A Balanced Approach to Health Information Evaluation: A Vocabulary-Based Naïve Bayes Classifier and Readability Formulas

Gondy Leroy and Trudi Miller
School of Information Systems and Technology, Claremont Graduate University, 130 E. Ninth Street, Claremont, CA 91730. E-mail: {Gondy.Leroy; Trudi.Miller}@cgu.edu

Graciela Rosemblat and Allen Browne
Lister Hill National Center for Biomedical Communications, U.S. National Library of Medicine, Bethesda, MD 20894. E-mail: grosemblat@mail.nih.gov; browne@nlm.nih.gov

Since millions seek health information online, it is vital for this information to be comprehensible. Most studies use readability formulas, which ignore vocabulary, and conclude that online health information is too difficult. We developed a vocabularly-based, naïve Bayes classifier to distinguish between three difficulty levels in text. It proved 98% accurate in a 250-document evaluation. We compared our classifier with readability formulas for 90 new documents with different origins and asked representative human evaluators, an expert and a consumer, to judge each document. Average readability grade levels for educational and commercial pages was 10th grade or higher, too difficult according to current literature. In contrast, the classifier showed that 70–90% of these pages were written at an intermediate, appropriate level indicating that vocabulary usage is frequently appropriate in text considered too difficult by readability formula evaluations. The expert considered the pages more difficult for a consumer than the consumer did.

Introduction

People have always searched for information to maintain or improve their health. This information used to come from healthcare providers (doctors, nurses) or from close family and friends. In recent years, the health information exchange has changed and millions now look online for information. Today's online consumers are not only people in poor health who want to get healthy but also healthy people who want to remain healthy. Baker et al. (2003) reported in 2003 that 40% of their 60,000 household sample looked for health information online. They found that for at least a third of their respondents, the online health information affected decisions about health, healthcare, and visits to a healthcare provider. Warner and Procaccino (2004) reported a stronger influence, as 80% of the women they interviewed claimed that online information affected their treatment decisions.

Consumers not only look to the Internet to get information but also, increasingly, claim a stake in providing information. Information providers now include many patients who communicate with one another and provide advice online. Johnson and Ambrose (2006) report that almost 30% of Internet users participated in medical or health-related groups. In addition to consumers themselves, there are the typical information providers, such as clinicians, hospitals, the government, and libraries, as well as pharmaceutical companies and other commercial enterprises.

When there are millions of online Web pages and millions of readers, the usability, trustworthiness, and readability of this information are no small matter. Human-computer interaction has focused on optimal Web site usability for average users and increasingly for groups with special needs. For example, Becker (2004) evaluated 125 Web sites based on guidelines provided by the National Institute on Aging and found that, counter to current recommendations (Morrell, 2005), too many homepages still used a small font, were lengthy, required scrolling, did not allow for font resizing, and used pull-down menus. Others focused more on optimal design for online health communities (Neal et al., 2006). In addition to usability, the trustworthiness of information also requires evaluation, and researchers such as Gaudinat et al. (2006) are trying to help consumers assess the credibility of online health information. Finally, the text itself cannot be ignored. This type of evaluation checks if all necessary content is included and if it is presented at an appropriate reading level. When there are mismatches, two approaches can be followed according to Parker and Kreps (2005): Community

Received November 7, 2007; revised January 2, 2008; accepted January 2, 2008.

 • Published online 28 April 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20837

programs can be developed to increase the health literacy of consumers, or information providers can provide easier, alternative versions to consumers.

Most ongoing research on the readability of healthcare information looks at the content or the readability of the online text. We discuss both in the next section. Our work also focuses on readability, but it differs from existing approaches in that it provides a second, complementary method for assessing readability. We use a vocabulary-based naïve Bayes classifier to categorize documents into three readability groups. This work advances evaluation of readability of online health information. Developing tools that help providers assess information allows them to tune the information so that it becomes more understandable. When consumers better understand the information, they become more knowledgeable and are able to ask more informed questions of their caregivers (Fox & Fallows, 2003). In contrast, consumers who do not understand the text are at a disadvantage. The Committee on Health Literacy for the Council on Scientific Affairs found that misunderstanding health information increases the risk of making unwise health decisions, leading to poorer health and higher healthcare costs (Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs—American Medical Association, 1999).

Current Approaches to Health Information Evaluation

There are currently three approaches to evaluate the health information provided online. Many have looked at the content itself, for example, whether the information is correct and complete. The usability is also often evaluated, e.g., many awards and tools are available to help judge this. A third group focuses on appropriate use of language to explain the material, which is measured with readability formulas.

Content and Accessibility

Several clinicians and librarians have looked at the content of Web sites and evaluated whether the information was complete and trustworthy. This is usually a page-by-page approach carried out by individual experts. For example, in their Web site evaluation of eight best-selling herbal products, Morris and Avorn (2003) found that most Web sites claimed to treat, prevent, or cure diseases and more than half omitted the standard federal FDA disclaimer. Hunter (2005) evaluated pamphlets intended to educate Mexican immigrant women about cervical cancer and found important information to be missing. Berland et al. (2001) evaluated 25 Web sites on breast cancer, asthma, depression, and obesity, using a count of required clinical elements to evaluate if the information was complete. They found that almost half of the pages provided only a minimal amount of information. However, in over 80% of the pages, the information provided was correct.

Other researchers have developed instruments to evaluate usability of Web sites without the need for an individual expert. These instruments have been popular for years and many different ones have been developed and used to bestow awards. In 1998, Jadad and Gagliardi (1998) reviewed 47 such rating instruments in the context of health information. More recently, Bernstam et al. (2005) searched for evaluation instruments and found 273 distinct ones, however, a large number (65%) were not meant to be used by consumers. Today, with increased multimedia information, accessibility evaluation has received renewed attention, especially for groups with special needs such as the blind or elderly. For example, Zeng and Parmanto (2003, 2004) focused on accessibility of online health information for people with disabilities. They developed a scoring tool based on a combination of the World Wide Web Consortium Web Content Accessibility Guidelines and the U.S. Access Board's Electronic and Information Technology Accessibility Standards. One of their conclusions was that all Web sites violated some guidelines. Based on more than 7,000 pages retrieved from 108 sites, they also concluded that governmental Web sites were the most accessible and the portal sites the least accessible for their user groups. Becker (2004) evaluated 125 popular sites and also noted many shortcomings based on the National Institute on Aging guidelines.

Readability

To evaluate if appropriate language is used, the readability of a text is usually measured by calculating a required "grade level" that the reader should have completed in school to be able to understand the text. These grade levels are assigned to a text based on several possible formulas of which the Flesch readability scores and the Flesch-Kincaid grade levels are probably the most popular. The formulas are based on sentence and word length to assign readability levels to predict level of difficulty. Friedman and Hoffman-Goetz (2006) offer a good review of readability measurements that are also suitable for health and medical text.

Most current research using the Flesch and Flesch-Kincaid formulas shows that health-related text is too difficult for average adults. Berland et al. (2001) found that required readability levels were higher for English compared to Spanish sites. More than 60% of the English sites required college or graduate school reading skills (13th-grade level or higher). Boulos (2005) collected pages on diabetes mellitus and found that only 7 out of 20 were written at the eighth-grade level or lower, a level they considered appropriate. Kusec et al. (2003) limited their evaluation to diabetes Web sites that display the Health On the Net Foundation Code of Conduct (HONcode) logo, which prescribes rules related to the presentation of information (www.hon.ch/HONcode). On average, these pages were written at a 10.8th-grade level, which was also too high according to the authors. Friedman et al. (2004) compared Web sites for three types of cancer and also concluded that the majority were written at college level. This problem is even found with information for pet owners (Murphy, 2006), with more than half the Web sites written at 11th-grade level or higher. Although there are many claims about the optimal reading level based on national reading

DOI: 10.1002/asi

level estimates, few studies evaluate the impact of lowering the readability levels. An exception is Pignone et al. (2005), who reviewed several studies focusing on text simplification and showed improved understanding in the low literacy group with simplified texts.

In response to these inappropriately high grade levels, several non-profit and government groups have published guidelines that, when followed, can help information providers write text at levels that are suitable for the average information consumer. The State of California Health Literacy Initiative (http://cahealthliteracy.org/) provides information and tools for educators, health professionals, and literacy students. The Health & Literacy Special Collection (http://lincs.worlded.org/) provides links to easy-to-read health information. MedlinePlus (http://www.nlm.nih.gov/medlineplus/etr.html) advocates writing easy-to-read versions, and The Plain Language Initiative at the National Institutes of Health (NIH) (http://execsec.od.nih.gov/plainlang/index.html) requires the use of plain language in all documents provided by the government. In addition to these government requirements, hospitals provide brochures developed with these guidelines. Clinicians are encouraged to help patients understand the written text, e.g., they can receive Continuing Medical Education credit for completing educational programs that teach them the importance of patients' health literacy and how they can help by changing their language (Weis, 2007).

To our knowledge, we were the first to look explicitly at vocabulary used in a health specific context. In previous work, we evaluated the concepts used and topics discussed online by different groups. We mapped terms and phrases in texts from different sources to the Consumer Health Vocabulary (CHV), a listing of terms commonly used by lay people (Q. T. Zeng & Tse, 2006; Q. T. Zeng et al., 2005) that also indicates how easy a term is to understand for this sector of the population (Keselman et al., 2006). We found that patients who blogged used easier to understand terms but also discussed easier concepts (Leroy, Eryilmaz, & Laroya, 2006). We continue this line of research with the current study and combine it with the readability research stream.

Research Goal

Even though many existing guidelines discuss the language that should be used, most ongoing readability research is based on the classical formulas. Since many of the guidelines discussed above emphasize the importance of using or avoiding specific language, we believe that an evaluation of online text should include this component. To this end, additional tools are needed to complement the existing ones and provide a balanced evaluation of text.

Our first goal in this project was to find an evaluation method complementary to readability formulas that can be used in an efficient and automated manner. The readability formulas have been shown to relate to understanding but they do not address vocabulary per se. We discuss here such a vocabulary-based assessment tool, its development, and evaluation. Our second goal was to apply this vocabulary-based evaluation to online health information and compare and contrast the results with readability formula outcomes and evaluations by both a representative expert and a consumer.

A Vocabulary-Based Naïve Bayes Classifier

We developed an automated document classifier that can distinguish between three difficulty levels in documents based exclusively on the vocabulary used. An improved and more thoroughly evaluated version of our first prototype (Miller, Leroy, Chatterjee, Fan, & Thoms, 2007) was used for this project. This newer version benefited from a more comprehensive and representative document set (250 documents) for training and testing, improved underlying vocabulary representation, and an improved smoothing algorithm. The combination of these elements increased the accuracy of the classifier.

Naïve Bayes Classifier Algorithm

Classification. Classification is a machine learning technique that requires an algorithm to internalize characteristics inherent in predefined classes or outcomes of interest. Once the algorithm has learned from a training set what characteristics represent a class, new elements can be labeled automatically. In this case, our classifier assigns labels to documents based on the vocabulary being used. We defined three levels (labels) of text difficulty—easy, intermediate, and difficult—and used representative texts for each.

Naïve Bayes approach. We chose a naïve Bayes approach because it is an efficient algorithm to train and run, and it is well-established for classification tasks. For example, Larsen (2005) used a naïve Bayes classifier to decide whether e-mail messages were spam. Sahami et al. (1998) used it to classify junk e-mail. For a more complete review of naïve Bayes for text classification, see Sebastiani (2002).

A naïve Bayes classifier is based on Bayes' theorem. It calculates the probability of a specific hypothesis (H) being true with (given) certain evidence (E) or $p(h|e)$. Bayes' theorem lets us convert this probability to components that can be calculated (see Equation 1) based on available data:

$$p(h|e) = \frac{p(e|h) * p(h)}{p(e)} \quad (1)$$

To use this approach to classify a set of items into several, mutually exclusive classes, a classifier calculates the probability that the item belongs to a specific class for each of the available classes. Those probabilities are then compared and the class that has the highest probability is selected by the classifier as the label for that item. Naïve Bayes assumes independence of the input parameters and although this assumption is not true in text, it often outperforms other machine learning approaches when using text.

DOI: 10.1002/asi

We developed our own classifier in Java. To calculate the base probabilities for the vocabulary evidence, we collected documents to represent each class. The easiest level was represented by 100 medically themed blog entries written by lay people. We collected these from www.blogger.com by using keywords such as "treatment" and "hospital." These 100 patient blogs provided 6,720 tokens (unique words) and their frequencies in these texts. The intermediate level was represented by documents written by professionals to educate consumers. We chose only documents that were tested in interaction with consumers or texts that received independent awards. We combined 50 documents provided to us by City of Hope (a comprehensive cancer center in Duarte, California) with 50 pages from FamilyDoctor.org, a site with several awards for appropriate content operated by the American Academy of Family Physicians. These 100 pages provided 5,384 tokens and their frequencies. Finally, the difficult level was represented by 50 journal articles from the Journal of the American Medical Association (JAMA). We chose JAMA because it is the most widely circulated medical journal in the world (JAMA, 2006) and it is not limited to one medical specialty or disease. Since these texts are longer, we collected only 50 articles in order to render a more balanced corpus. These 50 documents provided 8,499 tokens and their frequencies. We downloaded the relevant pages in HTML format and removed navigational and extraneous formatting, leaving only the content as raw text. The text was tokenized using the GATE tokenizer (Sheffield Natural Language Processing Group, 2005) and stored in a database. We removed all punctuation marks and literal numbers, leaving only word tokens. We did not use a stopword list; we did retain all words but not numbers.

Our classifier calculates the probability that a text belongs to the easy, intermediate, or difficult group. For each document, the three hypotheses (easy, intermediate, difficult) are calculated. The evidence consists of the vocabulary contained in the document. In our case, the comparison between the three required probabilities can be simplified. Because the evidence being evaluated (the document) does not change, the denominator can be dropped for the comparison. Moreover, because we do not know how many easy, intermediate, or difficult documents are presented on the Internet, we assume that the numbers are approximately equal, and we further simplify the calculations by ignoring $p(h)$. We found that even without this information, the classifier is very accurate (see results). The final probability to be calculated is $p(e|h)$, which is formally calculated (see Equation 2) for each category by multiplying for all the words in the document the probability of occurrence in that specific class or:

$$p(Doc|Cat_j) = \prod_i p(word_i|Cat_j) \qquad (2)$$

where:

Doc = document being classified
Cat_j = the category being tested: easy, intermediate, or difficult class
$Word_i$ = word in the document

Smoothing. When classifying a new text, words will be found in that test document that do not appear in the training corpus. This results in zero probabilities for those words and they decrease the accuracy of the classifier. To avoid these zero probabilities, we used add-lambda smoothing to approximate their frequency. Add-lambda smoothing uses a positive probability to recognize the likelihood of encountering a word unobserved in the training set. This approach was successfully used before by others, such as Dreyer and Eisner (Dreyer & Eisner, 2006) for adjustments during training and by Mann and Yarowsky (Mann & Yarowsky, 2005) to assign positive probabilities for words not in their vocabulary. We set the value of lambda to $l = 0.01$ based on English language estimates. Since Merriam-Webster (online) has 470,000 entries and the Oxford English Dictionary contains over 500,000 words while our classifier's training corpus has only 14,433 tokens, we assume that if the corpus were increased in size 100 times most of the words in the English language would be represented. Our lambda represents this adjustment. It is unlikely that a word absent in a corpus would require a corpus 1,000 times larger for that word to occur ($l = 0.001$). Testing different values also confirmed our assumptions. We evaluated six lambda values (0.1, 0.01, 0.001, 0.0001, 0.00001, and 0.000001) and found that with leave-one-out validation a lambda at 0.001 or higher correctly classifies only two additional documents, while lower lambda values led to lower accuracy.

Naïve Bayes Classifier Evaluation

We evaluated the classifier twice using a corpus of 250 documents. For the first evaluation, we used 10-fold cross validation. The approach divides the corpus into 10 equal parts of 25 documents each. The classifier is trained (calculation of probabilities) on 9 parts while the 10th is set aside for testing. Each section serves as a test set once and classification results for each test set are averaged. Overall the classifier was very accurate, with 98% of the documents correctly classified. We performed a second evaluation using leave-one-out validation, which uses all the documents for training except one which is reserved for testing. Each of the 250 documents is held out once as test document and the results are again averaged. The classifier performed equally well with 98.4% accuracy. Details are provided in Table 1.

TABLE 1. Classifier evaluation.

$N = 250$	Accuracy (%) Evaluation method	
Classification levels	10-fold	Leave-one-out
Easy level	99.0	99.0
Intermediate level	97.0	98.0
Difficult level	98.0	99.0
Overall	**98.0**	**98.4**

DOI: 10.1002/asi

Readability Evaluation Study of Online Documents

Corpus Development

Our second goal was to evaluate a set of documents representative of sets of documents a consumer might encounter when searching for health information online. These sets may differ because documents can be provided for different purposes and by different stakeholders. Among others, this approach has been followed by Zeng and Parmanto (2004), who evaluated information from e-commerce, corporate Web, portals, community Web, and government or education sites, and Becker (2004), who looked at commercial, non-profit, online newspapers, and state Web sites. Others have focused on different topics within specific fields: Gemoets et al. (2004) focused on allergies and celiac disease; Berland et al. (2001) looked at breast cancer, childhood asthma, depression, and obesity.

Our goal is not to evaluate as many Web sites as possible, but to investigate the complementary nature of vocabulary-based measures with readability formulas. We collected documents discussing three common conditions, melanoma, depression and prostate cancer, from commercial Web sites, government/educational Web sites, and those provided by consumer groups themselves. All three represent information sources that patients will encounter online. We searched with Google to find the Web sites. The commercial sites were only selected when they offered a treatment, drug, or therapy. We included alternative or complementary medicine sites because patients will read them when looking for information (Walji, Sagaram, Meric-Bernstam, Johnson, & Bernstam, 2005). The government/non-profit Web sites are those provided specifically to educate consumers. The consumer Web sites are those provided by consumers themselves and include texts from discussion boards and lists.

For each of the three topics (melanoma, depression, and prostate cancer) and for each source type (commercial, government/non-profit, and consumer) we selected five Web sites and downloaded two different pages, resulting in 90 documents ($3 \times 3 \times 5 \times 2$). We ensured that all documents were different from those used to develop our classifier. All Web pages were saved as text documents. Any final paragraph referring to more links (e.g., "return to top," "click on the links on the left for more information") or to author or copyright information was manually removed from the text. In some cases, saving in text format introduced extra spacing around bulleted or numbered lists and this was also corrected in all texts. We added semicolons to the end of bullets or lists when punctuation was missing.

Automated Evaluation: Readability Analyzer and Naïve Bayes Classifier

The successful development of our vocabulary-based classifier allowed us to evaluate online health information by looking at vocabulary. To provide a balanced approach in this evaluation, five readability formulas were selected for the complementary evaluation based on common formulas: the revised Fry Readability Graph, Flesch Reading Ease, Flesch-Kincaid, Gunning FOG, and New Dale-Chall formula. These readability measures are primarily based on syntactic (number of words in the sentence) and morphological (number of letters or syllables per word) factors. Different formulas assign different weights to these factors (see Table 2 for an overview). Regardless of how word complexity is measured, the core assumption is that word length and sentence length are directly correlated to the relative level of difficulty/ease with which a text can be read. The New Dale-Chall formula also incorporates word lists used to measure the difficulty of vocabulary on the underlying assumption that frequently used words will be more familiar and thus easier to understand.

The Readability Analyzer, a tool developed at the National Library of Medicine (NLM; Gemoets et al., 2004), implements a number of readability formulas, including those used in this study. The tool is written in Java with a Web-based front end. It uses tokenization and variant-generation software developed at NLM and publicly available syllable-counters. This tool provides an average score based on averaging the results given out by the application of these different formulas on a given text. It also provides other information such as sentence count, word count, words per sentence, and type-to-token ratio. Interestingly, the readability formulas provided as part of Microsoft Word do not assign grade levels higher than 12th. Thus, we did not deem this tool appropriate for our task.

Manual Evaluation: Expert and Consumer Judgments

To provide an additional evaluation that was not machine-based, we invited a representative expert and a consumer to evaluate each document in our set. The expert has spent over

TABLE 2. Overview of factors in readability formulas.

	Factors considered				
Formula	Sentence count	Word count	Long word count	Syllable count	Voc. freq.
New Dale-Chall (Chall & Dale, 1995)	X	X			X
Flesch-Kincaid (Flesch, 1948)	X	X	X		
Flesch Reading Ease (Flesch, 1979)	X	X	X		
Fry Readability Graph (Fry, 1977)	X	X	X		
Gunning FOG Index (Gunning, 1952)	X	X		X	

DOI: 10.1002/asi

TABLE 3. Overview of definitions provided to evaluate documents.

	Instructions provided to	
	Expert	Consumer
Document vocabulary		
Easy	medical vocabulary used by the average consumer	here are medical terms that you would use in conversation
Intermediate	medical vocabulary used in consumer health education	after reading the whole document or after asking for help with a few words, you understand the medical terms used
Difficult	medical vocabulary typically used by health professionals but not by consumers.	there are many medical terms you do not understand
Document structure		
Easy	a manner of speaking or syntactic constructions typically used by the average consumer	this has a structure that you would write
Intermediate	a manner of speaking or syntactic constructions typically used in consumer health education	this has a structure that you can understand
Difficult	a manner of speaking or syntactic constructions typically used by health professionals	this has a structure that health professionals would write
Overall evaluation		
Easy	understood by the average consumer without the need to consult reference sources or his/her network of friends/family	you can understand the document without help
Intermediate	understood as consumer health education	you can understand the document with the help of references or your network of friends/family
Difficult	understood by medical professionals but usually not by the "typical" consumer	difficult or impossible to understand; might be understood by medical professionals

25 years in Reference and Information Services departments of academic medical libraries that also served patients and the public. She chaired committees that prepared consumer pamphlets, taught medical terminology to staff and students, gave lectures to cancer survivors and families, and worked on projects that analyzed consumer Web site materials. We requested that she provide her expert opinion on the overall readability (in terms of appropriate audience) for the texts. Since research tells us that the average American reading level is only as high as the ninth grade, we asked her to make her determinations based on consumers with no higher than a ninth-grade reading level. We specifically asked her not to rely on readability formulas at all. Our consumer representative is a 55-year-old native English speaker without a medical or healthcare background. Her highest education level was High School, completed 37 years ago. She earned additional certifications but they are unrelated to medicine or healthcare. Both evaluators assessed the vocabulary, structure, and overall appearance of documents. They received an hourly compensation for their task. Table 3 provides an overview of how "easy," "intermediate," and "difficult" were defined for each in their respective instructions.

Corpus Evaluation

We used both the Readability Analyzer and the naïve Bayes classifier to evaluate the 90 documents. We report two measures from the Readability Analyzer: the Flesch-Kincaid grade level and the average grade level based on the five formulas in Table 2. From the classifier, we report the final classification (easy, intermediate, difficult) that the documents received. All 90 documents were also shown to both human evaluators. They were provided with all documents in text format and a spreadsheet to indicate their evaluations.

Results

We first describe the Readability Analyzer and classifier results for the documents by origin and topic. Then, we describe the evaluations by the expert and consumer. We complete the analysis by calculating and comparing the correlations between the different evaluations.

Grade levels. Overall, the Flesch-Kincaid metrics scored commercial documents at the 12th-grade level (12.1), the consumer documents at just under the eighth-grade level (7.8), and the governmental/non-profit documents at the 11th-grade level (11.0). The five-formula average provides a slightly higher estimate, with commercial documents at almost the 14th-grade level (13.7), consumer documents at the eighth-grade level (8.2), and government/non-profit documents at the 12th-grade level (12.4). Figure 1 shows a detailed overview of the readability of the texts based on the Flesch-Kincaid formula and on the Readability Analyzer five-formula average for the different types and topics of the documents.

To evaluate if the differences were significant, we performed two 3×3 ANOVAs with origin and topic as the independent variables, and either the Flesch-Kincaid Grade

DOI: 10.1002/asi

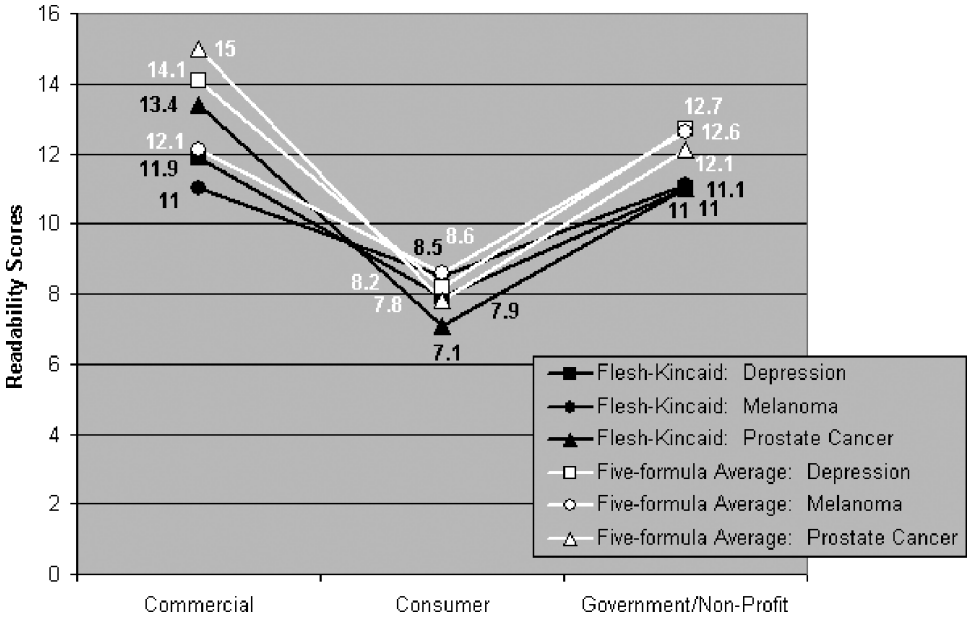

FIG. 1. Readability scores.

Level or the Readability Analyzer average as the dependent variable. The results were identical for both dependent variables. We found a significant main effect for origin for the Flesch-Kincaid Readability, $F(2,81) = 22.9$, $p < .001$, and for the Readability Analyzer average, $F(2, 81) = 43.7$, $p < .001$. There was no significant effect for topic or for the interaction between origin and topic. Post-hoc contrasts indicated that the differences between pages from consumer versus commercial sources ($p < .001$, Bonferroni adjustment) or government sources ($p < .001$, Bonferroni adjustment) were significant for both measures. The difference between pages from commercial versus government/non-profit sources was not significant.

Classifier levels. We then used the classifier to assign one of three levels (easy, intermediate, or difficult) to each document. Figure 2 provides an overview of the scores for documents according to their origin and topic. Commercial and government/non-profit pages scored on average at the intermediate level, with pages on prostate cancer slightly more difficult. In addition to average scores, we also looked at the distribution of labels. We found that of the commercial pages, 90% received an intermediate score and 7% a difficult score. From the government/non-profit pages, 70% received an intermediate score and 17% received a difficult score. The results do not differ for the different topics. All three topics have slightly more than half of the pages at the intermediate level. More specifically, 43.3%, 53.3%, and 3.3% of the melanoma pages, 40%, 53%, and 7% of the prostate cancer pages, and 33.3%, 53.3%, and 13.3% of the depression pages were easy, intermediate, and difficult. A 3×3 ANOVA for topic and origin confirmed these differences: There was only one significant main effect for source, $F(2,81) = 78.6$, $p < 001$. Post-hoc contrast showed that only the differences between consumer sources versus commercial ($p < .001$. Bonferroni adjusted) or government pages ($p < .001$, Bonferroni adjusted) were significant.

Expert and consumer evaluations. Figure 3 shows average scores for the overall evaluation by expert and consumer; details can be seen in Figure 4 and Figure 5. In most cases, the consumer evaluation is lower than the expert evaluation, indicating that the consumer finds the pages easier than the expert (the expert evaluated the pages on behalf of an average consumer). A paired-samples t-test confirmed this ($p < .001$).

Overall Evaluation by Expert and Consumer: We conducted a 3×3 ANOVA for the overall evaluations by expert and consumer. The results confirm that there is a slightly different pattern between expert and consumer in their evaluation of the pages. For the expert, we found a significant main effect for topic, $F(2, 81) = 4.0$, $p < .05$, and for source, $F(2, 81) = 46.2$, $p < .001$. There was only a strong trend for an interaction effect between the two ($p = .055$). The consumer evaluation showed three effects, a main effect for topic, $F(2, 81) = 5.8$, $p < .005$, and source, ($F(2, 81) = 9.3$, $p < .001$, and also a significant interaction effect, $F(4, 81) = 3.7$, $p < .01$.

Structure and Vocabulary Evaluation by the Expert: The expert found that almost all consumer pages were easy, with only 27% of them considered intermediate for their vocabulary use. The commercial pages were considered to be a lot more difficult with 23% being difficult overall, the vocabulary was considered difficult in 33% of the pages, and the structure was considered difficult in 23% of the pages. The government/non-profit pages were considered to have difficult vocabulary in 10% and difficult structure in 27% of the cases. We performed 3×3 ANOVAs for vocabulary and structure (overall scores are discussed above). The results

DOI: 10.1002/asi

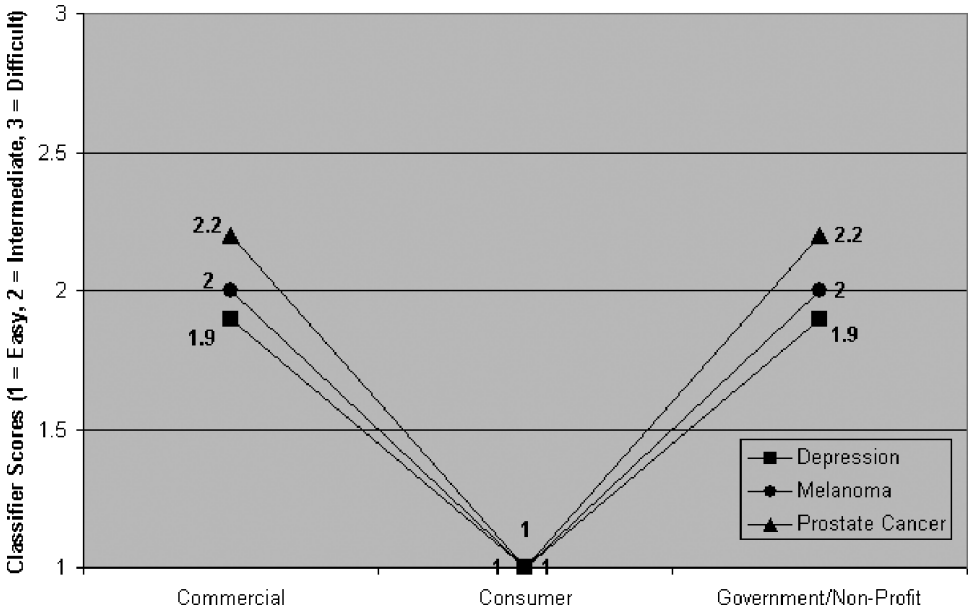

FIG. 2. Classifier levels.

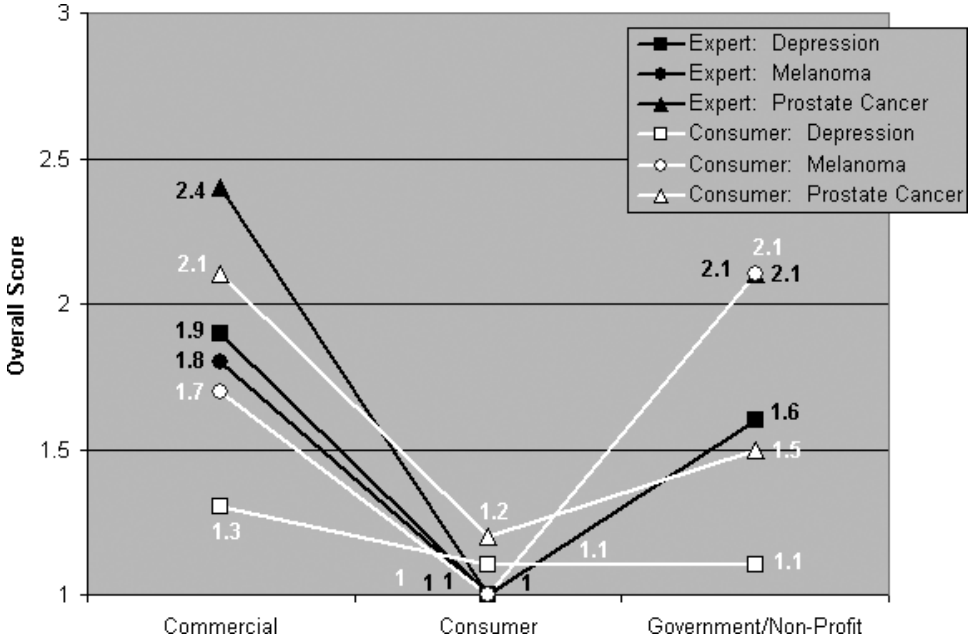

FIG. 3. Expert and consumer "overall" scores (1: easy; 2: intermediate; 3: difficult).

were very similar. For vocabulary, we found a main effect for source, $F(2, 81) = 3.9$, $p < .05$, and topic, $F(2, 81) = 21.6$, $p < .00$. For structure, we found similar main effects for source, $F(2, 81) = 3.9$, $p < .05$, and topic, $F(2, 81) = 32.0$, $p < .001$. Interactions were not significant.

Structure and Vocabulary Evaluation by the Consumer: The consumer did not consider any category completely easy. For example, the pages with a consumer origin were considered to use intermediate vocabulary in 17% of the cases and intermediate structure in 33%. The commercial pages had the most difficult texts (20%) and vocabulary (23%). For government pages, vocabulary was considered difficult in 7% of the cases. Comparable to the expert evaluation, we performed additional 3×3 ANOVAs for vocabulary and structure (overall scores are discussed above). The ANOVA for vocabulary shows a main effect for topic, $F(2, 81) = 5.1$, $p < .01$, for source, $F(2, 81) = 10.8$, $p < .001$, and a significant interaction effect,

DOI: 10.1002/asi

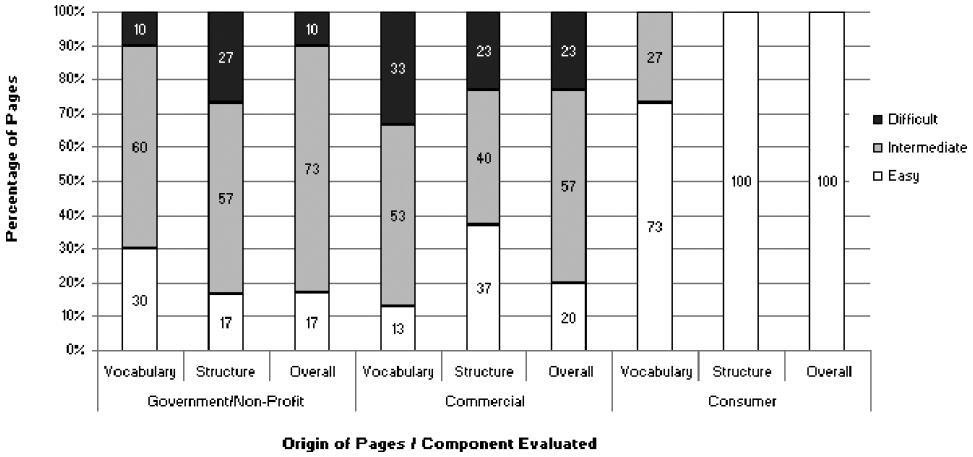

FIG. 4. All data for the expert evaluation.

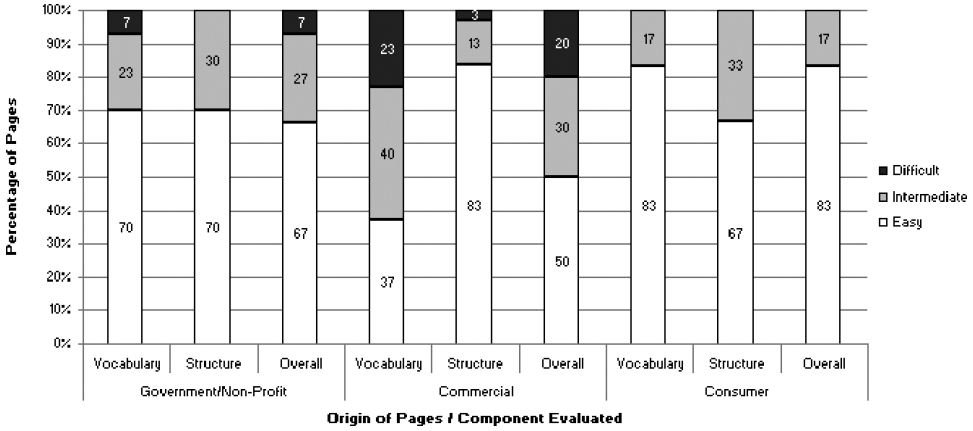

FIG. 5. All data for the consumer evaluation.

TABLE 4. Pearson's correlations ($N = 90$, 2-tailed).

				Expert			Consumer		
	Flesch-Kincaid	5-formula average	Classifier	Voc.	Struct.	Overall	Voc.	Struct.	Overall
Flesch-Kincaid	1								
5-formula average	.969**	1							
Classifier	.659**	.707**	1						
Expert voc.	.423**	.493**	.535**	1					
Expert struct.	.538**	.565**	.668**	.549**	1				
Expert overall	.569**	.620**	.747**	.758**	.873**	1			
Consumer voc.	.420**	.441**	.459**	.635**	.517**	.602**	1		
Consumer struct.	.194	.161	.258**	.248*	.434**	.348**	.440**	1	
Consumer overall	.468**	.474**	.485**	.483**	.558*	.623**	.888**	.634**	1

*is significant at the 0.05 level, **at the 0.01 level.

$F(4, 81) = 2.8$, $p < .05$. Similarly, the ANOVA for structure shows a main effect for topic, $F(2, 81) = 5.8$, $p < .01$, for source, $F(2, 81) = 3.1$, $p < .05$, and for the interaction, $F(4, 81) = 7.6$, $p < .001$.

Correlations. To evaluate how all scores relate to each other, we calculated Pearson's correlation coefficients (Table 4). With a few interesting exceptions, most scores are strongly correlated. The two readability metrics are strongly correlated with each other, as are the three evaluations of the expert and the three evaluations of the consumer.

The scores of the expert correlate strongly with both the readability formulas and the classifier scores. In general, the correlations are slightly higher for the expert-classifier

DOI: 10.1002/asi

compared to the expert-readability formulas. The scores of the expert correlate with those of the consumer. The consumer scores also correlate with the classifier. However, the consumer's structure evaluation does not significantly correlate with the readability formulas.

Discussion

Pages with a consumer-text origin were easier than pages with a commercial or government/non-profit origin. This difference was shown by the readability formulas and the classifier. Neither indicated a difference between commercial and government/non-profit pages. Basing conclusions only on readability formulas would indicate that these documents are too difficult. In contrast, most do not belong to the most difficult category according to the classifier results. The developers of the Readability Analyzer state that "formulas appear to serve as a reasonable 'first approximation' for predicting how well consumers might understand textual materials about health topics. . ." (Gemoets et al., 2004). They also recognized that incorporating a consumer health vocabulary component into the tool would render it more precise for this domain [personal communication, T. Tse, Nov. 7, 2003]. This is to be expected because medical and health vocabulary is very specific and influences how difficult a text is to understand. Moreover, readability formulas were devised for school-age materials, and even low-literacy adults will be familiar with terms rated difficult for a fourth grader, such as "hospitalization" or "diabetes." A text containing these words would result in an artificially high grade level when assessed by the Analyzer, which explains why many more pages were considered intermediate (or appropriate) by the classifier and also by the consumer. The two different methods thus appear to be complementary, and their fusion would result in a more precise evaluating tool for health-related text.

The results of the expert and consumer evaluations were somewhat unexpected. The expert scored many documents as too difficult for consumers to read. The consumer, however, found the documents generally at an appropriate level. Although both evaluations correlated with each other, it is possible that this shows one or more biases: The expert may be underestimating consumers or the consumer may be overestimating herself. More experimentation is needed that relies on actual explanation or actions based on the information in the text and on the assessments of many more consumers.

A further striking result was the lack of a correlation between the consumer's evaluation of document structure and the readability formulas. It was expected that documents scoring higher according to the readability formulas, due to longer sentences or longer paragraphs, would also present more structural challenges to a consumer. The results for the evaluation of the structure of the documents, however, did not correlate with these readability scores. This shows that what consumers consider difficult is not necessarily captured by readability formulas. In contrast, the expert's evaluations correlated with the scores of the readability formulas. This may be a result of the expert's training in applying rules based on such evaluations.

Conclusion and Future Directions

Most ongoing research on evaluation of online health information uses readability formulas to determine whether the text is written at an appropriate level for the general public. The formulas used are based on counts of syllables, words, and sentences, and ignore the vocabulary used. Our goal was to develop a complementary approach based on vocabulary. We used text representative of three difficulty levels—easy, intermediate and difficult—and developed a naïve Bayes classifier to assign these labels automatically to new text. The classifier was 98% accurate.

After development and evaluation of the classifier, we applied it to 90 new documents on melanoma, prostate cancer, and depression that were either commercial, educational (government and non-profit), or consumer-provided pages. We also used the Readability Analyzer and evaluated each document with five readability formulas. The results from the Readability Analyzer are consistent with the current literature. Except for the pages provided by consumers, the text was written on average at a 12th-grade level or higher, generally considered too difficult for the general public. According to the classifier, 90% of the commercial pages and 70% of the educational pages were written at an intermediate level. This intermediate level is based on documents written for and tested by consumers. We assume that this level is appropriate for the general public. We complemented this study with an evaluation by one expert and one consumer of health information. The expert considered more documents to be too difficult (for consumers) than the consumer did. This may indicate that the readability formulas can be an overestimation, especially when appropriate vocabulary is used. If we limited our evaluation to readability formulas, we would conclude that information not provided by consumers themselves is too difficult to read. The classifier results seem to indicate that the situation may not be as bleak as generally suggested. However, a limitation of our study is that we relied on one representative consumer and self-reporting for the consumer evaluation. More studies are needed that test multiple consumers for real text comprehension.

In the future, we hope to evaluate the importance of vocabulary versus writing style as measured by our classifier and readability formulas on degrees of understanding by laymen with different levels of education. Readability formulas have been validated with user groups, and we believe that the impact of medical vocabulary cannot be ignored. We will also make our test documents with lexicons available online for others to use.

Acknowledgements

The authors would like to thank Annette Mercurio and Benjamin T. Laroya at City of Hope (Duarte, California) for their help in acquiring the patient educational materials.

DOI: 10.1002/asi

G. Leroy and T. Miller's research was funded by a grant from the National Library of Medicine, R21-LM008860-01, http:// isl.cgu.edu/ConsumerHealth.htm. G. Rosemblat and A. Browne's research was supported by the Intramural Research Program of the NIH, National Library of Medicine/ Lister Hill National Center for Biomedical Communications.

References

Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs—American Medical Association. (1999). Health literacy: Report of the Council on Scientific Affairs. JAMA, 281, 552–557.

Baker, L., Wagner, T.H., Signer, S., & Bundorf, M.K. (2003). Use of the Internet and e-mail for health care information: Results from a national survey. Journal of the American Medical Association, 289(18), 2400–2406.

Becker, S.A. (2004). A study of Web usability for older adults seeking online health resources. ACM Transactions on Computer-Human Interaction, 11(4), 387–406.

Berland, G.K., Elliott, M.N., Morales, L.S., Algazy, J.I., Kravitz, R.L., Broder, M.S., et al. (2001). Health information on the Internet: Accessibility, quality, and readability in English and Spanish. JAMA, 285, 2612–2621.

Bernstam, E.V., Shelton, D.M., Walji, M., & Meric-Bernstam, F. (2005). Instruments to assess the quality of health information on the World Wide Web: What can our patients actually use? International Journal of Medical Informatics, 74(1), 13–19.

Boulos, M.N.K. (2005). British Internet-derived patient information on diabetes mellitus: Is it readable? Diabetes Technology & Therapeutics, 7(3), 528–535.

Chall, J., & Dale, E. (1995). Readability revisited: The new Dale-Chall readability formula. Cambridge, MA: Brookline Books.

Dreyer, M., & Eisner, J. (2006). Better informed training of latent syntactic features. Paper presented at the Conference on Empirical Methods in Natural Language Processing (EMNLP), Sydney, Australia.

Flesch, R. (1948). A new readability yardstick. Journal of Applied Psychology, 32, 221–233.

Flesch, R. (1979). How to write plain English: A book for lawyers and consumers. New York: Harpercollins.

Fox, S., & Fallows, D. (2003). Internet health resources—Health searches and email have become more commonplace, but there is room for improvement in searches and overall Internet access. Washington DC: Pew Internet & American Life Project.

Friedman, D., & Hoffman-Goetz, L. (2006). A systematic review of readability and comprehension instruments used for print and Web-based cancer information. Health Education & Behavior, 33(3), 352–373.

Friedman, D., Hoffman-Goetz, L., & Arocha, J. (2004). Readability of cancer information on the Internet. Journal of Cancer Education, 19(2), 117–122.

Fry, E. (1977). Fry's readability graph: Clarifications, validity, and extensions to level 17. Journal of Reading, 242–252.

Gaudinat, A., Ruch, P., Joubert, M., Uziel, P., Strauss, A., Thonnet, M., et al. (2006). Health search engine with e-document analysis for reliable search results. International Journal of Medical Informatics, 75(1), 73–85.

Gemoets, D., Rosemblat, G., Tse, T., & Logan, R. (2004). Assessing readability of consumer health information: An exploratory study. Presented at Medinfo, San Francisco, CA, 2004.

Gunning, R. (1952). The technique of clear writing. New York: McGraw-Hill.

Hunter, J.L. (2005). Cervical cancer educational pamphlets: Do they miss the mark for Mexican immigrant women's needs? Cancer Control, 12(Cancer, Culture and Literature Supplement), 42–50.

Jadad, A.R., & Gagliardi, A. (1998). Rating health information on the Internet: Navigating to knowledge or to babel? JAMA, 279(8), 611–614.

Johnson, G.J., & Ambrose, P.J. (2006). Neo-tribes: The power and potential of online communities in health care. Communications of the ACM, 49(1), 107–113.

Journal of the American Medical Association. (2006). JAMA—About JAMA. Accessed May 2007 from http://jama.ama-assn.org/misc/aboutjama.dtl

Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L., & Zeng, Q. (2006). Assessing consumer health vocabulary familiarity: An exploratory study. Presented at MEDNET, Toronto, Canada, 2006.

Kusec, S., Brborovic, O., & Schillinger, D. (2003). Diabetes Web sites accredited by the Health On the Net Foundation Code of Conduct: Readable or not? Studies in Health Technology and Informatics, 95, 655–660.

Larsen, K. (2005). Generalized naive Bayes classifiers. SIGKDD Explorations Newlsetter, 7, 76–81.

Leroy, G., Eryilmaz, E., & Laroya, B.T. (2006, November 11–15). Health information text characteristics. American Medical Informatics Association (AMIA) Annual Symposium, Washington DC.

Mann, G.S., & Yarowsky, D. (2005). Multi-field information extraction and cross-document fusion. 43rd Annual Meeting of the ACL, Ann Arbor.

Miller, T., Leroy, G., Chatterjee, S., Fan, J., & Thoms, B. (2007, January 3–7). A classifier to evaluate language specificity of medical documents. 40th Annual Hawaii International Conference on System Sciences (HICSS), Waikoloa, Big Island, Hawaii.

Morrell, R.W. (2005). www.nihseniorhealth.gov: The process of construction and revision in the development of a model Web site for use by older adults. International Journal: Universal Access in the Information Society, 4(1), 24–38.

Morris, C.A., & Avorn, J. (2003). Internet marketing of herbal products. Journal of the American Medical Association, 290(11), 1505–1509.

Murphy, S.A. (2006). Consumer health information for pet owners. Journal of the Medical Library Association, 94(2), 152–158.

Neal, L., Lindgaard, G., Oakley, K., Hansen, D., Kogan, S., Leimeister, J.M., et al. (2006). Online health communities. CHI'06 Conference on Human Factors in Computing Systems, Montréal, Québec, Canada.

Parker, R., & Kreps, G.L. (2005). Library outreach: Overcoming health literacy challenges. Journal of the Medical Library Association, 93(4), S81–S85.

Pignone, M., DeWalt, D.A., Sheridan, S., Berkman, N., & Lohr, J.N. (2005). Interventions to improve health outcomes for patients with low literacy: A systematic review. Journal of General Internal Medicine, 20(2), 185–193.

Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In proceedings of the AAAI Workshop on Learning for Text Categorization, Madison, Wisconsin.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 54.

Sheffield Natural Language Processing Group. (2005). General Architecture for Text Engineering (3.0 ed.).

Walji, M., Sagaram, S., Meric-Bernstam, F., Johnson, C., & Bernstam, E. (2005). Searching for cancer-related information online: Unintended retrieval of complementary and alternative medicine information. International Journal of Medical Informatics, 74(7–8), 685–693.

Warner, D., & Procaccino, J.D. (2004). Toward wellness: Women seeking health information. Journal of the American Society for Information Science and Technology, 55(88), 709–730.

Weis, B.D. (2007). Health Literacy and patient safety: Help patients understand. Manual for Clinicians (2nd ed.). AMA and AMA Foundation. ISBN#: 978-1-57947-982-4.

Zeng, Q.T., & Tse, T. (2006). Exploring and developing consumer health vocabularies. Journal of the American Medical Informatics Association, 13(1), 24–29.

Zeng, Q.T., Tse, T., Crowell, J., Divita, G., Roth, L., & Browne, A.C. (2005). Identifying consumer-friendly display (cfd) names for health concepts. AMIA 2005 Fall Symposium, Washington, DC.

Zeng, X., & Parmanto, B. (2003, November 8–12). Evaluation of Web accessibility of consumer health information. American Medical Informatics Association (AMIA) Annual Symposium, Washington, DC.

Zeng, X., & Parmanto, B. (2004). Web content accessibility of consumer health information Web sites for people with disabilities: A cross sectional evaluation. Journal of Medical Internet Research, 6(2).

DOI: 10.1002/asi