautiously. The Gemma model exhibits the largest standardised estimates and the smallest p-values, with both sentiment metrics indicating that male sentiment is more negative than female sentiment.

Table 11.22: Robust mixed effects model output

	Regard			SiEBERT		
Coef	Estimate	Std. Error	t	Estimate	Std. Error	t
(Intercept)	0.27000	0.0065	43.00	0.5900	0.0120	48.0
Model gemma	0.02300	0.0037	6.20	0.1400	0.0100	14.0
Model llama3	0.02800	0.0035	7.90	0.0510	0.0100	5.0
Model t5	-0.03500	0.0029	-12.00	0.1200	0.0110	11.0
gendermale	0.00410	0.0020	2.00	-0.0094	0.0039	-2.4
Max tokens 75	0.02100	0.0018	12.00	-0.0270	0.0034	-7.8
Max tokens 100	0.02900	0.0018	17.00	-0.0420	0.0034	-12.0
Max tokens 150	0.03500	0.0018	20.00	-0.0510	0.0034	-15.0
Max tokens 300	0.04200	0.0018	24.00	-0.0560	0.0034	-17.0
Max tokens None	0.04700	0.0018	27.00	-0.0790	0.0034	-23.0
Model gemma : Male	-0.01100	0.0029	-3.70	-0.0300	0.0055	-5.5
Model llama3 : Male	-0.00052	0.0029	-0.18	0.0130	0.0055	2.4
Model t5 : Male	0.00083	0.0029	0.29	0.0190	0.0055	3.5

Table 11.23: Robust mixed effects model: estimated marginal means (female - male)

	Regard				SiEBERT			
Model	Estimate		z	p	Estimate		z	p
bart	-0.0041	*	-2.0	0.0450	0.0094	*	2.40	0.016
gemma	0.0065	**	3.2	0.0014	0.0400	***	10.00	0.000
llama3	-0.0035	.	-1.7	0.0810	-0.0039		-0.99	0.320
t5	-0.0049	*	-2.4	0.0160	-0.0100	*	-2.60	0.010

11.3.2.4 Generalised Estimating Equations (GEE)

A Generalised Estimating Equations (GEE) model was also used to estimate population-averaged effects, adjusting for within-group correlation using robust sandwich estimators. This was implemented using the `geepack` R package [479]. The GEE model estimates population-averaged effects and can be more robust to misspecified correlation structures. The GEE model was specified as follows:

$$\begin{aligned} y_{ij} = \beta_0 &+ \beta_1 \text{model}_i + \beta_2 \text{gender}_j \\ &+ \beta_3 (\text{model}_i \times \text{gender}_j) + \beta_4 \text{max\_tokens}_i + \epsilon_{ij} \end{aligned} \tag{11.7}$$

The correlation structure of the residuals ϵ_i was modeled as exchangeable within groups defined by Document ID. No corrections were applied to the standard errors to reduce the risk of Type 1 error, as there are 617 document-level clusters, and with 100 or more clusters, such corrections are generally unnecessary [480]. The results of the GEE model are set out in Table 11.24. The estimated marginal means for the GEE model are presented in Table 11.25.

The point estimates obtained from the mixed-effects and GEE models were identical, indicating that the fixed effects are robust to the choice of modelling approach. However, the standard errors differed between the models. The mixed-effects model, which accounts for random effects, generally provided smaller standard errors compared to the GEE model. Attempts to fit a GEE model with an unstructured covariance matrix were unsuccessful, which may have contributed to the larger standard errors in the GEE model. As a result, significant differences in sentiment based on gender were not observed in the BART and T5 models. However, the Gemma model was not affected by these differences, and summaries about women remained significantly less negative than those about men.

Table 11.24: GEE model output

	Regard					SiEBERT				
Coef	Estimate		Std. Error	Wald	p	Estimate		Std. Error	Wald	p
(Intercept)	0.3000	***	0.0024	1.6e+04	0.0000	0.5800	***	0.0065	7800.0	0.0e+00
Model gemma	0.0250	***	0.0025	1.0e+02	0.0000	0.1500	***	0.0064	530.0	0.0e+00
Model llama3	0.0290	***	0.0025	1.3e+02	0.0000	0.0520	***	0.0067	61.0	0.0e+00
Model t5	-0.0330	***	0.0029	1.3e+02	0.0000	0.1000	***	0.0073	200.0	0.0e+00
gendermale	0.0036		0.0027	1.7e+00	0.1900	-0.0094		0.0072	1.7	1.9e-01
Max tokens 150	0.0050	*	0.0021	5.5e+00	0.0190	-0.0500	***	0.0059	71.0	0.0e+00
Max tokens 300	0.0130	***	0.0021	3.6e+01	0.0000	-0.0540	***	0.0059	85.0	0.0e+00
Max tokens 75	-0.0082	***	0.0023	1.3e+01	0.0003	-0.0240	***	0.0061	16.0	7.8e-05
Max tokens None	0.0180	***	0.0021	7.1e+01	0.0000	-0.0840	***	0.0060	200.0	0.0e+00
Model gemma : Male	-0.0110	**	0.0035	9.2e+00	0.0024	-0.0330	***	0.0090	13.0	2.8e-04
Model llama3 : Male	-0.0014		0.0036	1.6e-01	0.6900	0.0150		0.0095	2.5	1.2e-01
Model t5 : Male	0.0013		0.0041	1.1e-01	0.7400	0.0200	.	0.0100	3.6	5.9e-02

Table 11.25: GEE model: estimated marginal means (female - male)

	Regard				SiEBERT			
Model	Estimate		z	p	Estimate		z	p
bart	-0.0036		-1.30	0.1900	0.0094		1.3	0.19
gemma	0.0069	**	3.30	0.0011	0.0420	***	7.8	0.00
llama3	-0.0021		-0.92	0.3600	-0.0055		-0.9	0.37
t5	-0.0049		-1.60	0.1100	-0.0100		-1.4	0.17

11.3.2.5 Linear models

The mixed model includes an interaction term as well as both random intercepts and random slopes to account for variability between documents and within models. This specification is important because it reflects how document-level differences (random intercepts) and model-specific variability within documents (random slopes) can impact sentiment estimates across gender. However, while this specification makes theoretical sense, the sensitivity of the findings to the model specification was checked by splitting the data into separate tables for each combination of model (BART, Gemma, Llama 3, and T5) and metric (Regard and SiEBERT). A simple linear model was then fitted for each of these eight datasets. The linear model can be expressed as:

$$\text{sentiment}_i = \beta_0 + \beta_1 \cdot \text{gender}_i + \beta_2 \cdot \text{max\_tokens}_i + \beta_3 \cdot \text{doc\_id}_i + \epsilon_i$$

Where β_0 is the intercept, β_1, β_2, and β_3 are the coefficients for gender, maximum tokens, and document identifier, respectively, and ϵ_i is the residual error term. This model was run separately for each LLM, and the output for the Regard metric is presented in Table 11.26, and for SiEBERT in Table 11.27. The model also produced a coefficient for each Document ID, which is not of interest, so these were excluded from the tables. Similarly to the GEE model, the point estimates are close to those from the mixed-effects model, though with smaller standard errors in this case. The estimated marginal means by gender for each of the models are presented in Table 11.28, and they are consistent with the findings from the mixed model.

Table 11.26: Linear model (Regard)

Coef	Estimate		Std. Error	t	Pr(>|t|)
bart					
(Intercept)	0.2840833	***	0.0155566	18.2613205	0.0000000
gendermale	0.0035545	**	0.0012465	2.8515437	0.0043639
max_tokens150	0.0001643		0.0021590	0.0761052	0.9393376
max_tokens300	0.0001634		0.0021590	0.0756762	0.9396789
max_tokens50	-0.0307295	***	0.0021590	-14.2329516	0.0000000
max_tokens75	-0.0054155	*	0.0021590	-2.5083062	0.0121543
max_tokensNone	0.0001634		0.0021590	0.0756762	0.9396789
gemma					
(Intercept)	0.3048890	***	0.0204331	14.9213487	0.0000000
gendermale	-0.0069472	***	0.0016373	-4.2431642	0.0000223
max_tokens150	0.0009879		0.0028358	0.3483562	0.7275835
max_tokens300	0.0141746	***	0.0028358	4.9983907	0.0000006
max_tokens50	-0.0140059	***	0.0028358	-4.9388990	0.0000008
max_tokens75	-0.0069622	*	0.0028358	-2.4550780	0.0141103
max_tokensNone	0.0147827	***	0.0028358	5.2128392	0.0000002
llama3					
(Intercept)	0.3144663	***	0.0216083	14.5530579	0.0000000
gendermale	0.0021104		0.0017317	1.2187019	0.2229998
max_tokens150	0.0114336	***	0.0029989	3.8125689	0.0001387
max_tokens300	0.0167968	***	0.0029989	5.6009226	0.0000000
max_tokens50	-0.0399157	***	0.0029989	-13.3099939	0.0000000
max_tokens75	-0.0127653	***	0.0029989	-4.2566164	0.0000210
max_tokensNone	0.0185463	***	0.0030004	6.1812957	0.0000000
t5					
(Intercept)	0.2153391	***	0.0303610	7.0926341	0.0000000
gendermale	0.0048940	*	0.0024328	2.0117104	0.0442900
max_tokens150	0.0073932	.	0.0042137	1.7545669	0.0793786
max_tokens300	0.0191323	***	0.0042137	4.5405227	0.0000057
max_tokens50	-0.0229692	***	0.0042137	-5.4510934	0.0000001
max_tokens75	-0.0076979	.	0.0042137	-1.8268779	0.0677621
max_tokensNone	0.0372671	***	0.0042137	8.8443118	0.0000000

Table 11.27: Linear model (SiEBERT)

Coef	Estimate		Std. Error	t	Pr(>|t|)
bart					
(Intercept)	0.6601786	***	0.0412309	16.0117242	0.0000000
gendermale	-0.0093810	**	0.0033038	-2.8394946	0.0045320
max_tokens150	-0.0010080		0.0057223	-0.1761527	0.8601792
max_tokens300	-0.0008814		0.0057223	-0.1540313	0.8775896
max_tokens50	0.0324356	***	0.0057223	5.6682846	0.0000000
max_tokens75	0.0093005		0.0057223	1.6253194	0.1041410
max_tokensNone	-0.0008814		0.0057223	-0.1540313	0.8775896
gemma					
(Intercept)	0.7857799	***	0.0484271	16.2260289	0.0000000
gendermale	-0.0420346	***	0.0038804	-10.8325995	0.0000000
max_tokens150	-0.0118507	.	0.0067210	-1.7632239	0.0779078
max_tokens300	-0.0241479	***	0.0067210	-3.5928966	0.0003293
max_tokens50	0.0358635	***	0.0067210	5.3360130	0.0000001
max_tokens75	0.0115767	.	0.0067210	1.7224544	0.0850328
max_tokensNone	-0.0313662	***	0.0067210	-4.6668826	0.0000031
llama3					
(Intercept)	0.4881037	***	0.0503594	9.6924036	0.0000000
gendermale	0.0055138		0.0040358	1.3662261	0.1719133
max_tokens150	0.0129288	.	0.0069892	1.8498242	0.0643824
max_tokens300	0.0136312	.	0.0069892	1.9503233	0.0511788
max_tokens50	0.0283793	***	0.0069892	4.0604580	0.0000495
max_tokens75	0.0123671	.	0.0069892	1.7694619	0.0768618
max_tokensNone	0.0166275	*	0.0069926	2.3778786	0.0174401
t5					
(Intercept)	0.7611087	***	0.0698796	10.8917072	0.0000000
gendermale	0.0101714	.	0.0055993	1.8165436	0.0693312
max_tokens150	-0.0451223	***	0.0096983	-4.6525854	0.0000033
max_tokens300	-0.0504907	***	0.0096983	-5.2061143	0.0000002
max_tokens50	0.0582791	***	0.0096983	6.0091831	0.0000000
max_tokens75	0.0249979	**	0.0096983	2.5775492	0.0099713
max_tokensNone	-0.1642472	***	0.0096983	-16.9355987	0.0000000

Table 11.28: Linear models: estimated marginal means (female - male)

	Regard				SiEBERT			
Model	Estimate		t	p	Estimate		t	p
bart	-0.0036	**	-2.9	4.4e-03	0.0094	**	2.8	0.0045
gemma	0.0069	***	4.2	2.2e-05	0.0420	***	11.0	0.0000
llama3	-0.0021		-1.2	2.2e-01	-0.0055		-1.4	0.1700
t5	-0.0049	*	-2.0	4.4e-02	-0.0100	.	-1.8	0.0690

11.3.2.6 Conclusion of robustness checks

The robustness checks consistently indicated the reliability of the findings with regard to Llama 3 and Gemma. Across the linear mixed model, robust linear mixed models, Generalised Estimating Equations (GEE), and separate linear models, the point estimates for the fixed effects remained stable, and the direction of the effects was consistent for the Gemma model, as was the absence of an effect for Llama 3. However, the variance-structured mixed effects model and the GEE model did not find significant effects in the BART and T5 models. Similarly, the bootstrapped results indicated significant effects slightly less than half of the time. This suggests that the results for the BART and T5 models may be on the boundary of significance and should be interpreted with caution. However, as the older models were primarily included as benchmarks and are not currently being used in practice to summarise care records, their bias is of less concern for long-term care policy. The consistent results across the Llama 3 and Gemma models, particularly in terms of estimated marginal means, indicate that the conclusions regarding state-of-the-art models are not sensitive to model specification or the presence of outliers, validating the robustness of the model.

11.3.3 Appendix 3 Evaluation of themes word lists

The word lists for each individual theme are included below. These, along with the complete code, can also be found in the GitHub repository [419]. The Python `str.starts_with()` method [481] was used for these terms. This means that, for example, in the mental health list, the term `autis` would match words that start with these letters, such as `autism` and `autistic`, but not words containing these letters, such as `flautist`.

11.3.3.1 Mental health

```
alzheimer
anorexia
anxi
asperger
autis
behavio
bipolar
cognit
confus
deliri
delusion
dementia
depress
disorient
hallucinat
insight
mental
memory
mood
paranoi
personality disorder
power of attorney
psycho
ptsd
restlessness
schizoaffect
schizophreni
sectioned
therap
```

11.3.3.2 Physical health

```
activities of daily living
amputat
anaemia
angina
arthritis
aspirat
asthma
atrial fibrillation
balance
barrier cream application
bed bound
bed rails
bed-bound
bedbound
bilateral limb
bleeding
blood pressure
blood test
bowel
breath
cancer
care needs
cataract
catheter
cellulitis
chest rash
cholesterol
cirrhosis
commode
community acquired pneumonia
constipat
continen
copd
coronary
diabet
diarrhoea
disability
disable
dysphagia
dysphasia
dyspraxia
epilep
fall
fatigue
```

```
fractur
gallstone
glaucoma
gord
gout
hard of hearing
hearing and sight
hearing impair
heart attack
heart condition
heart disease
heart failure
heart problem
hemiplegia
hernia
hip replacement
hoist
house bound
house-bound
housebound
housework
hypercholesterolemia
hypertension
hypothyroid
idiopathic
immobile
incontinen
infarction
infect
influenza
injury
insulin
intravenous
ischaemic
ischemic
kidney
knee
leg clinic
leg ulcer
lung
macular
medication
melanoma
mobili
motor neuron
mrsi
```

myeloma
nutrition
obstructive sleep
oedema
oesophageal
osteo
pain
paralys
peg feed
personal care
physical deterioration
physical injur
pressure area
pressure relieving
pressure sore
pressure stockings
prostatic
psoriasis
pulmonary
puree
raised toilet
renal
reposition
rollator
sciatica
scoliosis
seizure
sleep apnea
slurred speech
spinal
standing tolerance
stiffness
stoma
stroke
surgery
swallowing
swollen
thickener
transfer
underweight
unsteady
urinary tract
urine retention
uti
vein
visual impairment

washing legs
weak
weight bear
weight loss
wheelchair
zimmer

11.3.3.3 Physical appearance

abdomen
appearance
appetite
bath
black eye
bmi
bruised
cloth
dental
dirty
discolouration
dishevelled
disshevelled
dress
drooling
dusty
faeces
fingernails
groom
hair
hygiene
kempt
messy
nails
naked
neglected
nude
odour
rubbish
scruffy
self neglect
self-neglect
shave
skin
slurred
smell

soil
spots
stained
teeth
tidy
tremors
trousers
unclean
underwear
underweight
unhygienic
unkempt
untidy
urin
vest
wear
weigh

11.3.3.4 Subjective language

abus
adamant
adjusted
adverse
aggress
agitat
agreeable
angry
annoy
appear
appropriate
argumentative
articulate
bad
behav
benefi
best
better
bored
bossy
breach
challeng
chatty
choose
chose

```
clean
clutter
coherent
concern
confine
conflict
confus
content
damage
demanding
dependent
deteriorat
difficult
dirty
dishevelled
dislike
disparaging
disruptive
distracted
distress
dusty
erratic
escalat
evasive
exacerbat
excessive
failed
feel
felt
fiercely
fixation
fluctuat
forgetful
frustrat
fuss
good
happier
happy
hard
harm
hate
high
ignor
illiterate
immens
impair
```

improv
impulsiv
inability
inappropriate
incoherent
increase
ineffective
insecure
insight
instrumental
insufficient
intense
invalid
involuntary
irk
irrita
isolat
issue
lack
less
likes
limited
loner
loudly
lovely
low
lucky
marked
massive
maverick
mess
mismanage
misses
misusing
mitigated
mood
more
muddle
needs
negative
neglect
nice
odd
oriented
paranoid
placid

pleasant
pleased
pointless
poor
prais
problem
proper
proud
racist
recommend
refus
relaxed
relentless
reliant
reluctan
resist
respect
restless
risk
rough
rude
sadly
safe
scared
scruffy
serious
settled
severe
shy
significant
silly
slow
small
smartly
smell
sociable
soil
strong
struggl
stupid
substantial
sufficient
suitable
suited
tearful
unable

```
unacceptable
unamenable
unaware
uncomfortable
uncontrollabl
uncooperative
under weight
underweight
unhygienic
unkempt
unreasonabl
unreliable
unsafe
unsatisfactory
unsettle
untidy
unwise
valid
verbal
vulnerab
wander
well
willing
wise
working
worried
worrying
worse
worst
```

11.4 Appendix to Chapter 9

11.4.1 Methodology for generative LLM classification

This section contains the methodology in Section 9.2.2.1. To investigate whether generative LLMs could effectively perform the classification task described in Chapter 6, I used three models: Llama 3, Gemma, and Llama 3.1. These models were provided with the same instructions that had been given to human annotators in the original study. The primary objective was to assess the models' ability to replicate the manual classification process without additional training. Llama 3 and Gemma both have a maximum prompt size of 8,192 tokens, which was approximately the length of the instructions provided to human annotators. Due to this limitation, it was not feasible to include additional training data or examples within the prompt for these models. Consequently, Llama 3 and Gemma were tested using the instructions alone.

In contrast, Llama 3.1, released on 23 July 2024 [430], supports a significantly larger context window of up to 128,000 tokens. This extended capacity allowed for the inclusion of additional classified case notes as examples within the prompt. For Llama 3.1, I conducted separate tests by providing the model with the same instructions plus 100 and 300 classified case notes, respectively. The aim was to determine whether the inclusion of examples would enhance the model's performance.

The full prompt used for each model, including the instructions and any additional examples, is detailed in Section 11.4.1.2. All analysis was conducted using a high-end Graphics Processing Unit (GPU) with 20 GB of memory. Despite this resource, memory constraints limited the number of examples that could be included in the prompt for Llama 3.1; including more than 300 examples exceeded the available memory capacity.

11.4.1.1 Technical considerations

Several technical considerations and limitations were encountered during the analysis. The prompt size limitations of Llama 3 and Gemma meant that only the instructions could be included without exceeding the maximum token limit. This restriction prevented the exploration of whether providing example case notes would improve the models' performance through few-shot learning. Although Llama 3.1 offered a significantly larger context window, practical limitations arose due to hardware constraints. Despite using a high-end GPU, I was unable to include more than 300 examples in the prompt without exceeding the available memory. This limitation underscores the challenges associated with working with large context windows in practice, as the computational resources required can be substantial. The inability of Llama 3.1 to improve performance with the addition of examples suggests potential limitations in the model's capacity for few-shot learning in this specific task. It may indicate that the models require fine-tuning on a labelled dataset rather than relying solely on prompt-based learning to achieve higher accuracy in classification tasks of this nature.

11.4.1.2 Prompt

The prompt used in the analysis is included below.

I am going to pass you some sentences in json form which are taken from long-term care case notes. I would like you to respond by telling me whether or not they indicate that the worker stated that the person is lonely or socially isolated. I will just use the term "lonely" from now on. Here are some rules which set out that a person is lonely:

- Statement that the person is lonely and socially isolated
- Statement that the person is socially isolated
- Statement the person feels lonely even if not socially isolated
- Statement that the person is socially isolated even it does not state that they are not lonely
- Statement that social isolation is a risk
- Statement that person lives alone if they do not like it
- Person expresses interest in attending a day centre with no reasons. If reasons given it must be for social reasons (not for carer respite or managing safety).
- Referral to a befriending service
- Receipt of befriending service
- Request for social support outside day centre
- Number of social contacts defined with subjective language
- Referral to a day centre.

Here are some rules that set out where a person is not lonely:

- Need for social support for purposes of managing safety
- Need for social support for purposes of carer respite
- Need for support because of risks associated with safety
- Need for support to manage practical tasks
- Generic request for support
- Need for support because of depression/anxiety
- Person is isolated in the sense of infection-control.
- Statement that person lives alone with no additional information
- Person has n social contacts per week without subjective language

```
- The fact that a person attends a day centre (i.e. not a referral,
    which happens at the point of isolation, but receipt of the service
    which suggests the need for social contact is being met)
- Offer of day centre refused
- Befriending for the purpose of carer support (e.g. sitting service to
    manage safety or for carer respite)

Example:

INPUT:

{
    "sentence_text": "She is socially isolated"
}

OUTPUT:

{
    "sentence_text": "She is socially isolated",
    "lonely": true
}

Please make predictions based on these rules. Please ensure that the output is
    valid json.
```

11.4.2 Methodology for evaluating bias in classification model

This section sets out the methodology used in Section 9.2.2.2. To assess potential gender bias in the loneliness classification model developed in Chapter 6, I used Llama 3 to generate gender-swapped versions of 1,000 case notes. The goal was to determine whether the model's predictions were influenced by gender-specific language, indicating bias in the classification algorithm.

The process began with the selection of 1,000 case notes from the dataset used in the original study. For the purpose of simplicity, I used a subset of case notes, only selecting those which mentioned one gender, and randomly selecting 1,000 case notes with a ratio of 700:300 not lonely to lonely. I automated the replacement of gendered pronouns and references within these texts using Llama 3, as set out in Chapter 8. This involved systematically changing words such as “he” to “she,” “his” to “her,” and altering any other gender-specific terms while preserving the overall context and meaning of each sentence. For example, the original sentence “He lives alone and has no regular visitors” was transformed into “She lives alone and has no regular visitors.”

Both the original and the gender-swapped case notes were then processed through the loneliness classification model. By comparing the model’s predictions on the two sets of texts, I aimed to identify any discrepancies that could be attributed to gender bias. If the model were biased, we would expect to see systematic differences in the predictions for male versus female versions of the same case notes.

The analysis of the model’s predictions revealed minimal differences between the original and gender-swapped case notes. Among the sentences indicating loneliness, there were 289 correct predictions for the male version and 290 for the female version. The total number of positive predictions (i.e., predictions indicating loneliness) was 303 for the male texts and 301 for the female texts. The slight differences observed in the number of true positives and total positive predictions are set out in Figure 9.1. While there are some gender-based differences, they are not of the magnitude that would explain the gender-based differences in loneliness found in the model’s results in Chapter 6. These findings support the conclusion that the observed gender differences in the original study reflect actual disparities in the data rather than bias introduced by the classification model.