Primer on Interpreting Surveys

To answer their research questions, investigators often need to ask questions of others. These questions may revolve around how people feel, what people know, and what people think. Some examples are given in the following table.

GENERAL QUESTIONS	EXAMPLES
How do people feel?	How do patients with lung cancer feel after having chemotherapy?
	How do physicians react to having their decisions reviewed?
	How much do healthy women fear breast cancer?
What do people know?	What do patients know about the benefit of chemotherapy in lung cancer?
	What do physicians know about the evidence supporting certain therapies?
	What do women know about their risk for heart disease?
What do people think?	Do patients with lung cancer think they should be told the average survival benefit?
	Do physicians think that there is a better way to change their behavior?
	Do women think that they are getting too much or too little information?

To address these questions, investigators must systematically question a defined group of individuals—in other words, administer a survey. This can be done in person, by mail, by phone, or over the Internet. Because surveys are increasingly common in the medical literature, readers need to be able to critically evaluate the survey method. Two questions are fundamental: 1) Who do the respondents represent? 2) What do their answers mean?

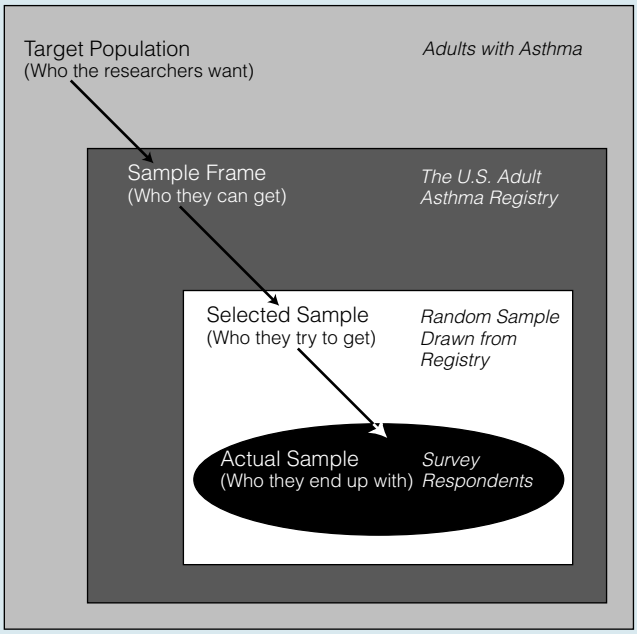
Who Do the Respondents Represent?

Like most types of research, surveys are useful only to the extent that they help us learn something about a defined population. The population we are interested in learning about is called the target population. Surveys are almost always based on a sample of the target population, and the respondents may not accurately represent this population.

Consider the following example. Suppose you are interested in how well adults with asthma are schooled in the use of spacers with inhalers. You question a sample of adults who are members of an asthma registry, and one third respond. You are surprised by how educated most of them are about the technique. You conclude that there is little need for further education.

What's wrong with this conclusion? Patients in the registry may be more motivated than patients in general. Furthermore, patients who received the survey and did not know the answers to the questions might have decided not to complete it. Therefore, it is possible that your conclusion is wrong and that, in fact, most asthmatic persons do not understand the use of spacers.

To avoid this general problem, readers need to ask themselves how well the respondents represent the target population. As shown in the following figure, there are three basic steps of selection between the target population (about which the conclusion will be drawn) and the actual sample (where the data come from). The reduction at each step potentially threatens a conclusion about the target population.



Target Population → Sample Frame

The sample frame is the portion of the target population that is accessible to researchers (e.g., persons who read newspapers, persons with phones). Often, the sample frame is some sort of list (e.g., a membership list). But individuals who are accessible may differ from those who are not. For example, persons with phones are different from persons without phones, and physicians who are members of professional organizations are different from those who are not. Readers should carefully judge how the sample frame might systematically differ from the target population.

Sample Frame → Selected Sample

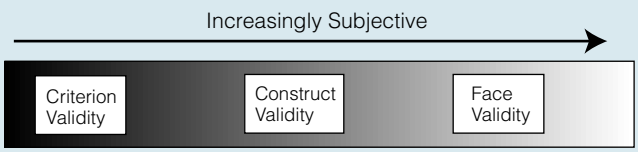
Although researchers may try to contact the entire sample frame, in many cases this would involve an unmanageable number of individuals. The selected sample is the portion of the sample frame that the researchers actually try to contact. If the selected sample is randomly selected from the sample frame, readers can be confident that this step does not seriously threaten generalizability. If it is selected by some other means, readers must be more circumspect. Suppose the selected sample is 100 patients who appear consecutively in an outpatient clinic (consecutive sample) or 100 persons who respond to a newspaper advertisement (convenience sample). Although both approaches are reasonable places to begin to learn about a topic, the first does not adequately represent patients coming to clinic (because it overrepresents persons who visit the clinic frequently) and the second does not adequately represent persons who read newspapers.

Selected Sample → Actual Sample

Not everyone who is contacted responds to a survey. The final sample is the portion of the selected sample that chooses to respond. However, the decision not to respond is usually not random—that is, respondents and nonrespondents usually differ. Patients who respond to questions about their disease may be more educated, have a smaller number of other problems, and care more about health. Physicians who respond to questions about guidelines may be more likely to believe that guidelines are important and more likely to be compliant. To judge these factors, readers need to consider the response rate. Whenever response rates are less than perfect (< 90%) and particularly when they are low (< 50%), readers should ask themselves how nonrespondents are likely to differ from respondents.

What Do Their Answers Mean?

Having decided who the respondents represent, readers can proceed to making judgments about their responses. The real challenge is to think about validity: How well do the survey questions do their job? Validity is the degree to which a particular indicator measures what it is supposed to measure rather than reflecting some other phenomenon. Although there are numerous kinds of validity (and even more names for each kind), it may be more useful for readers to consider validity as a spectrum, as in the following illustration.



Criterion Validity

At one extreme, readers can determine the extent to which researchers have compared the performance of their question with an external gold standard. Examples of criterion validity include comparing reported age with birth certificates, reported weight with measured weight, and reported eyesight with visual acuity. Although readers may be much more confident about a question that has been validated against an explicit criterion, they must also ask whether it may have been more accurate to simply apply the gold standard (e.g., why ask about weight when you can measure it?). Unfortunately, there is no criterion for many important questions (e.g., questions about what people think).

Face Validity

At the other extreme, readers need to consider for themselves whether the questions seem appropriate and reasonably complete "on the face of it." To really judge face validity, readers should look (and journals should publish) the exact language used in the question. Face validity has the disadvantage of being entirely subjective. At the same time, it may be the only type of validity that can be applied to the important subjective questions that survey researchers are trying to answer.

Construct Validity

Construct validity is somewhere between criterion validity and face validity. When the "gold standard" is not very objective but other data are available with which to judge a question's performance, we are in the realm of construct validity. The basic idea behind construct validity is that if your measurement does what you think it does, it should behave in certain ways. For example, the level of self-reported pain would be expected to decrease when respondents are given morphine. Wherever possible, readers should look for evidence that the pattern of responses is generally what would be expected given other data.

Interpreting Scores

It is increasingly common to see the answers for several questions aggregated into a single score ("The mean PDQ score for dentists was 2.5 points higher than for lawyers; $P = 0.03$"). If possible, readers should try to move beyond the score to consider the validity of individual questions. But because use of scores is increasing, readers also need to seek some grounding about what the scores mean ("Is 2.5 big or little?"). Sometimes this grounding can be achieved by knowing the mean score for groups with which one is familiar or by knowing how much a score changes after a familiar event. Knowing that the development of a new chronic disease translates to approximately a 5-point drop in the Physical Component Summary score of the SF-36, for example, helps give a sense for this measure of health status.

Conclusions

Survey research is an important way of learning what our patients understand and what they want. At the same time, it is often cluttered with unnecessary complexity and jargon. More important, false conclusions are a constant possibility. Simply figuring out what questions were asked and who the respondents were will go a long way toward avoiding these problems.