AN EVALUATION OF THE QUALITY OF CONSUMER HEALTH INFORMATION ON WIKIPEDIA

by

Daren T. Nicholson, M.D., C.M.

A CAPSTONE

Presented to the Department of Medical Informatics and Clinical Epidemiology

and the Oregon Health & Science University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Masters of Biomedical Informatics

March 2006

School of Medicine

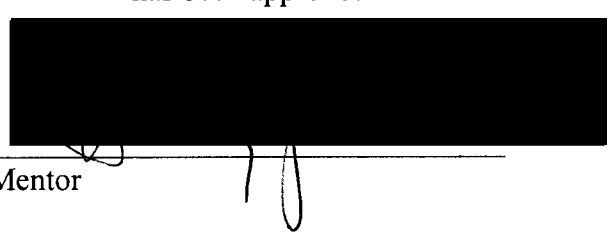
Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify the the Master's capstone of

Daren T. Nicholson

has been approved



Mentor

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my mentor, Holly Jimison, for supporting this project enthusiastically. Second, I must express sincere gratitude to three people who volunteered their time to help me complete this project: Aseem Kumar, Ken Guappone, and Greg Sicard. Finally, I would like to express warm appreciation for the general support that Michelle Lee provided me.

ABSTRACT

Background

Wikipedia is a multilingual, open-content, online encyclopedia that exists as a wiki. It is written collaboratively by people with varying degrees of expertise. Indeed, anyone who can access Wikipedia's Web site may alter any of its content. Wikipedia contains a significant number of articles on health-related topics, but the quality of this information is unknown.

Purpose of the Study

The accuracy and completeness of the information on Wikipedia is intuitively circumspect since experts and non-experts alike may contribute to the site's content. Thus, the intent of my research was to systematically evaluate the accuracy and completeness of a sample of health-related articles on Wikipedia.

Methodology

I selected a previously published methodology for use in my study. The advantage of this choice is that I could use the results from that study as control data for my evaluation of Wikipedia. As an additional control, I evaluated the content of the Healthwise® knowledge base (a collection of consumer health articles that is esteemed by many health professionals for its high-quality).

The articles reviewed concerned the following four health topics: breast cancer, childhood asthma, depression, and obesity. Evaluation criteria were defined a priori. For each health topic, several key elements (totaling 100 across all four topics) were identified as desirable components of the content. A panel of health professionals and consumer advocates developed this list of elements. For each element, a rater scored the

completeness of that element's coverage. Completeness scores were integers ranging from zero to two (0 = element not addressed, 1 = element minimally addressed, and 2 = element more than minimally addressed). For those elements that received a completeness score of one or two (in other words, those elements that were address by the given Web site), the raters gave accuracy scores ranging from zero to two (0 = element mostly incorrect, 1 = element mostly correct, and 2 = element entirely correct). Three physicians rated each article from Wikipedia and Healthwise. The raters were blinded as to the source of the articles.

Results

Wikipedia fared poorly with regard to completeness; however, all elements covered by Wikipedia were deemed either mostly or entirely correct. Healthwise was only slightly more complete than Wikipedia, but its content also proved to be accurate.

Discussion

Surprisingly, the results contradict the assumption that Wikipedia's health content would contain factual inaccuracies. However, Wikipedia's low coverage rates were not unanticipated. Wikipedia articles are works-in-progress that grow over time from "stub" status to more complete documents. One might argue that many of the clinical elements are missing from the Wikipedia articles because these articles are in fact still in the maturing process. In light of Wikipedia's poor completeness scores, consumers ought not rely solely on this Web site as their source of health information. However, consumers can be reasonably confident that the information found on Wikipedia is accurate.

INTRODUCTION

The quality of consumer health information on the Internet is a core issue in the field of consumer health informatics (CHI). Indeed, the medical literature is saturated with studies that assess the quality of online health information. (Given that this topic, the quality of consumer health information on the Internet, is necessarily wordy, I will use the acronym QCHII as a convenient alias when referring to this subject.) Fortunately, Eysenbach has recently published an excellent systematic review of the QCHII literature. The overall conclusion of this study was discouraging: "In our review, most authors who evaluated content found significant problems, criticizing lack of completeness, difficulty in finding high-quality sites, and lack of accuracy, in particular if 'accuracy' also implied 'completeness.'"[1] Explanations for the generally poor quality of online health information tend to focus on actors (i.e., the individuals and groups who publish the information) as opposed to processes (i.e., the means by which the information is generated). For example, three of the four Silberg criteria (a popular method for judging the quality of a health-related Web site) relate to actors:

- Authorship (those people who create the content).
- Attribution (other people who might be referenced within the content).
- Sponsorship (individuals or groups who finance or otherwise support the authors).

The medical informatics literature is woefully lacking in research that considers the process by which these actors generate online health information. Thus, the central theme of this study will be this question of process.

Health information permeates all media, and the Internet is no exception to this rule. Although there are many individuals who publish health information on the Web (via

blogs or other flavors of personal Web sites), various types of organizations sponsor the vast majority of health-related Web sites. These organizations include government institutions, non-governmental organizations, pharmaceutical companies, educational institutions, and disconcertingly, many purveyors of the latest nostrums. What unifies these disparate varieties of sponsorship is their shared model of information distribution. Each of these sponsors represent themselves as a small group of "experts" who are sharing their knowledge with the "lay" public. Indeed, this hierarchical model has dominated knowledge sharing throughout human history.

To the delight of some and the displeasure of others, the hierarchical model of knowledge transfer has been challenged by a new model, which masquerades itself with many fancy names such as peer collaboration and communal knowledge development. Perhaps a name more familiar to the average person would be "grass-roots development." Essentially, this method flattens the hierarchy into a single layer. Laypeople (and sometime experts as well) collaborate to generate and distribute information. The egalitarianism of the Internet has caused this grass-roots model to surge. One of the most intriguing results of this novel trend in knowledge distribution is a new Internet-based technology poetically named the wiki. This new and oddly named competitor has already entered the race to provide high quality consumer health information via the Web.

Wikis

In simplest terms, wikis are Web sites whose content may be altered (via additions, subtractions, and/or rearrangements) by any user. In more technical terms, users have both read- and write-access to a given wiki. The wiki's *raison d'être* is to foster a

collaborative authorship of Web content. Ward Cunningham created the first wiki to encourage collaboration among software programmers. Cunningham's invention has led to the development of all types of wikis that serve various communities from college classrooms to medical informatics enthusiasts. The technical capabilities and functions of each wiki vary, but one of the guiding principles behind wikis is that anyone with word processing skills should be capable of contributing. Most wikis provide a means of viewing a history of changes to the wiki. Usually, the wiki identifies each version of a given page with its author (either by user-defined aliases or IP addresses).

Until recently, wikis were the purview of geeks and mostly unknown to the common Web surfer. However, one particular wiki has introduced the concept to the Web masses: Wikipedia.[2] Over the course of five years, this online encyclopedia has emerged as a formidable opponent to giants like the *Encyclopædia Britannica* and Microsoft's *Encarta*. Originally, Wikipedia's founder, Jimmy Wales, sought to build a free online encyclopedia called Nupedia. Like its traditional competitors, *Encarta* and *Britannica*, Nupedia was to be written by "experts;" however, these experts were supposed to volunteer their authorship. Some did volunteer, but predictably, the project failed due to low participation. In January 2001, Wales launched Wikipedia to supplement the work done by Nupedia authors. Over the course of the past five years, the encyclopedia (now re-dubbed as Wikipedia) has rapidly expanded to include approximately 2.5 million articles in over 100 languages. The lengths of these articles range from a few sentences to several dozen paragraphs. Indeed, many of the articles have grown to be so long that sub-topics have spun off as separate articles. Already, the breadth of Wikipedia far exceeds that of traditional encyclopedias like *Britannica*.[3]

Wikipedia's popularity has skyrocketed over its short lifetime. Hitwise, an organization that tracks Internet usage, rates Wikipedia as the second-most visited reference Web site on the Net.[4] Traffic to the site has outpaced that of media giants such as the New York Times.[5] Alexa, a service that tracks Internet traffic, ranks Wikipedia as the 32nd most frequently visited Web site.[6] Perhaps the best measure of its popularity, however, is its ranking among Google searches. For many topics (especially names of famous people, places, and current events) Wikipedia consistently ranks in the top ten Google results. Although the results of health searches on Google continue to be dominated by traditional Web sources, Wikipedia's rank is steadily rising. Wikipedia articles on obscure medical topics tend to rank higher than common health subjects (on November 2, 2005 abasia ranked 10th and pseudocholinesterase deficiency ranked 8th); however, even some common health issues rank high (a search for "pneumonia" ranked Wikipedia 8th). If Wikipedia's popularity continues to grow at its current phenomenally rapid pace, it is highly likely that this Web site will become a major source of consumer health information.

The question that invariably arises in every Wikipedia-initiate's mind is: can one trust the quality of the information on Wikipedia? The argument for a hierarchical model of information distribution is intuitive; "experts" have accurate, current, detailed knowledge of a subject and are therefore the best source of information on that topic. There are two important assumptions underlying this argument: 1) we have a means of verifying expertise and 2) we trust that "experts" are honest when they share their information. Naturally, the first assumption does not apply to Wikipedia. There is no system in place to confirm the expertise, or lack of expertise, of any contributor to the wiki. (Wikipedia's

detractors, who noticeably tend to fall into the category of "experts," consistently bemoan this feature of Wikipedia.) One might argue that assumption two does not apply to Wikipedia either. We trust "experts" to supply bona fide information based on a simple premise: it is in the interest of "experts" to be honest since a dishonest expert risks losing the rewards and privileges of her expertise if she were discovered to be dishonest. Contribution to Wikipedia is virtually anonymous; thus, dishonesty may go unpunished. (There are, however, certain means of blocking users who repeatedly post misinformation on the wiki.) What these arguments reveal is that the concept underlying Wikipedia directly challenges a deeply-ingrained social system. We unquestionably accept the system that produces "experts" who then attempt to share their knowledge with non-experts. The same can not be said for Wikipedia's grass-roots model of knowledge distribution. Indeed, many argue that Wikipedia's usefulness is extremely deficient given the lack of accountability enjoyed by its authors.

What, then, are the arguments in favor of Wikipedia? The principal contention may be put simply: volume begets quality. Take, for example, the issue of accuracy. Striving for accuracy, most authors attempt to check the facts of their writing against trusted sources. One person may be able to adequately verify facts, but imagine if ten fact checkers were employed and their results were compared. It seems likely that the combined efforts of ten checkers would be superior to that of a sole checker. Now imagine tens or hundreds of fact checkers at work on the same piece of writing. This is the case for Wikipedia. As the argument goes, with so many people duplicating the fact checking effort, the accuracy of the article is bound to be better than the work of a solo checker. The underlying assumption here is that trusted reference sources are easily

accessible to anyone who wishes to look for them. In the networked world, this assumption may appear to be a foregone conclusion.

Now lets reexamine how the second assumption supporting the "expert" model applies to Wikipedia. How can we trust Wikipedia's contributors? Proponents of Wikipedia would argue that trust is not an issue. Some contributors may maliciously post misinformation on the wiki; however, the brigade of fact checkers will easily spot the misinformation and quickly correct it. The assumption here is that individuals may have good or bad intentions but that the results of the group's collaboration is always good. This is certainly debatable, but is widely accepted by Wikipedia proponents. Thus, once again the large number of contributors to Wikipedia ensures a high quality product.

A heated debate has developed between Wikipedia's zealots and their opponents. Arguments for and against the Web site abound on the Internet, but the debate has also been featured in traditional media sources such as National Public Radio's *Talk of the Nation*.[7] Indeed, the animated discussions about Wikipedia prompted the venerated journal *Nature* to conduct a study comparing the accuracy of Wikipedia to that of the *Encyclopædia Britannica*. *Nature* invited experts to judge the accuracy of 42 science entries (8 of which were related to health) from each encyclopedia. Wikipedia was found to be nearly as accurate as *Britannica* (Wikipedia articles averaged 4 inaccuracies per article while *Britannica* averaged 3 errors). This study was, to the author's knowledge, the first and only peer-reviewed assessment of the accuracy of information on Wikipedia.[8]

This vigorous debate over Wikipedia's quality and usefulness may soon be moot given the Web site's incredible inertia. People from all walks of life are beginning to use the site, not only for personal edification, but also for professional activities. Indeed,

some online media sources have begun to regularly site Wikipedia as background information sources (e.g., *Jurist*, an online news magazine from the University of Pittsburgh School of Law).[9] Wikipedia appears to be on the brink of joining the ranks of the Internet monopolies such as Google, eBay, and Amazon. The skeptics' complaints are slowly being drowned out by the millions of users' footfalls.

Assessing the quality of health care information on the Internet

The quality of consumer health information on the Internet has been the focus of numerous publications over the past decade. Indeed, Eysenbach and colleagues' recent meta-analysis identified 79 published studies that quantitatively evaluated the quality of health-related Web sites. Unfortunately, there is no consensus among these authors concerning the definition of quality nor the means by which one should judge quality. The Eysenbach meta-analysis compiled no less than 86 different measures of quality used in those 79 studies. Despite the diversity of quality criteria, several themes are oft-repeated in this large body of literature. A common definition of high quality may best be defined by four components: accuracy, completeness, currency, and relevancy. In the opinions of many authors, accuracy subsumes completeness. In other words, the content of a given piece of material may be entirely correct; however, if important information is missing from the content, the information may be misleading and therefore not accurate. Currency is closely related to accuracy. Medical knowledge is perpetually evolving. Frequently, health-related information published on Web sites is out of date. This may result in omissions of important, novel information or outright factual errors when new evidence contradicts previously held beliefs. Thus, lack of currency may lead to inaccuracy.

Finally, relevancy is an extremely important aspect of quality. Information on a Web site may be current, complete, and accurate, but it may be inconsequential to the site's intended audience. For instance, a consumer health Web site might properly explain how to stage breast cancer, but this type of information is more appropriate for health care practitioners as opposed to consumers. More relevant information for patients would, for instance, discuss the treatment options available for a given stage of cancer.

Having agreed on a basic definition of quality, one must define the means by which one will judge the quality of a given Web site. Despite the plethora of different rating tools (in 1999 Kim and colleagues identified 29 sets of criteria published on the Internet or in the medical literature),[10] rating methodologies can be separated into two broad categories: those that directly assess content and those that use proxy measures. The differences within the former group are numerous. First, the qualifications of the raters vary widely. Sometimes all raters are board-certified physicians. Other times clinicians such as nurses or physician assistants may act as raters. The number of raters employed for a study also varies widely. As previously noted, the criteria used to judge quality are numerous and inconsistent. Furthermore, QCHII studies differ based on when the quality criteria are defined. As Eysenbach notes, authors "either define clear criteria beforehand (a priori) or extract information from the Web first and then check these claims against [a reference source] (a posteriori)."[1] Indeed, the reference sources employed by each study also vary widely. Some studies judge accuracy based on the raters' expert knowledge only. Others use a gold standard reference (such as clinical guidelines) by which to judge the accuracy of Web site information. Clearly, the diversity of methods used in QCHII

studies makes it difficult to compare studies let alone make general inferences from this body of literature.

Quality assessments that include detailed analysis of the accuracy and completeness of a source's content can be labor intensive. Therefore, some researchers use proxy measures to judge the quality of health-related Web sites. A frequently cited list of core proxy measures was first published by Silberg and colleagues in 1997, shortly after the public emergence of the Internet. The Silberg criteria consist of four elements: authorship, attribution, disclosure, and currency. Thus a high quality health information Web site should: 1) list all authors with their respective credentials and affiliations, 2) cite all reference sources, 3) disclose all financially-related information such as Web site owners and sponsors, and 4) publish the "dates that content was posted and updated." Silberg and colleagues readily admit that these proxy measures "are no guarantee of quality in and of themselves;" however, the authors proposed these criteria as a benchmark that can easily be applied by both health care professionals and health care consumers to rapidly assess the quality of a given health-related Web site.[11] As noted by Kim and colleagues, many authors also consider the quality of a Web site's design and aesthetic elements as an important proxy measure of overall quality.[10] This meta-data approach to assessing the quality of a Web site is common to the rating instruments published on the Internet such as the Health on the Net Foundation's HONcode.

Purpose of this study

Despite the lack of accord among researchers concerning what constitutes a high-quality health Web site, the Eysenbach meta-analysis concludes that the quality of consumer

health information on the Internet is generally poor. One reason for the inadequacy of health-related content on the Web may be funding. Publishing high quality health information is costly (primarily because the experts who author the information are highly-paid). Given the abundance of free content already on the Web, consumers may be reluctant to pay for health information even if a fee-based product boasts of quality superior to that of the currently available free information. Wikipedia's grass-roots approach to information distribution promises to be a viable solution to this financial challenge. The open-source software movement is evidence that a collaboration of volunteers can provide consumers with high-quality products for free or minimal cost. The question we should ask ourselves is, does the health-related content on Wikipedia offer the same high standard of quality as do open-source stars such as Linux and Apache? This study will attempt to answer that question.

MATERIAL AND METHODS

Scope

A first-rate QCHII study requires an enormous amount of time to complete. The first step, designing an effective sampling technique, is necessarily time-consuming since the Web is littered with millions of health-related sites. Most often researchers use search engines to identify potential sites for review. Choosing appropriate search strings necessitates thoughtful planning. Since search engines usually return millions of results for a given search string, one must select a strategy to limit the sample to a reasonable size. Sample selection is further complicated by the fact that search results commonly include irrelevant sites. Identifying the relevancy of a given Web site may require someone to download and read (at least portions) of that site.

The second step of any QCHII process, rating the quality of the sampled Web sites, also entails a significant investment of time. Many, if not most Web sites, contain vast amounts of information, frequently comparable to that of a short book, and thus rating a single Web site may take anywhere from a few minutes to hours depending on the thoroughness of one's evaluation. Furthermore, the organization of content on a Web site often lengthens the review process. Reviewers may spend considerable time navigating through links to find relevant material. Rigorous QCHII studies employ multiple raters, thereby multiplying the length of the review process.

Given that the process of a QCHII evaluation is lengthy, and given my limited resources, I have elected to conduct a narrowly-scoped QCHII study (effectively a pilot study). First, for the purposes of my study, I have reduced the definition of quality to two of its elements: accuracy and completeness. Second, in this study I will only be evaluating

health-related content from two sources: Wikipedia and the Healthwise®[12] knowledge base. I will use data from a previously published study to compare the quality of Wikipedia to that of traditional Web sites. I have chosen to evaluate material from Healthwise because many experts consider it to be a gold-standard of high-quality consumer health information. Healthwise's quality score will be a benchmark by which to judge the quality of information on Wikipedia and the comparative Web sites. The content in the Healthwise knowledge base spans the full spectrum of consumer health information. Healthwise licenses the content in its knowledge base to other organizations, such as Kaiser Permanente® and WebMD®, for various consumer health information purposes.

Selection of comparative Web sites

As previously mentioned, a substantial number of researchers have already rated the quality of traditional health Web sites (i.e., those created by the hierarchical model of information transfer). Given the diversity of methodologies employed by QCHII studies, my choice of a comparative study was intentional. First, I reviewed the literature for studies that appeared highly rigorous. My criteria for rigor included: 1) a thoughtful explanation of the search strategy, 2) a large sample size of Web sites reviewed, and 3) a relatively large number of raters (three or more). Next, I narrowed down my options to studies whose methodologies were precisely explained. Finally, of the few remaining studies, I selected the one that I could repeat most easily given my limited resources. My final selection was a study published in JAMA in 2001 by Berland and colleagues.[13]

The Berland study was a complex evaluation of Web sites that included three phases: 1) an assessment of the relevancy of Web sites identified via Web search engines, 2) an evaluation of the accuracy and completeness of information on these Web sites, and 3) a review of the Web sites' readability. My study attempted to apply the methodology employed by Berland and colleagues for the second portion of their study.

Health Topic Selection

Berland and colleagues chose to focus their study on four health topics: breast cancer, childhood asthma, obesity, and depression. The authors selected these topics due to their "prevalence, clinical significance, and diversity of the affected populations."[13] Thus, I reviewed the articles from Wikipedia and Healthwise that address these four topics.[14]

Accuracy and Completeness Criteria

For each of the condition-related topics, Berland and colleagues assembled a panel of three to four people to identify *a priori* criteria by which to judge accuracy and completeness. The panels consisted of expert clinicians and representatives of patient-advocacy organizations. The criteria development process involved three stages. In stage one, the panelists identified important sub-topics for each of the four health topics (26 total). For example, they listed screening as one of the sub-topics for breast cancer. For each of these sub-topics, the panelists generated one or more consumer-oriented questions related to the sub-topic. For instance, for the sub-topic "breast cancer screening," the following question was generated: "Do I need [a mammogram] every year?" In the third phase of criteria development, the panelists identified key clinical elements that would answer the consumer-oriented questions generated in phase two. In the case of the

question concerning mammogram frequency, the corresponding element stated the following: "Women >50 years should have mammograms every 1-2 years." A range of one to eight clinical elements were written for each of the questions, with a total of 100 clinical elements across the four subjects areas. These clinical elements represent the specific criteria by which the reviewers in my study judged the accuracy and completeness of the selected Web sites.[13]

Web Site Rater Selection

In accord with Berland and colleagues' study design, I recruited physicians to act as raters. Of the three physicians that I recruited, two of them are board-certified in internal medicine. The third is beginning his first year of post-graduate training.

Web Site Retrieval

Using standard Internet search engines, Berland and colleagues identified 110 Web sites to review for accuracy and completeness.[13] As previously explained, our study seeks to compare the accuracy and completeness of articles from two sources, Wikipedia and Healthwise, to the accuracy and completeness of the Web sites identified by Berland et al. Pursuant with the Berland methodology, the relevant textual content from Wikipedia and Healthwise were copied into a word processor. Information that might identify either source was removed prior to review by the raters. The raters were provided with both a paper and electronic copy of the abstracted articles.

Evaluating Accuracy and Completeness

In accordance with the Berland methodology, I gave standardized evaluation forms to the reviewers to rate the accuracy and completeness of the Wikipedia and Healthwise articles. For each of the four condition-related topics, I created a form, which listed the sub-topics and associated clinical elements. The raters' task was to give two scores for each clinical element: one for completeness and one for accuracy. Completeness scores were integers ranging from zero to two. Clinical elements that were not addressed were scored with a zero. A score of one was assigned to those elements that were minimally addressed. Those elements that were addressed more than minimally received the top score of two. For those elements that received a completeness score of one or two (in other words, those elements that were addressed by the given Web site), the raters also gave accuracy scores. Like the completeness scores, the accuracy scores were integers ranging from zero to two. The raters assigned a two (the top score) to information that they deemed to be completely correct. For information that was not entirely correct but mostly correct, the raters designated a score of one. Finally, the raters gave a score of zero to information that was mostly incorrect.[13]

Analysis

For purposes of analyzing the accuracy and completeness scores, I averaged the scores across raters (using a median statistic) for each element. Thus, for each source (Wikipedia and Healthwise) 100 sets of median accuracy and completeness scores were calculated (one set of scores for each of the 100 elements). Statistical analysis of the data was conducted using Microsoft Excel 2003 and SPSS 13.0.

RESULTS

Coverage of topics

The majority (64%) of the 100 total clinical elements were not covered by Wikipedia. Healthwise fared much better than Wikipedia (41% of elements were not covered). By comparison, the average percentage of clinical elements that were not covered by the English-language Web sites reviewed by Berland et al. was 24. Interestingly, although Wikipedia's coverage of breast cancer, childhood asthma, and depression was much poorer than that of Healthwise, Wikipedia did cover slightly more elements related to obesity than did Healthwise (see Table 1). In terms of the percentage of elements that were covered more than minimally (coverage score = 2), both Wikipedia and Healthwise performed worse than the average Web site reviewed by Berland et al (see Table 3).

Table 1 – Percentage of elements not covered (coverage score = 0)

Topic	Wikipedia	Healthwise	Mean for sites studied by Berland et al.
Breast cancer	71	40	16
Childhood asthma	53	22	27
Depression	71	29	20
Obesity	67	70	35
Total	64	41	24

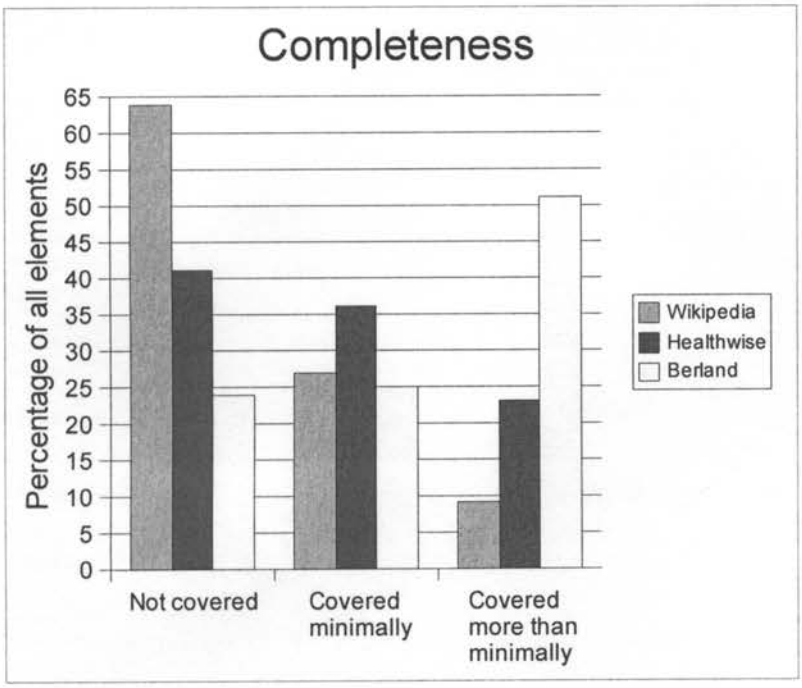
Table 2 – Percentage of elements covered minimally (coverage score = 1)

Topic	Wikipedia	Healthwise	Mean for sites studied by Berland et al.
Breast cancer	19	20	17
Childhood asthma	41	59	30
Depression	24	33	27
Obesity	16	20	25
Total	27	36	25

Table 3 – Percentage of elements covered more than minimally (coverage score = 2)

Topic	Wikipedia	Healthwise	Mean for sites studied by Berland et al.
Breast cancer	10	40	67
Childhood asthma	6	19	43
Depression	5	38	53
Obesity	17	10	40
Total	9	23	51

Figure 1



Accuracy of information

Overall, Wikipedia and Healthwise fared reasonably well on correctness scores (i.e., the accuracy of the clinical elements that were addressed at least minimally by each information source). For each of these two sources, slightly more than half of the covered elements were completely correct with the remaining elements being mostly correct. None of the elements covered by Wikipedia and Healthwise were rated as mostly incorrect (see Tables 4-6). The Berland-reviewed sites received much higher correctness scores than Wikipedia and Healthwise. On average 84% of the Berland-reviewed sites were rated as completely correct. However, on average, two percent of the elements covered by these sites were rated mostly incorrect (see Tables 4-6).

Table 4 – Percentage of covered elements that were mostly incorrect (correctness score = 0)

Topic	Wikipedia		Healthwise		Mean for sites studied by Berland et al.
Breast cancer	0	(0 of 5 elements)	0	(0 of 11 elements)	0
Childhood asthma	0	(0 of 15 elements)	0	(0 of 24 elements)	3
Depression	0	(0 of 6 elements)	0	(0 of 15 elements)	3
Obesity	0	(0 of 10 elements)	0	(0 of 9 elements)	3
Total	0	(0 of 36 elements)	0	(0 of 59 elements)	2

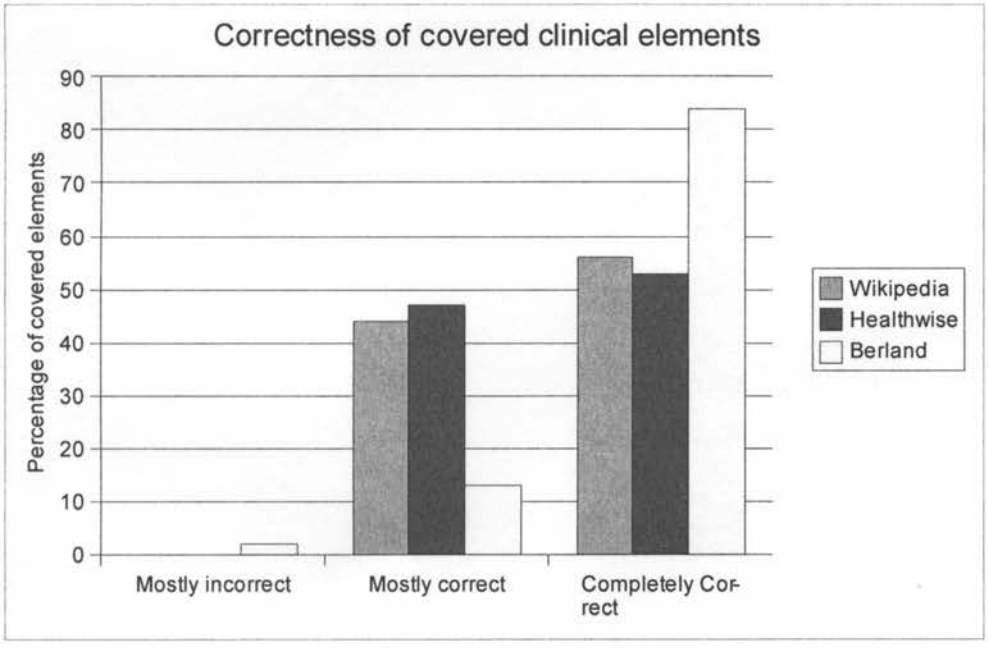
Table 5 – Percentage of covered elements that were mostly correct (correctness score = 1)

Topic	Wikipedia		Healthwise		Mean for sites studied by Berland et al.
Breast cancer	40	(2 of 5 elements)	45	(5 of 11 elements)	9
Childhood asthma	87	(13 of 15 elements)	75	(18 of 24 elements)	13
Depression	17	(1 of 6 elements)	33	(5 of 15 elements)	22
Obesity	0	(0 of 10 elements)	0	(0 of 9 elements)	11
Total	44	(16 of 36 elements)	47	(28 of 59 elements)	13

Table 6 – Percentage of covered elements that were completely correct (correctness score = 2)

Topic	Wikipedia		Healthwise		Mean for sites studied by Berland et al.
Breast cancer	60	(3 of 5 elements)	55	(6 of 11 elements)	91
Childhood asthma	13	(2 of 15 elements)	25	(6 of 24 elements)	84
Depression	83	(5 of 6 elements)	67	(10 of 15 elements)	75
Obesity	100	(10 of 10 elements)	100	(9 of 9 elements)	86
Total	56	(20 of 36 elements)	53	(31 of 59 elements)	84

Figure 2



Coverage and accuracy combined

Ideally, a high-quality consumer health information Web site would receive a score of 2 for both completeness and correctness for each clinical element. Wikipedia fell far below the ideal with only 9% of elements that were more than minimally covered and completely correct. Healthwise performed better with 26% of elements scored as both more than minimally covered and completely correct. On average, nearly half of the elements in the Berland-reviewed sites received a score of 2 for both completeness and correctness (see Table 7).

Table 7 – Percentage of elements that were more than minimally covered and completely correct (coverage score = 2 and correctness score = 2)

Topic	Wikipedia	Healthwise	Mean for sites studied by Berland et al.
Breast cancer	6	35	63
Childhood asthma	6	19	36
Depression	5	33	44
Obesity	17	23	37
Total	9	26	45

DISCUSSION

The average completeness of the Web sites reviewed by Berland et al. exceeded both Healthwise and Wikipedia. Healthwise did cover the majority of clinical elements, whereas Wikipedia only addressed a dismal 36% of elements. On the other hand, the content presented by both Wikipedia and Healthwise was mostly accurate. The results reveal an interesting finding: the obesity articles stand apart from the other three topics. Healthwise and the Berland-reviewed sources performed worse on coverage of obesity compared to coverage of the other three topics. Wikipedia's coverage of obesity, by contrast, was better than its coverage of the other three topics. Indeed, Wikipedia out-performed Healthwise with regard to the completeness of the obesity topic.

Healthwise's relatively poor coverage, as compared to the average Web site reviewed by Berland and colleagues, was not expected. As previously mentioned, Healthwise has a reputation of being an excellent source of consumer health information. Indeed, Healthwise supplies content to the Web sites of many large health-related organizations such as WebMD and Kaiser Permanente. To explain these surprising results, one might speculate that some of the clinical elements defined by the Berland study are not relevant to consumers and thus were not included in the consumer-focused content provided by Healthwise. However, this hypothesis appears baseless given that a panel of experts, which included members of patient advocacy organizations, compiled the list and that the panel derived the elements from consumer-oriented health questions. Thus, further investigation into the completeness of Healthwise's content seems warranted.

Wikipedia's extremely low coverage rates are much less surprising. Unlike Healthwise, Wikipedia's stated purpose is not to provide consumer health information.

Rather, Wikipedia endeavors to deliver general information that is accessible to a wide audience. Perhaps a better explanation for Wikipedia's incompleteness is the process by which the content is developed. Wikipedia articles are works-in-progress that grow over time from "stub" status to more complete documents. Unlike traditional Web sites that typically write a complete article before posting it on the Internet, Wikipedia publishes each step in the development of an article. Thus, one might argue that many of the clinical elements are missing from the Wikipedia articles because these articles are in fact still in the maturing process. A comparison of the lengths of the Wikipedia articles to those of Healthwise would support this argument (see Table 8).

Table 8 - Number of words per article

Topic	Healthwise	Wikipedia
Breast cancer	5999	2563
Childhood asthma	10007	4081
Depression	6120	5138
Obesity	6109	5803

The lengths of Healthwise's articles on breast cancer, childhood asthma, and depression were longer than the respective Wikipedia articles. As previously noted, Healthwise's coverage of these three topics exceeded that of Wikipedia. The length of Healthwise's article on obesity, on the other hand, is only slightly longer than Wikipedia's obesity article (the longest of the four Wikipedia articles). As already mentioned, Wikipedia's coverage of obesity surpassed that of Healthwise.

If indeed the above argument has merit, one could surmise that as time passes more content will be added to these Wikipedia articles and that correspondingly their completeness scores will increase. Given the current surge in contributions to Wikipedia,

it would be interesting to repeat the completeness analysis in the near future to determine the rate of development of these articles.

Limitations of study

Although my study design was meant to meticulously follow the methodology used by Berland and colleagues, it is conceivable that minor methodological differences between the two studies could lead to unintended consequences. For example, Berland and colleagues provided their raters with only printed text versions of the Web sites. By contrast, I gave both electronic and printed versions of the texts to my raters. Thus, in addition to reading the abstracted articles, my raters could perform automatic searches for relevant words or phrases in the text. On the one hand, it is possible that automatic text searches might discover relevant information that a typical reader would pass over unnoticed. On the other hand, over-reliance on automatic searching could result in undisciplined reading and thereby cause the rater to miss certain information. If either scenario were true, it would obviously affect the reliability of the ratings.

A second potential difference in methodologies between the two studies concerns the meanings behind the scoring categories (e.g., "minimally covered," "more than minimally covered," etc.). Given the ambiguity of these classifications, differing interpretations of meaning may exist between my raters and those of the Berland et al. study. If, indeed, such a difference existed, my raters' scores might be skewed in one direction or the other, thereby reducing the validity of comparing my results to those of the Berland et al. study.

Future work

This study of consumer health Web sites only evaluated two aspects of quality: accuracy and completeness. Although these two quality components are essential, a full evaluation of the quality of Wikipedia's health-related content is warranted. One aspect of quality that would be particularly interesting to study is the currency of Wikipedia's health content. Given the ease and rapidity by which Wikipedia may be updated, this novel information source may have the advantage of being more current than other health-related Web sites, which often fail to be completely current given the long lag-time between updates. A second important quality component that should be studied is Wikipedia's user-interface. This Web site is known for its simple, easy-to-navigate design. A full evaluation of the quality of Wikipedia's health content must include some appraisal of the site's design aspects.

CONCLUSION

In light of Wikipedia's unsatisfactory completeness scores, consumers ought not rely solely on this Web site as their source of health information. However, consumers can be reasonably confident that the information found on Wikipedia is accurate. As previously mentioned, the completeness of health-related articles in Wikipedia may rise to an adequate level as the project matures over the ensuing months and years. If this is the case, Wikipedia may become an excellent source of high-quality health information.

All things considered, the warning above is not likely to deter people from searching Wikipedia for health-related information. Wikipedia's popularity will most likely continue to grow, at least in the short term. The implications of this trend may be very

important to the ever-evolving physician-patient relationship. As I elucidated in the introduction, Wikipedia and other similar grass-roots initiatives pose a significant threat to traditional authority structures. A central theme of the physician-patient relationship is the authority of the physician, as Paul Starr notes in his Pulitzer Prize winning book "The Social Transformation of American Medicine." Few other individuals in the United States command a higher level of authority than physicians. Over the past century, doctors have held on to this authority tenaciously despite numerous threats.[15] Consumer reliance on Wikipedia as a source of health information is unlikely to change the power dynamic between patient and physician. However, as a marker of the general trend of individual empowerment--what Thomas Friedman refers to as the "flattening" of the world[16]--Wikipedia's rapid ascent is likely to cause anguish among health care "experts." Should this shift of power from health care providers to consumers continue, it is likely that the consumer will rise to the health care "throne," as she has in most other industries.

REFERENCES

[1] Eysenbach G, Powell J, Kuss O, Sa ER. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA. 2002 May 22-29;287(20):2691-700.

[2] Wikipedia [homepage on the Internet]. St. Petersburg, FL: Wikimedia Foundation; 2006. Available from: http://www.wikipedia.org.

[3] Wikipedia [homepage on the Internet. St. Petersburg, FL: Wikimedia Foundation; 2006 [cited on March 8, 2006]. History of Wikipedia; [approximately 10 screens]. Available from: http://en.wikipedia.org/wiki/History_of_Wikipedia.

[4] Hitwise [homepage on the Internet]. New York, NY: Hitwise Pty. Ltd.; 2005 [updated on April 25, 2005; cited on January 5, 2006]. Wikipedia Becoming A Magnet For Consumer Searches; [approximately three screens]. Available from: http://www.hitwise.com/press-center/hitwiseHS2004/wikipedia.html.

[5] Hitwise [homepage on the Internet]. New York, NY: Hitwise Pty. Ltd.; 2005 [updated on October 5, 2005; cited on January 5, 2006]. LeeAnn Prescott – Hitwise US: Wikipedia, The New York Times, and Hurricane Katrina; [approximately four screens]. Available from: http://weblogs.hitwise.com/leeann-prescott/2005/10/wikipedia_the_new_york_times_a.html.

[6] Ranking on January 5, 2006. See following URL: http://www.alexa.com.

[7] *Talk of the Nation*. [radio series]. Wikipedia, Open Source and the Future of the Web. Washington, DC: NPR. 2005 [program aired on November 2, 2005]. Audiofile available from: http://www.npr.org/templates/story/story.php?storyId=4986453.

[8] Giles J. Internet encyclopaedias go head to head. Nature. 2005 Dec 15;438(7070):900-1.

[9] Jurist [homepage on the Internet]. Pittsburgh, PA: Bernard J. Hibbitts; 2006. Available from: http://jurist.law.pitt.edu/index.php.

[10] Kim P, Eng TR, Deering MJ, Maxfield A. Published criteria for evaluating health related web sites: review. BMJ. 1999;318:647-649.

[11] Silberg WM, Lundberg GD, Musacchio RA. Assessing, Controlling, and Assuring the Quality of Medical Information on the Internet. JAMA. 1997 April 16;277(15):1244-1245.

[12] Healthwise [homepage on the Internet]. Boise, ID: Healthwise; 1996-2006. Available from: http://www.healthwise.org/a_legal.aspx.

[13] Berland GK. Health information on the Internet: accessibility, quality, and readability in English and Spanish. JAMA. 2001 May 23/30;285(20):2612-2621.

[14] Wikipedia does not have an article that focuses exclusively on *childhood* asthma, thus I have chosen to use its general asthma article, which covers both childhood and adult asthma.

[15] Starr P. The Social Transformation of American Medicine. Basic Books: New York, NY. 1982.

[16] Friedman TL. Reviews. The World Is Flat : A Brief History Of The Twenty-First Century. Thorndike Press: Waterville, ME. 2005.