Why Is It Difficult to Find Comprehensive Information? Implications of Information Scatter for Search and Design

Suresh K. Bhavnani
School of Information, University of Michigan, Ann Arbor, MI 48109. E-mail: bhavnani@umich.edu

The rapid development of Web sites providing extensive coverage of a topic, coupled with the development of powerful search engines (designed to help users find such Web sites), suggests that users can easily find comprehensive information about a topic. In domains such as consumer healthcare, finding comprehensive information about a topic is critical as it can improve a patient's judgment in making healthcare decisions, and can encourage higher compliance with treatment. However, recent studies show that despite using powerful search engines, many healthcare information seekers have difficulty finding comprehensive information even for narrow healthcare topics because the relevant information is scattered across many Web sites. To date, no studies have analyzed how facts related to a search topic are distributed across relevant Web pages and Web sites. In this study, the distribution of facts related to five common healthcare topics across high-quality sites is analyzed, and the reasons underlying those distributions are explored. The analysis revealed the existence of few pages that had many facts, many pages that had few facts, and no single page or site that provided all the facts. While such a distribution conforms to other information-related phenomena, a deeper analysis revealed that the distributions were caused by a trade-off between depth and breadth, leading to the existence of general, specialized, and sparse pages. Furthermore, the results helped to make explicit the knowledge needed by searchers to find comprehensive healthcare information, and suggested the motivation to explore *distribution-conscious* approaches for the development of future search systems, search interfaces, Web page designs, and training.

Introduction

An important objective of creating a quality Web site is to ensure it contains all relevant and important information about a chosen topic. For example, the National Cancer Institute's Web site attempts extensive coverage of more than 118 cancers distributed across hundreds of pages. A complementary objective of search engine developers is to help users easily find such extensive sites. For example, the goal of Google (in the words of its developers) is to "get you to the right site" (Thottam, 2001, p. 33). The development of such Web sites and search engines suggests that it is easy for a user to gain a comprehensive understanding of a topic. In domains such as consumer healthcare, finding comprehensive information about a topic (e.g., melanoma risk and prevention) is critical as it can help patients achieve important coping outcomes such as treatment compliance, reducing anxiety, and learning the language of their disease (Hinds, Streater, & Mood, 1995; Ream & Richardson, 1996; Sturdee, 2000; for a review, see Mills & Sullivan, 1999).

However, there is ample evidence that users often have difficulty in finding comprehensive information. For example, recent studies show that novice searchers of healthcare information typically use general purpose search engines to find relevant pages (Eysenbach & Kohler, 2002), go online without a definite search plan, find most sites accidentally (Fox & Fallows, 2003), and often end their searches prematurely with incomplete information (Bhavnani, 2001; Bhavnani et al., 2003a).

Why do novice users have difficulty in finding comprehensive information? One clue to the problem is given by the behavior of expert searchers like healthcare reference librarians. These experts know which combination of sites to visit in which specific order to obtain comprehensive information about a healthcare topic (Bhavnani, 2001, 2002). This search behavior implies that information is scattered across Web sites requiring users to know which sites to visit for what information. Understanding how information is scattered across Web sites therefore can reveal the knowledge required to find comprehensive information, and lead to better designs of systems and training.

However, while there have been several attempts to understand how research articles are distributed across journals and databases (e.g., Hood & Wilson, 2001), we found no studies of how facts about a topic are distributed across Web pages and Web sites, and the possible reasons for

Received March 26, 2004; revised July 9, 2004; accepted July 9, 2004

 ● Published online 3 May 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20189

those distributions.[1] This article begins by discussing the existing research that motivates the need to analyze the distribution of facts across relevant Web sites. Next, two experiments are presented: (a) Experiment 1 is aimed towards the identification of facts necessary for a patient's comprehensive understanding of five melanoma topics at different levels of importance; (b) Experiment 2 is aimed towards understanding the distribution of these facts across high-quality Web sites, and explore the possible causes of those distributions. The results of these experiments suggest the need for a *distribution-conscious* approach for the development of future search systems, interfaces, Web pages, Web sites, and training. This approach should help more users find comprehensive information in domains such as consumer health.

Motivation to Study the Distribution of Facts Across Web Sites

As described above, several studies have shown that novice searchers of healthcare information have difficulty in finding comprehensive healthcare information. These studies have shown that novice searchers begin their search by typing a few terms in search engines like Google (Eysenbach & Kohler, 2002; Fox & Fallows, 2003), access the resulting hits in the order presented (Bhavnani, 2001), do not check the reliability of their sources (Eysenbach & Kohler, 2002), and end their searches prematurely without accessing sources that in combination provide comprehensive information (Bhavnani, 2001).

In contrast, expert searchers know which sources to visit in which sequence (Bhavnani, 2001; Florence & Marchionini, 1995; Kirk, 1974). For example, in a recent study (Bhavnani, 2001) an expert healthcare searcher looking for flu shot information had a three-step search procedure: (a) Access a reliable healthcare portal to identify sources for flu shot information; (b) access a high-quality source of information to retrieve general flu shot information; and (c) verify that information by visiting a pharmaceutical company that sells flu vaccine. Such search procedures enabled experts to find comprehensive information quickly and effectively, compared to novices who were unable to infer such procedures by just using Google.

What motivates an expert to visit different sites to find information, and why is it difficult for novices to do the same? Perhaps the reason why experts had to visit many different sites was that the facts related to the information topic they were searching were scattered across the Web. However, as described below, there have been no studies that have analyzed how facts related to a topic were distributed across Web sites.

Pre-Web studies of content distribution (see Bates, 2002) include the classic works of Bradford who demonstrated the highly skewed distribution of articles about a topic across journals (1948), and Zipf who described the highly skewed distribution of different words across a book (1949). Recent studies of Web content have focused on the dynamic nature of online information. For example, Bar-Ilan and Peritz (1999) described how Web pages retrieved through search engines for the topic "informetric" disappeared, reappeared, or changed over the study period of several months, and Wormell (2000) studied how information about the topic "modern welfare state" spread and evolved through different forms of publication. Other studies of online content have focused on constructing typologies of the context in which query terms occur (Bar-Ilan, 1998, 2000a, 2000b; Cronin, Snyder, Rosenbaum, Martinson, & Callahan, 1998). For example, Cronin et al. (1998) identified 11 different source types (home page, conference page etc.) of pages retrieved from search engines that contained content about highly cited researchers; Bar-Ilan (1998) identified a range of different types of pages in which information about "Erdos" (a well-known mathematician) occurred.

Numerous studies of online content in different domains such as consumer health, and science, have analyzed the accuracy and completeness of online information (Allen, Burke, Welch, & Rieseberg, 1999; Beredjiklian, Bozentka, Steinberg, & Bernstein, 2000; Bichakjian et al., 2002; Biermann, Golladay, Greenfield, & Baker, 1999; Davison, 1997; Eng et al., 1998; Griffiths & Christensen, 2000; Impicciatore, Pandolfini, Casella, & Bonati, 1997; Jiang, 2000; McClung, Murray, & Heitlinger, 1998; Soot, Moneta, & Edwards, 1999; see Eysenbach, Powell, Kuss, & Sa, 2002 for a review). For example, Bichakjian et al. (2002) found that even the top healthcare sites had incomplete information about melanoma, and Allen et al. (1999) showed the presence of misleading, inaccurate, and unreferenced information in online science publications.

While the above Web content studies and related studies on Web links (e.g., Barabasi & Albert, 1999; Klienberg & Lawrence, 2001; Thelwall, 2001a; Vaughan & Thelwall, 2003) have begun to reveal the dynamic and complex nature of the Web, however, to the best of our knowledge none have attempted to analyze how facts related to a topic are distributed across relevant Web pages and Web sites. The following two experiments therefore fill an important gap in our understanding of information distributions. These experiments are not designed to reflect how users search the Web for healthcare information. Instead, the experiments are designed to analyze in detail the current distribution of healthcare information across high-quality sites. The goal is to understand how information is scattered across pages and sites, and to pinpoint the knowledge required to deal with such scatter. This understanding could suggest novel approaches that assist users in finding comprehensive information.

[1]Analysis of the distribution of content across sources has been studied in limited corpuses for purposes other than understanding the nature of the distributions and their causes. For example, van Halteran and Teufel (2003) use the presence or absence of factoids to evaluate automatic summarization, and aspects of a topic have been used to evaluate systems in the Interactive Track of the Text Retrieval Conference (e.g., Over, 1998).

Experiment 1: Identification of Facts About Melanoma

The goal of the first interrater experiment was to identify a set of facts that skin cancer physicians agreed was necessary for a patient's comprehensive understanding of each of five topics related to melanoma. The focus of the research was on melanoma (a deadly form of skin cancer) because we had access to two skin cancer physicians, both of whom had experience in studying the information needs of patients. Furthermore, because skin cancer is the most common type of cancer, there exists a large amount of information on the Web about this disease (Bichakjian et al., 2002). The following five melanoma topics (with their abbreviated notation in parentheses) were selected for detailed analysis:

1. Self-examination in the diagnosis of melanoma (*self-examination*)
2. Doctor's examination in the diagnosis of melanoma (*doctor's examination*)
3. Diagnostic tests used in the diagnosis of melanoma (*diagnostic tests*)
4. Disease stages used in the diagnosis of melanoma (*disease stage*)
5. Descriptive information related to melanoma risk and prevention (*risk/prevention*)

The above topics were selected from the most common question categories in a hierarchical taxonomy developed through the analysis of real-world skin cancer questions (Bhavnani et al., 2002). Topics 1–4 belonged to the category *Diagnosis of Melanoma* that contained the most number of questions, and Topic 5 belonged to another category of *Risk/Prevention of Melanoma* that contained the second highest number of questions. The last topic was included to ensure that the results were not specific to diagnosis. Therefore, the selected topics, besides being the most common categories of questions about melanoma on the Web, also came from two major categories in the skin-cancer taxonomy.

Method

A two-step method was used for identifying a list of facts that the skin cancer physicians stated were required for a comprehensive understanding of a melanoma topic. Facts are defined as statements about a topic agreed upon by experts in the field and can be claims or recommendations. In the first step, a list of facts for each of the above five topics were identified by analyzing all 38 links across high-quality sites on the melanoma page in MEDLINEplus (a leading healthcare portal developed by the National Library of Medicine at the National Institutes of Health, Bethesda, MD). The identification of facts about the five melanoma topics resulted in 14 facts for self-examination, 6 facts for doctor's examination, 6 facts for diagnostic tests, 13 facts for disease stage, and 15 facts for risk/prevention. Each fact consisted of a single sentence, with optional terms where required. For example, the following was a fact about risk/prevention:

Having many moles [or more than 50 moles] increases your risk of getting melanoma.

In the second step two experienced skin cancer physicians were asked to independently rate the importance of facts related to each of the five topics using a 5-point Likert scale of fact importance (1 = *Not important to know*, 2 = *Slightly important to know*, 3 = *Important to know*, 4 = *Very important to know*, 5 = *Extremely important to know*). The physicians were told that they should rate the importance of each fact keeping in mind a concerned user looking for the melanoma topic on the Web. Furthermore, they were free to modify the wordings of the facts, or to add new facts. After they had completed their ratings, the physicians independently discussed their ratings with the researcher to make any clarifications.

The above two-step method of first generating a list of the facts, and then using that list in the interrater experiment was based on our experience from an earlier pilot study (Bhavnani, 2003; Bhavnani, Jacob, Nardine, & Peck, 2003b). In that study, in the short time that the physicians could give to the project, they found it easier to critique a list of facts rather than to generate it. Cognitive psychologists believe that this is because humans, despite experience in a domain, have far superior cognitive capacities for recognition rather than for free recall (e.g., Anderson, 1995). However, because the physicians were encouraged to make any modifications to the content and number of facts during their ratings, the resulting list of facts accurately reflects what they wish health seekers should know about the different topics.

Results

One of the physicians made minor changes in the wordings of seven facts, and the other made minor wording changes to four facts. None of these changes altered the original overall meaning of the fact, and neither physician added any new facts.

Table 1 shows the high agreement between the two physicians across the five topics. For example, (as shown in the last column of Table 1) when rating facts for risk/prevention, the two physicians agreed completely on 11 facts (73%), but

TABLE 1. The level of disagreement based on the 5-point Likert scale of fact importance across the five melanoma topics. The mean disagreement shows high overall agreement between the judges.

Level of disagreement	Self-examination	Doctor's examination	Diagnostic tests	Disease stage	Risk/prevention
No disagreement	8	3	3	1	11
Disagreed by 1 point	5	2	3	5	1
Disagreed by 2 points	0	0	0	7	3
Disagreed by 3 points	0	1	0	0	0
Disagreed by 4 points	1	0	0	0	0
Mean disagreement	0.643	0.833	0.5	1.462	0.467

disagreed on 1 fact (7%) by 1 point, and 3 facts (20%) by 2 points. Therefore, there was a total of 7 points of disagreement over the 15 facts, resulting in a mean disagreement of 0.467 points. Across all the topics, the judges disagreed by more than 2 points on the Likert scale of importance for only two facts, and the highest mean disagreement for a topic was 1.462 points on the Likert scale of fact importance. The results therefore represent very high agreement between the judges.[2]

A final rating for each fact was calculated by averaging the two judge's scores. Facts that had an averaged score of 1 (*Not important to know*) were subsequently excluded from the analysis. This resulted in the exclusion of one fact from one topic (risk/prevention). The analysis therefore enabled us to identify a set of facts related to a comprehensive understanding of each of the five melanoma topics (see Appendix A for all identified facts across the five topics), at four levels of importance. This set of facts was used in the next experiment, which was designed to understand the distribution of these facts across relevant Web pages.

Experiment 2: Analysis of the Distributions of Facts About Melanoma

The goal of the second interrater experiment was to understand not only how facts about each of the five topics were distributed across relevant Web pages, but also the amount of such information in each page and site.

Material

There exists a large number of healthcare sources that are unreliable (see Eysenbach et al., 2002 for a review), hence the survey was focused on sites that were known to contain reliable melanoma information. A set of reliable sites with melanoma information was defined as the union of all the sites pointed to by the melanoma page in MEDLINEplus, and the top five most comprehensive sites identified in a recent study of online melanoma information (Bichakjian et al., 2002). This union resulted in 10 sites. The above method of gathering data from known sites (vs. customized crawls) has been used by other informetric researchers (Cui, 1999; Rosenbaum, 1998) for focused studies such as ours.

To compensate for the widely varying quality of internal search engines provided by these sites, we used Google to search within each of the 10 sites for pages related to the facts (identified from Experiment 1) for each topic, and for the general topic. For example, we generated queries for each of the 14 facts for risk/prevention, as well as 2 queries relating to risk/prevention in general. This resulted in 16 Google queries for each of the 10 sites (e.g., melanoma risk UV OR ultraviolet OR sun OR sunlight OR sunburn; site: cancer.gov) for a total of 160 queries just for risk/prevention.

[2]Neither Cohen's kappa nor Cohen's weighted kappa are relevant for these data because of the very high skew in the agreements. The data in Table 1 is therefore shown to provide direct evidence of the high interrater agreement.

TABLE 2. The number of queries and unique pages for each topic that were used in the interrater experiment.

Topic	Number of queries	Number of unique pages
Self-examination	150	192
Doctor's examination	70	110
Diagnostic tests	70	112
Disease stage	140	125
Risk/Prevention	160	189
Total	590	728

The second column of Table 2 shows the total number of queries for each topic, and the overall total of 590 queries used in our experiment.

Each of these queries was iteratively tested and refined by a group of 3 search experts until the best set of pages showed up in the top 10 hits. (See Appendix B for the entire list of 59 queries across all topics used in this experiment.) As stated in the Introduction, this approach of identifying the pages was not intended to reflect how users search the Web for healthcare information. Rather, it was intended to identify as many relevant pages as possible from the high-quality sites to understand how facts were distributed across them.

The highly targeted queries were used to retrieve the top 10 hits from each site. Subsequently, duplicates, news items, pages for health professionals, non-English pages, dictionary pages, personal home pages, and broken links were removed. The last column of Table 2 shows the total number of resulting unique Web pages for each topic, and the overall total of 728 pages that were analyzed.

The above page collection method of using a single search engine and selecting only the top 10 hits could have two possible page-omission errors. One type of page-omission error could occur because a relevant page does not show up in the top 10 hits provided by Google. The probability of this *not-in-top-10* omission error is very low for two reasons: (a) Separate queries were used for the different facts of the same topic, which produced highly redundant hits. Therefore, if a page did not show up in the top 10 hits for one fact, there were many other opportunities for that page to show up for other facts for the same topic. (b) Each query was iteratively tested by three search experts until the results produced highly relevant hits in the top 10 list. Another type of page-omission error could occur because a page was not indexed at all by Google, and therefore had no chance of showing up in the top10 list of hits. The ideal way to mitigate this *not-indexed* omission error (caused by forgotten or lost URLs as described by Bar-Ilan, 2002) would be to use multiple search engines as recommended by several authors (Bar-Ilan, 1999; Thelwall, 2001b). However, this was impractical for this study given the large number of queries already being issued. This led to the compromise of using only Google, which is considered the most comprehensive indexer of Web pages compared to other search engines (Sullivan, 2001; Thelwall, 2001a).

TABLE 3. The best-fitting curves, the type of the curve, and their goodness-of-fit likelihood ratio test (LR) statistic for the distributions of all facts, and for only very and extremely important facts, across the five topics. The *p* values with asterisks indicate that the curve was accepted by the likelihood ratio test for goodness-of-fit at the .05 level.

	All facts	Only very and extremely important facts
Self-examination	$Y = 22.172e^{-0.275x}$ **(discrete exponential)** LR = 26.058, $p = .011$	$Y = 15.291e^{-0.243x}$ **(discrete exponential)** LR = 23.247, $p = .010$
Doctor's examination	$Y = (e^{1.458} - 1)^{-1}(1.458^x/x!)$ **(truncated Poisson)** LR = 7.922, $p = .094$*	$Y = 100e^{-1.253x}$ **(discrete exponential)** LR = 4.794, $p = .029$
Diagnostic tests	$Y = (e^{1.726} - 1)^{-1}(1.726^x/x!)$ **(truncated Poisson)** LR = 8.745, $p = .068$*	$Y = (e^{0.550} - 1)^{-1}(0.550^x/x!)$ **(truncated Poisson)** LR = 0.645, $p = .422$*
Disease stage	$Y = 32.302e^{-0.419x}$ **(discrete exponential)** LR = 42.462, $p < .001$	$Y = 37.322e^{-0.471x}$ **(discrete exponential)** LR = 26.815, $p < .001$
Risk/Prevention	$Y = 33.308e^{-0.276x}$ **(discrete exponential)** LR = 20.976, $p = .051$*	$Y = 36.871e^{-0.301x}$ **(discrete exponential)** LR = 30.414, $p < .001$

Method

A printed version of the Web pages for each of the five topics was given to a rater who judged the extent to which the facts were covered in each page, using a 5-point Likert scale for fact coverage (0 = *Fact not covered on page*, 1 = *Fact covered in less than one paragraph*, 2 = *Fact covered in one paragraph*, 3 = *Fact covered in more than one paragraph*, 4 = *Web page mostly devoted to fact*, although other facts could also be covered on the same page). The reliability of the above rater was assessed by requesting a second rater to perform the same evaluation on a random selection of 25% of the Web pages for each of the five topics.

Analysis and Results

The raters had high agreement on whether or not a fact was present in a page, and the extent to which the fact was covered on that page. The kappa values (used to determine agreement for binary ratings) and the weighted kappa values (used to determine agreement for scaled ratings) ranged from .681–.867, which is considered good-to-very good agreement (Altman, 1990).

Distributions for each of the topics were plotted. As shown in Figure 1A–E, each plot shows the number of Web pages (shown on the *Y*-axis) that contain an ascending number of facts (shown on the *X*-axis). Pages with no facts[3] were dropped from the plots to limit our analysis to only relevant pages. As shown, all the distributions are skewed to the left, where there are many pages that contain a few facts, very few pages (toward the right tail) contain many facts, and no pages that contain all the facts.

To determine the shape of these distributions, three curves (power, discrete exponential,[4] and truncated Poisson[5]) were fitted to the data using maximum likelihood estimation (MLE), and tested for goodness-of-fit using the likelihood ratio test. As shown in the second column of Table 3, three of the distributions were best fit by discrete exponential curves, and two by truncated Poisson curves. Although the more important result is that the distributions were skewed (rather than the precise nature of their shapes), the above curves for goodness-of-fit were tested using the likelihood ratio test (LR). The second column of Table 3 shows that three of the five curves (shown with asterisked *p* values) were accepted by the likelihood ratio test.[6] Furthermore, the two distributions (doctor's examination and diagnostic tests), which were best fitted by truncated Poisson curves, also had exponential curves that were accepted by the likelihood ratio test. Therefore, although Poisson curves provided a slightly better fit for these two distributions, the discrete exponential curves cannot be rejected (see Appendix C for equations for the other curves that we attempted to fit to the distributions, and their respective likelihood ratio test scores).

[3]These pages were in the top 10 retrieved hits because although they contained terms in the query, those terms were used in a context that did not constitute a fact. For example, the top 10 hits retrieved for the query "melanoma risk site:jhu.edu" contained a page about genetics and pancreas cancer because it also contained the words "melanoma" and "risk." However, this page had information that was not relevant to our study.

[4]As exponential curves are typically used to describe continuous data, we describe our curve as a *discrete* exponential curve.

[5]As Poisson curves typically begin at zero, we describe our curve as a *truncated* Poisson curve.

[6]The null hypothesis in a likelihood ratio test states that the distribution fits the curve being tested. A curve therefore has acceptable fit when $p > 0.05$.

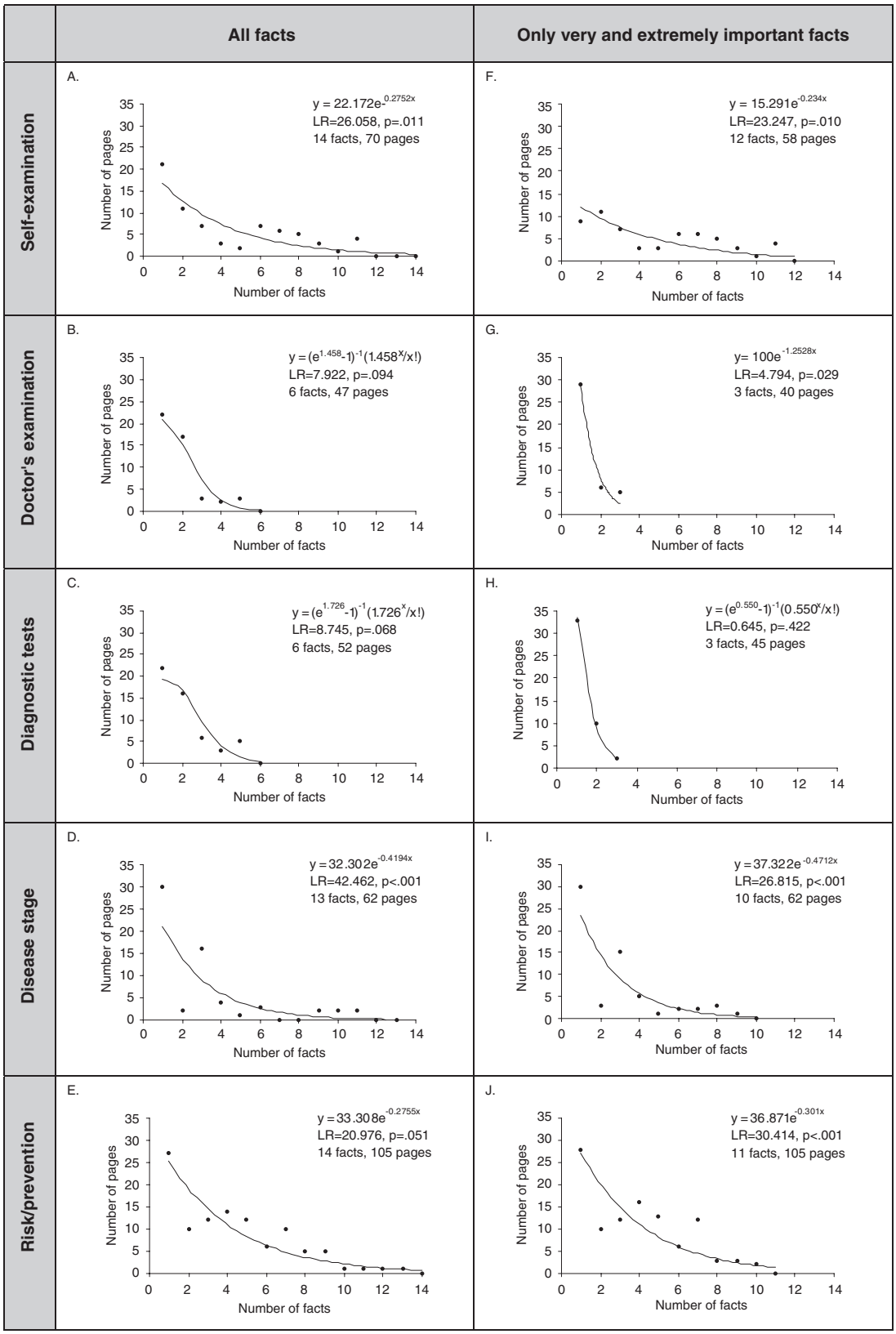

FIG. 1. The distribution of all facts across pages (A–E), and the distribution of only very important and extremely important facts across page (F–J). The curves shown are the best-fitting curves as determined by maximum likelihood estimation (MLE).

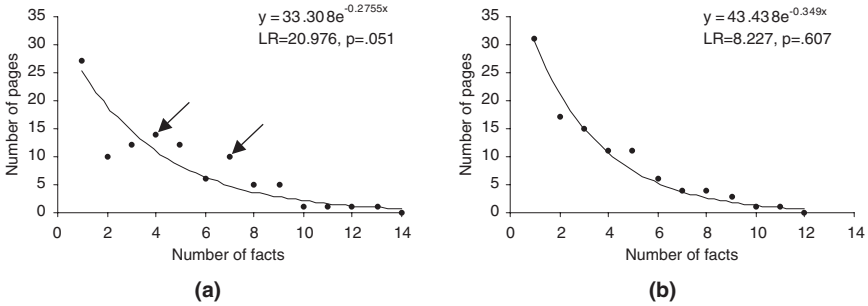

FIG. 2. The distribution for risk/prevention (a) smoothened out after collapsing three co-occurring facts into one fact (b).

To explore whether the importance of the facts made any difference in the distributions, the analysis was redone by including only those facts that the physicians stated were *very important*, and *extremely important* (the highest two levels on the 5-point Likert scale of fact importance used in Experiment 1). As shown in Figure 1F–J, even with this narrow set of facts, the distributions still remained skewed. Furthermore, three (F, I, and J) of the five distributions had no page that contained all the *very important* and *extremely important* facts. The two distributions (G and H), which had at least one page that contained *very important* and *extremely important* facts, were of topics that had only three facts. These facts were less than half of the average number of such facts across all the topics. While these skewed distributions are similar in spirit to the results of other information distribution studies of content (e.g., Bradford, 1948; Zipf, 1949) it does not explain why over 78% of the pages from *reliable* sites contained less than half of the facts. Furthermore, during the analyses of the 728 Web pages across the 10 sites, it was observed that some facts frequently co-occurred with other facts on many pages, pages had different concentrations of facts, and Web sites had different concentrations of different kinds of pages. To systematically probe these observations, analyses were performed to understand the relationship between: (a) facts and other facts, (b) facts and Web pages, and (c) facts, Web pages, and Web sites.

Relationship between facts. As described above, it was observed that for some topics there was a small set of facts that frequently co-occurred on different pages. Therefore, a fact-correlation matrix for each topic was created, which revealed that three topics (risk/prevention, self-examination, and disease stage) contained highly co-occurring facts. Highly co-occurring facts were defined as having an $r > .8$ (which was a natural break in the correlation numbers across all topics). For example, the following three facts about risk/prevention frequently co-occurred:

1. Wearing protective clothing can help to prevent melanoma.
2. Wearing sunscreen can help to prevent melanoma.
3. Avoiding UV rays [or avoiding peak sunlight hours; or seeking shade] can help to prevent melanoma.

These three highly co-occurring facts ($r_{1,2} = .92$, $r_{1,3} = .80$, $r_{2,3} = .84$) all are about UV-protection for melanoma prevention, and are therefore conceptually related. Conceptual relationships were also true for the highly co-occurring facts in the other two topics. For the topic self-examination, there were four highly co-occurring facts, which were about the *ABCD* technique (asymmetry, border irregularity, color, and diameter of moles) to detect melanoma. For the topic disease stage, there were two sets of highly co-occurring facts. One set was about high-level definitions for melanoma stages, and the second set of facts was about detailed definitions for melanoma stages.

To explore whether these co-occurring facts had any affect on the distributions, each set of co-occurring facts was collapsed into a single averaged fact. This averaged fact was considered to be on a page if that page contained more than 50% of the original set of facts. After collapsing these co-occurring facts, the new distributions were plotted, and in two cases (risk/prevention and self-examination) the distributions smoothened out and provided a superior fit to curves. For example, as shown in Figure 2A, the original distribution for risk/prevention has two prominent bumps, one at four facts and another at seven facts (as indicated by the arrows). After collapsing the highly co-occurring facts, these two bumps smoothened out as shown in Figure 2B, with an improved fit to an exponential curve, which was accepted by the likelihood ratio test. However, there was little effect on the distribution for the topic disease stage, where the bumps did not smoothen. This implies that some bumps may be caused by factors other than co-occurring facts. A deeper exploration of the different manifestations of co-occurring facts in a distribution is currently being conducted, which could suggest automated ways for detecting clusters of conceptually-related facts in Web pages.

Relationship between facts and Web pages. An exploratory analysis of pages at both ends of the distribution for risk/prevention revealed that pages with many facts appeared to provide information in not much detail, while pages with a few facts appeared to provide a lot of detail about a few facts. A more rigorous analysis revealed that pages with a maximum detail level of 2 or 3 (on the Likert scale described earlier), had a significantly higher number of facts ($p < .001$, Mean number of facts = 5.89, $SD = 2.63$) compared to pages that had a maximum detail level of 4 (Mean = 2.87, $SD = 2.12$), or a maximum detail level of 1

TABLE 4. With the exception of two cases (shown in bold italic font), pages with a maximum detail level of 2 or 3 have a higher number of mean facts than those pages with a maximum detail of 1 or a maximum detail of 4. This suggests the existence of general, sparse, and specialized pages, which appear to be the cause of the skewed distributions.

	Mean number of facts (*SD*)			Significance (two-tailed)
	Max detail = 1	Max detail = 2 or 3	Max detail = 4	
Self-examination	3.17 (2.32) $n = 24$	6.68 (2.96) $n = 28$	1.56 (1.25) $n = 18$	2 or 3 > 1, $p < .001$ 2 or 3 > 4, $p < .001$
Doctor's examination	1.35 (0.49) $n = 20$	2.12 (1.27) $n = 25$	4 (0) $n = 2$	2 or 3 > 1, $p < .05$ ***2 or 3 = 4, p = .0502***
Diagnostic tests	1.28 (0.46) $n = 18$	2.68 (1.35) $n = 31$	1 (0) $n = 3$	2 or 3 > 1, $p < .001$ 2 or 3 > 4, $p < .05$
Disease stage	1.85 (1.11) $n = 39$	4.85 (3.45) $n = 20$	4 (5.20) $n = 3$	2 or 3 > 1, $p < .001$ ***2 or 3 = 4, p = .71***
Risk/Prevention	1.86 (1.21) $n = 28$	5.89 (2.63) $n = 54$	2.87 (2.12) $n = 23$	2 or 3 > 1, $p < .001$ 2 or 3 > 4, $p < .001$

(Mean = 1.86, *SD* = 1.21).[7] This suggests that fact breadth (measured by number of facts), and fact depth (measured by maximum detail of any fact on the page) can be used to characterize the pages. Furthermore, it suggests the existence of general pages that cover many facts in a medium amount of detail, specialized pages that cover few facts in a high level of detail, and sparse pages that contain few facts in very little detail. To test the generality of this observation, the same analysis was conducted across the other four topics.

Table 4 shows that the overall pattern is suggestive of the existence of general, specialized, and sparse pages. The two cases (marked bold and italic in the last column of Table 4) where this pattern did not hold had only marginal significance. Our future analysis will probe our informal observation that these two cases are caused by the existence of a set of pages that have both general and specialized information. However, overall we believe that the current analysis does suggest that the skewed distribution is being caused by many specific and sparse pages (at the head of the distributions), and a few general pages (at the tail of the distribution). Page authors therefore appear to be making a trade-off between depth and breadth of fact coverage when designing the content of different pages.[8]

Relationship between facts, Web pages, and Web sites. As described earlier, no single page contained all relevant facts for a topic. This situation might exist because authors typically design an entire site, not just a single page. It is therefore likely that the authors might spread the facts over many pages within their site. If this were the case, there could be a combination of pages within a single site that contained all the facts. We therefore analyzed fact occurrence along two dimensions: (a) source granularity at two levels (page and site), and (b) fact importance at two levels (all facts, and only *very important* and *extremely important* facts).

Table 5 shows a summary of the above analyses. The second and third columns show the previously discussed results (where there was no page that contained all the facts for any topic), and a few pages that contained only those facts rated as *very important* and *extremely important* for two topics. As shown in the last two columns of Table 5, the site analysis

[7]These inferential statistics were done for exploratory purposes mainly to generate a hypothesis for future testing.

[8]The trade-off between breadth and depth of fact coverage on a page is similar to that observed in traditional document genres such as review articles versus detailed journal articles.

TABLE 5. An analysis of fact coverage along two dimensions (source granularity and fact importance) revealed that while no one page contained all relevant facts for a particular topic, there exist sites that cover only very and extremely important facts for most topics.

	Number of pages		Number of sites	
	All facts	Only very and extremely important facts	All facts	Only very and extremely important facts
Self-examination	0	0	0	1
Doctor's examination	0	5	0	2
Diagnostic tests	0	2	2	3
Disease stage	0	0	0	0
Risk/Prevention	0	0	0	3

TABLE 6. The minimum number of sites that a user must visit in order to get comprehensive information for a topic.

	Minimum number of sites necessary for comprehensive information	Example sites
Self-examination	2	American Academy of Dermatology (aad.org) & melanoma.com
Doctor's examination	No combination of sites yields comprehensive information	No sites
Diagnostic tests	1	Skincarephysicians.com
Disease stage	No combination of sites yields comprehensive information	No sites
Risk/Prevention	2	American Academy of Dermatology (aad.org) & American Cancer Society (cancer.org)

revealed that only two sites contained all the facts for one topic (diagnostic tests), whereas there was at least one site for four of the five topics that contained the *very important* and *extremely important* facts.

If a user cannot get comprehensive information from a single Web site, how many of the high quality sites that we analyzed should she visit? Table 6 shows the minimum number of high-quality sites that a user must visit to have access to all of the facts for each topic. As shown, a user can get comprehensive coverage of all the facts from a single site for only one topic (diagnostic tests). For two other topics (self-examination and risk/prevention), users must visit at least two sites. Finally, for two topics (doctor's examination and disease stage), no combination of sites yielded comprehensive coverage. Thus, the number of sites that a user must visit to get comprehensive coverage depends on the topic. Furthermore, as shown in Table 6, the particular sites that users must visit to get comprehensive coverage are also topic-dependent. These results are similar to those described by Hood and Wilson (2001), who reported that the number of databases necessary for comprehensive coverage of journal articles about a topic, was dependent on the topic being searched.

The situation remains similarly complex when only the *very important* and *extremely important* facts are considered. While most topics have at least one site that covers all the *very important* and *extremely important* facts, the most comprehensive sites for a particular topic may not be the most comprehensive sites for a different topic. For example, as shown in Table 7, Harvard's Web site contains all the *very important* and *extremely important* facts for risk/prevention, but it does not contain all the *very important* and *extremely important* facts for any other topic. Three sites (cancer.org, skincancer.org, and melanoma.com) contain only *very important* and *extremely important* facts for two topics, but no site contains all such facts for more than two topics. Therefore, although it may be possible for searchers to access the *very important* and *extremely important* facts for a single topic in a single site, they will most likely have to visit a different site to find such comprehensive coverage for a different topic. Our current research is analyzing the number and rank of Google hits that a user must visit to get comprehensive information about a topic. In general, the analyses show that pages with many facts are typically not ranked high, and users need to visit many different hits to get comprehensive information. However, a more detailed discussion of these results is beyond the scope of this article.

TABLE 7. The sites that cover only very and extremely important facts for each topic. No single site provides complete coverage for more than two topics.

	Sites with only very and extremely important facts
Self-examination	Skincancer.org
Doctor's examination	American Cancer Society (cancer.org) Skincancer.org
Diagnostic tests	American Cancer Society (cancer.org) Melanoma.com Skincarephysicians.com
Disease stage	No sites
Risk/Prevention	American Academy of Dermatology (aad.org) Harvard University (harvard.edu) Melanoma.com

Discussion

The distribution of facts about melanoma across high-quality Web pages presents a complex picture for users searching for comprehensive information about a topic. Furthermore, this situation is also problematic for physicians who are under increasing pressure to advise their patients on which healthcare sources to visit. Through a rigorous data collection method: (a) the facts were identified for five topics at different levels of importance from experienced physicians; (b) pages were collected from only the top 10 melanoma sites; (c) queries were hand-crafted and iteratively tested for each fact to ensure high relevance of hits; and (d) the best search engine was used. The results showed that all the distributions were skewed towards fewer facts, with no single page from any of the high-quality sites that contained all the facts. Furthermore, entire sites also did not provide comprehensive coverage, confirming the study by Bichakjian et al. (2002). Only one topic had two sites with all the facts.

The above situation improved slightly when the analysis included only those facts that were rated as *very important* and *extremely important*. The analysis showed that only two topics (Doctor's exam and Diagnostic tests), had a few pages that had all the facts rated as *very important* and *extremely important*. These two topics had only six facts each, which was less than the average number (10.6) of facts over all the five topics. Finally, although there was at least one site that covered only very and extremely important facts for most topics, none of them provided full coverage for more than two of the five topics.

Analysis of the different levels of detail of facts within pages suggested the existence of general pages that had many facts in medium amount of detail, specific pages that had a few facts in a lot of detail, and sparse pages that had few facts in little detail. Because there were overall many more sparse and specific pages, this pattern provides an explanation for the skewed distribution. Finally, sets of facts co-occurred frequently and appear to leave telltale bumps in the frequency distributions, a phenomenon that also needs to be explored in future research.

The above results pinpoint the kind of knowledge that users must have when searching for comprehensive information about healthcare. Users must know that some pages have breadth information spanning many facts with medium levels of detail (general pages), while others have few facts in high detail (specific pages). In addition, users also need to know that they have to visit more than one general page to get all the relevant facts. Because conventional search tools like Google and MEDLINEplus do not provide this kind of information about relevant pages, the lack of such knowledge often leads users to end their searches early, leading to the retrieval of incomplete information (Bhavnani, 2003). For example, a user must visit at least two sites to obtain breadth information of all the facts about risk/prevention (e.g., Cancer.org and Harvard.edu,), and at least four sites to obtain depth information about each fact (e.g., AAD.org, Skincarephysicians.com, Cancer.org, Skincancer.org). A more detailed analysis across all topics at the page level is in progress. Admittedly, users visit other sources of healthcare information such as chat rooms, support groups, etc. However, while such sources often do provide valuable information such as alternate medicine and therapies, they are also fraught with misinformation and incomplete information (Culver, Gerr, & Fumkin, 1997).

One might argue that content providers must strive harder to make sure that the information they provide on relevant pages is complete. However, such an argument does not acknowledge the nature of information, especially as provided on the Web. Information on the Web (even in the best sites) is created by different authors, with different intentions (Eysenbach et al., 2002), and targeted to different audiences resulting in high variability along many dimensions. This analysis delves deeper into these issues to understand what sets of facts are provided in different kinds of sites (e.g., sites that are focused on cancer vs. melanoma). While there might be pages that comprehensively cover topics that have a small number of facts, the facts related to a vast number of topics will often have a scattered and complex distribution. This is the nature of most information on the Web. Hence, we must understand it, and design for it.

One might also argue that these results are an artifact of the facts used in the study, as they were particular to the set of physicians interviewed. The method of fact identification in Experiment 1 greatly reduces the probability of this external validity problem for three reasons. First, the initial list of facts was identified by visiting different high-quality Web sites pointed to by MEDLINEplus. Second, two doctors were interviewed independently, both of whom were active members of skin cancer professional societies where such patient education issues are regularly discussed. Third, because not all facts are of equal importance, the doctors rated the facts at different levels of importance. If another set of doctors went through this fact identification and rating exercise, they would most probably identify a similar list of facts for each topic. Furthermore, while not all searchers might be interested in finding comprehensive information about a topic all the time, it is an important goal in healthcare especially from a treatment compliance perspective. Getting a comprehensive understanding of a healthcare problem has important consequences, and hence easy access to such information is critical.

Finally, these experiments were not immune to the limitations of collecting and analyzing information on the Web. These limitations have been reported by other researchers (Bar-Ilan, 1999, 2001; Björneborn & Ingwersen, 2001, Thelwall, 2001b). This study focused on the distribution of facts across Web pages and Web sites. However, because there were no customized tools available, the time-consuming tasks of fact identification, page collection, and fact rating were manually performed. Such a process severely constrained the range of topics that could be analyzed, and the range of search engines that could be used. In our current research, we are exploring automated ways to identify the facts in a page using techniques such as latent semantic analysis (Dumais, Furnas, Landauer, & Deerwester, 1998), which is expected to trade-off some accuracy for speed, with the ultimate goal of making such analyses more practical.

Implications for Developers, Authors, and Trainers

As described above, the analysis of the distributions helped to pinpoint the knowledge required by users who wish to get a comprehensive understanding of a topic. This understanding has direct implications for search engine developers, page authors and designers, and trainers.

For search engine developers, the results provide the justification to explore new search engine paradigms other than the current dominant paradigm, espoused by search engines like Google. The current paradigm for search engines, as discussed in the Introduction, is to "get you to the right site" (Thottan, 2001, p. 33). However, our studies (Bhavnani, 2001, 2002; Bhavnani et al., 2003a) have shown that such an approach is more appropriate for questions that have specific answers (e.g., What is a melanoma?) rather than for a comprehensive under-

standing of a topic. For search tasks that require a comprehensive understanding, we believe search engine developers need to begin exploring ways that embrace the notion that information is scattered across Web pages and sites, and that in most cases, there is no single page or site that contains all the information.

There are a few attempts to explore the above idea. For example, similar to Carbonell and Goldstein (1998), researchers in the Novelty Track sessions (Soboro & Harman, 2003) within the Text Retrieval Conference (TREC) have focused on how to identify a small set of pages that together cover a topic completely with little overlap. While the above research has focused on a domain-independent approach to deal with information scatter, our own work has focused on a domain-specific approach (Bhavnani et al., 2003a). For example, we have built and tested a prototypical domain portal called the *Strategy Hub for Healthcare* that attempts to address the distribution of healthcare information across sources. When a user selects a topic such as descriptive information about melanoma risk/prevention, the Strategy Hub responds by suggesting a search procedure that first guides the user to visit a combination of general pages (that together provide an overview of all the relevant facts about melanoma risk/prevention), followed by specialized pages (that focus on specific facts about melanoma risk/prevention such as the danger of tanning booths). A pilot study (Bhavnani et al., 2003a) has shown that such a distribution-conscious system does improve a user's ability to search for comprehensive information.

Page authors and Web site designers might consider making more explicit which pages are general overviews, and which pages are more detailed, through the use of metadata, through better design of menus, or by adding appropriate text in the page itself. Furthermore, new tools could be developed to find pages within large Web sites that help to distinguish between pages with different densities of facts.

Trainers who teach how to perform effective searches could also benefit from the results of our analyses. The analyses suggest that trainers should include in their instruction *declarative* knowledge (concepts), and *procedural* knowledge (how to perform a task) about distributions. Declarative instruction should include how facts tend to be scattered across pages and Web sites in different amounts of detail, and how such a distribution makes searching for comprehensive information different from searching for a specific fact. Procedural instruction should go beyond teaching how to find relevant sources of information, but also how to visit relevant sources in a particular order. For example, such instruction could suggest that users read several general pages that provide an overview of the topic, before reading pages that are more specific. When alternate search engines that take into account information scatter become more practical, then trainers must steer students to those search engines for finding comprehensive information.

Summary and Conclusions

This research was motivated by two observations: (a) Novice searchers have difficulty finding comprehensive information, and (b) expert searchers know which combination of sources to visit in which specific order to obtain comprehensive information about a topic. Given the importance of finding comprehensive information by patients, we focused on the consumer health domain. We hypothesized that understanding the distribution of facts related to common healthcare topics across sources could shed light on the knowledge required to perform comprehensive searches, and lead to novel approaches to help users perform such tasks. However, we found no studies that had analyzed the distribution of facts across Web pages, and had explored the possible reasons for those distributions.

Therefore, two experiments were conducted to understand the distribution of melanoma facts across 10 high-quality sites, and the reasons for those distributions. The results of the experiments showed that: (a) the distributions of facts across pages for all five topics were skewed towards pages having few facts, (b) no page in any site had all the facts for any topic, (c) no combination of pages within a site contained all the facts for four of the five topics, and (d) the distribution of facts that were rated as *very important* and *extremely important* was only marginally different from the above results, and full coverage of those facts was inconsistent across the sites. Further analysis suggested the existence of general pages (that cover many facts in a medium amount of detail), specialized pages (that cover few facts in a high level of detail), and sparse pages that contain (few facts in very little detail). The skewed distributions therefore appear to occur because there were many more specialized and sparse pages compared to general pages. The analyses therefore provide an explanation for the skewed distribution of facts across Web pages.

The above results also pinpoint the kind of knowledge needed by a searcher to find comprehensive information when information is scattered across pages and Web sites. As such knowledge is not easily inferred from current general-purpose search engines or domain portals, the results provide direct implications for a *distribution-conscious* approach to the development of search systems, Web pages and sites, and training.

The results also provide more justification for the behavior of search experts like healthcare librarians. Such search experts visit a combination of select sources in a specific order when searching for comprehensive information because they have acquired an inherent understanding of the complexities in the distribution of healthcare information across sources. While much research has focused on identifying strategies for *finding* sources of information (e.g., Bates, 1979; Belkin, Cool, Stein, & Thiel, 1995; Drabenstott, 2000), far less is known about how experts *select* and *order* known and relevant sources of information. The results of this study suggest that a large part of search expertise must emerge from the complexities inherent in the types and distribution of the information within relevant pages. These complexities therefore need much more scrutiny than they have received in the past, and should be the focus of future research.

Acknowledgments

Many thanks to M.J. Bates for guiding my initial observations towards the formal study of information distributions. I appreciate the diligent statistical analysis done for this project by F. Peck. I also thank A. Abernathy, N. Belkin, C. Bichakjian, G. Furnas, T. Johnson, R. Little, J. Nardine, F. Reif, V. Strecher, R. Thomas, and G. Vallabha, for their contributions to the data collection and analysis.

References

Allen, E.S., Burke, J.M., Welch, M.E., & Rieseberg, L.H. (1999). How reliable is science information on the web? Nature, 402, 722.

Altman D.G. (1990). Practical statistics for medical research. London: Chapman and Hall.

Anderson, J.R. (1995). Learning and memory. New York: Wiley.

Bar-Ilan, J. (1998). The mathematician, Paul Erdos (1913–1996) in the eyes of the Internet. Scientometrics, 43(2), 257–267.

Bar-Ilan, J. (1999). Search engine results over time—A case study on search engine stability. Cybermetrics, 2/3, 1. Retrieved March 12, 2005, from http:// www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

Bar-Ilan, J. (2000a). The web as information source on informetrics? A content analysis. Journal of the American Society for Information Science, 51(5), 432–443.

Bar-Ilan, J. (2000b). Results of an extensive search for S&T indicators on the web—A content analysis. Scientometrics, 49(2), 257–277.

Bar-Ilan, J. (2001). Data collection methods on the Web for informetric purposes—A review and analysis. Scientometrics, 50(1), 7–32.

Bar-Ilan, J. (2002). Methods for measuring search engine performance over time. Journal of the American Society for Information Science and Technology, 53(4), 308–319.

Bar-Ilan, J., & Peritz, B.C. (1999). The life span of a specific topic on the web: The case of 'Informetrics': A quantitative analysis. Scientometrics, 46(3) 371–382.

Barabasi, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 509–512.

Bates, M.J. (1998). Indexing and access for digital libraries and the internet: Human, database, and domain factors. Journal of the American Society for Information Science, 49(13), 1185–1205.

Bates, M.J. (1979). Information search tactics. Journal of the American Society for Information Science 30(4), 205–214.

Bates, M.J. (2002). Speculations on browsing, directed searching, and linking in relation to the Bradford Distribution. In H. Bruce, R. Fidel, R. Ingwersen, & P. Vakkari (Eds.), Emerging Frameworks and Methods: Proceedings of the Fourth International Conference on Conceptions of Library and Information Science (CoLIS4) (pp. 137–150). Greenwood Village, CO: Libraries Unlimited.

Belkin, N., Cool, C., Stein, A., & Thiel, U. (1995). Cases, scripts, and information-seeking strategies: on the design of interactive information retrieval systems. Expert Systems with Applications, 9(3), 379–395.

Beredjiklian P.K., Bozentka D.J., Steinberg D.R., & Bernstein J. (2000). Evaluating the source and content of orthopedic information on the Internet: The case of carpal tunnel syndrome. Journal of Bone and Joint Surgery, 82, 1540–1543.

Bhavnani, S.K. (2001). Important cognitive components of domain-specific search knowledge. In E.M. Voorhees and D.K. Harman (Eds.), NIST Special Publication 500-250: The Tenth Text Retrieval Conference (TREC'01) (pp. 571–578). Washington, DC: NIST.

Bhavnani, S.K. (2002). Domain-specific search strategies for the effective retrieval of healthcare and shopping information. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves (CHI'02) (pp. 610–611).

Bhavnani, S.K. (2003). The distribution of online healthcare information: A case study on Melanoma. In Proceedings of the AMIA 2003 Annual Symposium (AMIA'03) (pp. 81–85).

Bhavnani, S.K., Bichakjian, C.K., Schwartz, J.L., Strecher, V.J., Dunn, R.L., Johnson, T.M., et al. (2002). Getting patients to the right healthcare sources: From real-world questions to Strategy Hubs. In Proceedings of the AMIA 2002 Annual Symposium (AMIA'02) (pp. 51–55).

Bhavnani, S.K., Bichakjian, C.K., Johnson, T.M., Little, R.J., Peck, F.A., Schwartz, J.L., et al. (2003a). Strategy hubs: Next-generation domain portals with search procedures. In Proceedings of the Conference on Human Factors in Computing Systems (CHI'03) (pp. 393–400).

Bhavnani, S.K., Jacob, R.T., Nardine, J., & Peck, F.A. (2003b). Exploring the distribution of online healthcare information. In Proceedings of the Conference on Human Factors in Computing Systems (CHI'03) (pp. 816–817).

Bichakjian, C., Schwartz, J., Wang, T., Hall J., Johnson, T., & Biermann, S. (2002). Melanoma information on the internet: Often incomplete-a public health opportunity? Journal of Clinical Oncology, 20(1), 134–141.

Biermann, J.S., Golladay, G.J., Greenfield, M.L., & Baker, L.H. (1999). Evaluation of cancer information on the internet. Cancer, 86(3), 381–390.

Björneborn, L., & Ingwersen, P. (2001). Perspectives on webometrics. Scientometrics, 50, 1, 65–82.

Bradford, S.C. (1948). Documentation. London: Crosby Lockwood.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98) (pp. 335–336).

Cronin, B., Snyder, H., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the web. Journal of the American Society for Information Science and Technology, 49(14),1319–1328.

Cui, L. (1999). Rating health web sites using the principles of citation analysis: A bibliometric approach. Journal of Medical Internet Research, 19(1), e4. Retrieved July 2003, from http://www.jmir.org/1999/1/e4/ index.htm

Culver, J., Gerr, F., & Fumkin, H. (1997). Medical information in the internet: A study of an electronic bulletin board. Journal of General Internal Medicine, 12(8), 466–471.

Davison, K. (1997). The quality of dietary information on the world wide web. Clinical Performance and Quality Health Care, 5, 64–66.

Drabenstott, K. (2000). Web search strategies. In W.J. Wheeler (Ed.), Saving the user's time through subject access innovation: Papers in honor of Pauline Atherton Cochrane (pp. 114–161). Champaign, Ill: University of Illinois.

Dumais, S.T., Furnas, G.W., Landauer, T.K., & Deerwester, S. (1988). Using latent semantic analysis to improve information retrieval. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'88) (pp. 281–285).

Eysenbach, G., & Köhler, C. (2002). How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. British Medical Journal, 324, 573–577.

Eysenbach, G., Powell, J., Kuss, O., & Sa, E-R. (2002). Empirical studies assessing the quality of health information for consumers on the world wide web: A systematic review. Journal of the American Medical Association, 287(20), 2691–2700.

Eng, T.R., Maxfield, A., & Gustafson, D. (1998). Access to health information and support: A public highway or a private road? Journal of the American Medical Association, 280(5), 1371–1375.

Florance, V., & Marchionini, G. (1995). Information processing in the context of medical care. In Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95) (pp. 158–163).

Fox, S., & Fallows, F. (2003). Health searches and email have become more commonplace, but there is room for improvement in searches and overall internet access. Pew Internet and American live project: Online life report. Retrieved May 13, 2005, from http://www.pewinternet.org/reports/ toc.asp?Report = 95

Griffiths, K.M., & Christensen, H. (2000). Quality of web based information on treatment of depression: Cross sectional survey. British Medical Journal, 321, 1511–1515.

Hinds, C., Streater, A., & Mood, D. (1995). Functions and preferred methods of receiving information related to radiotherapy: Perceptions of patients with cancer. Cancer Nursing, 18, 374–384.

Hood, W., & Wilson, C. (2001). The scatter of documents over databases in different subject domains: How many databases are needed? Journal of the American Society for Information Science, 52(14), 1242–1254.
Impicciatore, P., Pandolfini, C., Casella, N., & Bonati, M. (1997). Reliability of health information for the public on the world wide web: Systematic survey of advice on managing fever in children at home. British Medical Journal, 314, 1875–1879.
Jiang, Y.L. (2000). Quality evaluation of orthodontic information on the world wide web. American Journal of Orthodontics and Dentofacial Orthopedics, 118, 4–9.
Kirk, T. (1974). Problems in library instruction in four-year colleges. In: J. Lubans, Jr. (Ed.), Educating the library user (pp. 83–103). New York: R.R. Bowker.
Kleinberg, J., & Lawrence, S. (2001). The structure of the web. Science, 294, 1849–1850.
Lancaster, F.W. (1968). Information retrieval systems: Characteristics, testing, and evaluation. New York: Wiley.
McClung, H.J., Murray, H.D., & Heitlinger, L.A. (1998). The internet as a source for current patient information. Pediatrics, 101, 1–4.
Mills, M.E., & Sullivan, K. (1999). The importance of information giving for patients newly diagnosed with cancer: A review of the literature. Journal of Clinical Nursing, 8, 631–642.
Over, P. (1998). TREC-6 Interactive track report. In E.M. Voorhees & D.K. Harman (Eds.), NIST Special Publication 500-242: The Seventh Text Retrieval Conference (TREC'98).
Ream, E., & Richardson, A. (1996). The role of information in patients' adaptation to chemotherapy and radiotherapy: A review of the literature. European Journal of Cancer Care, 5, 132–138.
Rosenbaum, H. (1998). Web-based community networks: A study of information organization and access. In Proceedings of ASIS'98 (pp. 516–530).
Soboro, I., & Harman, D. (2003). Overview of the TREC 2003 Novelty Track. In E.M. Voorhees & L.P. Buckman (Eds.), NIST Special Publication 500-255: The Twelfth Text Retrieval Conference (TREC'03).
Soot, L.C., Moneta, G.L., & Edwards, J.M. (1999). Vascular surgery and the internet: A poor source of patient-oriented information. Journal of Vascular Surgery, 30, 84–91.
Sturdee, D.W. (2000). The importance of patient education in improving compliance. Climacteric, 10(2), 9–13.
Sullivan, D. (2001). Search engine sizes. Search engine watch (Online). Retrieved March 13, 2005, from http://www.searchenginewatch.com/reports/ article.php/2156481
Thelwall, M. (2001a). Extracting macroscopic information from web links. Journal of the American Society for Information Science and Technology, 52(13), 1157–1168.
Thelwall, M. (2001b). The responsiveness of search engine indexes. Cybermetrics, 5. Retrieved March 13, 2005, from http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html
Thottam, J. (2001, November). Search smarter. On Magazine, 33–37.
Van Halteren, H., & Teufel, S. (2003). Examining the consensus between human summaries: Initial experiments with factoid analysis. In Proceedings of the HLT-NAACL 2003 Text Summarization Workshop and Document Understanding Conference (DUC'03) (pp. 57–64).
Vaughan, L., & Thelwall, M. (2003). Scholarly use of the web: What are the key inducers of links to journal web sites? Journal of the American Society for Information Science and Technology, 54(1), 29–38.
Wormell, I. (2000). Critical aspects of the Danish welfare state—As revealed by issue tracking. Scientometrics, 48(2), 237–250.
Zipf, G.K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Cambridge, MA: Addison-Wesley.

Appendix A

Facts Identified for the Five Melanoma Topics

Self-Examination

1. Self-examination is used to find a potential melanoma on the skin.
2. During a self-examination, you should follow a prescribed method to check the entire skin surface.
3. During a self examination, you use the ABCDs [or ABCs; or size, shape, color] of melanoma to identify moles that might be melanoma.
4. During a self-examination, you should be looking out for itching [or bleeding; or tender] moles.
5. During a self-examination, you should look for dysplastic nevi [or irregular moles].
6. During a self-examination, you should use body maps to mark existing moles.
7. During a self-examination, you should check the entire skin surface [or entire body, or everywhere on body].
8. In the ABCDs of melanoma, "A" stands for asymmetry [also acceptable: look for asymmetrical moles; or look for moles where one half is different from the other half].
9. In the ABCDs of melanoma, "B" stands for border irregularity [also acceptable: look for irregular border; or look for a scalloped border].
10. In the ABCDs of melanoma, "C" stands for color [or color variance, or multicolored, or two or more colors, or mentions multiple colors]. [Also acceptable: look for moles with color variance; or look for moles with two or more colors; or look for moles with multiple colors.]
11. In the ABCDs of melanoma, "D" stands for diameter $>$ 6.0 mm [or width of a pencil eraser]. [Also acceptable: look for moles with diameter $>$ 6.0 mm; or look for moles with a width of a pencil eraser.]
12. Experts [or it is] recommend that you should perform a self-examination every month.
13. Confirm your self-diagnosis by consulting a local healthcare provider [or doctor, or dermatologist, or nurse practitioner, or physician's assistant].
14. Locate a local dermatologist (source for locating a dermatologist is provided).

Doctor's Exam

1. A doctor's [or healthcare professional's, or nurse practitioner's] examination determines if a biopsy should be done to test a mole for melanoma.
2. During a doctor's [or healthcare professional's, or nurse practitioner's] examination, the doctor will examine the entire skin surface [or entire body, or everywhere on body].
3. During a doctor's [or health care professional's, or nurse practitioner's] examination, the doctor will ask about your history of exposure to UV rays [or tanning beds, or sun exposure].
4. During a doctor's [or healthcare professional's, or nurse practitioner's] examination, the doctor will check your family history [or personal medical history] for cancer.
5. A doctor's [or healthcare professional's, or nurse practitioner's] examination should be done when there is a change in the size [or shape; or color; or feel] of an existing mole [or when you find a mole that matches one of the ABCDs].
6. Experts [or it is] recommended that people have regular [or every 3 years for 20–40-year-olds, or every year for above 40-year-olds] doctor's examination [or skin examination].

Diagnostic Tests

1. A skin biopsy is done to remove all or part of a suspicious growth to test if it is a melanoma.

2. A skin biopsy is the only way to be certain if a mole is a melanoma.
3. A skin biopsy is either a punch biopsy, saucerization [or deep shave] biopsy, or incisional/excisional biopsy.
4. The tissues removed during a biopsy are examined by a pathologist for melanoma cells.
5. A sentinel lymph node biopsy determines if melanoma has spread to the lymph nodes [or glands].
6. Scans of the liver [or bones, or brain, or chest] may be ordered based on findings of a history and physical examination.

Disease Stage

1. After melanoma is diagnosed, the doctor will determine the stage of the disease.
2. Staging is a method of determining the risk of spread.
3. Staging is used to help in the prognosis of melanoma.
4. Staging is used to help determine treatment options.
5. Melanoma is staged according to the TNM system [or tumor, lymph nodes, metastasis].
6. Stage 0 is when the melanoma is in the outer layer of skin only.
7. Stage I is when the melanoma tumor is less than 2 mm thick with no ulceration, or less than 1 mm thick with ulceration.
8. In stage I, the melanoma has not spread beyond the skin.
9. Stage II is when the melanoma tumor is between 1–2 mm thick with ulceration, or greater than 2 mm thick.
10. In stage II, the melanoma has not spread beyond the skin.
11. Stage III is when the melanoma has spread to the lymph nodes.
12. Stage IV is when the melanoma has spread to distant organs in the body.
13. Calculate the stage of your melanoma (method for calculation is provided).

Risk/Prevention

1. Having fair skin [or type I or II skin, or white skin, or tendency to burn, not tan, or green or blue eyes, or red or blond hair] increases your risk of getting melanoma [or skin cancer].
2. High UV exposure [or sunburn] increases your risk of getting melanoma [or skin cancer].
3. Having many moles [or more than 50 moles] increases your risk of getting melanoma.
4. Having dysplastic nevi [or atypical moles] increases your risk of getting melanoma [or skin cancer].
5. Having a giant [or $>$ 20 cm] congenital mole (or mole present at birth) increases your risk of getting melanoma [or skin cancer]. [Must mention "giant" and "congenital" or "mole present at birth."]
6. Having a family history of melanoma [or members of your family who have had melanoma] increases your risk of getting melanoma [or skin cancer].
7. Having a personal history of melanoma increases your risk of getting melanoma [or skin cancer].
8. If you have a weakened immune system [or immune deficiencies] and melanoma [or skin cancer], your risk of faster growth may be increased. [Must mention "have melanoma" and "increased risk of growth."]
9. If you have Xeroderma Pigmentosum, your risk of getting melanoma [or skin cancer] is increased.
10. Calculate your personal risk of getting melanoma (source of calculator is provided).
11. Wearing protective clothing can help to prevent melanoma.
12. Wearing sunscreen can help to prevent melanoma.
13. Avoiding UV Rays [or avoiding peak sunlight hours; or seeking shade] can help to prevent melanoma.
14. Examining your body for suspicious moles [or changing moles, or itching moles, or moles that match the ABCDs] can help to prevent melanoma from becoming a more advanced melanoma.

Appendix B

All 59 Queries[A1] Used to Identify Relevant Pages for the Five Melanoma Topics

Self-Examination

1. melanoma self examination
2. melanoma self examination find
3. melanoma self examination method
4. melanoma ABC OR ABCD OR size OR shape OR color
5. melanoma itching OR bleeding OR tender
6. melanoma dysplastic OR "irregular mole"
7. melanoma "body map" OR "body maps"
8. melanoma self examination entire OR everywhere
9. melanoma asymmetry
10. melanoma border
11. melanoma C color
12. melanoma diameter OR "pencil eraser"
13. melanoma self examination month
14. melanoma self examination doctor OR dermatologist OR "healthcare provider"
15. melanoma locate OR find doctor OR dermatologist OR "healthcare provider"

Doctor's Examination

1. melanoma doctor's OR doctor exam OR examination
2. melanoma doctor's OR doctor exam OR examination biopsy
3. melanoma doctor's OR doctor skin exam OR examination
4. melanoma doctor's OR doctor ask exam OR examination
5. melanoma doctor's OR doctor exam OR examination family OR genetic
6. melanoma change OR changing
7. melanoma exam OR examination regular OR "every year" OR "every three years"

Diagnostic Tests

1. melanoma diagnostic test
2. melanoma biopsy
3. melanoma biopsy certain OR sure

[A1]Each of these queries was used with the *site*: operator to search within a specific site. For example, Query 1 for self-examination was extended to melanoma self-examination site: cancer.gov when it was used to search within the cancer.gov site.

4. melanoma biopsy OR punch OR saucerization OR "deep shave" OR incisional OR excisional
5. melanoma pathologist examines OR examine
6. melanoma sentinel node OR SNB OR SLND
7. melanoma liver OR brains OR bones OR chest

Disease Stage

1. melanoma stage OR staging
2. melanoma diagnosis stage OR staging
3. stage OR staging
4. melanoma prognosis stage OR staging
5. melanoma treatment stage OR staging
6. melanoma TNM OR "tumor, lymph nodes, metastasis"
7. melanoma stage zero OR 0 OR "in situ"
8. melanoma stage + I OR 1 OR one
9. melanoma spread skin stage + I OR 1 OR one
10. melanoma stage II OR 2 OR two
11. melanoma spread skin stage II OR 2 OR two
12. melanoma stage III OR 3 OR three
13. melanoma stage IV OR 4 OR four
14. melanoma stage OR staging calculator OR "determine your stage"

Risk/Prevention

1. melanoma risk
2. melanoma prevention
3. melanoma fair OR "green eyes" OR "blue eyes" OR "red hair" OR blonde
4. melanoma risk UV OR ultraviolet OR sun OR sunlight OR sunburn
5. melanoma "many moles" OR "multiple moles" OR "lots of moles"
6. melanoma risk dysplastic OR irregular OR atypical
7. melanoma giant OR large congenital
8. melanoma family OR genetic
9. melanoma risk personal OR medical history
10. melanoma risk immune
11. melanoma "xeroderma pigmentosum" OR XP
12. melanoma risk personal OR estimate OR my
13. melanoma clothing OR clothes
14. melanoma sunblock OR sunscreen
15. melanoma prevent OR prevention UV OR ultraviolet OR sun OR sunlight
16. melanoma prevent OR prevention self examination

Appendix C

Three Curves Tested for All Distributions

	Power	Discrete exponential	Truncated Poisson
Risk/Prevention			
All facts	$Y = 47.019x^{-1.62}$ LR = 97.703, $p < .001$	**$Y = 33.308e^{-0.276x}$** **LR = 20.976, $p = .051$***	$Y = (e^{4.087} - 1)^{-1}(4.087^x/x!)$ LR = 61.270, $p < .001$
Very and extremely important facts	$Y = 48.069x^{-1.64}$ LR = 107.941, $p < .001$	**$Y = 36.871e^{-0.301x}$** **LR = 30.414, $p < .001$**	$Y = (e^{3.766} - 1)^{-1}(3.766^x/x!)$ LR = 49.440, $p < .001$
Self-exam			
All facts	$Y = 32.046x^{-1.64}$ LR = 62.949, $p < .001$	**$Y = 22.172e^{-0.275x}$** **LR = 26.058, $p = .011$**	$Y = (e^{4.092} - 1)^{-1}(4.092^x/x!)$ LR = 81.473, $p < .001$
Very and extremely important facts	$Y = 24.145x^{-1.56}$ LR = 76.060, $p < .001$	**$Y = 15.291e^{-0.243x}$** **LR = 23.247, $p = .010$**	$Y = (e^{4.753} - 1)^{-1}(4.753^x/x!)$ LR = 44.625, $p < .001$
Doctor's exam			
All facts	$Y = 30.273x^{-2.11}$ LR = 23.439, $p < .001$	$Y = 53.881e^{-0.764x}$ LR = 7.962, $p = .093$*	**$Y = (e^{1.458} - 1)^{-1}(1.458^x/x!)$** **LR = 7.922, $p = .094$***
Very and extremely important facts	$Y = 31.024x^{-2.65}$ LR = 9.273, $p = .002$	**$Y = 100e^{-1.253x}$** **LR = 4.794, $p = .029$**	$Y = (e^{0.715} - 1)^{-1}(0.715^x/x!)$ LR = 5.056, $p = .025$
Diagnostic tests			
All facts	$Y = 31.429x^{-1.99}$ LR = 29.307, $p < .001$	$Y = 47.437e^{-0.648x}$ LR = 8.863, $p = .065$*	**$Y = (e^{1.726} - 1)^{-1}(1.726^x/x!)$** **LR = 8.745, $p = .068$***
Very and extremely important facts	$Y = 36.158x^{-2.81}$ LR = 7.987, $p = .005$	$Y = 144.648e^{-1.439x}$ LR = 1.650, $p = .199$*	**$Y = (e^{0.550} - 1)^{-1}(0.550^x/x!)$** **LR = 0.645, $p = .422$***
Disease stage			
All facts	$Y = 34.218x^{-1.85}$ LR = 51.077, $p < .001$	**$Y = 32.302e^{-0.419x}$** **LR = 42.462, $p < .001$**	$Y = (e^{2.705} - 1)^{-1}(2.705^x/x!)$ LR = 92.631, $p < .001$
Very and extremely important facts	$Y = 34.956x^{-1.88}$ LR = 41.150, $p < .001$	**$Y = 37.322e^{-0.471x}$** **LR = 26.815, $p < .001$**	$Y = (e^{2.472} - 1)^{-1}(2.472^x/x!)$ LR = 54.221, $p < .001$

Note. Curves were fit using maximum likelihood estimation (MLE). Bolded text represent the best-fitting curve for each distribution. Asterisks indicate that the curve was accepted by the likelihood ratio (LR) test for goodness-of-fit at the .05 level.