International Journal of Medical Informatics (2004) xxx, xxx–xxx

ELSEVIER

International Journal of Medical Informatics

www.elsevier.com/locate/ijmedinf

Instruments to assess the quality of health information on the World Wide Web: what can our patients actually use?

Elmer V. Bernstam[a,*], Dawn M. Shelton[a], Muhammad Walji[a], Funda Meric-Bernstam[b]

[a] *School of Health Information Sciences, The University of Texas Health Science Center at Houston, 7000 Fannin Street, Suite 600, Houston, TX 77030, USA*
[b] *Department of Surgical Oncology, The University of Texas M. D. Anderson Cancer Center, Houston, TX 77030, USA*

Received 14 July 2004; received in revised form 4 October 2004; accepted 22 October 2004

KEYWORDS
Internet;
Medical informatics;
Patient education

Summary
Objective: To find and assess quality-rating instruments that can be used by health care consumers to assess websites displaying health information.
Data sources: Searches of PubMed, the World Wide Web (using five different search engines), reference tracing from identified articles, and a review of the of the American Medical Informatics Association's annual symposium proceedings.
Review methods: Sources were examined for availability, number of elements, objectivity, and readability.
Results: A total of 273 distinct instruments were found and analyzed. Of these, 80 (29%) made evaluation criteria publicly available and 24 (8.7%) had 10 or fewer elements (items that a user has to assess to evaluate a website). Seven instruments consisted of elements that could all be evaluated objectively. Of these seven, one instrument consisted entirely of criteria with acceptable interobserver reliability (kappa $\geq$ 0.6); another instrument met readability standards.
Conclusions: There are many quality-rating instruments, but few are likely to be practically usable by the intended audience.

1. Introduction

Searching for health information online using general-purpose search engines such as Google is the third most common use of the Internet following email and product research [1]. Consumers

* Corresponding author. Tel.: +1 713 500 3901; fax: +1 713 500 3929.
E-mail address: Elmer.V.Bernstam@uth.tmc.edu (E.V. Bernstam).

1386-5056/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.ijmedinf.2004.10.001

are satisfied with their online experience and are making choices based on the information that they encounter [2,3]. Increasingly, clinicians are faced with patients who have been informed (or misinformed) by the Internet. As a result, clinicians, researchers and health care consumers are concerned about the quality and accuracy of online health information [2,4–6]. Surveys show that a physician's recommendation carries a great deal of weight, but few patients are using specific sites recommended by their physician [7,8]. Therefore, consumers who search for health information online, do so without professional guidance.

Physicians may be unwilling to recommend specific sites because content and web addresses change quickly. Recommendations may be out of date or even wrong, unless physicians are willing to spend a great deal of effort reviewing and evaluating online content. A better approach would be to empower patients to evaluate online content for themselves. Therefore, multiple organizations developed quality rating instruments intended to be used by healthcare consumers to evaluate websites that display health information. As a result, there are hundreds of instruments intended to be used by consumers to evaluate online health information. However, it is not known which, if any, instruments can be practically used by consumers searching for health information online.

To be usable, an instrument must at least: (1) be available to consumers; (2) require a limited number of elements to be assessed; (3) all elements must be assessable and (4) the instrument must be readable. These criteria are necessary but not sufficient to ensure usability. In other words, an instrument can satisfy the above criteria, yet fail in another way. However, an instrument that does not satisfy these criteria will be difficult for consumers to use.

Recent studies found that online consumer-oriented health information was often above the expected reading ability of a significant proportion of the US population [9,10]. Many different formulas, including the Flesch Reading Ease Score (FRES) used in this study, have been developed to assess readability. Formulas generally rely on counting the number of syllables per word and the number of words per sentence; lower values suggest that the text is easier to read. The FRES is one of the most common and validated measures of readability which is accepted by the insurance industry for evaluating documents intended to be read by consumers [9]. A lower FRES suggests that the material is more difficult to read. A score of 60 or greater, corresponding to secondary school reading level, is considered to be minimally acceptable for consumer-oriented information.

The goal of this study was to identify quality-rating tools that can be practically used by consumers to evaluate online health information for themselves without professional supervision. Motivated by a desire to help clinicians advise their patients, we performed a systematic review of available quality assessment instruments to identify those instruments that can be used by healthcare consumers to evaluate websites that display health information.

2. Methods

We define an instrument as any evaluative tool for rating website quality. An instrument is composed of one or more criteria. Each criteria consists of one or more elements. An element is an item of information that must be assessed in order to evaluate compliance with the criterion. For example, the instrument commonly known as the "JAMA" benchmarks [11] consists of four criteria: authorship, attribution, disclosure and currency. "Attribution" is defined as "references and sources for all content should be listed clearly, and all relevant copyright information noted". Therefore, in order to assess whether a given website complies with the JAMA benchmarks, a user has to determine whether the site complies with "attribution" which requires answering at least the following questions: (1) "does the site clearly list references and sources for all content?" and (2) "does the site display relevant copyright information?" In other words, to evaluate the single criterion "attribution", a user would have to assess two elements.

2.1. Search methodology

Search strategies were adapted from multiple previous systematic reviews of instruments used to evaluate the quality of online health information [5,12,13]. These strategies included:

- A review of the literature by using PubMed (http://www.ncbi.nlm.nih.gov/PubMed/) from January 1995 through May 2003 using the terms: "Internet Health", "Quality Rating Instruments", "Consumer Health Informatics", "Internet Health Information", and "World Wide Web Consumer Health"
- A search in July 2003 of the first 100 results from five search engines: Google (www.google.com), Lycos (www.lycos.com), Yahoo (www.yahoo.com), Excite (www.excite.com)

and Web Crawler (www.webcrawler.com). Two different search strings were used (only English language results pursued):
  a. {(rate OR rank OR top OR best) and (health)} [5,12],
  b. {(evaluate OR award OR assess) (internet health information quality)} [13].
- A search of American Medical Informatics Association's 1998–2003 annual symposiums for mention of Internet rating instruments [12].
- Connections to relevant articles and author links identified by the web searches [12].
- A search of references of relevant printed articles.
- Sources were examined if they were freely available and did not require subscription or login.

2.2. Data extraction

For all instruments, we identified:

- availability of criteria,
- number of criteria as listed by the instrument,
- number of elements for each criterion.

All instruments were reviewed by the same investigator (DMS). Instruments were included if they had discoverable criteria and appeared to be used for rating health information. We set a maximum of 10 elements at the most that a motivated consumer is likely to be able to practically assess. We based this on the classic observation that few experimental subjects can remember more than nine chunks of information, the famous 7 ± 2 [14].

Lastly, eligible instrument's criteria were evaluated for objectivity and readability. Instruments with 10 or fewer elements were examined by two reviewers to determine their objectivity and adherence to the definition of technical criteria: general, domain-independent criteria (i.e., criteria referring to how the information is presented or what meta-information is provided) [15]. Because the definition of technical criteria is itself open to interpretation, two reviewers independently evaluated each instrument, and differences were resolved by consensus.

In previous work, we determined the inter-observer reliability of 22 commonly used technical quality criteria [16]. In the present study, criteria were considered "objective" if they were associated with acceptable inter-observer reliability score (kappa ≥ 0.6), after appropriate operational definitions were agreed upon. If an instrument contained criteria not previously studied, those criteria were not assigned kappa values, but we retained the instrument for further analysis. Instruments containing one or more criteria associated with (kappa < 0.6) were considered to have poor inter-rater reliability. For readability, we measured the Flesch reading ease and the Flesch-Kincaid reading level using functions provided by Microsoft Word 2000.

3. Results

We found 273 unique instruments. One hundred and seventy eight (65%) were some type of award or kitemark (trustmark or seal of approval) whose criteria were never intended to be applied by Internet users. Only 80 of 273 (29%) instruments publicly disclosed their criteria (Fig. 1). Of these, the number of criteria per instrument ranged from 1 to 53 (mean = 10.25), and the number of elements ranged from 1 to 153 (mean = 24.9). Fifty-four instruments had 10 or fewer criteria and only 24 had 10 or fewer elements. Most instruments had more than 10 elements that had to be assessed and were considered to be too long for routine use. Notably many instruments had a small number of criteria, but required multiple questions be answered for each criterion (i.e., few criteria, many elements to be assessed).

Of the 24 instruments that had 10 or fewer elements, three were eliminated: International Association of Business Communicators Washington (http://www.iabcdc.org/inkwell/rules.html, provides insufficient detail to evaluate criteria), MedIndex Electronic Dictionary (no longer exists, site retained in a temporary web archive) and NeoVizion (http://www.neovizion.com/review, rates design, not content). Therefore, 92% of instruments were either not available or had too many elements to be usable by consumers. This left 21 websites eligible for further review (Table 1).

To assess objectivity, two expert reviewers independently evaluated each of the 21 instruments and agreed that six consisted entirely of objective technical quality criteria (Table 2). Fourteen instruments contained elements that were determined to be subjective (e.g., "author qualified") and therefore could not be expected to be reliably assessable by consumers. The reviewers independently agreed regarding their assessment of all instruments except one. For this one instrument, reviewers disagreed regarding the objectivity of a single element ("Doesn't push a single point of view or sell miracle cures") this instrument was retained for further analysis. Inter-observer reliability (kappa) scores for commonly cited criteria were available from our previously published work [16]. As shown

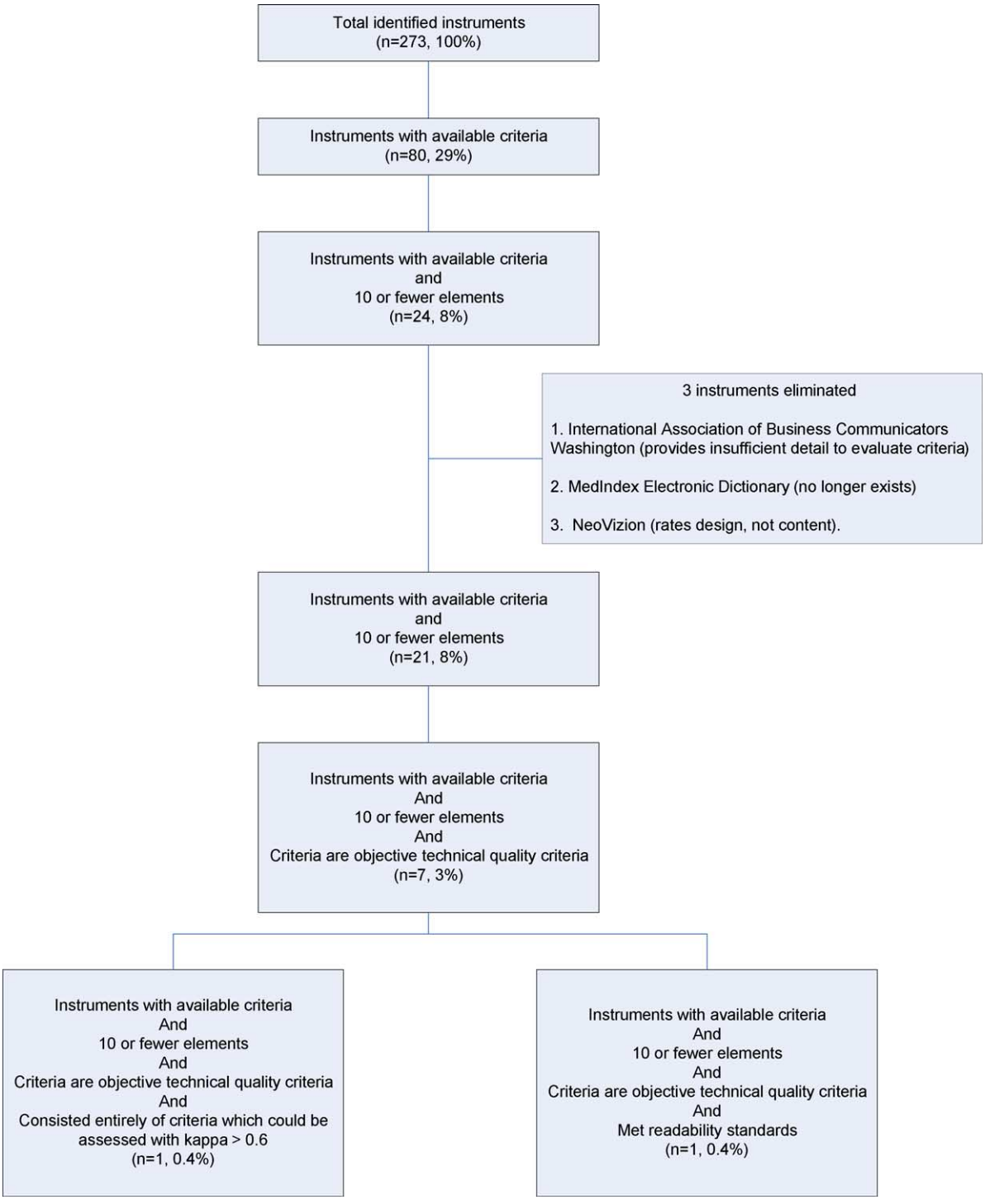

Fig. 1. Number of instruments that met usability goals.

in Table 2, only the Mayo clinic instrument consisted of elements which could be reliably assessed.

Finally, the seven instruments were evaluated for readability. The higher the 'Flesch reading ease' score, the easier a document is to read. The recommended score for the average reader is 60 or greater. The maximum recommended 'Flesch–Kincaid reading level' for consumer-oriented materials is 8th grade [9]. The World Health Organization instrument was the only one to comply with both readability guidelines (Table 3).

Table 1 Description of quality rating instruments with less than 10 elements

Name or source	Web address	Instrument type	Criteria	Elements
Alexa Ranking	www.curezone.com/websites/default.asp	Top traffic	1	1
Steliart Award	www.steliart.com/page2.html	Award	1	1
What is this?	www.whatisthis.com/award.htm	Award	2	2
Consumer's Choice Award	www.consumerschoiceaward.com/aboutus.cfm	Award	1	4
Houston Chronicle	n/a	Tips	3	4
Basic Publishing elements	n/a	Criteria	6	6
Select Surf	www.selectsurf.com/help/#howeval	Award portal	5	5
Advanced Programming Concepts Web Development Excellence Award	apcweb.com/webaward.htm#Judging%20Criteria	Award	5	5
WHO	www.who.int/medicines/library/qsm/who-edm-qsm-99-4/medicines-on-internet-guide.html	Guidelines	5	5
Healing Well.com	www.healingwell.com/editoraward.aspx	Resource	3	6
Best Health Web Ring	www.riverflow.com/besthealth/ring.html	Top 100 Portal	6	6
Hardin MD Clean Bill of Health Site	www.lib.uiowa.edu/hardin/md/submit.html	Award	6	6
Mayo Clinic	www.mayoclinic.com/invoke.cfm?id=HQ00805	Article	3	6
FDA	www.fda.gov/fdac/features/596_info.html	Recommend	6	6
Randifino, Ph.D.	www.imt.net/~randolfi/HealthyWebs.html	Guidelines	4	6
Nephron Information Center	www.nephron.com/goldennephron.html	Award	5	7
Awesome Library editor's Choice	awesomelibrary.org	Award	7	7
Med Rocket Health	www.medrocket.com/health_site_award/health_web_site_excellence_award.html	Award	7	7
JAMA Benchmarks	n/a	Guidelines	4	8
Clark's Summary	n/a	Guidelines	4	8
Nutrition Navigator among the best	navigator.tufts.edu/ratings.htm	Rating Tool	4	9

Table 2 Objective tools with kappa scores

Criteria elements (items)	Mayo	Publishing	WHO	JAMA	Alexa	FDA	Clark	Kappa[a]
Common elements: final 7 instruments								
Date of creation or update disclosed (currency)	X	X	X	X		X	X	1
Disclosure of physician's credentials							X	1[b]
Editorial review process	X					X		0.95[b]
References provided	X	X		X				0.90
Disclosure of ownership	X			X		X		0.86[b]
Sources clear	X	X		X				0.81
Feedback mechanism: web site contact information			X				X	0.81[b]
Copyright notice		X		X				0.79
Disclosure of author's credentials				X				0.78
Disclosure of authorship		X						0.77
Statement of purpose			X					0.58
Disclosure of advertising							X	0.58
Links provided						X		0.53
Disclosure of sponsorship		X	X	X			X	0.52
Disclaimer re: can not substitute for physician's care				X			X	0.52
Disclosure of privacy policies							X	Na
Payment policy/procedures for secure transactions							X	Na
Graphics and multimedia present						X		Na
Does the site charge a fee						X		Na
Higher rank = higher quality					X			Na
Doesn't push a single point of view or sell miracle cures	X							Na

Na = No Kappa available. Please see Table 1 for URLs of the final seven instruments (Publishing = Basic publishing elements).
[a] Source: Sagaram, S., et al., Inter-observer agreement for quality measures applied to online health information. MedInfo, 2004.
[b] Kappa statistic not calculated due to zero variability. Percentage agreement between two raters reported.

Table 3 Readability scores of final instruments

Instrument name or source	Flesch–Kincaid grade (Goal < 8)	Flesch reading ease (goal > 60)
WHO	7.2	62.9
Basic Publishing	9.9	44.3
Mayo Clinic	10.3	39.9
FDA	11	47.4
Alexa	11.5	43.8
JAMA Benchmarks	12	28.6
Clark Summary	12	2.7

4. Discussion

Although many quality assessment instruments have been published by a variety of organizations, few can actually be used by health care consumers. At the time of our study, many instruments did not make their criteria available, likely because they are intended as awards and kitemarks, rather than instruments to be used by consumers. Of the 273 instruments we found, only 80 (29%) disclosed their criteria. Since only 24 instruments (8%) had 10 or fewer elements, we concluded that the majority of published instruments have too many elements to be used routinely.

We make special note of Alexa (http://www.alexa.com), which rates the popularity of websites based on measures of user traffic. Although Alexa may have met our very loose definition of technical quality criteria, our previous work shows that popularity and quality as represented by domain-independent technical quality criteria are not related [17]. Therefore, we do not consider Alexa to be a quality assessment tool that can benefit consumers.

Our study is limited by the fact that a single reviewer (DMS) determined the eligibility for inclusion and the number of elements in each quality assessment instrument. Both of these may, in some cases, be subjective. To mitigate this limitation, we were liberal in including instruments (e.g., Alexa as a quality rating instrument even though it actually rates popularity). Consequently, we discovered more instruments (273), compared with recent reports such as Jadad (2002) who found 98 quality rating instruments [12].

With respect to determining the number of elements/criterion, if the number of questions that need to be answered in order to establish compli-

ance with a criterion is subjective, then that criterion is likely subjective and not reliably assessable by consumers. However, we picked a very conservative limit of 10 elements as discussed above.

Most studies of online health information, including ours, are limited by the constantly changing nature of the Internet. Therefore, if our study were repeated, the findings may be different. For this reason, it is more important to focus on generalizable knowledge (e.g., quality assessment instruments), rather that describe the state of the Internet (e.g., quality of specific sites). For example, just a few months after our study, a disease-specific instrument for rating information about diabetes was published [18]. However, this instrument contains at least 17 elements which must be assessed and was validated using trained reviewers. We hope that our findings will encourage research regarding quality assessment tools that can be used by consumers without requiring undue effort, time, or specialized training.

A strength of our study is that we cast a very broad net to capture many available quality assessment instruments that may be of use to consumers. We combined multiple methodologies and identified more instruments than previous studies. In addition, we focused on usability, which to our knowledge, has not previously been assessed. Further, to mitigate any potential bias and inaccuracy in assessment, we picked very conservative thresholds for eliminating an instrument from subsequent evaluation.

Multiple previous studies found that Internet users do not assess quality when searching for health information online [17,19]. Future research should focus on developing instruments and/or techniques that allow consumers to evaluate online health information. If we are to empower consumers to evaluate online health information for themselves, we can no longer ignore the issue of whether consumers can actually use the tools that we place at their disposal.

Acknowledgements

This work is supported in part by a grant from the Robert Wood Johnson Foundation Health-e-Technologies initiative (E.V.B, F.M.-B) and in part by a training fellowship from the Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (NLM Grant No. 5T15LM07093) (M.W.).

References

[1] S. Fox, D. Fallows, Internet health resources. Pew Internet and American Life Project, Washington, DC, 2003.
[2] S. Fox, L. Rainie, The online health care revolution: How the Web helps Americans take better care of themselves, Pew Internet and American Life Project: Online, Washington DC, 2000, pp. 3—7.
[3] P.R. Helft, et al., Hope and the media in advanced cancer patients, in: American Society of Clinical Oncology 36th Annual Meeting, New Orleans, LA, 2000.
[4] J.S. Biermann, et al., Evaluation of cancer information on the Internet, Cancer 86 (3) (1999) 381—390.
[5] A.R. Jadad, A. Gagliardi, Rating health information on the Internet: navigating to knowledge or to Babel? JAMA 279 (8) (1998) 611—614.
[6] S.L. Price, W.R. Hersh, Filtering Web pages for quality indicators: an empirical approach to finding high quality consumer health information on the World Wide Web, Proc. AMIA Symp. (1999) 911—915.
[7] P.C. Tang, C. Newcomb, Informing patients: a guide for providing patient health information, JAMIA 5 (6) (1998) 563—570.
[8] S. Reents, Impacts of the Internet on the doctor-patient relationship: The Rise of the Internet Health Consumer, Cyber Dialogue, Inc., 1999.
[9] M.A. Graber, D.M. D'Allessandro, J. Johnson-West, Reading level of privacy policies on Internet health Web sites, J. Family Practice 51 (7) (2002) 642—645.
[10] G.K. Berland, et al., Health information on the Internet: Accessibility, Quality, and Readability in English and Spanish, JAMA 285 (20) (2001) 2612—2621.
[11] W.M. Silberg, G.D. Lundberg, R.A. Musacchio, Assessing, controlling, and assuring the quality of medical information on the Internet: caveant lector et viewor—Let the reader and viewer beware (editorial), JAMA 277 (15) (1997) 1244—1245.
[12] A. Gagliardi, A.R. Jadad, Examination of instruments used to rate quality of health information on the internet: chronicle of a voyage with an unclear destination, BMJ 324 (7337) (2002) 569—573.
[13] P. Kim, et al., Published criteria for evaluating health related web sites: review, BMJ 318 (7184) (1999) 647—649.
[14] G.A. Miller, The magical number seven, plus or minus two: some limits on our capacity for processing information, Psychol. Rev. 63 (1956) 81—97.
[15] G. Eysenbach, et al., Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review, JAMA 287 (20) (2002) 2691—2700.
[16] S. Sagaram, et al., Inter-observer agreement for quality measures applied to online health information, MedInfo (2004).
[17] F. Meric, et al., Breast cancer on the world wide web: cross sectional survey of quality of information and popularity of websites, BMJ 324 (7337) (2002) 577—581.
[18] J.J. Seidman, D. Steinwachs, H.R. Rubin, Design and testing of a tool for evaluating the quality of diabetes consumer-information web sites, J. Med. Internet Res. 5 (4) (2003) 30.
[19] G. Eysenbach, C. Kohler, How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews, BMJ 324 (7337) (2002) 573—577.