Issues of Representation in eHealth Research (with a Focus on Web Surveys)
Mick P. Couper, PhD

Abstract: The growth of eHealth research has been both rapid and sustained over the past several years. As part of this growth, the Internet is increasingly being used, both to deliver online interventions, and to collect data from a wide variety of groups of people. This paper reviews some of the issues around the latter use of the Internet, and in particular, the use of web surveys to produce results intended to be generalizable to a broader population. The paper provides an overview of issues of sample representation and the challenges faced in using Internet surveys. Various methods of sample selection, and their strengths and weaknesses are reviewed, and the threats to valid inference are discussed.

Introduction

The Internet has already had an enormous impact on many fields of human endeavor. Health research is certainly one of them. The original intent of the Internet was to facilitate collaborative research among scientists. More recently, however, attention has turned to the use of the Internet as a tool for connecting with research subjects or participants, for online experiments, randomized controlled trials, and surveys. The primary focus of this paper is on survey applications, but the lessons being learned about the issue of representation also apply to other research methods.

Interest in the Internet as a survey research tool came at a time when the survey industry was being assailed by declining response rates, legislation aimed at curbing telemarketers, increasing numbers of “bad polls” that threatened to give surveys a black eye, and—as a consequence of all this—rapidly rising costs for traditional survey methods. None of these pressures have abated in recent years. Consequently, some view online surveys as a replacement technology—one that would make other modes of data collection extinct.1

The Internet offered access to vast numbers of potential subjects, often those with rare characteristics, and research that could be conducted rapidly and cheaply, using complex designs driven by computer logic. In the headlong rush to adopt the new methods of eHealth research, sight should not be lost of the fundamentals of research design. This paper is not arguing for an abandonment of the Internet as a research tool, but urges caution, particularly in replacing existing research methods with online approaches. The Internet offers many exciting opportunities for health research, and is a valuable supplement to the methods currently used. Indeed, there are many things that cannot be done as well, or at all, without the Internet. These include reaching rare and hard-to-study populations, and delivering interventions and using measurement approaches that are hard to do using other methods. The paper focuses on a particular issue—that of representation—in the context of eHealth research, and specifically surveys, an area where representation has traditionally been of special concern.

Representation

Kish2 noted that research designs can be evaluated along three dimensions: representation, randomization, and realism. Research designs often compromise on one dimension in order to optimize on others, and eHealth research is no exception. A key feature of sample surveys is that they are designed to make inference or generalize to a broader population—that from which the sample is selected. 3 As Kish2 noted, “The aim of representative sampling is to make the sample a miniature so as to mirror and to represent the population with similar distributions.” In this sense, sample surveys are generally best on representation, but weak on realism.

Experiments gain their strength from randomization, a mechanism for controlling potentially confounding factors. The Internet may be ideal for randomized controlled trials, where the inference is to Internet users, and the intervention is Internet based. Controlled observational studies, including case—
control designs, are relatively weak on randomization and often on representation, but they are of relatively low cost and convenient, and typically high on realism in terms of studying phenomena in natural settings. With the computerization of surveys (as on the web), randomization is an increasing part of survey design, often blurring the distinction between surveys and experiments. But the primary objective of sample surveys remains generalization or inference to a broader population.

The strength of a sample survey is derived from probability theory, in which members of a target population are selected with known positive (i.e., nonzero) probabilities of selection. As Levy and Lemeshow\textsuperscript{4} note, “In probability sampling . . . the reliability of the resulting population estimates can be evaluated. This gives the user of the survey estimates insight into how much value can be placed on the estimates.” The same cannot be said of nonprobability methods. In practice, the word “survey” has often come to mean any data collected in a systematic manner from a relatively large number of people. This makes it hard to distinguish between those based on sound statistical theory, and those whose inference is based on less solid footing.

Furthermore, a probability sample is a necessary but not sufficient condition for inference. In theory, a key feature of probability sampling is the unbiasedness of the estimates, but this rests on two additional assumptions: (1) all members of the target population have a chance of being included in the sample (i.e., there is no coverage error), and (2) all persons sampled completed the survey (i.e., there is no nonresponse error). Sampling, coverage, and nonresponse are errors of non-observation, in that all result in a failure to observe, or obtain data from, a sampled member of the population. Even if all members of the population were observed (i.e., there was no sampling, no noncoverage, and no nonresponse), the estimates could still be different than the “true value,” mainly as a result of measurement error. The wording of the questions and the design of the survey instrument are key sources of measurement error in surveys. For an overview of survey errors, see Groves et al.; for an in-depth discussion, see Andersen et al.\textsuperscript{6} and Groves.\textsuperscript{7}

In practice, ideal designs are rarely if ever achieved. Survey design, like all other types of research design, involves compromise. Quality is not an absolute. When evaluating the quality of a particular research method or design, one important consideration is how the research design compares to alternative approaches. For example, a self-selected web survey would likely hold up well with regard to representation when compared to a mall intercept approach, flyer-based subject recruitment, or the use of a college subject pool. On the other hand, the same web survey would likely not be able to make the same claims regarding representation as would a probability-based sample survey method. Another consideration concerns how the quality of the research should be evaluated relative to the stated purpose of the study or the claims that it makes. In other words, is the inference, whether explicit or implicit, justified?

With this in mind, I review different methods for selecting samples for web surveys and their implications for survey quality. I then review what is known about errors of noncoverage and nonresponse in the context of eHealth research and web surveys.

**Sampling and Coverage Error**

Sampling error derives from the fact that a sample is drawn to represent the population that is not measured. Sampling errors form the basis of inferential statistics such as standard errors, confidence intervals, and significance tests. All other things being equal, the size of sampling error is inversely related to the sample size; in other words, large samples usually mean smaller sampling errors or narrower confidence intervals. But the estimation of sampling errors is based on an assumption of probability sampling. Large samples alone do not imply that there is no or little error.

The process of selecting a sample requires a “frame,” which can be a list of all members of the target population or a procedure for identifying such members (as in the case of random-digit-dial [RDD] samples). Coverage error arises through mismatches between the frame and the target population, and is a function both of the rate of noncoverage (the proportion of target population members not on the frame) and the differences between those covered and those not covered on the variable of interest. Thus, in the case of Internet surveys, members of the target population without access to the Internet are not covered by the Internet frame, and to the extent they differ from those who have access to the Internet, the resulting estimates may be biased. The levels of sampling and coverage error thus depend on the methods used to select the sample and characteristics of the target population and the sampling frame.

**Nonprobability Samples**

“Samples” for web surveys are obtained in many different ways (see Couper\textsuperscript{1} for a review). In many of these, little effort is made to ensure representativeness, and in fact such surveys are nothing more than very large groups of volunteers. In some cases, efforts are made to balance the sample to some defined population in terms of selected demographic variables, with the implication that the respondents mirror the population of interest on all other variables. This is akin to quota sampling. Of course, this is not meant to dismiss these approaches out of hand, but to argue that the conclusions (i.e., inferences) drawn on the basis of such
samples should be tempered by an understanding of the selectivity of such volunteers.

A brief review of the main approaches to identifying and selecting “samples” for web surveys follows. The two main types of nonprobability or self-selected methods used in eHealth research are (1) open, unrestricted surveys, and (2) opt-in panels. Both methods rely on volunteers. The former involves the wide advertisement of a URL where interested persons can complete a survey. Usually, there is no control over who completes the survey, or how many times they do so. A key concern with such surveys is that of “ballot stuffing,” with multiple completions by the same person or, in the extreme case, by a program designed to produce responses. The use of cookies or IP addresses may reduce most of such duplication but does not eliminate it.

The use of open-access surveys is no different than subject recruitment via flyers or word of mouth, and no worse than requiring research participation of undergraduates in psychology.\(^8\)\(^–\)\(^10\) In fact, online recruitment is likely to yield a larger and more diverse group of participants than college subject pools or mall intercept approaches, and at lower cost. Further, if the invitations go to sites visited by people with particular characteristics, it may be an efficient way of finding subjects with certain attributes or interests. But it is not a representative sample, and to claim that the views of those who complete the survey represent any broader group is to make unjustified inferential leaps.

Table 1 lists a set of published examples of health surveys based on open, unrestricted surveys. While not exhaustive, this list provides a flavor of the type and variety of topics covered by such surveys, and the range of journals in which they are published. Clearly, posting a web survey on the Internet and advertising its presence broadly through various electronic means is a popular way to collect data. This is not to suggest that such approaches are a waste of time, but that appropriate caution ought to be exercised about generalizations made on the basis of such samples.

Another type of nonprobability approach used frequently in market research, but less so in the health domain, is that of opt-in panels. There are a large number of companies operating such panels, with claims of numbers of panel members ranging from the hundreds of thousands to millions. All members of these panels are volunteers in the same way that those who do unrestricted online surveys are, but they may be even more selective in that they have agreed to participate in a number of surveys in exchange for a variety of rewards, financial or otherwise. In fact, a recent claim (www.comscore.com/custom-research/sample.asp) suggests that more than 30% of all online surveys are completed by less than 0.25% of the population, and that panelists in this group complete an average of 80 surveys over a 90-day period. While these claims may be exaggerated, it would be a stretch to argue that such panel members represent the broader Internet population. On the other hand, one advantage of the opt-in panel over the unrestricted approach is that profile data are often available on panel members, permitting the selection of samples based on desired characteristics. Another advantage is greater control over multiple submissions by the same person, which is a concern in unrestricted surveys.

It could be argued that the second stage of selection in such opt-in designs is a probability sample, but only in a technical sense is this true, as the inference is not

<table>
<thead>
<tr>
<th>Author (year)</th>
<th>Journal</th>
<th>Topic</th>
<th>Number of respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alexander (1996)</td>
<td>Urology</td>
<td>Chronic prostatitis</td>
<td>163</td>
</tr>
<tr>
<td>Anderson (2001)</td>
<td>Behav Modif</td>
<td>Dyspnea and panic attacks</td>
<td>205</td>
</tr>
<tr>
<td>Berman (2003)</td>
<td>Reprod Endocrinol</td>
<td>Sexual function complaints</td>
<td>3807</td>
</tr>
<tr>
<td>Conboy (2001)</td>
<td>Maturitas</td>
<td>Menopause and hormone replacement therapy</td>
<td>448</td>
</tr>
<tr>
<td>Corless (2002)</td>
<td>Respir Med</td>
<td>Public knowledge of tuberculosis</td>
<td>564</td>
</tr>
<tr>
<td>Eigenmann (2001)</td>
<td>Allergy</td>
<td>Anaphylactic reactions to food</td>
<td>264</td>
</tr>
<tr>
<td>Erwin (2004)</td>
<td>J Anxiety Disorder</td>
<td>Social anxiety disorder</td>
<td>434</td>
</tr>
<tr>
<td>Klapper (2000)</td>
<td>Headache</td>
<td>Cluster headache</td>
<td>789</td>
</tr>
<tr>
<td>Rodgers (2001)</td>
<td>Hum Psychopharmacol</td>
<td>Effects of Ecstasy and cannabis on self-reported memory</td>
<td>488</td>
</tr>
<tr>
<td>Ross (2004)</td>
<td>AIDS Educ Prev</td>
<td>Latino men who have sex with men</td>
<td>1026</td>
</tr>
<tr>
<td>Soetikno (1997)</td>
<td>J Am Med Informatics Assoc</td>
<td>Ulcerative colitis</td>
<td>463</td>
</tr>
</tbody>
</table>
to the set of panel members but presumably to some broader population from which they are drawn. Similarly, response rates can be calculated (i.e., the proportion of invited panel members who completed the survey), but these are useful only in an operational sense and add little information on the quality of the achieved set of responses.

Few studies using opt-in panels have been published in the health literature. One exception is Reed et al. Female members of an opt-in panel were invited to participate in a survey on “women’s health issues” to reduce the likelihood of self-selection on the topic of the survey (vulvar pain). The authors caution against making broad generalizations from the findings, but nonetheless suggest that a survey of this type and size is valuable, given the paucity of other broad-based studies on the topic. This study can be contrasted with that of Gordon et al. who examined web-based survey data from a group of “highly educated, Internet-savvy women who frequented various vulvar pain discussion lists.” Another example of opt-in panel use is found in Cameron et al.

While both nonprobability approaches are based on self-selected volunteers, they differ in important ways. In the first, people are attracted to the site or survey because they have an interest in the topic or engage in the behavior in question. In other words, their participation is based directly on the focus of the study. Selection bias may be directly related to the variable of interest in a particular survey, the behavior in question. In the second approach, the decision to join a panel may be orthogonal to the topic of the study. While those who join such online panels are likely to differ from the Internet population or the general population in many respects, they may not necessarily do so on the variable of interest in a particular survey (e.g., a health behavior or medical condition). The selection mechanism in the second case may be unrelated to the focus of the study.

There is another case that could lead to bias, in which the topic of the survey is revealed to the panel members at the time of invitation. Recent examples from one such panel are “Whiskey drinkers—A new survey for you!” or “Share your thoughts on home, lawn, and garden care,” or even more specific, “Attn: Parkinson’s Disease Patients or Caregivers.” Those attracted to the topic are more likely to start the survey. As the survey progresses, the content becomes clear to the respondent, and may lead to break-off as well. But this is generally thought of as a problem of nonresponse rather than sample selection, and will be addressed later.

The difference between the two approaches has implications for the choice of methods. If one is trying to locate members of a rare group, the use of an open access site may be more effective. If the goal is to get a broader range of participants, where the selection mechanism is unrelated to the topic being studied, the opt-in panel approach may make more sense.

In summary, there are two major inferential assumptions underlying many of these self-selected approaches: (1) that the Internet population (if such can be defined) resembles the broader population on the variables of interest, and (2) that those who volunteer for web surveys resemble Internet users in general. Evidence in support of both claims is generally lacking. Without empirical support for these assumptions, caution must be taken in interpreting the results from nonprobability samples of Internet users.

**Probability-Based Samples**

Despite the challenges facing those attempting to use web surveys to make broad-based inferences to the entire U.S. population—or that of any other country—there are several ways in which probability samples of more restricted populations can be drawn. Two general approaches can be described. The first restricts the sample frame (and hence the population) to active Internet users. The second uses alternative methods to identify and select the sample, switching the selected sample to the Internet for data collection, or using mixed-mode designs to counter the limits of Internet access. Each of these two broad approaches is reviewed in turn.

In terms of sample restriction, even if the entire population of interest were to have Internet access, there is no way to draw a probability sample without a list or a mechanism for identifying members of the population. The appeal of RDD telephone surveys is that a list is unnecessary, as a sample of all telephone numbers can be generated. The format of e-mail addresses makes the development of a similar approach for Internet samples unlikely. For this reason, the two approaches for sampling persons from an online frame used most often are (1) intercept-based approaches and (2) list-based samples.

Intercept sampling is often used when construction of a frame is impractical. Probably the best-known examples are the exit polls conducted on election day, with a sample of voters being intercepted as they leave a polling place. Intercept samples are popular on the Internet, particularly if the population of interest comprises visitors to a website. The key challenge for this approach is that of nonresponse, as shall be seen shortly. Many health-related websites (including the Centers for Disease Control and Prevention and the National Library of Medicine) use intercept surveys for customer feedback and site evaluation.

Turning to list-based samples, there is a growing number of populations of interest in their own right, for which researchers may have a near-complete list of members and their e-mail addresses. For example, many college campuses are fully wired. Students are not
only automatically assigned e-mail addresses when enrolled, but are increasingly required to use the Internet for all aspects of their education. College students are thus an ideal group for Internet-based survey methods. Other groups include members of professional associations, such as the American Medical Association, American Medical Informatics Association, and the like. While not all members may have Internet access, membership lists often include an e-mail address if one is available, as well as an alternative form of contact (telephone or mailing address), thereby permitting follow-up using an alternative mode. Yet another group of interest to eHealth researchers is those who use the Internet for a variety of health activities. This may require registering online to access a particular site or service, thereby providing e-mail addresses that form the sampling frame. The number and variety of services provided online by health providers means that a growing number of member databases contain information about Internet access.

This is not to suggest that HMOs, for example, will soon be able to survey all of their members using the Internet. But both the number and proportion of members with online access makes this a potential resource for studying the online behavior and use of eHealth resources of those with access. It is primarily these kinds of populations that are driving the interest in eHealth research.

A recent paper cautions us about the limitations of this approach. Hsu et al. conducted a records-based analysis of members of Kaiser Permanente–Northern California. To use eHealth services, members must register by requesting a secure password-protected account. Of the 3.4 million members in 2002, only 9.3% had documented eHealth access. While this is a dramatic increase over the 1.6% with access in 1999, it is still a minority of members. Of the subjects with known access, 36% used eHealth services at least once during 2002. Further, they found that those with access differed systematically from those without on a number of demographic and health-related variables. They conclude that “there is a significant and growing digital divide with respect to eHealth services across racial/ethnic and SES groups.” Thus, even in this case it is important to remember that the inference is at best to those HMO members who use the Internet, and not by extrapolation all members of the HMO.

There are two other approaches to obtaining probability samples using the Internet that deserve a brief mention: (1) mixed-mode designs and (2) recruitment using alternative modes. In the first, respondents are sent questionnaires in the mail, with the option of completing the survey online if they so choose. The Census Bureau uses this approach for the decennial census and the American Community Survey, among others. This approach realizes some cost savings in terms of back-end processing, but such savings are generally only realized on very large survey operations.

Another approach to probability-based samples of Internet users begins with other modes of data collection. For example, Knowledge Network’s online panel begins with RDD telephone recruitment. Those without Internet access are provided with access in exchange for participation in the panel. Similar approaches are being used in telephone and face-to-face surveys. Again, the biggest challenge of these approaches, aside from cost, is that of nonresponse. Much more research is needed on how respondents recruited this way may differ from the general population.

Nonresponse Error

As with coverage error, nonresponse error is a function of both the rate of nonresponse (the proportion not responding over the total eligible for the survey) and the differences between those who respond and those who do not on the variables of interest.

Given the inferential weaknesses of nonprobability designs, the calculation of response rates or the estimation of nonresponse error makes little sense, other than to evaluate the operational aspects of the survey. The focus here is thus on probability-based approaches. While coverage and sampling errors are of less concern for such samples, nonresponse may be of more concern. Intercept- and list-based approaches are likely to have different implications for nonresponse error. In the first case, visitors to a website often leave few traces (other than an IP address, which is insufficient for identification). This means that if visitors fail to comply with the request to complete the survey, there is little opportunity to follow up, and little is known about them other than their browsing behavior during that session. In other words, there is little opportunity to increase response rates, or to understand possible differences between respondents and nonrespondents.

In general, response rates to intercept surveys are very low. For example, Dodge and Cucchi report a 0.26% response rate to a survey of visitors to a poison center website. Similarly, the National Library of Medicine conducted an intercept survey of visits to the website (www.nlm.nih.gov) in July 2002. Of the 92,288 users invited to complete the survey, 4163 did so, for a response rate of 4.6%. The claim that “this is a representative sample of site visitors for a 10-day period and can be projected to represent an average profile of site visitors” thus rests on the assumption that the small fraction of visitors who completed the survey are like the balance of site visitors with respect to their answers to the survey questions. Given response rates this low, and the likelihood that those who choose to respond are different from those who do not, nonresponse error is likely to be a serious threat to inferential accuracy in intercept surveys. Nonetheless, they serve a useful pur-

May 2007

pose, and remain a popular approach for giving users an opportunity to voice their opinions and concerns.

In the case of list-based samples, by contrast, the opportunity to follow up with nonrespondents exists. In addition, more information than just an e-mail address is typically available. If other contact information is available, nonrespondents can be followed up using another mode (e.g., mail or telephone). If background information is available on the list member, indicators of potential nonresponse error can be examined. In other words, the richer the frame—the more information available about the sample members—the more can be learned about the potential biasing effects of nonresponse on the survey estimates. Response rates to list-based surveys vary greatly, depending on factors such as the population, topic, and length of the survey, method of invitation, use of incentives, and so on. Several experimental comparisons suggest that web survey response rates may be lower than those for comparable mail surveys, but some exceptions are also found in the literature.41–45 It is probably fair to say that while decades of research have been done on other modes of data collection, much work remains to be done to identify the best techniques for increasing response rates in web surveys.46

Similar problems with nonresponse are encountered when using other modes of data collection (face-to-face to telephone) for recruitment, even if coverage is addressed by restricting the sample to Internet users or providing sample persons with Internet access. For example, a current project is using an ongoing telephone survey to recruit potential participants to complete a web survey. About 60% of telephone respondents report using the Internet; of these, about 55% express willingness to do an Internet survey and, of these, 74% provide an e-mail address. Given that the response rates to the telephone survey are running around 53%, this means a combined coverage/response rate of around 13% (0.53 × 0.60 × 0.55 × 0.74 = 0.13), before the invitations to the web survey have even been sent. Even if the population of interest is restricted to Internet users, the cumulative response rate is still under 30%. Schlenker et al.36 report a cumulative response rate of around 30% for their survey using the Knowledge Networks panel. The advantage, however, is that a lot of information on respondents and nonrespondents alike is available from the initial recruitment, permitting the use of informed nonresponse adjustment strategies.47

As Groves40 review notes, however, there appears to be no direct relationship between response rates and nonresponse error, and much more work is needed here. Focusing solely on response rates is a risky strategy. Whatever methods are used for recruiting participants to web surveys, researchers ought to incorporate special studies to explore the nature of the potential selection bias, both in terms of descriptive statistics (e.g., means, proportions) and analytic measures (e.g., regression coefficients).

Summary
To summarize this brief review, while the Internet offers a lot of promise for eHealth research, it also suffers from some notable limitations. Key among these are the challenge of drawing representative samples of the general population, of dealing with the issue of people without Internet access, and of minimizing the potential for nonresponse bias in web surveys. Together, these threaten the inferential value of web surveys, particularly those focused on broad and diffuse populations. On the other hand, the cost of web surveys relative to alternative modes, the speed with which they can be conducted, and the ability to combine the power of computerized survey instruments with the advantages of self-administration, make this a valuable research tool under certain conditions.

The Internet is certainly not a panacea for all that aids survey research in the 21st century. It is a tool and, like all other tools in a researcher’s toolkit, is better suited to some tasks than to others. Much work remains to identify barriers to broader use of this method, and to overcome potential disadvantages. Potential threats to valid inference need to be overcome. Methods need to be developed to guide researchers in using this tool appropriately and successfully.

No financial conflict of interest was reported by the author of this paper.

References