The Low Availability of Metadata Elements for Evaluating the Quality of Medical Information on the World Wide Web

John Shon MD, Mark A. Musen MD, PhD
Stanford Medical Informatics, Stanford University School of Medicine,
Stanford, CA 94305-5479

ABSTRACT

A great barrier to the use of Internet resources for patient education is the concern over the quality of information available. We conducted a study to determine what information was available in Web pages, both within text and metadata source code, that could be used in the assessment of information quality. Analysis of pages retrieved from 97 unique sites using a simple keyword search for "breast cancer treatment" on a generic and a health-specific search engine revealed that basic publishing elements were present in low frequency: authorship (20%), attribution/references (32%), disclosure (41%), and currency (35%). Only one page retrieved contained all four elements. Automated extraction of metadata elements from the source code of 822 pages retrieved from five popular generic search engines revealed even less information. We discuss the design of a metadata-based system for the evaluation of quality of medical content on the World Wide Web that addresses current limitations in ensuring quality.

INTRODUCTION AND BACKGROUND

In recent years, patients have become active participants in the management of their illnesses and have looked to the Internet as a source of medical information. There are multiple advantages of using the Internet for patient education and its popularity among patients attests to its power. One study suggested that patients even prefer using Internet-based sources over physicians for certain types of information.[1] Perhaps the largest barrier to the widespread adoption of the Internet for patient education by health-care providers is the concern for ensuring quality. Several formal studies have clearly demonstrated problems with the accuracy of medical information on the World Wide Web.[2,3]

Rating systems and instruments

Patients often lack the medical knowledge to assess accurately the quality and appropriateness of medical content in Web pages. A large number of third-party "rating" systems and awards have proliferated on the Internet to assist consumers, but a systematic review of these instruments revealed that most were incompletely developed and did not have established validity or reliability.[4] Nonetheless, there are simple publishing standards proposed by Silberg[5] and Wyatt[6] that perhaps constitute a minimum set of criteria by which to evaluate generic information sources on the Internet: authorship, attribution, disclosure, and currency. A recent review of Internet rating tools for the evaluation of health-related Web sites found that these were among the most commonly used criteria in current instruments.[7]

Quality criteria can be easily thought of as metadata, or "data about data." Generic standards for metadata include the W3 Consortium's XML and the Dublin Core Metadata Element set of 15 terms.[8] Brennan[9] has used metadata for indexing medical concepts, and Munoz has developed a tool for the creation of "Medical Core Metadata" for use in Web pages.[10] The Platform for Internet Content Selection (PICS) metadata standard has been used with some success in rating and filtering pornographic material on the Internet.[11] Eysenbach and Diepgen have proposed the use of a similar system, med-PICS, for rating medical content on the Internet.[12] Although the validity, value, and desirability of such a system remain to be evaluated, the use of metadata at least offers a potential technical solution for ensuring quality for patients who will undoubtedly continue to seek medical information on the Internet.

Before a filtering and retrieval system can be implemented, it is important to establish the presence of "filterable" elements. Little is known about the presence of quality criteria elements, such as authorship, on Web sites. The Health on the Net Foundation (HON) has established a "code of conduct" which prescribes eight principles of publishing on the Internet,[13] but many Web pages do not adhere to this standard. Hersh *et al* did report on the low applicability and poor quality of information of Web pages for answering clinical questions using detailed searches performed by a skilled librarian.[14] Whether these findings are relevant to pages retrieved by a patient performing a simple search is unknown.

In the descriptive study reported here, we sought to assess the availability of quality criteria, in both

1091—8280/99/$5.00 © 1999 AMIA, Inc.

human and machine-readable form, in individual Web pages and Web sites. We discuss the results and their implications for the design of a metadata-based system to be used in rating, filtering and retrieving medical information on the Internet.

METHODS

Search Methodology:

We chose to search for information on breast cancer treatment because therapy for breast cancer can entail a large amount of patient education. To simulate the information and sites which a "typical" patient would encounter, we performed a simple key-word search, consisting of "breast cancer treatment" (without quotes) from a commonly available and widely used search engine, AltaVista (www.altavista.com). The search retrieved a total of 2736 references to Web pages, and we analyzed the first 100 consecutive URLs. We then performed the same search on the Health On The Net (HON) foundation MedHunt search engine. The first 100 consecutive "HONoured" sites (sites which were reviewed by HON) were used for analysis. If we retrieved a Web page that consisted solely of hyperlinks, or non-relevant information, we followed the first link leading to potentially relevant information on the same site until we encountered a page consisting of more than just hyperlinks. We then evaluated the page for the presence of specific criteria.

To evaluate metadata in the source code of pages, the same search for "breast cancer treatment" was performed on five popular engines: AltaVista, Excite, Hotbot, Infoseek and Lycos. Approximately 200 URLs were retrieved from each site, and metadata was extracted using a Perl based robot which we developed. Sites that responded to HTTP requests were then used for automatic extraction of metadata.

Quality Criteria and Data Extraction

We chose to evaluate a minimal set of elements from Web pages consisting of authorship, attribution, currency, and disclosure, based on previous editorials,[5,6] a review of currently used criteria,[7] and rating instruments in development.[15] We also evaluated a supplementary set of criteria listed in Figure 1. If we did not observe an element on the page retrieved, we searched the Web site for the information manually where appropriate. An element was considered to be part of the page if it was directly retrievable through a hyperlink on the page. When we retrieved multiple pages from the same site, we used the median element characteristic for the analysis to avoid overrepresentation of a single site in the analysis. We considered references to be applicable to a page if the page contained statements that were more than just definitions. One of the authors (JS), who is Board-certified in internal medicine, performed all data extraction and analysis.

RESULTS

Manual Extraction of Metadata

85 of 100 pages retrieved from the AltaVista search engine were "relevant" in that they addressed some aspect of treatment of breast cancer. This term was applied to be broadly inclusive of a wide variety of materials. Of the 85 relevant pages, 47 were from unique Web sites. Similarly, 50 unique and relevant pages were retrieved from the "HONoured" sites, 12 of which had the HON logo. Tables 1 and 2 and Figure 1 refer to these 47 and 50 respective pages.

Automated Extraction of Metadata

822 out of 1050 sites responded to HTTP requests and were used in the analysis. The mean number of metadata elements per page ranged from .975 (Lycos) to 1.53 (Hotbot) with an overall mean of 1.23 (std dev 1.72). Of the pages with at least one element, there was an average of only 2.18 elements/page. The most frequently encountered metadata elements are listed in Figure 2.

Table 1 Document Characteristics

	Altavista		HON		Total
Characteristic	No.	(%)	No.	(%)	(%)
Funding					
Nonprofit	23	(49)	31	(62)	(56)
Commercial	19	(40)	17	(34)	(37)
Unclear	5	(11)	1	(2)	(6.2)
Target Audience					
Patients	40	(85)	47	(94)	(90)
Providers	3	(6.4)	4	(8)	(7)
Researchers	2	(4.3)	0	(0)	(2)
Other	2	(4.3)	16	(32)	(19)

Table 2 Document Type

	Altavista		HON		Total
Document Type	No.	(%)	No.	(%)	(%)
Treatment summary or review	17	(36)	34	(68)	(53)
Annotated Links	6	(13)	6	(12)	(12)
Questions and Answers	5	(11)	1	(2)	(6.2)
Advertisement	4	(8.5)	2	(4)	(6.2)
Research abstract summary	3	(6.4)	5	(10)	(8.2)
Research abstracts	2	(4.3)	1	(2)	(3.1)
Testimonies/stories	2	(4.3)	2	(4)	(4.1)
Other	8	(17)	0	(0)	(8.2)

Figure 1

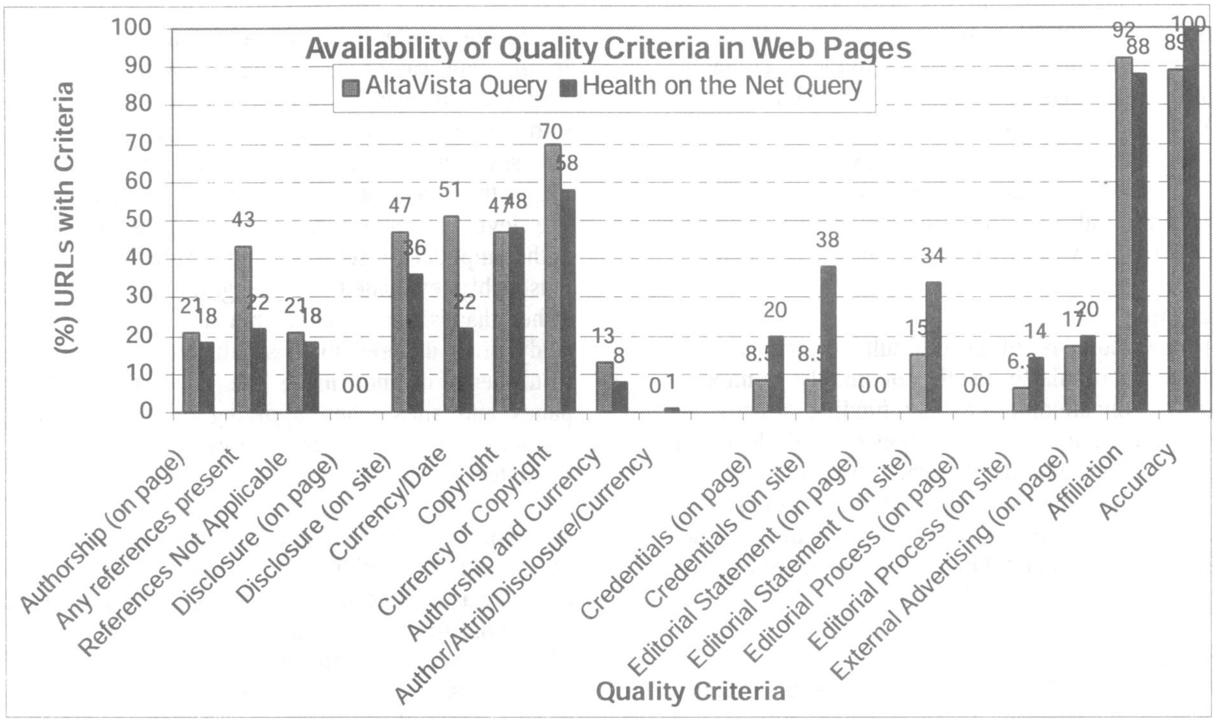
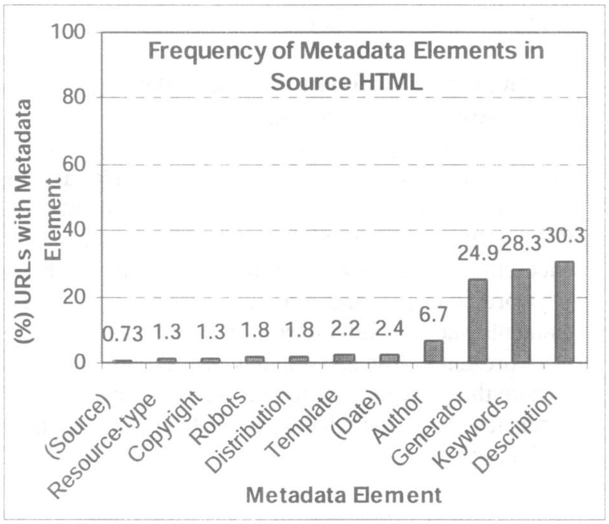


Figure 2



DISCUSSION

We report the two queries separately in Tables 1 and 2 and Figure 1, but we discuss the combined results in the following analyses.

Document and Organization Characteristics

55% of the documents were associated with commercial organizations, and 38% were associated with non-profit or academic groups (Table 1). A majority (90%) of the documents targeted a patient audience. Most (53%) pages consisted of summary treatment information, although a wide variety of document types were encountered.

Availability of Information

Authorship

Surprisingly, the presence of an author's name on the page was found in only 20% of the documents. One explanation for this finding is that "institutional" authorship may be implied with the presence of a logo, copyright, or disclaimer. For example, although many pages analyzed within Oncolink (a Web site for patients with cancer) had no specific author, most had a copyright notice, followed by "The Trustees of the University of Pennsylvania" prominently displayed at the top of the page (data not shown). Whereas this clearly implies ownership and responsibility, it is unclear whether it is sufficient for most readers. Only 10% of pages had no clear affiliation, suggesting that affiliation may be acting as a surrogate for authorship.

Attribution

Copyrights notices were present on 47% of pages. Of the documents where references were applicable (contained more than just definitions), only 32% had any reference to an article in the biomedical literature (Figure 1). Of the pages with references, most were very sparsely annotated. Referencing standards appeared to vary depending on the audience for the page. For example, in two pages retrieved from the National Cancer Institute, one page, directed toward consumers, contained no references, whereas a similar page, directed toward providers, contained detailed references and an extensive bibliography. The dichotomy may arise from a belief that patients are not interested in "evidence-based" information, but this may be a false assumption; approximately 30% of the traffic to free Medline services from the National Library of Medicine is from the lay public.[16]

Disclosure

Full disclosure is often difficult to achieve and evaluate, particularly in the minimally contextual Internet. Any attempt to explain funding, advertising, or conflicts of interest was accepted as disclosure in this study. Despite this minimal definition, only 41% of sites contained this information. The reason for this lack of disclosure is not clear. Without the presence of authors' names and dates, it is even more difficult to evaluate this quality criterion.

Currency

A copyright notice, which may serve as a rough estimate of the currency of the document, was present on 47% of pages. Many documents (36%), however, had neither a date nor copyright notice. Although many dates are indirectly available through browser functions, these dates are not easily accessible for most users. Again, it is not clear why this information is lacking on most pages; it may be related to the relative recent introduction of most information on the Internet.

Supplemental Criteria

In this study only 38% of pages/sites listed credentials of authors or editors, only 32% of pages/sites contained any editorial statement, and only 13% contained information stating an editorial process existed. Most pages (95%) were judged not to contain misleading or inaccurate information (accuracy in Figure 1). This finding is encouraging, given the results of previous studies.[2,3] External advertising (advertising on the page related to another entity) was present in 19% of sites.

Metadata elements

Only 46% of pages in the two searches contained any metadata, and only one page contained all of the Dublin Core elements (data not shown). When metadata were present, they generally did not contain authorship, attribution, disclosure, or currency information. Many commercial sites exploited the "keywords" element by inserting extremely long lists of words, presumably to attract search engines.

Information on Web pages vs. Web sites

Although some sites did contain information on disclosure, credentials, and editorial processes, this information was generally not available on the page retrieved. It would be impractical to require certain information such the editorial policy, site purpose, and site mission, to be present on each and every page; these criteria are perhaps best evaluated at the site level. Conversely, certain elements such as authorship, date, disclosure, and accuracy can only be reasonably evaluated at the granularity of pages, rather than sites; pages within a single site in our study often differed in these attributes. Most rating instruments do not make the distinction between pages and sites when applying criteria, but it is arguably necessary to do so given the heterogeneity of content on sites.

Surprisingly, only one page encountered met all four criteria of authorship, attribution, disclosure, and currency as defined in this study. Overall, our results are similar to those presented in a study by Hersh *et al.*, who found low percentages of quality criteria on Web pages: authorship (30.8%), sources (12.2%), disclosure (11.1%), and date (17%). The very lack of these elements in Web pages may explain why only 2 of 47 instruments evaluated by Jadad used these elements as quality criteria.[4]

Automated Extraction of Metadata Elements From Multiple Search Engines.

Of the 822 pages retrieved, 376 (46%) had no metadata element present in the source code, and the average number of elements per page (1.23) was very low. Overall, a total of 74 different elements were encountered, with the most frequent elements being keyword, content, and generator. Basic publishing elements such as author, source, copyright, and date were present in low percentages (see Figure 2). Thus, despite the potential use for metadata, it appears most publishers have included sparse, if any, metadata in Web pages.

Our study has limitations related to the design of the search strategies and data extraction. Variability and bias in results can occur due to the search engine, keyword search, medical subject, and interpretation of criteria used. It is thus difficult to assess whether the samples are representative of pages about breast cancer treatment, or of all pages that contain medical information on the Internet. Many of the pages retrieved may not be relevant to patients with breast cancer interested in making decisions about therapy. The criteria used in the study have not been validated, and there are other important criteria not evaluated, such as literacy, navigability, use of multimedia, and interactivity, which have an obvious impact on quality.[6] In addition, the use of only one reviewer limits the reliability of the analysis.

CONCLUSIONS AND FUTURE WORK

Our study suggests that most Web pages contain accurate information, but do not have adequate standard publishing quality criteria present in the text that could serve as a proxy for quality. The most prevalent element on pages was an institutional affiliation: Trust in a branded medical institution and its ownership of pages may be a *de facto* substitute for trust obtained using traditional publishing criteria. The analysis of publishing elements also leads us believe that rating systems must operate at the granularity of pages (not sites) to ensure an adequate evaluation of quality for users. In addition, our study suggests that systems that rely solely on automated extraction of metadata for evaluation of quality measures will be of limited utility given the small subset of pages containing such data.

We are using the results of this study for the design and implementation of a metadata-based system for evaluating medical information on the Internet. The system is very similar to the med-PICS[12] system in the following respects: It relies on evaluation of sites using quality criteria in the form of metadata, it relies on a distributed, post-publication evaluation process, and it relies on third party acquisition, maintenance and distribution of metadata. We believe metadata criteria supplied by known and trusted organizations may serve in lieu of metadata supplied by the original content providers. The leverage of external information resources exists currently in the form of hyperlinks. A metadata evaluation system using quality criteria simply provides a formal methodology for providing quality links for the benefit of patients, and we plan to evaluate the use and impact of such a system by consumers of health information.

Acknowledgements

This work was supported in part by the National Library of Medicine Training Grant NLM LM 07033.

References

1. Ferguson T, Kelly WJ. The Ferguson Report; 1999 Jan/Feb. Report No.: Vol 1, No. 1.
2. Impicciatore P, Pandolfini C, Casella N, Bonati M. Reliability of health information for the public on the World Wide Web: systematic survey of advice on managing fever in children at home [see comments]. BMJ 1997;314(7098):1875-9.
3. McClung HJ MR. The Internet as a Source for Current Patient Information. Pediatrics 1998;101(6)(Jun 1):E2.
4. Jadad AR, Gagliardi A. Rating health information on the Internet: navigating to knowledge or to Babel? JAMA 1998;279(8):611-4.
5. Silberg WM, Lundberg GD, Musacchio RA. Assessing, controlling, and assuring the quality of medical information on the Internet: Caveat lector et viewor--Let the reader and viewer beware [editorial] JAMA 1997;277(15):1244-5.
6. Wyatt JC. Commentary: measuring quality and impact of the World Wide Web [comment] [see comments]. BMJ 1997;314(7098):1879-81.
7. Kim P, Eng TR, Deering MJ, Maxfield A. Published criteria for evaluating health related web sites: review. BMJ 1999;318(7184):647-649.
8. Dublin Core Metadata Element Set. http://purl.oclc.org/dc/about/element_set.htm
9. Brennan PF, Caldwell B, Moore SM, Sreenath N, JonesJ. Designing HeartCare: custom computerized home care for patients recovering from CABG surgery. Proc AMIA Symp 1998:381-5.
10. Munoz F, Hersh W. MCM generator: a Java-based tool for generating medical metadata. Proc AMIA Symp 1998:648-52.
11. Resnick P, Miller J. PICS: Internet access controls without censorship. Communications of the ACM 1996 Oct:87.
12. Eysenbach G, Diepgen TL. Towards quality management of medical information on the Internet: evaluation, labelling, and filtering of information. BMJ 1998;317(7171):1496-1500.
13. Boyer C, Selby M, Scherrer JR, Appel RD. The Health On the Net Code of Conduct for medical and health Websites. Comput Biol Med 1998;28(5):603-10.
14. Hersh WR, Gorman PN, Sacherek LS. Applicability and quality of information for answering clinical questions on the Web [letter]. JAMA 1998;280(15):1307-8.
15. Ambre J, Guard R, Perveiler FM, Renner J, Rippen H. Criteria for Assessing the Quality of Health Information on the Internet: Health Information Technology Institute, Mitretek systems; October 14, 1997.
16. Lindberg DAB. Fiscal Year 1999 President's Budget Request for the National Library of Medicine.; March 18, 1998.