International Growth Centre

Authors

Jeffrey Hammer

Charles and Marie Robertson Visiting Professor in Economic Development

Princeton University

Published on: 3 Feb 2016

Countries: Pakistan

Research themes: Inclusive Growth, Public Sector and Tax

Is 'big data' over-hyped? The importance of good data for improving health policy in Punjab

Despite the hype around 'big data', caution should be urged. Efforts to improve data collection require greater coordination and planning in order to ensure greater quality and usability of data. Overcoming the inertia around data collection needs a longer-term view on the application of the data. Improvements to collection and organisation of data sets can then facilitate integration of different kinds of data to better inform experimental policy designs.

Health is determined by multiple factors. As such, the health sector represents a good test case for exploring data collection and organisation. As the Government of Punjab thinks about overhauling and restructuring its data management systems, efforts to improve policy making call for a rethinking of the types of data collected, applications of it and way to organise the data that allow it to be readily merged with other data types.

Most rich countries have undergone a veritable revolution in data management in recent years. The availability of 'big data' has enabled entirely new avenues of research geared towards policymaking. For example, the availability of complete tax records enables public finance researchers to examine behavioural responses to tax code changes in finer detail than ever before. Similarly, hospital exit records, insurance claims and demographic data have been linked to answer questions of facility use, overuse, re-admittance and responsiveness to reimbursement rules. Answers to all these questions have informed policy design and changes. But all of these require systems in place to capture and use it productively.

A long-term vision: Data collection and usage

Information should be collected with an eye to its ultimate application, and put in a form that addresses policy questions. This requires systems to be in place that efficiently capture data. This process is certainly not automatic, especially in Pakistan.

Analysing contributing factors to health would be easiest through one comprehensive data set. At present, data

consolidation has not been feasible. A **recent analysis** of health outcomes in Punjab using two sets of population-based surveys (Demographic and Health Surveys (DHS) in 2006 and 2012 and the Punjab Multiple Indicator Cluster Surveys (MICS) of 2008 and 2011) was unable to answer any serious questions on correlations between policy inputs and health status.

Using the above-mentioned data sets, some basic findings do emerge. Economic status and mother's education remain the most important variables in determining the health status of children (using heights, weights and mortality rates as indicators of health status[1]). In fact, the relationship between economic and health status seems to strengthen over time. While it is also a clearly good idea, based on clinically proven effects, for children to be fully immunised against specific diseases[2], the impact of full immunisation on overall health status is largely undetectable.

It is clear from the outset, that existing surveys are inadequate for answering essential policy questions. Beyond these meagre results, there is little that is completely certain. Insufficient data on water quality limits our understanding of its impact on health. Proxies such as 'piped' water are distinct to 'safe' water. Similarly, a lack of evidence on the correlation between public primary health care facilities and population health outcomes is an obstacle to justifying a policy to expand primary health care facilities.

Understanding how direct and indirect variables drive outcomes requires being able to integrate different types of data

Data capturing indirect variables is also useful. Non health-care variables, such as roads (e.g. a measure of accessibility of health care), the disease environment, dietary patterns, educational attainment for children and parents alike, can differentially impact health outcomes.

> "Information should be collected with an eye to its ultimate application, and put in a form that addresses policy questions"

Data does already exist on soil moisture, well water cleanliness, and other vital statistics such as the road networks and construction, school enrolment and achievement. However, these are not in forms that can be merged. School data is by school, clinic data by clinic, water data by watershed (or village) and so forth. A synoptic treatment of data that can be integrated as needed.

Long-term panel data collection lays the foundation for more rigorous policy assessments

One organising principle can be to create a 'panel' dataset – a set of repeated observations over time across specific geographical units. If each village, for example, had reliable measures of health status that also included coverage of roads and their quality, of the educational status of the girls of the village and of weather or other environmental conditions of the

village, then this panel could enable 'before' and 'after' assessments as well as 'with' and 'without' comparisons for any policy change. Such data can also provide a 'rolling baseline' for new policy initiatives.

Policy analysts in every sector bemoan the lack of data available. This sometimes results in piecemeal attempts to correct the situation with one-shot or uncoordinated attempts to fill information gaps. The long run ideal will be to develop a comprehensive approach to data collection that builds upon the initial baseline of data over time. The challenge is that the full usefulness of the data may not be apparent at the point of collection, but instead accrues and becomes more apparent over a longer time horizon.

Poor data feeds a vicious cycle: Finite budgets and competing interests demand greater emphasis on evidenced-based policymaking

At present much the data is collected as a one-shot or uncoordinated attempt to fill information gaps. It tends to be relatively low quality, which produces a vicious cycle where data is seen to be unusable, and therefore little care is paid to its collection. This again results in poor quality data production which is ultimately an unconvincing input for policy conclusions.

> “ Policy analysts in every sector bemoan the lack of data available. This sometimes results in piecemeal attempts to correct the situation with one-shot or uncoordinated attempts to fill information gaps. ”

Incentives and pressures by different stakeholder may also complicate more effective data collection for certain policy challenges. Take doctor absenteeism as an example. Without accurate data, healthcare facilities can deny the problem exists, which may perpetuate disincentives to support more accurate data collection.

Getting more precise evidence on the role of public intervention on health is of utmost priority. If the intention is to rely on public primary healthcare to serve the people of Punjab, the fact that such provision has, to date, no discernible impact on health status needs to be explained. Ideally *before* large sums of additional money are allocated to it. In any case, when public budgets are limited, as in Punjab, comparing alternative uses for funds provides critical comparisons. More money spent on any of the following: health care, roads, water, sanitation, income support and education means less is available for the others. A fact often ignored by specific interests. Selecting which projects or initiatives the government should put its money into, depends on the relative value of each.

[1] Height relative to a standard established by the WHO for children of each age has been found to be a useful measure of nutritional status and overall health. It has also been connected

to cognitive development in children and susceptibility to disease. Weight relative to similar standards is not considered as good an indicator of long run health status as height but is much easier to measure during survey interviews. While the underlying concept is not as good an indicator, the pure measurement error is less. Therefore we examine both. Mortality is, tragically, self explanatory.

[2] Deaths from diseases such as measles are relatively "rare" in a statistical sense, that is, where very large samples would be necessary to detect incidence at the population level. Vaccine preventable diseases, of course, are all-too-common to families affected and tragic when they are, indeed, preventable.