

3.4 Security, privacy, and confidentiality issues

127

Grant Kelly and Bruce McKenzie

In this chapter we introduce the issues around protecting information about patients and related data sent via the Internet. We begin by reviewing three concepts necessary to any discussion about data security in a healthcare environment: privacy, confidentiality, and consent.

Privacy

'Privacy' is a vaguely defined term that, in an online context, includes the right of an individual to:

- Determine what information is collected about them and how it is used. Sometimes we are not aware what data are being collected about us (e.g. via 'cookies' on a Web site—see Glossary) or how it may be used. Registering with a Web site (i.e. giving your name, e-mail address, medical registration number, etc.), for example, may enable that site to keep track of what you—a readily identifiable individual—view or spend online. Such information could be passed on to third parties. Some sites publish 'privacy policies' in an attempt to inform users and reduce the chances of patients or healthcare professionals placing their privacy at risk.
- Access information held about them and know that it is accurate and safe.
- Anonymity (e.g. not having your Web-browsing habits tracked).
- Send and receive e-mail messages or other data (e.g. credit card numbers) that will not be intercepted or read by persons other than the intended recipient(s). Encryption (discussed below) is one way of ensuring this.

For more information about privacy on the Internet, see Box 1.

Box 1 Privacy resources on the Internet

Platform for Privacy Preferences Project (W3C):

<http://www.w3.org/P3P/>

Understanding security and privacy (Netscape):

<http://home.netscape.com/security/basics/>

Privacy and security fundamentals (Microsoft):

<http://www.microsoft.com/privacy/safeinternet/>

e-Health Code of Ethics (Internet Healthcare Coalition):

<http://www.ihealthcoalition.org/ethics/ehcode.html>

Statutory and professional considerations**Confidentiality**

The ethical duty of confidentiality is defined by the British Medical Association as 'the principle of keeping secure and secret from others, information given by or about an individual in the course of a professional relationship' [1]. In the UK the legal duty of confidentiality is underpinned by the Data Protection Act (1998), regulating the processing of information ('data') that could lead to the identification of individuals—including its collection, storage, and disclosure [2]. To ensure the protection of confidentiality in an electronic environment the General Medical Council (GMC) recommends that doctors should [3]:

- Make appropriate security arrangements for the storage and transmission of personal information.
- Obtain and record professional advice given prior to connecting to a network.
- Ensure that equipment, such as computers, is in a secure area.
- Note that Internet e-mail can be intercepted.

Consent

'Consent' for our purposes is the means by which we are authorized by an individual to process information about them based on their informed understanding of what we intend. To include identifiable patient information in an e-mail message or on a Web site in the absence of a patient's express consent would constitute a breach of confidentiality. Obtaining consent should involve making the patient aware of any risks to his or her privacy and the arrangements in place to protect it. Identifiable patient

information could therefore be transmitted via the Internet with the informed consent of the patient, and with regard for the advice of the GMC (or equivalent professional body) and established principles such as those of Caldicott (see Box 2) and the Data Protection Act (see Box 3).

Box 2 Caldicott Principles

In relation to identifiable patient information:

- Justify the purpose(s) for using confidential information.
- Only use it when absolutely necessary.
- Use the minimum that is required.
- Access should be on a strict need-to-know basis.
- Everyone must understand their responsibilities.
- Understand and comply with the law.

For further information, see:

<http://www.doh.gov.uk/nhsexipu/confiden/report/index.htm>

Box 3 Data Protection Act Principles

Personal data must be:

- fairly and lawfully processed
- processed for limited purposes
- adequate, relevant, and not excessive
- accurate
- kept for no longer than necessary
- processed in accordance with the data subject's rights
- secure
- not transferred to countries without adequate protection.

For further information, see:

<http://www.hms.gov.uk/acts/acts1998/19980029.htm>

Information that cannot result in identification of an individual may have been 'anonymized' (where identifiers are removed) or 'aggregated' (where data from a number of individuals are summed). The requirement for consent to transmit or place such information online in this event is less certain, but perhaps prudent, although such non-personal data are not subject to legal restriction (i.e. the Data Protection Act).

Where is the enemy?

Security tends to be the progeny of scandal. A few years ago, a bank in the Midwest USA purchased a hospital along with its medical records. It coolly compared the records against its personal bank accounts, and foreclosed on the loans of all account holders with a diagnosis of cancer. It was business-like, simple, ignorant, cruel, and an example of the damage that medical data can do in the wrong hands. Today computer 'security' is typically perceived to mean keeping hackers (those attempting unauthorized computer access) and other troublemakers from your private data. But what if such troublemakers are part of the system, or even own it?

Clearly, a simple 'cops and robbers' model does not offer enough protection, highlighting the need to ensure data security at multiple levels. The risks are internal, external, and random, and can result in data damage, falsification, loss, or leakage. It is helpful to imagine your connected system as resembling a data stream right from your keyboard to that of the recipient, and to consider the risks along the way.

Protecting local data

Even before you connect, your data is at risk. Clearly you don't want your Internet-linked clinical system or home computer to be burnt, flooded, stolen, hit by lightning, damaged by third party software, or accessed by untrained staff or inappropriate people. You will need to back it up properly, look after the backups, and periodically reconstitute the system from backups so that you know it will work if you ever need it.

Ensure that your terminal or PC is left logged out when you are apart from it for a reasonable length of time. Most systems can be set to log out automatically by default under these circumstances and this makes good sense. Make sure that your screen shows information only to people who are entitled to see it.

If you connect to the Internet at work (e.g. via NHSnet—p. 16) you may wish to ensure that your e-mail server (p. 9) has central control over a shared address book, with limited access rights to alter it and to reply to external addresses. Doing so prevents staff from using e-mail at work to converse with friends—which not only reduces working efficiency, but also provides a means of access for viruses (see below) and other unwelcome material.

Appropriate advice and countermeasures are detailed elsewhere [4,5], enabling you to develop robust protocols to preserve the integrity of your local system. Further NHS-specific guidance is available from the NHS Information Authority Web site:

<http://www.standards.nhsia.nhs.uk/sdp/>

The risks of connecting

Open systems: the Internet

Linking computers together means that you can access other people's data, but it inevitably follows that this allows others to access data on your own system. Until such time as individual computers or networks are linked together they resemble 'islands' of electronic data. Security on a data island is simple: reassuringly firm borders trap all unauthorized entrants. However, when you build bridges by creating a network link this approach on its own is inadequate. When a computer connects to the Internet, it loses its island status by compromising the integrity of its 'borders'. Any potential benefits of connecting must be weighed against the risks to your own data. In a healthcare environment, this data is often of a highly sensitive nature. Even connecting a home computer may expose data, such as banking details, which you would prefer to remain private.

Closed systems: the intranet

Why connect in such an open way? Why not restrict the connection to 'friends' only? In other words, why don't we connect only to trusted computers over trusted network links, thus extending our own trusted computing base? Enter the intranet, introduced in Chapter 1.2. Intranets are suited to smaller organizations with enforced security policies and strict personnel control—something not always attainable within a large health service. They are by nature restrictive, as security through exclusion conflicts with the potential of a network to enhance medical communications in a connected world. Intranets may provide a false sense of security: as the electronic thief attacks the weakest link in the chain, security measures must reflect this. A properly secured intranet therefore demands such things as locked rooms for terminals, physiological checks for terminal access, and armoured, pressurized cables to detect cable tapping.

Virtual private networks

Blurring the divide between public and private networks, a virtual private network (VPN) uses a 'tunnelling protocol' and encryption (see below) to

send private data through public networks such as the Internet. Although communicating parties do not need to invest in a private network infrastructure, they have no control over the network used and no guaranteed standard of service. The lack of interoperable implementations has been the main impediment to the deployment of VPNs to date [6].

Firewalls

Just as you wouldn't allow anybody to listen in to your telephone conversation, so you need to care for your Web browsing sessions and e-mail exchanges. For this purpose you need a firewall, designed to prevent damage to your system. These software or hardware devices operate by recognizing the IP address (p. 10) that a message or system query comes from, and only allowing past those that are recognized as 'good' or trusted. With the advent of higher-risk 'always on' Internet connections, firewall solutions of varying complexity are readily obtainable.

Protecting data in transit

Whether you are connected to NHSnet or the Internet the security threats to your data in transit are the same; data may be subject to loss, late delivery, damage, or attack. Against loss or lateness, there is little the individual can do, but damage or attack can be dealt with. You should assume the wires (or other network infrastructure) could be got at—as indeed they can—and thus must give your data a metaphorical envelope to maintain its integrity and privacy. This is precisely what cryptography can do.

Message encryption

A popular technique for protecting messages in transit is so-called asymmetric public-key infrastructure (PKI) cryptography. Alice and Bob (who wish to exchange messages) each use an algorithm based on very large prime numbers to develop two separate but related numbers, by way of typing in a pass-phrase. Both end up with an alphanumeric code that forms their 'public' key (which they publish), and an alphanumeric code that forms their 'private' key (known only to themselves and represented by their pass-phrase). If Alice wishes to send a message to Bob, she finds his public key (typically from a directory), writes her message, and encrypts (addresses) the data to Bob's public key, thus producing a unique set of digital data. Bob receives this in encrypted form and uses his private key to extract the data back into Alice's original text message. This process is illustrated in Fig. 1.

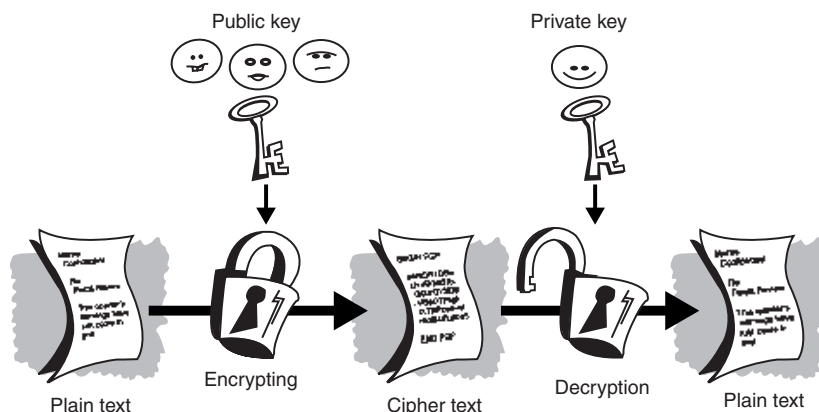


Fig. 1 Using a public/private key pair to encrypt messages helps ensure protection during transit.

In use, this is easier than it sounds, and confers integrity (the data haven't been manipulated), authenticity (the identity of the sender is known), non-repudiation (the data can't be disowned) and privacy on the data. Any attempt to interfere or damage the contents messes up the mathematics, and the message becomes unintelligible, thus warning the recipient not to trust it. Provided the verification of the identity of the key-holders is carried out in a dictatorial fashion, the origin authentication of the message is also assured. If only Alice knows the private phrase key to make an exchange work, then only Alice can have sent the message.

Authentication and privacy of e-mail via encryption is offered by Pretty Good Privacy (PGP) and Secure Multipurpose Internet Mail Extensions (S/MIME), both proposed Internet standards.

Pretty Good Privacy (PGPi Project):

<http://www.pgpi.org/>

S/MIME (RSA Security Inc.):

<http://www.rsasecurity.com/standards/smime/>

Browser encryption

As we move towards a browser-accessible type of electronic patient record there will arise a need to protect the exchange of data from leakage and attack. A precedent has been set by the widespread practice of Internet banking and commerce, which out of necessity involves transmitting confidential information. The *de facto* Internet standard for encrypting Web-based information interchanges is Secure Sockets Layer (SSL), more

recently known as Transport Layer Security or TLS [7]. SSL/TLS can also be used to encrypt e-mail messages. It uses a symmetrical one-time electronic key that works between the browser and the server for as long as the connection is open. When the session ends, the encryption dies with it, and thus it depends largely on its length of key structure and short time of operation for its safety. SSL/TLS is more demanding on server resources than non-encrypted connections, so secured Web pages are often slow to display.

Assurance of identity (authentication) on the Web presently requires the use of a certificate supplied by a third party Certificate Authority, such as VeriSign Inc.:

<http://www.verisign.com/>

UK readers should note that the NHS has its own cryptography strategy:

<http://www.doh.gov.uk/nhsexipu/strategy/crypto/index.htm>

Receiving data

Digital signatures

There is a simpler PKI process using the same algorithms referred to above to 'sign' a message whereby the private key of an individual can be used to 'hash' the message. This can then be verified against the sender's public key. This ensures the data's authenticity and origin without conferring privacy, and is called a 'digital signature'. The process is illustrated in Fig. 2. In the UK the Electronic Communications Act 2000 provides the legal framework for the recognition of digital signatures [8].

What about viruses?

Viruses are small segments of code that have been inserted into computer files, often with malicious intent. An infected file may cause annoyance or the loss of data. In theory, any file you download from the Internet is a potential vector. Viruses may also be present in files attached to e-mail messages (but cannot be transmitted via a text-only e-mail itself). There are a number of antiviral programs available (some are free) that will screen for and help you neutralize infected files on your computer—before they are activated or have a chance to 'replicate'. Some viruses are activated when you use an infected program; others merely require you to view an infected document. Antiviral programs act like the body's immune system in that they are always on the lookout for 'foreign' material—in this case, foreign program code. However, even if your software is regularly

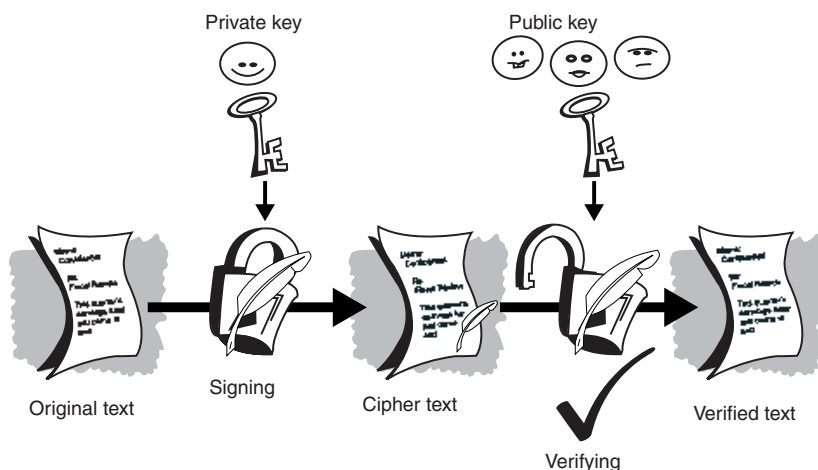


Fig. 2 Using a public/private key pair to verify a digital signature.

updated it won't catch all viruses (especially new ones). Security should be based on the sound sense of not opening e-mails from unknown sources or those containing unusual message headers.

Conclusion

The protection of personal data in a connected world defaults not so much to high-tech applications or hardware, as to careful management of staff and relatively common techniques to ensure the simple, frequent risks are catered for. The determined criminal or government agency will get access somehow, but what matters to doctors is making sure that we take care of the data we collect about patients in a manner appropriate to the twenty-first century.

References

1. British Medical Association (UK). Confidentiality and disclosure of health information. 1999 October [cited 2001 Apr 19]. Available from: URL: <http://web.bma.org.uk/public/ethics.nsf/webguidelinesvw?openview>
2. Her Majesty's Stationery Office (UK). The Data Protection Act (1998). 1998 [cited 2001 Apr 19]. Available from: URL: <http://www.hmsso.gov.uk/acts/acts1998/19980029.htm>
3. General Medical Council (UK). Confidentiality: Protecting and Providing Information. 2000 September [cited 2001 Apr 19]. Available from: URL: <http://www.gmc-uk.org/standards/secret.htm>

4. NHS Executive's Security and Data Protection Programme. Ensuring security and confidentiality in NHS organisations (E5501 v1.1). 1999 [cited 2001 Sep 22]. Available from: URL: <http://194.101.83.13/library/cards/c0000365.htm>
5. British Standards Institution (UK). *BS ISO/IEC 17799:2000 (BS 7799-1:2000) Information technology: code of practice for information security management*. London: BSI; 2000. Available from: URL: <http://www.bsi-global.com/>
6. Gleeson B, Lin A, Heenan J, Armitage G, Malis A. A framework for IP based virtual private networks (RFC 2764). 2000 Feb [cited 2001 Jun 5]. Available from: URL: <http://www.rfc-editor.org/>
7. Dierks T, Allen C. The TLS protocol (RFC 2246). 1999 Jan [cited 2001 Jun 5]. Available from: URL: <http://www.rfc-editor.org/>
8. Her Majesty's Stationery Office (UK). The Electronic Communications Act (2000). 2000 [cited 2001 Jun 5]. Available from: URL: <http://www.hmso.gov.uk/acts/acts2000/20000007.htm>

6.3 Facilitating research

Gunther Eysenbach and Jeremy Wyatt

This chapter concerns the use of the Internet in the research process, from identifying research issues, through using the Web for surveys and clinical trials, to pre-publishing and publishing research results. Although literature searches using databases such as MEDLINE are obviously an important and integral part of every research process, this is considered in Chapter 6.1.

Identifying issues for qualitative research

As the most comprehensive archive of written material representing our world and people's opinions, concerns, and desires (in industrialized countries), the Internet can be used to identify 'issues' for qualitative (descriptive) research and to generate hypotheses. Material published on the Internet may be a valuable resource for researchers desiring to understand people and the social and cultural contexts within which they live—outside of experimental settings—with due emphasis on the interpretations, experiences, and views of 'real world' people. Reviews of information posted by consumers on the Internet may help to identify health beliefs, common topics, motives, information, and emotional needs of patients, and point to areas where research is needed. Comparing recommendations found on the Web against evidence-based guidelines is one way to identify areas where there is a gap between opinion and evidence, or where there is a need for clinical innovation.

The accessibility of information for analysis and the anonymity of the Internet allow researchers to analyse text and narratives on Web sites, to use newsgroups as global focus groups, and to conduct interviews and surveys via e-mail, chat rooms, Web sites, or newsgroups. Topics suited to qualitative research include:

- Analysis of interactive communications (e.g. e-mail).
- Study of online communities (virtual self-help groups, newsgroups, mailing lists).

- Investigation of communication processes between patients and professionals.
- Study of consumer preferences, patient concerns, and information needs.
- Exploration of the 'epidemiology of health information' on the Web [1,2].

The Internet population is unrepresentative of the general population, restricting the use of the Internet for quantitative studies (i.e. studies focusing on measurement). Qualitative studies, however, do not require representative samples: 'In qualitative research we are not interested in an average view of a patient population, but want to gain an in-depth understanding of the experience of particular individuals or groups; we should therefore deliberately seek out individuals or groups who fit the bill' [3]. Three different research methodologies for qualitative research on the Internet may be distinguished:

- **Passive analysis:** For example, studying information on Web sites or interactions in newsgroups, mailing lists, and chat rooms—without researchers actively involving themselves.
- **Active analysis:** Also called participant observation; the researcher participates in the communication process, often without disclosing their identity as researcher. For example, they may ask questions in a patient discussion group implying that she or he is a fellow patient. Such studies often involve elements of deception, unless the researcher is a sufferer him- or herself.
- **Interviews and surveys:** See below.

Examples of these three types of qualitative research on the Internet are available elsewhere [1].

Using the Internet for surveys

Using the Internet for surveys requires an awareness of methodologies, selection bias, and technical issues.

Methodological issues

Internet-based surveys may be conducted by means of interactive interviews or by questionnaires designed for self-completion. Electronic one-to-one interviews can be conducted via e-mail or using chat rooms. Questionnaires can be administered by e-mail (e.g. using mailing lists), by posting to newsgroups, and on the Web using fill-in forms.

When e-mail is used to administer questionnaires, messages are usually sent to a selected group with a known number of participants, thus allowing calculation of the response rate. Surveys posted to newsgroups may request that the completed questionnaire is posted back to the researcher, but it is impossible to know who and how many people read the questionnaire. If Web-based forms are used, questionnaires can be placed in a password-protected area of a Web site (i.e. participation by invitation or registration only), or alternatively they may be open to the public (i.e. any site visitor can complete the survey). The latter option makes calculation of a response rate more difficult but not impossible: the number of people who access (without necessarily completing) the questionnaire is counted and used as the denominator. Web-based surveys have the advantage that the respondent can remain anonymous (as opposed to e-mail surveys, where the e-mail address of the responder is revealed). Furthermore, they are very convenient for the researcher, as responses can be directly stored in a database where they are immediately accessible for analysis.

Electronic interviews and surveys ('e-surveys') are emerging scientific research methodologies, pioneered by communication scientists, sociologists, and psychologists, although their use for health-related research is still in its infancy [4–10]. Examples of health-related research include:

- A Web-based survey on the effects of ulcerative colitis on quality of life [11].
- Collection of clinical data from atopy patients [12].
- A Web-based survey looking at complementary and alternative medicine use by patients with inflammatory bowel disease and Internet access [13].
- A survey of dentists regarding the usefulness of the Internet in supporting patient care [14,15].

E-surveys may be part of a qualitative research process, but results can be analysed quantitatively as long as researchers are aware of potential bias (see below). In addition to gathering data, the Internet may also be used in the course of developing questionnaires, as it allows rapid prototyping and pilot testing of instruments, e.g. to evaluate the effect of framing the questions differently [16].

Several studies have checked the validity of Web-based surveys by comparing the results of studies conducted on the Web with identical studies in the real world. These seem to suggest that the validity and

reliability of data obtained online are comparable to those obtained by classical methods [4,5,17–19]. However, issues of generalizability (mainly due to selection bias, discussed in detail below) remain important considerations, and the researcher should select his or her research question and interpret the results with care. The benefits and problems of Web-based surveys have been summarized by Wyatt, who suggests guidelines for when they may be appropriate (see Box 1) [20].

Box 1 Guidelines for Web-based surveys

Scenarios that may be suitable for a Web-based survey

Respondent features:

- Respondents are already avid Internet users; e-mail addresses known for reminder messages.
- Respondents are enthusiastic form fillers; will not require monetary incentives.
- Need for respondents covering a wide geographical area (e.g. rare clinical specialties, diseases).
- Respondents are known to match non-respondents and even non-Internet users on key variables.

Survey features:

- Need for complex branching, interactive questionnaire or multimedia as part of the survey instrument.
- Survey content will evolve fast (e.g. Delphi method surveys use repeating rounds of revised questionnaires delivered over a short period, incorporating aggregate results from previous rounds until convergence is achieved).
- Intent is to document bizarre, rare phenomena whose simple occurrence is of interest.
- No need for representative results: collecting ideas vs. hypothesis testing.

Investigator features:

- Limited budget for mailing and data processing, but good in-house Web skills.
- Precautions can be taken against multiple responses by same individual, password sharing.
- Web survey forms have been piloted with representative participants and demonstrate acceptable validity and reliability with most platform, browser, and Internet access provider combinations.
- Data is required fast in a readily analysed form.

Scenarios that are unsuitable for a Web-based survey

Respondent features:

- Target group is under-represented on Internet; e.g. the underprivileged, elderly people.

continued

Box 1 continued

- Target group is concerned, however unreasonably, about privacy aspects.
- Target group requires substantial incentives to complete the survey.
- Need for a representative sample.

Survey features:

- Need for very accurate timing data on participants (inaccuracies in the range of seconds are added due to network transmission times, unless JavaScript or Java applets are used; see Glossary) or observational data on participants.
- An existing paper instrument has been carefully validated on target group.
- Need to capture qualitative data or observations about participants.
- Wish to reach the same group of participants in the same way months or years later.

Investigator features:

- Limited in-house Web or Java expertise but existing desktop publishing and mailing facility.

Selection bias

In 'open' surveys conducted via the Internet where Web users, newsgroup readers, or mailing list subscribers are invited to participate by completing a questionnaire, selection bias is a major factor limiting the generalizability (external validity) of results. Selection bias occurs due to:

- The non-representative nature of the Internet population.
- The self-selection of participants, i.e. the non-representative nature of respondents, also called the 'volunteer effect' [21].

The non-representative nature of Internet demographics was briefly considered in Chapter 1.4. Considering whether the topic chosen for study is suitable for the Internet population is the first and probably the most important step in minimizing bias, thus maximizing response rates and increasing the external validity of the results [20]. For example, targeting elderly homeless alcoholics is unsuitable for an Internet survey and the results are likely to be heavily skewed by hoax responses.

Self-selection bias originates from the fact that people are more likely to respond to questionnaires if they see items which interest them, e.g. because they are affected by the items asked about, or because they are attracted by the incentives offered for participating. As people who respond almost certainly have different characteristics than those who do not, the results are likely to be biased. This kind of selection bias is more serious than the bias arising from the non-representative nature of the

population, because the researcher deals with a myriad of unknown factors and has little opportunity to interpret his or her results accordingly. Such bias may be exacerbated via loaded incentives (e.g. typical 'male' incentives such as computer equipment). Evidence suggests women are generally more interested in health topics and exhibit more active information-seeking behaviour [22], so are more likely to volunteer participation in health questionnaires. For Web surveys, the potential for self-selection bias can be estimated by measuring the response rate, expressed as the number of people completing the questionnaire divided by those who viewed it (cf. the participation rate, expressed as the number of site visitors viewing the questionnaire divided by the total number of site visitors).

Technical issues

Although a detailed analysis is beyond the scope of this chapter, a synopsis of important techniques and tips for implementing Web-based surveys provides some insight into the difficulties faced by survey designers (see Box 2).

Box 2 Technical issues in implementing Web-based surveys

Use of 'cookies'

Cookies can assign a unique identifier to every questionnaire viewer, useful for determining response and participation rates (see text), and for filtering out multiple responses by the same person. As cookies may be regarded with suspicion, we recommend that researchers openly state that cookies will be sent (and the reasons for this); set the cookie to expire on the day that data collection ceases; and publish a privacy policy (p. 127).

Measuring response time

The time needed to complete a questionnaire can be readily calculated by subtracting the time a form was called up by the browser from the time it was submitted using an automatic time-stamp. The response time may be used to exclude respondents who fill in the questionnaire too quickly: this may identify hoax responses, where respondents don't read the questions.

Avoiding missing data

Forms can be configured to automatically reject incomplete questionnaires and point out missing or contradictory items. Checks can be made on the client (p. 9) prior to submission, or following submission to the server (where incomplete responses can also be analysed, e.g. during a questionnaire pilot).

continued

Box 2 continued

Maximizing response rate

The number of contacts, personalized contacts, and contact with participants before the actual survey are the factors most associated with higher response rates in Web surveys [23]. Incentives increase the risk of selection bias (see text), but less so if cash is offered. Perhaps the best incentive (and the easiest to deliver via the Internet) is the promise of survey results or personalized answers (e.g. a score). The option to complete questionnaires anonymously avoids wariness associated with requests for personal information (e.g. an e-mail address), but increases the risk of hoax responses. Researchers should be open about who is behind the study, what the aim is, and provide opportunities for feedback. Although postal surveys are superior to e-mail surveys with regard to response rate, online surveys are much cheaper [24,25]. Schleyer [15] estimated that the cost of their Web-based survey was 38 percent less than that of an equivalent mail survey and presented a general formula for calculating break-even points between electronic and hard-copy surveys. Jones gave figures of 92 p per reply for postal surveys, 35 p for e-mail, and 41 p for the Web [24].

Randomizing items

Scripting languages may be used to build dynamic questionnaires (as opposed to static forms) that look different for certain user groups or which randomize certain aspects of the questionnaire (e.g. the order of the items). This can be useful to exclude possible systematic influences of the order of the items upon responses.

Ethical issues

The ethical issues involved in any type of online research should not be forgotten [1,26–31]. These include informed consent as a basic ethical tenet of scientific research on human populations [32], protection of privacy, and avoiding psychological harm.

In qualitative research on the Web, informed consent is required when:

- Data are collected from research participants through any form of communication, interaction, or intervention.
- Behaviour of research participants occurs in a private context where an individual can reasonably expect that no observation or reporting is taking place, except when researchers do research 'in public places or use publicly available information about individuals (e.g. naturalistic observations in public places, analysis of public records, or archival research)' [33].

The question therefore arises of whether researchers analysing newsgroup postings enter a 'public place', or whether the space they invade is perceived as private. In the context of research, the expectation of the individual (whether he/she can reasonably expect that no observation is taking place) is crucial. Different Internet services have different levels of perceived privacy (in decreasing order of privacy: private e-mail; chat rooms; mailing lists; newsgroups; Web sites). The perceived level of privacy is a function of the number of participants, but also depends on other factors such as group norms established by the community to be studied. For example, in a controversial paper, Finn studied a virtual self-support group where the moderator was actively discouraging interested professionals who were not sexual abuse survivors from joining the group [34]. In those cases, obtaining informed consent (or seeking an ethical waiver, if the research could not practicably be carried out were informed consent to be required) is mandatory.

In practice, obtaining informed consent, especially for passive research methods, is difficult, as researchers usually cannot post an announcement to a mailing list or newsgroup saying that it will be monitored and analysed for the next few months, as this may greatly influence or even spoil the results, and because the mere posting of such a request may disrupt the community, and therefore be considered unethical. Researchers should therefore first obtain consent from a group moderator in order to explore whether even a request for permission is felt to be disruptive to the group process. If the moderator or person responsible for the list has no objections, one may then post a message to a newsgroup or mailing list explaining the purpose of the research, explaining that one will observe the community, assuring all participants of anonymity, and giving them the opportunity to withdraw from the newsgroup or mailing list or to exclude themselves from the study by writing to the researcher. The fundamental problem is that this may influence the communication process and may even destroy the community. Besides, participants who later join the group need to get the same information. An alternative would be to analyse the communication retrospectively and to write individual e-mails to all participants whose comments were to be analysed or quoted, asking for permission to use them; this technique has been used by Sharf [35].

In any case, researchers should make themselves familiar with the virtual community they are approaching; i.e. read the messages in a newsgroup for some time ('lurking'). Under no circumstances should researchers blindly

spam (p. 31) or cross-post requests for research participation to various newsgroups.

Informed consent may also play a role when researchers report aggregate (collated and hence anonymous) data on usage patterns, such as a log-file analysis (reporting data on what Web sites have been accessed by a population). Crucial here is an appropriate privacy statement stating that these data may be analysed and reported in aggregate [28]. Note that aggregate data are exempt from the registration requirements of the UK's Data Protection Act of 1998 (see Chapter 3.4).

In conducting surveys researchers may obtain informed consent by declaring the purpose of the study; disclosing which institutions are behind the study; explaining how privacy will be assured; and detailing with whom data will be shared and how it will be reported, before participants complete the questionnaire.

When reporting results, it is obvious that the total anonymity of research participants needs to be maintained. Researchers have to keep in mind that, by the very process of quoting the exact words of a newsgroup or mailing list participant, the confidentiality of the participant may already be broken as Internet search engines (see Chapter 6.2) may be able to retrieve the original message, including the e-mail address of the sender. It is essential, therefore, to ask participants whether they agree to be quoted whenever there may be a retrievable archive, pointing out the risk that they may be identifiable. Problems can also potentially arise from just citing the name of the community (e.g. of a newsgroup), which may damage the community being studied.

Finding methods, protocols, and instruments

For laboratory 'bench work', researchers often need a protocol for a specific assay method. In addition to the possibility of searching literature databases, there are also specialized services on the Web that can assist in this research, such as MethodsFinder and the 'Technical tips online' database at BioMedNet:

MethodsFinder (BIOSIS):

<http://www.methodsfinder.org/>

BioMedNet:

<http://www.bmn.com/>

Sometimes asking a specific question on the right newsgroup or mailing list is also very effective. Clinical researchers may be more interested in

instruments to measure patient outcomes. An excellent guide to selecting quality-of-life instruments is the Quality of Life Instruments Database at the Mapi Research Institute:

<http://www.qolid.org/>

Online statistical analysis tools are available at the Simple Interactive Statistical Analysis (SISA) Web site, while background information is available within the online book *Statistics at square one*:

SISA (Daan Uitenbroek):

<http://home.clara.net/sisa/>

Statistics at square one (British Medical Journal Publishing Group):

<http://www.bmj.com/collections/statsbk/>

Protocols of clinical trials, which may be useful for researchers developing their own protocols, can be found in some of the clinical trial databases available on the Web, as described below.

Clinical trials and the Web

The Web is being used to assist in the identification and conduction of clinical trials.

Identifying trials

To prevent unintended duplication of clinical research, detect under-reporting of research, and ease the work of systemic reviewing, it has been suggested that we should prospectively register clinical trials [36–39]. It is, however, unlikely that there will ever be one complete centralized multinational database. Instead, multiple resources set up by numerous different organizations will exist [40]. Internet technology will play a central role in linking these databases and making this information available to researchers and patients. Some scenarios in which a search of trial databases may be useful:

- A researcher wants to conduct a randomized controlled trial and wants to know whether anyone else is already running one on the same topic.
- A physician has a patient who is asking about available trials.
- A patient is looking for ongoing trials.
- A researcher is looking for possible participants for his trial.
- A researcher doing a systematic review is looking for unpublished trials.

Information about ongoing and completed clinical trials is increasingly being published on the Internet, and searches on the Web may be a useful

means of complementing traditional bibliographic searches if authors of systematic reviews wish to find ongoing or unpublished trials [41].

Researchers use their personal or department home pages to announce their interest in a certain research area or to recruit patients [42]. Journals like *The Lancet* have begun to publish research protocols on their Web site [43], and more and more researchers will also publish 'pre-prints' (p. 239) of their findings on the Web [44].

Consumers and patient organizations also have an interest in disseminating information about ongoing trials; e.g. the National Alliance of Breast Cancer Organizations:

<http://www.nabco.org/>

Government and funding agencies react to this need by establishing trial databases for consumers; e.g. the US National Institutes of Health searchable database [45]:

<http://ClinicalTrials.gov/>

Commercial enterprises also help researchers to recruit patients, or help patients to find clinical trials. For example:

CenterWatch Clinical Trials Listing Service (CenterWatch, Inc.):

<http://www.centerwatch.com/>

ClinicalTrialFinder.com (Clinical Data Technologies Ltd):

<http://www.clinicaltrialfinder.com/>

Current Controlled Trials (BioMed Central):

<http://www.controlled-trials.com>

Pharmaceutical companies and industry associations have likewise begun to recognize that openness and access to information on clinical trials and new drug developments can improve patient care and are part of social responsibility [46]. For example:

Clinical Trials Register (GlaxoSmithKline):

<http://ctr.glaxowellcome.co.uk/>

Search for Cures (Pharmaceutical Research and Manufacturers of America):

<http://www.phrma.org/searchcures/>

Finally, information or databases on ongoing clinical trials can often also be found on disease-specific sites. For example:

Canadian HIV Trials Network:

<http://www.hivnet.ubc.ca/ctn.html>

CancerNet (National Cancer Institute):

<http://cancernet.nci.nih.gov/>

Conducting trials on the Web

The Web is increasingly being used in the course of conducting large-scale multi-centre clinical trials (e.g. for remote randomization and data entry), and in the distribution of information on trial progress or protocols [47–48]. Trial centres may enter patient data using Java applets (see Glossary) that encrypt data and send it to the data centre via the Internet [49–52], where the data are stored and randomized, returning for example a study number and randomization information.

Pre-publishing and publishing research

As discussed in Chapter 7.1, traditional publication is a well-defined event, whereas 'publication' in the electronic age is much more of a continuum [53], reflecting and occurring during the entire research process from hypothesis formulation to data gathering, interpretation, and the presentation and discussion of the final results. In order to distinguish online collaborative 'work in progress' from 'final' peer-reviewed publication we may term the former 'Type 1' and the latter 'Type 2' electronic publication [54]. Here, peer review is not the distinguishing characteristic: in Type 1 publication a 'post-publication' peer review process takes place. Type 2 publication will ordinarily take place in online journals (see Chapter 7.1). The following scenarios illustrate how researchers might use Type 1 electronic publication on the Internet:

- Sending and discussing preliminary results on mailing lists.
- Publishing drafts of scientific papers on pre-print/e-print sites (p. 239) in order to solicit comments and to improve the manuscript.
- Publishing data and information in databases; e.g. nucleotide sequences in the EMBL/Genbank databases.
- Publishing clinical trial protocols and raw data in a 'trial bank' [55].

Current awareness services

Electronic editions of paper journals and 'stand alone' e-journals typically offer subscriptions to 'TOC alerts', where users receive a table of contents by e-mail as soon as a new issue appears. The more sophisticated systems allow users to specify their interests using a controlled vocabulary, enabling the system to screen each newly published article for certain keywords or

citations. Examples of current awareness services include:

Customised @lerts (*British Medical Journal*):

<http://bmj.com/cgi/customalert/>

JournAlert (*Doctors.net.uk*):

<http://www.doctors.net.uk/>

Journal Watch (*Massachusetts Medical Society*):

<http://www.jwatch.org/>

References

1. Eysenbach G, Till JE. Ethical issues in qualitative research on the Internet. *British Medical Journal* 2001; 323: 1103–5.
2. Eysenbach G, Sa ER, Kuss O, Diepgen TL. A framework for evaluating e-health: systematic review of empirical studies assessing the quality of health information and services for patients on the Internet. *Journal of the American Medical Association* In press 2001.
3. Greenhalgh T, Taylor R. Papers that go beyond numbers (qualitative research). *British Medical Journal* 1997; 315: 740–3.
4. Buchanan T, Smith JL. Using the Internet for psychological research: personality testing on the World Wide Web. *British Journal of Psychology* 1999; 90(1): 125–44.
5. Buchanan T, Smith JL. Research on the Internet: validation of a World-Wide Web mediated personality scale. *Behavioural Research Methods in Instrumental Computing* 1999; 31: 565–71.
6. Schmidt WC. World-Wide Web survey research: benefits, potential problems, and solutions. *Behavioural Research Methods in Instrumental Computing* 1997; 29: 274–9.
7. Pealer LN, Weiler RM. Web-based health survey research: a primer. *American Journal of Health Behavior* 2000; 24: 69–72.
8. Zhang Y. Using the Internet for survey research: a case study. *Journal of the American Society of Informatic Sciences* 2000; 51: 57–68.
9. Lazar J, Preece J. Designing and implementing Web-based surveys. *Journal of Computer Information Systems* 1999; 39: 63–7.
10. Kaye BK, Johnson TJ. Research methodology: Taming the cyber frontier: techniques for improving online surveys. *Social Science Computer Review* 1999; 17: 323–37.
11. Soetikno RM, Mrad R, Pao V, Lenert LA. Quality-of-life research on the Internet: feasibility and potential biases in patients with ulcerative colitis. *Journal of the American Medical Informatics Association* 1997; 4: 426–35.
12. Eysenbach G, Diepgen TL. Epidemiological data can be gathered with world wide web [letter]. *British Medical Journal* 1998; 316: 72.
13. Hilsden RJ, Meddings JB, Verhoef MJ. Complementary and alternative medicine use by patients with inflammatory bowel disease: an Internet survey. *Canadian Journal of Gastroenterology* 1999; 13: 327–32.
14. Schleyer TK, Forrest JL, Kenney R, Dodell DS, Dovgy NA. Is the Internet useful for clinical practice? *Journal of the American Dental Association* 1999; 130: 1501–11.

15. Schleyer TK, Forrest JL. Methods for the design and administration of web-based surveys. *Journal of the American Medical Informatics Association* 2000; 7: 416–25.
16. Suchard MA, Adamson S, Kennedy S. Netpoints: piloting patient attitudinal surveys on the web. *British Medical Journal* 1997; 315: 529.
17. Nathanson AT, Reinert SE. Windsurfing injuries: results of a paper- and Internet-based survey. *Wilderness & Environmental Medicine* 1999; 10: 218–25.
18. Senior C, Phillips ML, Barnes J, David AS. An investigation into the perception of dominance from schematic faces: a study using the World-Wide Web. *Behavioural Research Methods in Instrumental Computing* 1999; 31: 341–6.
19. Krantz JH, Ballard J, Scher J. Comparing the results of laboratory and World-Wide Web samples on the determinants of female attractiveness. *Behavioural Research Methods in Instrumental Computing* 1997; 29: 264–9.
20. Wyatt JC. When to use web-based surveys [editorial]. *Journal of the American Medical Informatics Association* 2000; 7: 426–9.
21. Friedman CP, Wyatt JC. *Evaluation methods in medical informatics*. New York: Springer-Verlag; 1997.
22. Fox S, Rainie L. *The online health care revolution: how the Web helps Americans take better care of themselves*. Washington: The Pew Internet & American Life Project; 2000 [cited 2001 Sep 20]. Available from: URL: <http://www.pewinternet.org/reports/toc.asp?Report=26>
23. Cook C, Heath F, Thompson RL. A meta-analysis of response rates in Web- or internet-based surveys. *Educational and Psychological Measurement* 2000; 60: 821–36.
24. Jones R, Pitt N. Health surveys in the workplace: comparison of postal, e-mail and World Wide Web methods. *Occupational Medicine (London)* 1999; 49: 556–8.
25. Mavis BE, Brocato JJ. Postal surveys versus electronic mail surveys. The tortoise and the hare revisited. *Evaluating Health Professions* 1998; 21: 395–408.
26. Polzer JC. Using the Internet to conduct qualitative health research: methodological and ethical issues [dissertation]. University of Toronto; 1998.
27. Cho H, LaRose R. Privacy issues in Internet surveys. *Social Science Computer Review* 1999; 17: 421–34.
28. Thomas J. The ethics of Carnegie Mellon's 'cyber-porn' study, 1995 [cited 2001 Jan 12]. Available from: URL: <http://sun.soci.niu.edu/~jthomas/ethics.cmu>
29. Till JE. Research ethics: Internet-based research. Part 1: on-line survey research. 1997 [cited 2001 Jan 12]. Available from: URL: <http://members.tripod.com/~ca916/index-3.html>
30. King SA. Researching Internet communities: proposed ethical guidelines for the reporting of results. *The Information Society* 1996; 12: 119–28.
31. Karlinsky H. Internet survey research and consent. *M.D. Computing* 1998; 15: 285.
32. World Medical Association. Declaration of Helsinki: ethical principles for medical research involving human subjects (as amended 2000 Oct). 2000 [cited 2001 Jan 12]. Available from: URL: http://www.wma.net/e/policy/17-c_e.html
33. American Sociological Association. American Sociological Association's Code of Ethics. 1997 [cited 2001 Jan 12]. Available from: URL: <http://www.asanet.org/members/ecoderev.html>
34. Finn J. An exploration of helping processes in an online self-help group focusing on issues of disability. *Health and Social Work* 1999; 24: 220–31.

35. Sharf BF. Communicating breast cancer on-line: support and empowerment on the Internet. *Women and Health* 1997; 26: 65–84.
36. Simes RJ. Publication bias: the case for an international registry of clinical trials. *Journal of Clinical Oncology* 1986; 4: 1529–41.
37. Chalmers I, Dickersin K, Chalmers TC. Getting to grips with Archie Cochrane's agenda [editorial]. *British Medical Journal* 1992; 305: 786–8.
38. Chalmers I, Gray M, Sheldon T. Handling scientific fraud. Prospective registration of health care research would help [letter]. *British Medical Journal* 1995; 311: 262.
39. Horton R, Smith R. Time to register randomised trials. The case is now unanswerable [editorial]. *British Medical Journal* 1999; 319: 865–6.
40. Tonks A. Registering clinical trials. *British Medical Journal* 1999; 319: 1565–8.
41. Eysenbach G, Tuische J, Diepgen TL. Evaluation of the usefulness of Internet searches to identify unpublished clinical trials for systematic reviews. *Medical Informatics and the Internet in Medicine* 2001; 26(3): 203–18.
42. Wilmoth MC. Computer networks as a source of research subjects. *Western Journal of Nursing Research* 1995; 17: 335–8.
43. Chalmers I, Altman DG. How can medical journals help prevent poor medical research? Some opportunities presented by electronic publishing. *Lancet* 1999; 353: 490–3.
44. Delamothe T, Smith R, Keller MA, Sack J, Witscher B. Netprints: the next phase in the evolution of biomedical publishing [editorial]. *British Medical Journal* 1999; 319: 1515–16.
45. McCray AT, Ide NC. Design and implementation of a national clinical trials registry. *Journal of the American Medical Informatics Association* 2000; 7: 313–23.
46. Sykes R. Being a modern pharmaceutical company: involves making information available on clinical trial programmes [editorial]. *British Medical Journal* 1998; 317: 1172.
47. Santoro E, Nicolis E, Franzosi MG, Tognoni G. Internet for clinical trials: past, present, and future. *Controlled Clinical Trials* 1999; 20: 194–201.
48. Kelly MA, Oldham J. The Internet and randomised controlled trials. *International Journal of Medical Informatics* 1999; 47: 91–9.
49. Sippel H, Eich HP, Ohmann C. Data collection in multi-center clinical trials via Internet. A generic system in Java. *Medinfo* 1998; 9(1): 93–7.
50. Sippel H, Ohmann C. A web-based data collection system for clinical studies using Java. *Medical Informatics (London)* 1998; 23: 223–9.
51. Eich HP, Ohmann C. Generalisation and extension of a web-based data collection system for clinical studies using Java and CORBA. *Studies in Health Technology Information* 1999; 68: 568–72.
52. Keim E, Sippel H, Eich HP, Ohmann C. Collection of data in clinical studies via Internet. *Studies in Health Technology Information* 1997; 43(A): 57–60.
53. Smith R. What is publication? [editorial] *British Medical Journal* 1999; 318: 142.
54. Eysenbach G. Challenges and changing roles for medical journals in the cyberspace age: electronic pre-prints and e-papers. *Journal of Medical Internet Research* [serial online] 1999 Dec [cited 2001 Jan 12]; 1(2): e9. Available from: URL: <http://www.jmir.org/1999/2/e9/>
55. Sim I, Owens DK, Lavori PW, Rennels GD. Electronic trial banks: a complementary method for reporting randomized trials. *Medical Decision Making* 2000; 20: 440–50.

