Obtaining Sensitive Data Through the Web: An Example of Design and Methods

Atar Baer,[1] *Stefan Saroiu,*[2] *and Laura A. Koutsky*[1]

Abstract: Several studies have suggested that the quality of coital data from diaries is superior to that collected by retrospective questionnaires. By collecting data over short intervals of time, diaries can present a more comprehensive picture of exposure, while minimizing the potential for recall bias. Despite these advantages, paper diaries have limited use because of their expense and difficulty of implementation. Web-based data collection offers the opportunity to make improvements to the quality of epidemiologic exposure measurement by providing privacy and convenience to study participants while reducing costs associated with questionnaire administration and allowing for real-time data processing. We adapted coital diaries for Web-based data collection in a study of transmission rates of genital human papillomavirus infection among young adults. University women complete an online sexual behavior questionnaire ("diary") every 2 weeks over a 3-year follow-up period; men complete a single online sexual behavior questionnaire ("journal"). In this paper we describe the design, methodology and implementation issues that emerge in conducting a Web-based epidemiologic study. We also discuss compliance, as well as methods for assuring appropriate security, confidentiality and privacy.
(EPIDEMIOLOGY 2002;13:640–645)

Key words: survey methodology, data collection, questionnaires, Internet, epidemiologic research design, computing methodologies.

Epidemiologic studies rely on valid, reliable and efficient methods of data collection. Systematic differences in soliciting, recording, or interpreting information from study participants can bias any type of epidemiologic study. Studies that collect information about sensitive topics, such as sexual exposure and illicit drug use, may be especially prone to bias if study participants give distorted accounts of their exposure histories to avoid embarrassment or to safeguard privacy.[1–3]

The increasing popularity and availability of the Internet offers researchers an opportunity to administer questionnaires published as web pages. This methodology has several advantages over traditional means of data collection. Computer-assisted interviewing techniques provide a greater degree of anonymity than face-to-face interviews and may yield more truthful answers to sensitive questions compared with either paper questionnaires or face-to-face interviews.[1–3] Web-based administration also offers study participants the comfort of completing questionnaires at a time and location that is convenient for them. Furthermore, Web-based questionnaires provide a means for data to be returned to the investigator rapidly, without the need for data entry and validation, thereby allowing for statistical processing in real time. Finally, Web-based questionnaires can eliminate costs associated with printing, mailing, keying and interviewing. Although Web-based data collection introduces some unique security challenges, tools are available for protecting study participants and the data they provide.

Our ongoing prospective cohort study was designed to estimate transmission rates of genital human papillomavirus (HPV) infection among young adults. In this paper we describe the design, methodology and implementation issues that emerge in conducting a Web-based epidemiologic study in a university population. We also discuss compliance with our study protocol among the female study participants, as well as methods of assuring security, confidentiality and privacy.

From the [1]Department of Epidemiology, School of Public Health and Community Medicine, and [2]Department of Computer Science and Engineering, University of Washington, Seattle WA.

Address correspondence to: Laura Koutsky, HPV Research Group, University of Washington, 1914 North 34th Street, Suite 300, Seattle, WA 98103; kouts@u.washington.edu.

Supported in part by the STD/AIDS Predoctoral and Postdoctoral Training Program (National Institute of Allergy and Infectious Diseases [NIAID] Grant AI0714P) and by NIAID Grant AI38383.

Submitted 24 July 2001; final version accepted 23 July 2002.

DOI: 10.1097/01.EDE.0000032360.92664.BC

Methods

Study Population

Two groups of study participants are involved in the Web-based component of the study: (1) 18- to 20-year-old female university students who either have never had vaginal intercourse or have had vaginal intercourse for the first time with only one partner within 4 months of their first study visit and (2) a random sample of male university students, representative of the population of potential male sex partners of the enrolled women. The anticipated sample size is 300 women and 500 men.

Sample Selection

Most participants are recruited via letters sent to names and addresses made available through the university registrar's office. Additional sources of participants include referrals from contraceptive counselors at the university health clinic, word of mouth, and advertisements posted on campus. The Institutional Review Board of the University of Washington approved the study protocol.

Those who call the research coordinator are asked specific questions to determine eligibility and ability to comply with examinations. Women who are not residents of Washington, are currently pregnant, have had a hysterectomy, are suffering from a serious medical condition, or are unable to provide informed consent are not eligible to participate. Men who have never had vaginal or anal intercourse, are younger than 18 years of age, have a serious medical condition, or are unable to provide informed consent are also ineligible.

Study Design

Women who consent to participate are traced for a period of 2 years, during which they complete a sexual behavior questionnaire ("diary") administered on our web site every 2 weeks. They are also examined by a clinician every 4 months. Men are asked to attend one clinic visit at which they have a physical examination and complete a single Web-based sexual behavior questionnaire. A face-to-face medical and sexual history interview is administered by the nurse practitioner at the start of every clinic visit. We collected information using standard forms that have been extensively tested in studies of HPV infection. Women are compensated $25 per study visit ($75 per year), and men are compensated $75 for their single visit.

Throughout the duration of their study participation, women receive an automatic e-mail message every other Monday reminding them to log into our web site to complete a diary. To protect confidentiality, these e-mails do not contain private information, and e-mail recipients cannot see who else is on the mailing list. The research coordinator receives a weekly report listing those women who have submitted a diary, so that compliance with the study can be monitored. Women who have missed a diary entry are asked to complete a make-up diary for every week that was missed (these make-up diaries cover 1-week periods, whereas the usual diaries cover 2-week periods). Women are paid $2 for every diary that they submit online, up to $4 per month.

Diary Design

We adapted a questionnaire about sexual behaviors, which we had used previously for face-to-face interviewing, for the online diary. Several published recommendations were followed for designing the Web-based questions.[4–6] The diaries were written in HTML using a combination of an authoring tool (Microsoft Front Page Express) and Microsoft Windows WordPad. Diaries were designed to accommodate direct entry of data through the selection of check boxes and radio buttons, pulldown menus and text boxes, making interaction with the Web simple even for the novice user (Figure 1). We pilot-tested the diaries among the research study staff using a variety of browsers (*eg*, versions of Microsoft Internet Explorer and Netscape), screen sizes, operating systems (Windows and Macintosh) and connections (*eg*, Ethernet and modem dial-up).

To access our web site, study participants are required to use one of several browsers available at all University of Washington computing labs and by free download. Participants are given a page of instructions on how to download the browser and how to set the required preferences.

Participants complete their first (and, for men, only) online diary while at the clinic, so that problems relating to usage of the site can be resolved. The research assistant provides an orientation to the system and then remains available but does not directly observe the respondent during questionnaire completion. Later, the female participants have the option of connecting to the site from on or off campus.

Men and women have separate web sites with distinct URLs. The index (start) page of each site introduces the study and provides the research coordinator's contact information. By clicking on hypertext key words (to skip from one point of the web site to another), participants are able to (1) complete their "diary" (women) or "journal" (men), (2) learn more about HPV infection, or (3) read a privacy statement. Users are able to scroll up and down a page or jump to a previous page using the mouse or keyboard.

Depending on frequency of sexual activity, the questionnaires take between 2 and 15 minutes to complete. For female study participants, three versions of the diary are available (first-time, follow-up and make-up). Participants are asked about sexual behaviors for each day over a 2-week period, with questions about each sex

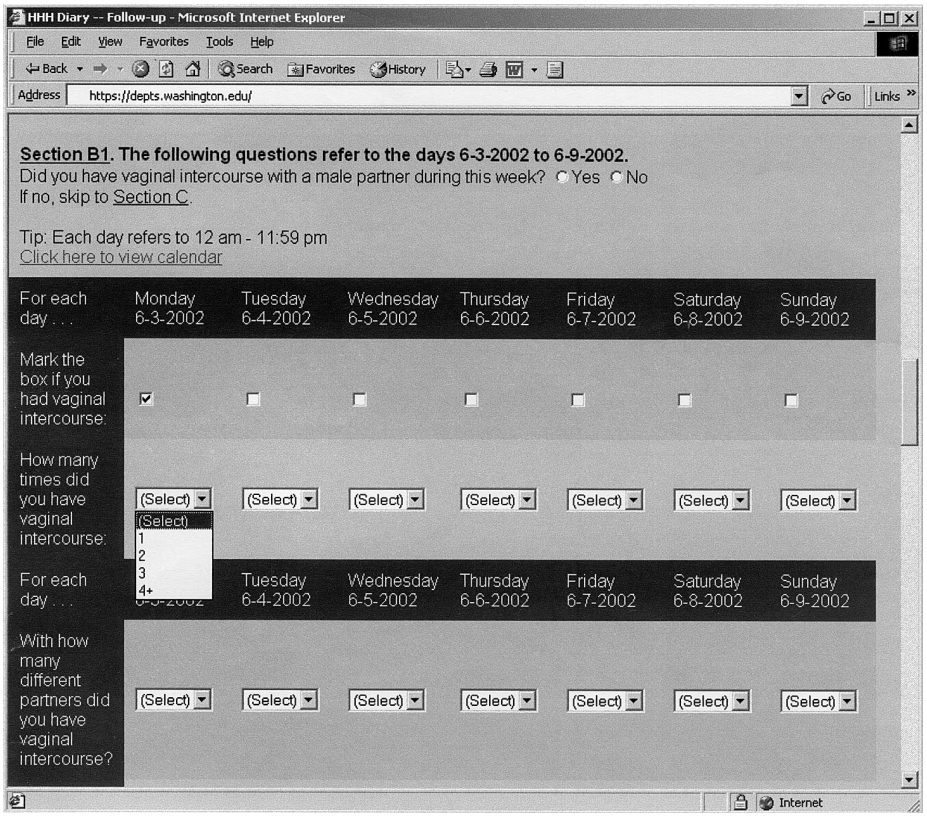

FIGURE 1. A section of the women's diary. Respondents answer questions by selecting radio buttons, check boxes, and options from pull-down menus. Dates are updated automatically every week using Javascript.

partner. The dates corresponding to each day are automatically updated using Javascript. Every diary and journal within the site is contained within a single page and arranged in a table, with each column corresponding to a day of the week (or alternately, a sex partner) and each row representing a question (Figure 1). This format ensures consistency of layout for different browsers, operating systems and window sizes. Using Javascript, fields are dynamically validated at the time of data entry, and skip patterns are in place to minimize user errors. Respondents who complete a field incorrectly are informed before moving to the next question. Validation includes checks for data length, data type and data range. To assure the right of participants to skip questions they are not comfortable answering, incomplete or blank responses are allowed. To protect their privacy, users are instructed to exit all Web browser sessions and windows when they finish.

Security

We were concerned with ensuring that our web site would provide appropriate security, confidentiality and privacy. We identified four security goals: authentication, patient confidentiality, data privacy and data integrity. The first goal, authentication, refers to having the system validate a user's identity, thereby preventing attackers from impersonating legitimate participants. The second goal, confidentiality, ensures that the identity of a participant is revealed only to the system administrators. The third goal, data privacy, guarantees that data entered by a participant can be accessed only by the system administrators. Participants cannot review their previously entered records, which eliminates the danger of a "password leak," in which an attacker could impersonate a participant in the system and peruse previously entered records. The final security goal, data integrity, ensures that no third party can alter a record entered by a user.

At the University of Washington, each student receives an e-mail account on the university web site along with a login name and a password. The student uses this name/password combination to authenticate to the web site. The authentication scheme is Kerberos-based[7] and certified by the Thawte Certification Authority (http://www.thawte.com). The web site is hosted by the university web server and is continuously maintained and monitored by a professional staff.

We leveraged this mechanism to authenticate the participants in our system. Individuals provide their login ID on their consent form when they agree to participate in the study. This ID is added to a list of individuals who have been authorized to gain access to the site. Upon loading our web site's URL, a user is redirected to the university site, where authentication takes place. If successful, the university site redirects the user to our site and passes an encrypted record of the user's credentials in the form of a "cookie," which proves the user's authenticated credentials. Once the authentication is complete, our web site verifies whether a user is authorized to gain access to our site. If the login ID and encrypted credentials do not match the identity of an enrolled patient, an error page is presented.

The benefits of using the university authentication mechanism are two-fold. First, the participants do not

need to remember additional passwords. Second, our site's authentication mechanism is as strong as the university's, which is professionally designed and continuously monitored and maintained.

The server that currently hosts the study web site is the same physical server that hosts all the official university department web pages. Accounts on this server are limited to faculty and staff, and the server is continuously serviced and monitored by professional staff. All remote access to this server must use encrypted authentication. In particular, regular versions of Telnet and file transfer protocols that use plain-text authentication cannot be used. The high level of security offered by the university official web server greatly reduces the likelihood of server intrusion.

Every communication between the web server and the web site users occurs over a connection with secure socket layer (SSL version 3.0) enabled. SSL is a standard cryptographic protocol for secure Web communication; data entered by a participant are encrypted on the user's computer and then transferred to the web server. In general, SSL connections can be "server-sided" only, in which the server authenticates itself to the user's browser, or "double-sided," in which both the server and the user authenticate each other. Because user authentication is separately established through the university web site, our web server uses server-sided SSL connections.

SSL defeats most attempts to eavesdrop or to forge or otherwise tamper with data while in transit. However, SSL does not prevent authenticated users from acting maliciously. Security breaches can occur if an authenticated user enters unexpected input, which can "crash" the web site application or result in unauthorized access. Our web site application uses a simple approach to deal with this scenario. Only questionnaire entries are accepted as user input; the data are never parsed or interpreted, but instead are saved to stable storage (*ie*, files on hard disk) for later analysis.

A common gateway interface (CGI) script (written in Perl language) collects the entered data ignoring any personal identifiers. It then assigns a unique identification number to the form input. Before being saved to a flat-text, delimited file, data are encrypted using the "Pretty Good Privacy" (PGP version 2.6.2) application based on public-key cryptography. Data are encrypted using a public key, which can be known by anyone, and are decrypted by a private key, which is known only to the key owner. Without the private key, it is computationally infeasible to decrypt an encrypted message. Our application uses a standard secure technique (Unix-based pipes) to transfer unencrypted data to the PGP application, which encrypts it before writing it to a file. We use a 1,024-bit key to encrypt the data (1,024 bits should be an appropriate key length for ensuring several decades of security). The web server stores only the public key and has no knowledge about the private key.

Once a week, the encrypted data are transferred to a local computer where they are decrypted for analysis. Data are never decrypted while the local computer is online. The private key is stored on a floppy disk, which is inserted in the floppy drive only after the computer is disconnected from the network. This encryption/decryption mechanism offers a high degree of data privacy and integrity. Even in the extremely unlikely event of a web server intrusion, the attacker has no possibility of decrypting the previously collected data or of undetectably altering it. In addition to the continuous monitoring by the server's technical staff, we also monitor our site for suspicious activity by gathering log entries. The web site administrator reviews these logs weekly using the wwwstat program[8] to check for suspicious activity, such as repeated unsuccessful attempts to access a password-protected document.

Only the web site administrator has access to the password of the web server account on which the data are stored, to the private PGP key for data encryption, and to the key to the room where data are stored on a floppy disk. The account password is changed on a monthly basis, or at the first sign of suspicious activity. Together with SSL, PGP, and removal of personal identifiers from transmitted data, these security measures attempt to ensure that the data collected in the server's database are as secure as databases kept on individual client office computers.

Results

We began Web-based diary collection in December 2000 and have results available through the end of January 2002. This report focuses on the compliance of the female study participants. All female participants during this period were between the ages of 18 and 22 years (40% were 18 years of age). The majority (61%) were white, 26% were Asian, 2% were African-American, 1% were Hispanic, and 9% were of other race/ethnicity. Of the 85 female study participants, 71 were virgins at enrollment. Entry into the study was staggered over the 62-week period, as not all study participants were enrolled on the same date. Some women chose to submit a diary more frequently than at the scheduled 2-week intervals; these extra diaries (N = 125) were retained in the database. An additional 47 diaries from 24 women were deleted because they were identified as duplicate entries (*ie*, submitted for the same dates) or they were make-up diaries that could not be matched to the dates of a missing entry.

Compliance

If all post-enrollment diaries were submitted on time (*ie*, one diary entry submitted by each woman every other week), we would have expected to receive 1,014 post-enrollment diaries (covering a total of 2,028 weeks). Of the diaries expected, 829 (82%) were submitted on time, covering a total of 1,658 weeks. Of the remaining 370 weeks not covered on time (185 diaries), 83 weeks were covered by diaries submitted out of sequence, and 287 weeks were missed. Of these 287 missing weeks, 206 (72%) were made up. Including diaries submitted on time, out of sequence, or as make-up entries, 96% of the 2,028 expected weeks of coverage were received; the total percentage of weeks covered by diaries submitted on time or out of sequence was 86%, and the total percentage of weeks covered by make-up diaries was 10%.

The frequency of missing data for selected questions in the women's diaries is presented in the table. For items relating to first partner's demographic characteristics as well as frequency of intercourse and condom use, only one question (asking women to provide partner initials) had more than 10% missing data.

Excluding make-up entries, the average interval between post-enrollment diary submissions was 14.3 days. Of the diaries received on time, the majority (66%) were submitted on a Monday (N = 548/829), the day on which our reminders were e-mailed to study participants, and 17% (N = 141) were submitted on a Tuesday. The mean time between entry of a make-up diary and the first day of the week it was supposed to cover was 56 days.

We did not find an association between length of study enrollment and whether or not a diary entry was skipped (odds ratio = 1.0; 95% confidence interval = 0.99–1.01). Of the total number of diary entries that were initially missing (N = 287), we have received 218 (76.0%) make-up diaries.

Frequency of Missing Data Among Women's Diary Entries

Item	Missing N	Missing %
If vaginal intercourse was reported for a given week (N = 153), for each day of the week:		
Marked whether or not vaginal intercourse occurred	1	<1
Recorded number of times of vaginal intercourse	2	1
Recorded number of different partners for vaginal intercourse	14	9
Provided initials of each partner	17	11
Reported whether or not a condom was used for vaginal intercourse	5	3
If a condom was used for vaginal intercourse during a given week (N = 157), for each day of the week:		
Recorded number of times condom was used for vaginal intercourse	1	<1
Reported whether or not partner's penis ever touched opening of vagina without a condom	2	1
Among nonvirgins (N = 23), reported:		
Initials of first partner	2	9
Age of first partner	1	4
Circumcision status of first partner	1	4
Sexually transmitted disease history (yes/no/don't know) of first partner	1	4
Condom use with first partner (yes/no)	1	4

Discussion

By collecting prospective data at set short intervals of time, diaries can present a comprehensive picture of exposure, while minimizing the potential for recall bias. Diaries are especially useful in studies examining sexual behavior, in which exposure measurement using biomarkers or direct observation may not be possible. Several studies have suggested that the quality of coital diary data is superior to that collected by retrospective questionnaires.[9,10] Despite their advantages, paper diaries have limited use as a data collection method because of their expense and difficulty of implementation.[11]

Evidence suggests that electronic diaries could be a valid, viable alternative to paper diaries. First, studies have reported that compliance with electronic diaries appears to be as good as, if not better than, that achieved by paper diaries.[12–14] Compliance in our study was excellent: 82% of all entries were submitted within 1 week of receiving the automatic reminder by e-mail, and 96% of diary entries were complete after make-up diaries were submitted. We did not find an association between length of study enrollment and whether or not a diary entry was skipped. Second, computer-based interviews have been judged by respondents to be more private than face-to-face interviews.[1] Several studies have reported that computer-assisted interviews yield more valid measures of sensitive or stigmatized behaviors (*eg*, for reporting of stigmatized sexual behaviors[3] and HIV-related risk behaviors[1]) compared with paper-and-pencil or face-to-face interviews.

Another advantage is that study participants can complete questionnaires at a time and location that is comfortable for them. This arrangement also frees up time and potentially reduces costs for investigators, who would otherwise schedule appointments and hire interviewers to administer questionnaires. Moreover, Web-based diaries can be submitted with a time stamp, whereas there are no guarantees that participants completing paper diaries actually do so in a timely manner.

Despite the advantages of computer-assisted interviewing techniques, methodologic and technical issues unique to the Internet can present obstacles. Resources are widely available on the Internet and in the published literature to assist with the design of a Web-based data collection system; hiring a professional programmer or using established

services for Web interviewing are alternatives for those investigators who do not wish to design a system on their own. Regardless of the method chosen to develop a system, careful consideration must be given to ensure the security and confidentiality of data and the privacy of study participants. Also, investigators must be mindful of selection bias when choosing a study population, as the current online population is not necessarily representative of the general population.[4]

In conclusion, Web-based data collection may offer an opportunity to improve the quality of exposure measurement in epidemiology studies. However, use of the Internet for data collection poses some unique challenges. Methods for ensuring the reliability, security and accuracy of data captured by Web-based questionnaires have improved substantially in recent years. As more epidemiologists become familiar with the capabilities of Web-based data collection and as access to the Internet increases among the general population, we anticipate that this technology will play an increasingly prominent role in addressing questions of public health importance.

Acknowledgments

We thank Sandra O'Reilly (study coordinator) and Bethany Weaver for assistance with the design of the Web-based questionnaires. For the face-to-face questionnaires, we thank Diane Adam and Ellen Cassen for interviewing, Debi Hertel for data entry, and Shu-Kuang Lee for data management. Finally, we thank Steve Gribble for providing feedback and suggestions for improving the security of our Web-based data collection.

References

1. Locke SE, Kowaloff HB, Hoff RG, *et al*. Computer-based interview for screening blood donors for risk of HIV transmission. *JAMA* 1992;268:1301–1305.
2. Navaline HA, Snider EC, Petro CJ, *et al*. An automated version of the risk assessment battery (RAB): enhancing the assessment of risk behaviors. *AIDS Res Hum Retroviruses* 1994;10:S281–S283.
3. Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL. Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. *Science* 1998;280:867–873.
4. Dillman DA, Tortora RD, Bowker D. Principles for constructing web surveys. 1998. Available at: http://survey.sesrc.wsu.edu/dillman/papers.htm. Accessed 4/19/2001.
5. Dillman DA, Bowker DK. The web questionnaire challenge to survey methodologists. 2001. Available at: http://survey.sesrc.wsu.edu/dillman/papers.htm. Accessed 4/29/2001.
6. Marshall WW, Haley RW. Use of a secure Internet Web site for collaborative medical research. *JAMA* 2000;284:1843–1849.
7. Neuman BC, Tso T. Kerberos: an authentication service for computer networks. *IEEE Commun* 1994;32:33–38.
8. wwwstat. HTTPd Logfile Analysis Software. Available at: http://www.ics.uci.edu/pub/websoft/wwwstat/. Accessed April 9, 2001.
9. Coxon APM. Parallel accounts? Discrepancies between self-report (diary) and recall (questionnaire) measures of the same sexual behaviour. *AIDS Care* 1999;11:221–234.
10. Leigh BC, Gillmore MR, Morrison DM. Comparison of diary and retrospective measures for recording alcohol consumption and sexual activity. *J Clin Epidemiol* 1998;51:119–127.
11. Armstrong BK, White E, Saracci R. *Principles of Exposure Measurement in Epidemiology*. New York: Oxford University Press, 1992.
12. Jamison RN, Raymond SA, Levine JG, Slawsby EA, Nedeljkovic SS, Katz NP. Electronic diaries for monitoring chronic pain: 1-year validation study. *Pain* 2001;91:277–285.
13. Johannes C, Woods J, Crawford S, Cochran H, Tran D, Schuth B. Electronic *vs* paper instruments for daily data collection. *Ann Epidemiol* 2000;10:457.
14. Rabin JM, McNett J, Badlani GH. A computerized voiding diary. *J Reprod Med* 1996;41:801–806.