JAN RESEARCH METHODOLOGY

Methodological issues in online data collection

Mary Ann Cantrell & Paul Lupinacci

Accepted for publication 3 August 2007

Correspondence to M.A. Cantrell:
e-mail: mary.ann.cantrell@villanova.edu

Mary A. Cantrell PhD RN
Associate Professor
College of Nursing, Villanova University, Villanova, Pennsylvania, USA

Paul Lupinacci PhD
Associate Professor
Department of Mathematical Sciences, College of Arts and Sciences, Villanova University, Villanova, Pennsylvania, USA

CANTRELL M.A. & LUPINACCI P. (2007) Methodological issues in online data collection. *Journal of Advanced Nursing* **60**(5), 544–549
doi: 10.1111/j.1365-2648.2007.04448.x

Abstract

Title. Methodological issues in online data collection

Aim. This paper is a report of a study to evaluate the use of an online data collection method to survey early survivors of childhood cancer about their physical and psychosocial characteristics and health-related quality of life.

Background. A major advantage in conducting web-based nursing research is the ability to involve participants who are challenging to study because of their small numbers or inaccessibility because of geographic location. As paediatric oncology patients and early survivors of childhood cancer are often not easily accessible because of their small numbers at single institutions, web-based research methods have been proposed as a potentially effective approach to collect data in studies involving these clinical populations.

Method. Guided by published literature on using the Internet for data collection, an online protocol was developed; this included construction of a website, development of a homepage and interactive HyperText Markup Language pages and the posting of the study link on various websites. Data collection occurred over a 6-month period between December 2005 and May 2006.

Findings. Despite using strategies in conducting online research cited in published literature, the recruitment of subjects was very prolonged and the volume of missing data among many respondents excluded them from the study and created bias within the study's results.

Conclusion. Web-based, online data collection methods create opportunities to conduct research globally, especially among difficult to access populations. However, web-based research requires careful consideration of how the study will be advertized and how data will be collected to ensure high quality data and validity of the findings.

Keywords: childhood cancer survivors, empirical research report, methodological issues, nursing, online questionnaire, web-based research

Introduction

Literature that discusses the advantages, challenges and specific methodological considerations in conducting web-based research continues to emerge. In an excellent article on using the Internet to conduct research, Ahern (2005) summarizes published literature on the advantages that online research offers for nurse researchers and study participants. Ahern identified the advantages of online data collection for researchers as: (1) being less expensive, (2) reaching a larger pool of potential study participants, (3) increasing access to study sensitive issues, cultural groups

and 'hidden populations', (4) decreasing data collection time, (5) increasing methodological rigour and control, (6) increasing accuracy and efficiency of data entry and analysis, and (7) having the ability to follow-up with participants. Advantages for study participants involved in web-based research are: (1) increased anonymity, (2) ability to provide information at their own pace, (3) increased sense of control, (4) increased willingness to participate because of it being a novel approach to research, and (5) convenience and ease of use (Ahern). Despite these advantages, several authors have cautioned researchers to consider the unique methodological issues that web-based research can pose (Duffy 2002, Duffy 2000). Duffy (2002) suggests that while this methodology is potentially useful in nursing research, care and attention to the nature of the sample, testing environment and environmental factors, privacy and confidentiality, and response rates must be carefully considered and addressed. If attention is not given to these methodological issue, serious negative consequence can occur that affects the validity of study findings (Duffy).

Background

Childhood cancer patients and survivors of childhood cancer are not easily accessible in sufficient numbers to conduct on-site or even limited-site studies (Hinds *et al.* 2006). In addition, researchers involved with the Childhood Cancer Survivor Study, a longitudinal cohort study funded by the National Cancer Institute, have recognized the need for using technology to meet the long-term needs of childhood cancer survivors (Dalton 2005). Programs such as Passport for Care, an Internet-based portable care summary of treatment history with individualized guidelines for follow-up care allow survivors to be active participates in their follow-up care. Web-based health-related research, including both survey and intervention studies, has increased exponentially in the past 15 years. Non-experimental web-based studies have received the most attention, but Internet-based clinical trials are now feasible (McAlindon *et al.* 2006). Computer-based research can address the prohibitive factors of participating in research because of geographic distant locations, travel time, weather and road conditions (Hill *et al.* 2006). Hinds *et al.* acknowledge that adolescents and young adults with cancer are similar to most individuals in this age group in that they are active users of web-based technology, and use chat rooms and other discussion forums to describe their experiences with cancer. These researchers point out that discussion forums and electronic bulletin boards have not been used for research purposes related to HRQOL among paediatric oncology patients but may be innovative, valuable future research tools.

The study

Aim

The aim of this study was to evaluate the use of an online data collection method to survey early survivors of childhood cancer about their physical and psychosocial characteristics and their health-related quality of life (HRQOL).

Design

A web-based online data collection protocol via the Internet was chosen for the study, based on the advantages cited by Ahern (2005) – specifically the nature of the population being studied and from the suggestions by Hinds *et al.* (2006) identified previously. The online protocol included construction of a website, the development of a homepage, interactive HTML (HyperText Markup Language) pages and the posting of the study link on various web sites. The development of the protocol was guided by published literature on using the Internet for data collection (Dillman 2000, Hanscom *et al.* 2002, Klein 2002, Lakeman 1997, Thomas *et al.* 2000). The methodological considerations cited in these sources are summarized in Table 1. These recommendations differ from those identified by Ahern (2005) in that they include specific methodological techniques and suggestions on how to construct a web-based protocol. For example, in the development of the study's web pages that contained the online questionnaires the decision as to whether to construct the web pages with a graphic background (pages with colored images) vs. plain pages was guided by recommendation by Dillman (2000). Dillman suggests that using web pages that do not contain graphics requires less transmission time and can provide better results. Likewise, the font size of the writing on the web pages required consideration. Hanscom *et al.* (2002) recommend a sizeable, easily read format with large, clearly delineated buttons for each response to make questions less difficult to read for persons with visual difficulties. Finally, all seven survey instruments, including the demographic/health profile questionnaire, were formatted in individual active HTML web pages using the FRONT PAGE programming system (Batagelj & Vehovar 1998). The instruments to measure participants' physical and psychosocial characteristics and their HRQOL were the General Health Rating Index, Affect Balance Scale, Nowotny Hope Scale, Personal Resource Questionnaire, Nowotny Hope Scale, and Minneapolis-Manchester Quality of Life Instrument.

To advertize the study, Klein (2002) and Lakeman (1997) recommend posting an active link to a study's homepage on established web sites frequented by potential study

Table 1 Suggested techniques for web-based research

Batagelj and Vehovar (1998)	1 month is widely accepted as the amount of time required to collect data via online surveys
	An online study should include both static and active web pages. Static pages are the same for all respondents vs. active web pages that capture individualized data such as demographic data, web address from which responses came from, and time of completion of survey.
	Data should be immediately transferred to a sender browser called a CGI (Common Gateway Interface) script (computer program) to store it.
	In using plain HyperText Markup Language (HTML) Forms, questions should be ordered one after another on a single page. However, one weakness with this approach is that respondents are asked to view the whole questionnaire. Online interviews that work on the principle of the HTML forms are in essence electronic versions of mail survey interviews – they must be short, simple and without complex jumping patterns.
	CAWI – Computer Assisted Web Interviewing. In this mode, the questions (or groups of questions) are ordered one after another on several pages. CGI support at the beginning and at the end of each HTML form can have controls over missed questions. CAWI is most frequently used among professional research firms.
	Layout of web pages could be one long scrolling page for the whole questionnaire or a layout where each question block was put on its own page (CAWI); the next page appears only when the previous page was finished.
	Completion rate is highly influenced by the type of browser used. The average length of an interview is higher when text browsers are used and when multiple page layouts are used.
	ICDC – Integrated Computer-Assisted Data Collection is software that enables easy transformation of the questionnaire to the surveys of different modes. The basic idea is to create the questionnaire once and let the software create the final layouts for all modes of data collection. Such a solution is in all major packages for computer-assisted data collection.
Hanscom *et al.* (2002)	Missing value rate for a computer survey is approximately half the missing value rate for a paper survey.
	Missing and inconsistent data are relatively rare in both paper and computer surveys, online surveys can improve on data completeness and consistency.
	If one question is posted on a full computer screen, it is harder to skip a question than a paper copy with 20 questions on a page.
	A large font and clearly delineated buttons for each response make questions easier to read for persons with visual difficulties.
	Online surveys eliminate the need for data entry and less error is expected.
	Requiring responses to all questions and re-asking skipped questions may decrease response rates because some questions may be sensitive to answer or may not be understood by respondents.
Klein (2002)	Plain web pages without colour and HTML tables provide better results than fancier versions.
	Acknowledgement of respondents participation in a study should occur seconds after they have completed the survey.
	To ensure anonymity, online surveys should be e-mailed to a neutral party, such as an established website.
	Internet data must be properly secured when stored on a server or computer.
	Informed consent must be provided and respondents made aware of the risks and benefits of electronic records.
Lakeman (1997)	Advertize study on world wide web (WWW) pages, established web sites or web pages that focus on special interest groups. These online sites frequently provide guests (researchers) to describe their work.
	When recruiting participants online a posting of a précis of the research proposal, an outline of how the results will be used and invitation to participate should be included in the advertizement of the study.
	Include information about where individuals can obtain more information, for example the researcher's e-mail address, or WWW page should be provided.
	A questionnaire embedded in a WWW page is easier for a participant to complete.

participants. This promotes awareness of the study among targeted groups and promotes anonymity among study participants, as participants do not have to directly sign on to a study's web site. A link to the study's homepage was posted on six web sites that are used by survivors of childhood cancer. Permission to post a link was granted by each web site's administrator. The posting on these established web sites included an introduction to the research protocol, a pledge of anonymity, the researcher's email address, information about the investigator's research and clinical nursing experience background and a direct link to the study, as suggested by Lakeman (1997).

The online protocol was pilot tested with 10 healthy young adults, ages 20–21 years. After these individuals linked onto

the study's homepage and completed the questionnaires, they were asked how long it took to complete all questionnaires and if they encountered any difficulties in accessing the study's homepage or the online web pages. These participants were also queried if they had suggestions to improve the ease in completion of the questionnaires. The average time to complete all survey instruments was approximately 15 minutes and no one in the pilot group expressed difficulty in accessing the study's homepage and link to the online questionnaires. Modifications to the social support scale were made as some of the questions were to be answered if a respondent only answered yes to the first question in the grouping. A non-applicable response was included to these questions' response set if the previous question was answered with a no response.

Participants

Eligibility criteria for subjects to enter the study were: (1) diagnosed with cancer during the adolescent period (12–18 years of age), (2) currently experiencing young adulthood (22–28 years of age), (3) not diagnosed with a brain-related tumour, (4) off treatment for at least 1 year with no recurrence of cancer and (5) no major health or developmental problems. Self-reported demographic data, such as age at diagnosis and type of cancer, were screened and monitored to ensure that eligibility criteria were followed. However, verification of the accuracy of these data was not possible as all data were supplied via online self-report. Data collection occurred over a 6-month period from December 2005 through May 2006.

The average age for respondents in the study was 23; the majority were white, single and female. Most survivors had been off treatment for 1 year and were diagnosed with cancer at age 15·15 years. The majority had graduated from high school and had earned college credits (68·5%) and 50% were employed. Of those employed (n = 27), 55·5% earned an income of < $10,000 yearly. No data were collected that identified whether respondents simultaneously went to school and/or worked or if workers were employed on a part- or full-time basis.

Data collection

At the start of data collection, completion of all fields was required so that every item on each questionnaire had to be completed by the respondent before he/she could move to the next survey instrument. After 16 participants had completed the questionnaire and at the request of the administrator of one of the posting web sites, the decision was made to allow respondents to skip one or more questions on one questionnaire and go onto the next one. The rationale for the allowing optional fields to be completed was supported by Hanscom *et al.* (2002), who suggest that requiring responses to all questions may decrease response rates because some questions may be sensitive to answer or may not be understood by respondents. This decision to allow fields to be optional created the circumstances of having a significant amount of missing data and many incomplete responses.

Ethical considerations

Institutional review board approval was obtained prior to the start of data collection. Before respondents began to complete the questionnaires, they were made fully aware that procedures to ensure anonymity were in place and were informed about the potential risks that exist in data security violation associated with providing online information. Respondents were made aware that submission of completed questionnaires implied their consent to participate in the study. Once respondents' submitted their responses, they were immediately acknowledged and thanked for their participation in the study.

Data analysis

The data were downloaded, recorded and stored in an ACCESS program on a secure server. Ninety respondents who reported to be childhood cancer survivors entered the site and began completing the questionnaires; however, many did not respond to all items on each questionnaire. Because of the existence of these missing data, extensive data cleaning was conducted and guidelines on how to address missing data in each questionnaire were developed. To address the missing data, the statistical procedure of data imputation was conducted. This is a procedure for entering a value for a specific data item where the response is missing or unusable. This data editing procedure was carried out manually by identifying those records containing missing data. The goal of the imputation process was to obtain reliable information from the online survey while trying to minimize the consequences of allowing subjects to skip to another questionnaire before entirely finishing the preceding. In most cases, if the childhood cancer survivors failed to answer more than two questions on an individual questionnaire, they typically answered fewer than half of the questions on the questionnaire. It was decided that respondents with more than two missing values in one scale were not eligible for data imputation. Given the dichotomous nature of the questions on the Affect Balance Scale and the Self-Esteem Inventory, no imputations were made for missing values on these two

Table 2 Data set sample sizes and imputations for study instruments

Instrument	Number of respondents who viewed the web page containing a survey instrument	Number of surveys used in the analysis	Number of surveys requiring imputations
Affect Balance Scale	76	41	0*
Coopersmith Self-Esteem Inventory	44	28	0*
The General Health Rating Index	42	27	4
Nowotny Hope Scale	48	28	3
The Personal Resource Questionnaire	55	31	2
The Minneapolis-Manchester Quality of Life Instrument	35	26	5

*No imputations performed on missing data as questions in scale were dichotomous.

scales. Those survivors who had missing data for these two scales were eliminated from the analysis. However, for the General Health Rating Index, Nowotny Hope Index, Personal Resource Questionnaire and the Quality of Life Survey of Health, imputations for a missing value were made by imputing the average score of the rest of the respondents for that particular question. This method of data imputation is referred to as the mean substitution method, in which the mean value of the total sample for a variable is substituted for all the missing values for that variable (Saunders *et al.* 2006). Table 2 shows the total number of childhood cancer survivors who viewed each web page containing a survey instrument, the number of questionnaires used in the analysis, and the number questionnaires requiring of imputations.

Discussion

Collecting data via online surveys has several advantages, yet the experiences encountered in this study suggest that traditional methods of conducting research require careful review and adaptation to be relevant in web-based environments. The issues encountered in this study reflect Duffy's (2000, 2002) perspective that web-based research requires a refocusing among researchers with respect to research methodology. Specifically, considerations, as to how study participants are recruited and how data are collected, require a change in usual research methods.

Despite these considerations, online data collection has several advantages for nurse researchers. The anonymity that online surveys offer has positive implications for data quality. This anonymity is thought to foster a greater sense of confidence among participants to respond to sensitive questions more freely, and thus it decreases social response bias and researcher-influenced bias and thereby enhances the truthfulness of the data. Its ease of use in allowing questionnaires to be answered at home without requiring time to travel to another location complements expectations among many young adults to have access to information quickly and without significant effort. In addition, it offers researchers the opportunity to collect data without the constraints of geographical location and inaccessibility of study participants. Thus, data can include representation of participants around the world.

However, because of the anonymity online collection offers, it is nearly impossible to follow-up with individuals if data are missing. As experienced in this study, allowing completion of fields to be optional creates the strong possibility of missing data and gives little opportunity for follow-up. We therefore recommend that all fields be designated as required fields and that each questionnaire must be completed and submitted before a respondent can proceed to the next one. If respondents choose not to complete all items on one questionnaire, they terminate their participation in the study. These individuals would then be encouraged to contact the principal investigator to give feedback about the reasons why they did not respond to certain items. According to Hanscom *et al.* (2002), participants may chose not to respond to certain questions because they do not understand them or find some questions too sensitive. Unfortunately, this strategy was not incorporated into the protocol for our study.

Another possible explanation for the number of missing data may be related to the total number of questions to be answered. This was 167 in our study. While this was not an overwhelming number for the healthy young adults who pilot tested the online questionnaires, perhaps it was too many for the childhood cancer survivors. In addition, the nature of the questions may have been too thought-provoking and may have caused them to experience emotional distress; this, in turn, may have influenced many to leave numerous questions unanswered.

Posting notices of a study on websites that are frequently used by childhood cancer survivors theoretically appears to be a reasonable approach to gaining the desired sample size for a study. Despite having a live link posted on the homepage of six websites for easy access, the response rate

What is already known about this topic

- Online methodology is a feasible approach to conducting research because of its relative speed and low cost in data collection and ease of use for nurse researchers and study participants.
- With online data collection, there are few geographical limitations.
- If constructed with the proper controls, participant anonymity can be enhanced, thus decreasing social response set and researcher-influenced bias.

What this paper adds

- Web-based research requires review of traditional approaches and relevant adaptation to online environments.
- Considerations in web-based research include how the study will be advertized on established websites, whether data fields will be optional or not, and the total number of questions to be answered.
- Support is needed from a website administrator in advertizing the study to increase the response rate.

among childhood cancer survivors in our survey was very low. The study's link may have been overlooked or ignored by users of the site as it was posted among other links and information. Additionally, the reported time needed to collect data via the Internet is estimated to be typically 1 month (Batagelj & Vehovar 1998). Because of the poor response and extent of missing data, we decided to suspend data collection after 6 months. Our experiences suggest that posting a link on an established website for advertizement to enroll participants in a study may need to be accompanied by other opportunities to advertize the study; for example, an e-mail by the website administrator to subscribing patrons might promote increased interest and increase response rate.

Conclusion

Web-based research can be useful for nurse researchers worldwide in studying many different clinical as well as healthy populations with relative ease and without compromising rigour. Using the Internet to collect data on sensitive topics and with vulnerable populations offers the opportunity to capture data without the constraints of geographical location; however, posting the study on websites used by these people needs careful attention, and support from website administrators in advertizing the study to increase the response rate is encouraged. Having optional data fields within an instrument in online data collection methods should be given careful consideration with respect to how missing data will be addressed.

Author contributions

MAC was responsible for the study conception and design and the drafting of the manuscript. MAC performed the data collection and MAC and PL performed the data analysis. MAC obtained funding and provided administrative support. MAC made critical revisions to the paper. PL provided the statistical expertise. MAC supervised the study.

References

Ahern N.R. (2005) Using the Internet to conduct research. *Nurse Researcher* **13**(2), 55–69.

Batagelj Z. & Vehovar V. (1998) *Technical and Methodological Issues in WWW Surveys*. Retrieved from http://www.ris.org/ris98/stlouis on 20 March 2006.

Dalton W.S. (2005) The "total cancer care" concept: linking technology and health care. *Cancer Control* **12**(2), 140–141.

Dillman D.A. (2000) *Mail and Internet Surveys: The Tailored Design Method*. Wiley, New York.

Duffy M. (2000) Web-based research: an innovative method for nursing research. *Journal of Canadian Oncology Nursing* **10**(2), 45–49.

Duffy M.E. (2002) Methodological issues in web-based research. *Journal of Nursing Scholarship* **34**(10), 83–88.

Hanscom B., Lurie J.D., Homa K. & Weinstein J.N. (2002) Computerized questionnaires and the quality of survey data. *Spine* **27**(16), 1797–1801.

Hill W., Weinert C. & Cudney S. (2006) Influence of a computer intervention on the psychological status of chronically ill rural women: preliminary results. *Nursing Research* **55**(1), 34–42.

Hinds P.S., Burghen E.A., Haase J.E. & Phillips C.R. (2006) Advances in defining, conceptualizing, and measuring quality of life in pediatric patients with cancer. *Oncology Nursing Forum* **33**(Suppl. 1), 23–29.

Klein J. (2002) Issues surrounding the use of Internet for data collection. *American Journal of Occupational Health* **56**(3), 340–343.

Lakeman R (1997) Using the Internet for data collection in nursing research. *Computers in Nursing* **15**(5), 269–275.

McAlindon T., Formica M., Kabbara K., LaValley M. & Lehmer L. (2006) Conducting clinical trials over the internet: feasibility study. *BMJ* **327**, 484–487.

Saunders J.A., Morow-Howell N., Spitznagel E., Dore P., Proctor E.K. & Pescarino R. (2006) Imputing missing data: a comparison of methods for social work researchers. *Social Work Research* **30**(1), 19–31.

Thomas B., Stamler L.L., Lafreniere K. & Dumala R. (2000) The Internet: an effective tool for nursing research with women. *Computers in Nursing* **18**(1), 13–18.

