Reading level of privacy policies on Internet health Web sites

MARK A. GRABER, MD; DONNA M. D'ALESSANDRO, MD; AND JILL JOHNSON-WEST, MSW
Iowa City, Iowa

■ OBJECTIVE Most individuals would like to maintain the privacy of their medical information on the World Wide Web (WWW). In response, commercial interests and other sites post privacy policies that are designed to inform users of how their information will be used. However, it is not known if these statements are comprehensible to most WWW users. The purpose of this study was to determine the reading level of privacy statements on Internet health Web sites and to determine whether these statements can inform users of their rights.

■ STUDY DESIGN This was a descriptive study. Eighty Internet health sites were examined and the readability of their privacy policies was determined. The selected sample included the top 25 Internet health sites as well as other sites that a user might encounter while researching a common problem such as high blood pressure. Sixty percent of the sites were commercial (.com), 17.5% were organizations (.org), 8.8% were from the United Kingdom (.uk), 3.8% were United States governmental (.gov), and 2.5% were educational (.edu).

■ OUTCOMES MEASURED The readability level of the privacy policies was calculated using the Flesch, the Fry, and the SMOG readability levels.

■ RESULTS Of the 80 Internet health Web sites studied, 30% (including 23% of the commercial Web sites) had no privacy policy posted. The average readability level of the remaining sites required 2 years of college level education to comprehend, and no Web site had a privacy policy that was comprehensible by most English-speaking individuals in the United States.

■ CONCLUSIONS The privacy policies of health Web sites are not easily understood by most individuals in the United States and do not serve to inform users of their rights. Possible remedies include rewriting policies to make them comprehensible and protecting online health information by using legal statutes or standardized insignias indicating compliance with a set of privacy standards (eg, "Health on the Net" [HON] http://www.hon.ch).

■ KEY WORDS Readability levels; Internet; privacy. (*J Fam Pract 2002; 51:642–645*)

Approximately 33 million individuals in the United States have used the Internet to access medical information.[1,2] Even though most people would like to maintain the privacy of their medical and other information,[3–6] few users of the Internet take steps to do so.[7] Commercial vendors develop profiles of individual users of the Internet. The information tracked includes Web sites visited; terms entered into search engines (including medical terms); goods or services bought online; and participation in forums, chat rooms, and e-mail lists (eg, listservs). The text of any postings in forums and e-mail lists can also be tracked. This information is sold to anyone willing to pay for it, including advertisers, employers, and insurance companies. Commercial vendors use this information to offer goods and services targeted to a user's needs, including medical needs. For example, an individual who visits Web sites dedicated to the care of diabetes mellitus will receive advertising about new diabetes medications and glucose monitoring devices. However, the information can and has been used in other ways, leading to job termination and arrest.[8] A user who repeatedly visits a breast-cancer-related Web site, for example, could be discriminated against by a potential employer or insurance company because she is suspected of being afflicted with the disease.

Unauthorized access to an individual's personal information also occurs. Doubleclick.com, a corporation that collects user information, has had several high-profile breaches of computer security, leaving individuals' information vulnerable to exploitation.[9]

The importance of the confidentiality of medical information has been underscored by the recent publication of the new "Standards for Privacy of Individually Identifiable Health Information" by the Department of Health and Human Services.[10] In part,

From the Departments of Family Medicine and Emergency Medicine (M.A.G.) and Pediatrics (D.M.D., J.J.W.), University of Iowa, Iowa City, IA. Funding for this study was made possible through the Robert Wood Johnson Generalist Faculty Scholars Program. The authors report no competing interests. Address reprint requests to Mark Graber, MD, Department of Family Medicine, University of Iowa, 200 Hawkins Drive, Iowa City, IA 52246. E-mail: mark-graber@uiowa.edu.

TABLE

Domains of all Web sites examined and percentage with privacy statements

Top level domain	All Web sites examined (n = 80)	Web sites with a privacy policy (n = 54)
Commercial (.com)	48 (60%)	37 (68.5%)
Organizations (.org)	14 (17.5%)	8 (14.8%)
United Kingdom (.uk)	7 (8.8%)	5 (9.2%)
Government (.gov)	3 (3.8%)	3 (5.6%)
Network (.net)	2 (2.5%)	0 (0.0%)
Educational (.edu)	2 (2.5%)	1 (1.8%)
Web site did not exist*	2 (2.5%)	0 (0.0%)
South Africa (.za)	1 (1.2%)	0 (0.0%)
Numerical (no domain listed)	1 (1.2%)	0 (0.0%)
Total	80 (100%)	54 (100%)

*Although the list of the top 25 health-related WWW sites was current, 2 of the sites had gone out of business and were no longer available.

these guidelines are designed to "protect the privacy of individually identifiable health information."[10] Although Internet use does not generate a formal medical record, online profiling allows the collection of detailed medical information about a user's diagnoses, medications, etc, which essentially creates "individually identifiable health information" when associated with their names.

One proposed solution to maintaining Internet privacy has been the voluntary posting of privacy statements. These statements serve to inform users of the privacy policies of the Web site, such as what user information is collected and with whom this information will be shared. Three recent studies have shown that the readability level of much of the patient information on the Internet is beyond the comprehension of many individuals in the United States.[11–13] For voluntary privacy statements to be useful, they need to be written at a level understood by most individuals using the Internet. The purpose of this study was to determine the readability level of privacy statements on Internet health Web sites.

METHODS

A total of 80 Internet health Web sites were examined in May 2001 to see if they included privacy policies. To emulate the way a consumer might find information on the Internet, 55 of the sites were selected by entering search terms for common conditions into a widely used Internet search engine (http://www.google.com). The terms searched for were "high blood pressure," "fever," "cough," and "wellness." The other 25 Web sites analyzed represent the most commonly visited health information Web sites on the Internet.[14] For Web sites identified by Google, the Web pages represented by the top 10 results for each term were viewed, and any links on those pages were followed until 55 health information Web sites were identified. We did not limit the Web sites to only those identified by the search engines because in many cases, users will follow the links on a page identified by the search engine. Links that led to medical school lectures, nonhuman diseases, online journal articles aimed at health care professionals, or the contents of e-mail or listserv summaries were omitted.

The Web sites were then examined by 1 researcher (J.J.W.) to see if they included privacy statements. This review included looking at the page that the link led to, the Web site's homepage, any Web site policies, "about us" type of material, etc. Links to privacy foundations such as Health on the Net were considered to represent privacy statements. If a privacy statement was found, the text was copied into Microsoft Word 98 for Macintosh (Microsoft Corporation, Redmond, WA) and the documents' Flesch Reading Ease score was generated using the built-in software in Microsoft Word 98 (Macintosh). The same privacy statements were printed and then analyzed using 2 hand-calculation methods: the Fry Readability formula and the SMOG method (simplified measure of gobbledygook).[15] If a Web site was found to have no privacy statement, this was confirmed by a second researcher (M.G.). Web sites on which no privacy policy was found after 2 searches were not contacted to determine if a policy exists. It is not likely that end users would contact a site to determine the existence of a privacy policy. Thus, there would be the de facto absence of a privacy policy with respect to the end user. Ten percent of the readability levels were analyzed by a second researcher (D.M.D.) and found to have accurate coding. All data were entered into Microsoft Excel spreadsheet and analyzed using the built-in statistical formulas. We did not seek to evaluate the content of the privacy statements themselves.

Approximately 40 different readability formulas are available for use, all of which will give a reasonably accurate grade level (generally plus or minus 1 grade level with 68% confidence: an acceptable standard in the field).[16] Most of these formulas rely on number of syllables in a word and sentence length to judge readability. In general, readability is easier when there are a low number of words per sen-

tence, characters per word, and syllables per word, and a low percentage of passive sentences.[16] Illustrations and tables also improve the readability. The Flesch Reading Ease score is one of the most widely used and validated systems for scoring readability. It is the standard used by the insurance industry for consumer documents and contracts.[17,18] Documents scoring 70 or above are described as "easy" and are written at the grade school level. A score of 60 to 70 is described as "standard" and written at approximately at the high school level. Scores below 60 are described as "fairly difficult," "difficult," or "very difficult" as the score decreases.[19] The Fry formula is a hand-calculation method that is recommended by experts in the field for use with patient education materials. The SMOG is another hand-calculation method commonly used for evaluation of health information.[16] Results from the Fry formula and SMOG methods are expressed as standard United States grade levels.

RESULTS

Of the 80 health Web sites in the initial sample, 2 had recently closed down and an additional 24 (30%) (including 23% of the commercial Web sites) did not have a privacy statement. Thus, the privacy statements of 54 Web sites were analyzed using the 3 readability formulas. The Web site domains for the 80 Web sites as well as the 54 with privacy statements are shown in the Table. The average Flesch Reading Ease score of privacy statements was 39 or "difficult." This level is similar to reading a corporate annual report. The Flesch Reading Ease score range was 24.4 to 54.2. This range is described as "very difficult" (eg, similar to reading legal contracts) to "fairly difficult" (eg, similar to reading novels such as the Henry James novel, *The Ambassadors*).[19] The Fry formula had a readability level equal to 14.6 or 14th grade 6th month (range, 10–17). The overall SMOG readability level was 14.7 (range, 11–19).

DISCUSSION

The goal of this study was to determine the readability level of privacy statements on Internet health Web sites. Privacy statements are meant to protect the individual, a particularly important endeavor when dealing with medical information because of its potentially sensitive nature. Only 70% of the health Web sites examined in this study even contained a privacy statement. The readability levels of the privacy statements found were considered difficult and would require approximately 2 years of university training to be comprehensible. In contrast, only 60% of diabetic patients in 1 study could understand information written at the 6th grade reading level.[20] Most studies find that only a minority of patients can comprehend information written at a 9th grade reading level.[21–24] Overall, for 90% of adults to be able to read and comprehend written materials, the materials should be written at less than the 8th grade reading level.[16] Although current Internet users may be a more educated group than the general population, this trend is changing. A recent study found that 21% of individuals with less than a high school education have Internet access, as do 43% of high school graduates.[25] Because an individual's actual reading level is usually 2 to 5 grades below the grade completed,[26] many of these individuals will not be able to comprehend a Web site's privacy policy and thus may not be protected or understand their options pertaining to protection.

Many patients are beginning to use the Internet as a source of information and many are unaware that their every move is being tracked. Physicians can help to protect patients by making them aware of this problem and suggesting that patients take steps to protect their information.

A number of possible solutions can be applied to this problem. Rewriting privacy policies to be comprehensible to most Internet users should be the initial step. Resources are available to Web site developers that help to maximize the readability of information.[27–29] More widespread use of the "Health on the Net" (HON, http://www.hon.ch) privacy criteria would reassure individuals that their data are secure. Consideration should also be given to extending the same protection to medical information gleaned from the Internet by Web sites and advertising companies as is given to the formal medical record because, in effect, they can reflect the same type of information. For example, on several sites, users can enter medications they are taking and diagnoses (http://www.drkoop.com and others).

The first limitation of this study is that the Web sites studied represent a minority of the health Web sites on the Internet. However, the study set includes the top 25 visited health Web sites and did seek many different types of Web sites for evaluation. Also, none of the Web sites had a policy that was written at a level comprehensible to most of the English-speaking people in the United States. It is unlikely that including more Web sites would substantially change the outcome. Finally, given the large (and unknown) number of health Web sites on the Internet, it is not practical to obtain a large enough sample to be representative of all health Web sites.

The second problem is readability levels themselves. Readability calculations have been criticized because they depend mainly on sentence length and

the number of syllables per word. If an individual is familiar with the sentence terminology, it may be possible to interpret the information correctly using context clues. However, most individuals do not have a working knowledge of legal terminology (eg, "indemnify") and thus may not be able to use these clues to determine the meaning of a sentence.

CONCLUSIONS

Thirty percent of health Web sites do not have privacy statements, and those that are posted are beyond the reading level of most adults. For this reason, current privacy statements do not function to adequately inform users of a Web site's privacy policy and do not protect users' privacy rights.

REFERENCES

1. Pew Internet and American Life Project Survey. Daily Internet activities. Available at: http://www.pewinternet.org/reports/chart.asp?img=6_daily_activities.jpg. Accessed April 10, 2001.
2. Pew Internet and American Life Project Survey. Internet activities. Available at: http://www.pewinternet.org/reports/chart.asp?img=6_internet_activities.jpg. Accessed April 10, 2001.
3. Cheng TL, Savageau JA, Sattler AL, DeWitt TG. Confidentiality in health care. A survey of knowledge, perceptions, and attitudes among high school students. JAMA 1993; 269:1404–7.
4. Merz JF, Spina BJ, Sankar P. Patient consent for release of sensitive information from their medical records: an exploratory study. Behav Sci Law 1999; 17:445–54.
5. Weiss BD. Confidentiality expectations of patients, physicians, and medical students. JAMA 1982; 247:2695–7.
6. Fox S, Rainie L, Horrigan J, Lenhart A, Spooner T, Carter C. Trust and privacy online: why Americans want to re-write the rules. The Internet Life Report. Pew Internet and American Life Project; August 20, 2000. Available at http://www.pewinternet.org/reports/toc.asp?Report=19. Accessed April 19, 2001.
7. WebSideStory. Cookie rejection less than 1 percent on the Web According to WebSideStory. Available at: http://www.websidestory.com/cgibin/wss.cgi?corporate&news&press_2_124. Accessed April 19, 2001.
8. The Privacy Foundation. EEOC v. Burlington Northern Santa Fe Railroad. Available at: http://www.privacyfoundation.org/legal/case/show.asp?id=25&t=2. Accessed May 5, 2002.
9. Junk Busters. DoubleClick reaches settlement on class action suits. Available at: http://www.junkbusters.com/new.html#March. Accessed May 21, 2001.
10. Department of Health and Human Services. Standards for privacy of individually identifiable health information. Department of Health and Human Services, Federal Register, 2000 45 CFR Parts 160 and 164. Available at: http://aspe.hhs.gov/admnsimp/final/pvcguide1.htm. Accessed May 6, 2002.
11. Graber MA, Roller CM, Kaeble B. Readability levels of patient education material on the World Wide Web. J Fam Pract 1999; 48:58–61.
12. D'Alessandro DM, Kingsley P, Johnson-West J. The readability of pediatric patient education materials on the World Wide Web. Arch Pediatr Adolesc Med 2001; 155:807–12.
13. Berland GK, Elliott MN, Morales LS, et al. Health Information on the Internet: accessibility, quality, and readability in English and Spanish. JAMA 2001; 285:2612–21.
14. Lee RD, Conley DA, Preikschat A. Wit's wisdom on eHealth Wit SoundView. Available at: http://us.mediametrix.com/home.jsp_2000. Accessed June 6, 2001.
15. McLaughlin G. SMOG grading: a new readability formula. J Reading 1989; 12:639–46.
16. Doak CC, Doak LG, Root JH. Teaching Patients With Low Literacy Skills. 2nd ed. Philadelphia: JB Lippincott; 1995.
17. Kincaid JP, Fishburne RP, Rogers RL, Chissom BS. Derivation of New Readability Formulas (Automated Reliability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Millington, TN: Naval Air Station; 1975. Branch report 8-75.
18. Flesch R. A new readability yardstick. J Appl Psychol 1948; 32:221–33.
19. Roberts JC, Fletcher RH, Fletcher SW. Effects of peer review and editing on the readability of articles published in Annals of Internal Medicine. JAMA 1994; 272:119–21. Also available at:http://www.amaassn.org/public/peer/7_13_94/pv3083x.htm. Accessed January 11, 2001.
20. Overland JE, Hoskins PL, McGill MJ, Yue DK. Low literacy: a problem in diabetes education. Diabet Med 1993; 10:847–50.
21. Foltz A, Sullivan J. Reading level, learning presentation preference, and desire for information among cancer patients. J Cancer Educ 1996; 11:32-8.
22. Davis TC, Mayeaux EJ, Fredrickson D, Bocchini JA Jr, Jackson RH, Murphy PW. Reading ability of parents compared with reading level of pediatric patient education materials. Pediatrics 1994; 93:460–8.
23. Wilson FL. Measuring patients' ability to read and comprehend: a first step in patient education. Nursingconnections 1995 Winter; 8(4):17–25.
24. Estey A, Musseau A, Keehn L. Patient's understanding of health information: a multihospital comparison. Patient Educ Couns 1994; 24(1):73–8.
25. Pew Internet and American Life Project Survey. Internet and specific population groups. Available at: http://www.pewinternet.org/datadump/demo.asp?img=6_specific_pop_groups.jpg. Accessed May 22, 2001.
26. Lee PP. Why literacy matters. Links between reading ability and health. Arch Ophthalmol 1999; 117:100–3.
27. Gaston N, Daniels P. Guidelines: writing for adults with limited reading skills. Available at: http://www.cyfernet.org/research/writeadult.html. Accessed September 18, 2000.
28. Smith S. Readability testing health information. Available at: http://www.prenataled.com/story9.htm. Accessed September 18, 2000.
29. Lynch PJ, Horton S. Web Style Guide: Basic Design Principles for Creating Web Sites. New Haven, CT: Yale University Press; 1999. Also available at: http://info.med.yale.edu/caim/manual/. Accessed March 1, 2000.

JFP