Bias in surveys

Huw TO Davies

The simple survey is a regular tool in health services research. However, like any research method, surveys can be flawed in design, execution, analysis or interpretation. This short article outlines the basis of good survey design and advises how bias in published studies can be assessed.

INTRODUCTION

Surveys are ubiquitous in health care: surveys to assess health-care need, use or demand; surveys to assess attitudes, beliefs, knowledge and opinions (of health-care professionals as well as patients); and surveys to assess the patients' experience of health-care encounters. The simple survey is an invaluable tool for gaining insight into patients, health-care services or the interactions between the two.

Nonetheless, for all their apparent simplicity, surveys have the power to mislead as well as to inform. Like any research technique, surveys have pitfalls — they can be used inappropriately (for example, to answer questions to which they are ill suited), or they can be badly designed, poorly executed or erroneously interpreted. This short paper discusses the key design issues upon which surveys stand or fall, and advises on how to assess any flaws in published studies.

THE RIGHT QUESTIONS

Surveys have a design that is described as 'cross-sectional', i.e. they take a slice of the world at a given time point. They are ideally suited to quantifying how many individuals have what sort of characteristics and to what extent. Problems arise when some notion of temporality is introduced into the research questions, when we begin to ask about what happened in the past or about how different characteristics are linked. For example, suppose we took a sample of patients with diabetes who are being managed on a new form of insulin. Surveying these patients to discover what they think of the new formulation, and how it compares to their previous experience, may seem like a reasonable thing to do. However, such a survey will fail to capture the experience of individuals who tried the new formulation for only a brief time and then reverted back to their old approach.

A second common pitfall with surveys is the presentation of associations between different characteristics in the hope of shedding light on causal linkages. A previous paper in this series demonstrated how unsatisfactory it is to make these kinds of assertions from cross-sectional data (Davies and Williams, 1999).

Further problems may arise when analyses of subgroups within survey data are used to draw conclusions for which longitudinal data are needed. For example, consider data gathered from a population-based survey on blood cholesterol levels. Suppose these data show that the average blood cholesterol level is lowest in the youngest age group, higher in successive older age groups, and then lower again in those aged over 65 years. One might be tempted to conclude from this that blood cholesterol level increases as we age, but falls off after retirement. This may of course be true, but these data, gathered from a survey, are unable to demonstrate this. It may be, for example, that confounding factors explain the apparent rise (middle-aged people may, for example, have a quite different diet from younger people). Alternatively, it may be the case that people who are in mid years now had a very different upbringing and have had consistently elevated cholesterol levels all along. Perhaps even the apparent fall-off of cholesterol in the oldest group can be explained by the selective early death of those with abnormally high levels of cholesterol.

Thus what we do not know, and would need to know to draw some conclusions about how cholesterol levels vary with age, is what happens to the surveyed individuals as time unfolds. Following up the individuals in the survey and looking at individual changes in cholesterol level would help to determine which scenario will be played out in practice: will cholesterol fall, rise or remain the same? As soon as we begin to be interested in questions with a temporal component (e.g. what happens next and why? Or what happened in the past and how does it relate to the present?), then the more appropriate study designs are longitudinal ones, such as cohort studies, case control studies and trials.

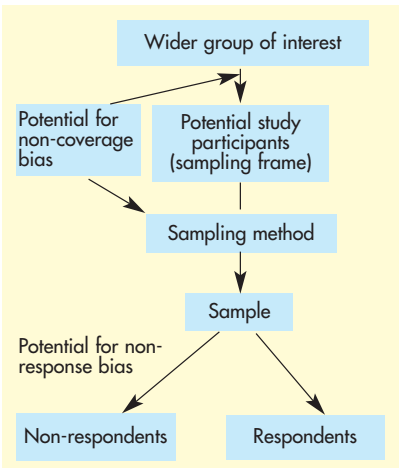

Figure 1. From group of interest to survey data: the potential for bias.

Dr Huw TO Davies is Reader in Health Care Policy and Management, Department of Management, University of St Andrews, St Andrews, Fife KY16 9AL

REPRESENTATIVE SAMPLES

Assuring that a survey is the appropriate design for the questions being asked is only the first stage in critically appraising a published study. Whatever data are actually gathered and presented in a survey, it is rare indeed that those individuals included in the study are the only individuals of interest. Far more usual is the situation when we want to use the data that we have on some selected individuals to say something interesting and relevant about some other wider group of interest. Sometimes, this wider group of interest has a clearly defined existence (e.g. a survey of general practitioners might want to infer something about general practitioners as a whole). At other times, the wider group of interest is more notional — for example a survey of patients attending for a flu shot might want to draw some conclusions about all such patients, both now and in the immediate future.

That surveys want to make some inferences about wider groups than the individuals actually studied raises some important questions about whether the data collected can support such inferences. In order to be confident that drawing wider conclusions is reasonable, we must first be convinced that the study group are representative of the wider group of interest; that is, they must be similar in all possible ways to the target group. Confirming that this is so can be difficult or even impossible. However, having some knowledge about how the survey was designed and conducted can allow judgments about representativeness to be well founded.

There are two ways in which representativeness can be compromised. First, the survey may be designed in such a way that certain individuals in the wider group of interest have no chance of being included in the study sample. This is called non-coverage. Second, those selected for inclusion in the study can fail to deliver any information. This is called non-response. *Figure 1* illustrates where these deficiencies may arise: both may introduce bias.

NON-COVERAGE

The design of the survey begins with some definition (even informally) of the target group of interest — the notional group about whom data from the survey are designed to provide insight. Next some kind of sampling frame is set up — a list or means of contact from which study subjects will be selected for inclusion. In assessing bias in surveys, the initial questions to ask are:

- Which members of the target group might be missing from the sampling frame?
- How might those missing differ from those who are included?

Of course, getting onto the sampling frame is no guarantee that a representative sample will be drawn. To avoid bias requires that a sampling method be chosen which ensures that each member of the sampling frame is equally likely to be selected in the final sample. Simple random sampling is the simplest way of achieving this, but other methods (such as multistage sampling) may also be acceptable. Failure at either of these two stages (drawing up the sampling frame, or selecting the sample from the sample frame) may introduce non-coverage bias.

NON-RESPONSE

Once a fair sample has been selected the problem still remains of non-response. Individuals contacted may fail to provide data for a range of reasons: they may have died, moved away or even just refuse to participate. As some of these reasons may be not unconnected with the questions being asked in the survey, it is a safe assumption that those who do not reply will differ in many ways from those that do. Thus non-response may introduce serious bias into the survey findings. For example, those who respond to a survey may well be those who hold strong views on the issue in question — thus presenting a false picture.

There is no level at which a response rate becomes respectable — and conversely, no level beneath which the findings are valueless. If the survey is carried out to discover a ballpark figure (many or few? often or rarely?) then even quite low response rates (less than 50%) may still provide useful information. For example, consider a survey with a response rate of 50%, which discovers that 60% of respondents have suffered a hangover. This survey at least places the likely true proportion as being somewhere between 30% and 80% (depending on whether all of the non-respondents have had a hangover or none of them, and subject to additional uncertainty through measurement bias and chance variation — see below). This may be sufficiently precise — or hopelessly imprecise — depending on what was previously known and what use is intended to be made of the survey information.

In contrast, a survey that seeks an accurate estimate of a relatively rare phenomenon may be hopelessly flawed even with a response rate of over 90%. Thus an acceptable level for the response rate depends crucially on the questions being asked.

MEASUREMENT PROBLEMS

Having ascertained that the study sample is not unreasonably biased, the next major consideration is of the quality of the data collected. The questions asked are about the questionnaire or other data-gathering instrument, and any measures used. In essence we are interested in the data reliability and the data validity. Reliability is concerned with reassurance that data gathering by different individuals, or at a different time or in a different context, would not yield different findings given no true change in the study participants. Validity, a more elusive concept, is concerned with ensuring that the data presented are in fact measuring what it is intended that they measure.

Simple physical measurements such as height and weight, or factual data gathering such as age, gender and educational attainment, may be relatively unproblematic (although even here there is scope for systematic biases to creep in). However, substantial difficulties arise when attempts are made at measuring complex con-

cepts such as quality of life, psychosocial functioning, or patient satisfaction. Even seemingly relatively straightforward issues such as pain, mobility or symptoms can hide measurement difficulties.

Measures may be inherently inadequate or flawed, or they may be particularly prone to bias arising from the settings in which the data are gathered. For example, data collected in the home setting may tell a different story from ostensibly the same data collected in a clinical setting. Convincing surveys will go to considerable lengths to reassure that any measurement difficulties have been understood and overcome.

SIZE MATTERS

All of the above considerations have been about bias in surveys: systematic deviations from the truth introduced during the design, execution, analysis or interpretation. One further problem remains: that of chance variability. As surveys almost invariably present data from samples, any given set of findings is just one of many possible instances. Thus any findings must be placed in the context of an understanding about how data might vary given the play of chance. This is usually done by creating '95% confidence intervals' around any sample values.

Table 1 shows the confidence intervals for a given sample percentage and different sizes of samples. There are two things to note from this table. First, samples of under a hundred or so have relatively wide confidence intervals (perhaps ±10% around the sample estimate, and even as much as ±20% for samples of just 25). Second, in order to increase the precision of an estimate from a survey much larger samples are needed. In fact, in order to reduce the confidence interval by half, four times as many individuals are required in the sample. It is only with sample sizes of around 1000 that the confidence interval around a sample estimate is as narrow as ±3%. Again the level of sample variability acceptable depends on the questions that the survey is being used to answer. Small samples may be acceptable for rough estimates, but precision requires the use of large samples.

In particular, if a survey finds no instances of a particular feature (a zero percent) this does not necessarily mean that the feature is absent. The zero finding could have been produced by chance alone. A useful statistical approximation is that the confidence interval for a zero finding is from 0% to approximately 300/*n* % (where *n* is the survey sample size). This useful approximation holds for samples of about 20 or more (Eypasch et al, 1995). Thus a zero percent from a sample of 30 gives an upper confidence interval of about 10%; and a zero from a sample of size 300 still has an upper confidence interval of 1%.

TABLE 1.
Confidence intervals (95%) for sample proportions

	Survey sample size				
Proportion in survey	*n*=25	*n*=75	*n*=100	*n*=200	*n*=1000
20%	4–36%	11–29%	12–28%	14–26%	17–23%
50%	30–70%	39–61%	40–60%	43–57%	47–53%
80%	64–96%	71–89%	72–88%	74–86%	77–83%

KEY POINTS

- Surveys answer questions about how many, how much and what sort? They are not readily able to say much about the interrelationships between variables over time.
- The three key questions to ask about a published survey are: Is the sample representative? What was the quality of the measures used in data gathering? Was the sample of sufficient size?
- In assessing representativeness, we should be asking what are the implications for the study findings of both non-coverage and non-response?
- In assessing the quality of the measures used the key issues are reliability (does the measure produce stable answers?) and validity (is the measure providing a true picture of the attributes of interest?).
- Small surveys provide imprecise estimates. Doubling the precision of these estimates requires a four-fold increase in sample size.

CONCLUSION

Surveys are easy to design and relatively quick to execute. They are also easy to appraise for flaws. In essence we are interested in four things:

- Are the research questions appropriate for a survey design?
- Is the sample group representative?
- Are the measures reliable and valid?
- Was the sample of sufficient size?

Providing clear answers to these questions should expose any major deficiencies in published studies.

More detailed guidance for assessing survey design and analysis can be found in a number of published texts, both for generic surveys (Crombie, 1996) and for surveys with specific design intentions such as assessing patient satisfaction (Fitzpatrick, 1991a,b). Full explication of the design issues can be found in standard textbooks (Streiner and Norman, 1989; Oppenheim, 1992; Schuman and Presser, 1996). HM

Crombie IK (1996) *The Pocket Guide to Critical Appraisal.* BMJ Publishing, London

Davies HTO, Williams FLR (1999) Confounded by confounding: separating association from causation. *Hosp Med* **60(8):** 599–601

Eypasch E, Lefering R, Kum CK, Troidl H (1995) Probability of adverse events that have not yet occurred: a statistical reminder. *Br Med J* **311:** 619–20

Fitzpatrick R (1991a) Surveys of patient satisfaction: I - Important general considerations. *Br Med J* **302:** 887–9

Fitzpatrick R (1991b) Surveys of patient satisfaction: II - Designing a questionnaire and conducting a survey. *Br Med J* **302:** 1129–32

Oppenheim AN (1992) *Questionnaire Design, Interviewing and Attitude Measurement.* Pinter Publishers Ltd, London

Schuman H, Presser S (1996) *Questions and Answers in Attitude Surveys.* Sage, Thousand Oaks, CA

Streiner DL, Norman GR (1989) *Health Measurement Scales: A Practical Guide to their Development and Use.* Oxford University Press, Oxford