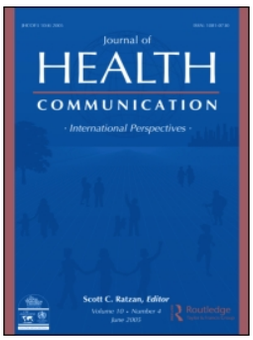

Taylor & Francis
Taylor & Francis Group

Journal of Health Communication
International Perspectives

ISSN: 1081-0730 (Print) 1087-0415 (Online) Journal homepage: http://www.tandfonline.com/loi/uhcm20

The Effect of Limited Health Literacy on How Internet Users Learn About Diabetes

Elad Yom-Tov, Barbara Marino, Jennifer Pai, Dawn Harris & Michael Wolf

To cite this article: Elad Yom-Tov, Barbara Marino, Jennifer Pai, Dawn Harris & Michael Wolf (2016) The Effect of Limited Health Literacy on How Internet Users Learn About Diabetes, Journal of Health Communication, 21:10, 1107-1114, DOI: 10.1080/10810730.2016.1222033

To link to this article: http://dx.doi.org/10.1080/10810730.2016.1222033

Published online: 29 Sep 2016.

Journal of Health Communication, 21: 1107–1114, 2016

ISSN: 1081-0730 print/1087-0415 online
DOI: 10.1080/10810730.2016.1222033

Routledge
Taylor & Francis Group

The Effect of Limited Health Literacy on How Internet Users Learn About Diabetes

ELAD YOM-TOV[1], BARBARA MARINO[2], JENNIFER PAI[2], DAWN HARRIS[3], and MICHAEL WOLF[4]

[1]*Microsoft Research Israel, Herzeliya, Israel*
[2]*Merck, Boston, Massachusetts, USA*
[3]*Merck, Upper Gwynedd, Pennsylvania, USA*
[4]*Division of General Internal Medicine and Geriatrics, Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA*

The Internet continues to be an important supplemental health information resource for an increasing number of U.S. adults, especially for those with a new or existing chronic condition. Here we examine how people use the Internet to learn about Type 2 diabetes and how health literacy (HL) influences this information-seeking behavior. We analyzed the searches of approximately 2 million people who queried for diabetes-related information on Microsoft's Bing search engine. The HL of searchers was imputed through a community-based HL score. Topics searched were categorized and subsequent websites were assessed for readability. Overall, diabetes information–seeking strategies via the Internet are similar among adults with limited and adequate HL skills. However, people with limited HL take a longer time to read pages that are quickly read by people with adequate HL and vice versa. Information seeking among the former is terminated prematurely, as is evident from a Hidden Markov Model of the search process. Our findings indicate that the reading level required to understand the majority of diabetes-related information is high. Especially on government websites, more than 80% of information requires a reading level corresponding to 7th grade or higher. Our results indicate that individuals with lower HL may disproportionately struggle with Internet searches and fail to get an equivalent benefit from this information resource compared to users with greater HL. Future interventions should target the quality and ease of navigation of health care websites and find ways to leverage other relevant professionals to encourage and promote successful information access on the Web.

For the past two decades, there has been increased attention directed at how well individuals are informed about their own or a loved one's health. As a result, the availability and quality of written and spoken health information from health care providers and systems has come under great scrutiny by those in the now established fields of health literacy and shared decision making (Berkman, Sheridan, Donahue, Halpern, & Crotty, 2011; Parker, Ratzan, & Lurie, 2003). Studies have repeatedly found that physicians often miss opportunities to adequately counsel patients on their disease and treatment during medical encounters. Similarly, most print materials distributed by health systems and pharmacies to date are not easily understood by patients, especially those with limited health literacy skills, and thus not relied on or even read (Williams et al., 1995). Given these many communication breakdowns, it is unclear how patients are gathering sufficient information to learn about their condition(s) and treatment.

The Internet continues to be one important supplemental health information resource for an increasing number of U.S. adults, especially for those with a new or existing chronic condition. According to a 2014 survey by the Pew Research Center (Fox & Rainie, 2014), access to and use of the Internet among adults is very high. Among the vast majority (87%) of Americans who use the Internet on any platform, two of three report using the Web to seek out health information. In the most basic steps, the majority of health information seekers begin their Internet inquiry with search engines. Thus, researchers have examined search patterns on these sites; findings suggest that individuals' information-seeking behaviors evolve rapidly and differ depending on whether they are looking for themselves or for a loved one (Ofran, Paltiel, Pelleg, Rowe, & Yom-Tov, 2012). Furthermore, Cartright, White, and Horvitz (2011) concluded that people tend to switch between evidence-based and hypothesis-directed information-seeking behaviors on the Internet. Given the immense amount of content available on health topics, it can be quite challenging for one to seek out, retrieve, and critically evaluate information. The quality of medical information online is also known to vary. Just as there are socioeconomic disparities in Internet access—in terms of both educational attainment and household income—there may be differences by health literacy skills among those using the Internet and having the proficiency to find valid and accurate health information that meets their needs. Thus, as the Internet continues to become an ever more valued and timely source for learning about one's health, patients with lower health literacy may become further marginalized if they are less able to access health information.

Address correspondence to Elad Yom-Tov, Microsoft Research Israel, 13 Shekar Street, Herzeliya 46733, Israel. E-mail: eladyt@microsoft.com

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/UHCM.

Health literacy is correlated with general literacy, but there are also differences between the two. The complexity of the health care system and its specific language, as well as the fact that health literacy is frequently utilized in times of stress, means that people's health literacy skills do not completely correspond to their general literacy (Martin et al., 2009).

To date, several studies have been published examining the relationship between health literacy and use of the Internet to learn about a health condition (e.g., Katz, Jacobson, Veledar, & Kripalani, 2007; Wolf et al., 2004). However, these investigations used self-report to evaluate the participants' trust in and use of the Internet. None of these studies examined actual use of the Internet to retrieve information related to a health condition. Here we sought to detail Web-based information-seeking behaviors of U. S. consumers and specifically narrowed the context to diabetes in order to focus the investigation on a manageable number of search terms while also providing a common case example. In addition, potential differences in content searches were explored by the reading level of the information and health literacy level. To do the latter, we used a modified RAND predictive model to determine consumers' probability of having limited health literacy skills. We hypothesized that individuals with more limited health literacy may have less effective health information–seeking behaviors, which could therefore explain the well-known disparities in disease and treatment knowledge and overall informed decision-making capacity by health literacy.

Materials and Methods

We extracted all English-language queries submitted to the Microsoft Bing search engine by users in the United States for the 3-month period from May 2014 to July 2014 (inclusive). For each query, we extracted the query text, the time and date, a list of pages visited by the user as a result of the query, the time spent on each page (referred to as *dwell time*), the approximate location of the user (given as his or her zip code), and an anonymized user identifier. We note two intrinsic limitations of the data used in the study: (a) We cannot distinguish between multiple users on the same machine; and (b) if a searcher used the search service on multiple devices, they would appear in logs with separate identifiers, one per device.

In order to maintain user privacy, data were first anonymized by hashing before we had access to them. Thus, each user identifier was replaced by a pseudorandom set of characters that made it possible to identify that two searches were made by the same user but not the identity of that user. Data were then aggregated prior to analysis, and no individual-level user datum was examined by us. The research was reviewed by the Microsoft Research institutional review board (IRB9672) and was deemed institutional review board exempt.

Filtering Queries to Diabetes-Related Terms

Queries were filtered to include only those that contained one of a list of 62 terms related to Type 2 diabetes. This list included various names of the disease, drugs and devices used to treat it, and symptoms of diabetes. The complete list is given in Appendix A.

The list of terms was developed as follows. First we found all users who used a small set of terms clearly related to diabetes at least 10 times during the data period. These terms were as follows:

1. AODM (adult-onset diabetes mellitus)
2. DMII (diabetes mellitus Type 2)
3. NIDDM (non-insulin-dependent diabetes mellitus)
4. T2DM (Type 2 diabetes mellitus)
5. Adult AND diabetes
6. Type 2 AND diabetes
7. Type II AND diabetes

We then found the pages clicked after these queries were made and kept those pages that were clicked 10 or more times. Finally, we found all those queries that appeared 10 or more times and led people to click on these pages. We filtered this list to remove irrelevant terms.

Scoring Pages by Their Reading Scores

Pages clicked by the users were accessed in November 2014, and each of the pages was given a reading level score (Collins-Thompson, Bennett, White, De La Chica, & Sontag, 2011). This score denotes an estimate of the reading proficiency needed to understand the page and is given as an integer between 1 and 12, referring to the minimal school grade level that a user needs to have completed to comprehend the document. The score is based on a statistical model generally derived from the percentage of the population that should be familiar with the words in the document.

All data were analyzed as individual searches except in the last section (state transitions, see below), in which data were analyzed at the individual user level.

Health Literacy

Finally, we imputed a community-based health literacy (CBHL) score of users from the zip code of users in a similar manner as Martin and colleagues (2009) and Hanchate, Ash, Gazmararian, Wolf, and Paasche-Orlow (2008). The CBHL score is the percentage of adults with limited literacy proficiency in the county in which they reside. This percentage was based on the health literacy component of the National Assessment of Adult Literacy survey (Hanchate et al., 2008; Martin et al., 2009). Martin and colleagues used gender, age, race/ethnicity, educational attainment, poverty status, marital status, language spoken in the home, metropolitan statistical area, and length of time in the United States to create a regression model to predict individual health literacy scores as recorded in the National Assessment of Adult Literacy. These variables were then applied to census data to create county-level information on average health literacy proficiency, which are the CBHL scores used in this work.

Imputation of search engine users by zip code was shown previously to be a relatively accurate way to estimate user-level parameters of interest, for example, demographics such as income, education, race, and the likelihood that English is the first language spoken at home (Weber & Castillo, 2010); voting patterns (Borra & Weber, 2012; Yom-Tov, Dumais, & Guo,

2013); and attitudes toward vaccination (Yom-Tov & Fernandez-Luque, 2014). Although the imputation of search engine users with zip code–level data cannot be correct for all users in the zip code, it is correct on average and is useful (as shown in the aforementioned work) with large sample sizes such as the one analyzed herein (see Appendix B).

We refer to users with CBHL lower than the median as having adequate health literacy and to users with CBHL higher than the median as having limited health literacy.

Information Seeking

We categorized the 1,000 most popular pages (by clicks) into one of seven categories, as follows:

1. Medical information (pages devoted to reporting medical information)
2. General search engines (e.g., Bing, Google)
3. Social media (any page that has as its primary purpose supporting dialogue between a user and others with his or her illness; a few pages allow dialogue between the user and experts as well as nonexperts; all allow the user to choose the discussion topic)
4. Diet and fitness (pages devoted to providing diet and/or fitness information)
5. E-magazines for diabetics (online content in the form of a magazine with a broad range of information related to diabetes, including diet and fitness, reviews of diabetic supplies, discussion of the disease and treatments)
6. Nonspecific e-magazines (online content in the form of a magazine or book or newspaper with content that is not specific to diabetes)
7. Other

The first 200 pages were manually categorized by one of the authors. The remaining 800 pages were labeled by five crowdsourcing workers on the CrowdFlower website (www.crowdflower.com). The majority of workers (three or more of five) agreed on the same label for 90.5% of the pages. Therefore, exactly one of the categories fit the vast majority of pages.

Modelling the Transitions Between Hidden States

We examined the order in which users browsed pages in different categories. The need for information was previously shown (Ofran et al., 2012) to be dynamic, changing as a function of physical and mental changes (Butow et al., 1997). Changing patterns of the type of information requested may reflect a transition between mental states common to individuals who share clinical and psychological states. These unobservable states are hidden but may be identified from the search patterns using Hidden Markov Models (HMMs; Eddy, 2004), which estimate the hidden states of a system, the probability of transition between them, and the likelihood of each observable signal given the hidden state. We used HMMs to group searches into states and to predict the sequence of state changes based on the sequence of observations. In this case we were interested in predicting when the search pattern results in usable, diverse knowledge (as evident by state changes) based on the patterns in the individual searches (sequence of observations).

To find the most likely number of hidden states users went through during the time we observed them, we first trained HMMs of a varying number of hidden states. Visible categories were defined by the category of pages and further stratified by reading level score (1–6, 7–12; i.e., health pages with a 7 or greater reading level score).

Results

Basic statistics on queries are in Table 1. Following Yom-Tov, Borsa, Hayward, McKendry, and Cox (2015), we ascertained whether most queries referred to people with diabetes (or who suspected they had it) versus people inquiring for family members. We tested for expressions that referred to self or to others (either "I" or "my wife/husband/spouse/son/daughter/boyfriend/girlfriend"). Approximately 94% of users who queried with these expressions referred to themselves rather than to family members. This percentage rose to 95% in the population of users who queried five or more times for diabetes-related information. The Spearman correlation between the number of people making a diabetes-related query in each state and the age-adjusted 2012 incidence of diabetes by state (Centers for Disease Control and Prevention, n.d.) was .51 (p = .0001), again suggesting that many of the people asking diabetes-related queries may likely have been patients. The average value of CBHL was 13.6 (SD = 7.3), and the median was 12.1.

Users with adequate health literacy searched for information on diabetes on average for a period of 2.8 days (times between the first and last diabetes-related queries of a user) compared to 2.7 for users with limited health literacy. Among users who searched five or more times, the corresponding search times were 12.7 and 12.1 days, respectively.

Site Variability and Dwell Times

The average dwell time for the searches of all people across the pages they visited was 97 seconds. People with limited health literacy had a slight tendency for longer dwell times, for those pages that they read a long time. We differentiate between pages that were read for less than 120 seconds and those that were read for longer, as we assume that medical information pages would require longer dwell times for comprehension and, conversely, that lower dwell times imply that the users did not spend enough

Table 1. Basic statistics

	All users who made diabetes queries	Only users with five or more diabetes queries
Number of users	~2 million	281,954
Average number of days on Bing during data period	47	56
Average length of period asking about diabetes (days)	3	14
Average number of all queries	254	406
Average number of diabetes queries	2.8	12.4

time to acquire the information therein. The median dwell time for pages with dwell times of 120 seconds or more was 178.5 seconds for people with limited health literacy and 177.8 for people with adequate health literacy (a 0.4% difference, rank sum, $p = .002$). This indicates that people with limited health literacy spend longer reading those pages they decide to focus on.

For 14,174 pages that were viewed by five or more people with adequate health literacy and five or more with limited health literacy, we computed the median dwell time for the two groups of people at regular intervals of dwell time. The result is shown in Figure 1. It is interesting that pages that required reading times up to approximately 90 seconds for people with adequate health literacy required longer reading times by people with limited health literacy. However, above a dwell time of 90 seconds the picture reversed, and people with limited health literacy typically read pages for a shorter time than people with adequate health literacy.

Table 2 shows the percentage of pages from each category. Diet and fitness pages accounted for 69% of pages visited. Table 1 also shows the percentage of clicks made by users in response to diabetes-related queries on these pages. Although pages classified as being predominantly medical information reflected 4% of visited sites, they generated 80% of clicks, indicating more in-depth search patterns for that information. This was followed by diet and fitness, with 16% of total clicks.

We note that all pages viewed by people who queried referring to others (see above) were in the medical information category compared to the pages viewed by people referring to themselves, which were distributed among all categories.

We tested the dependence of the fraction of clicks to each category by CBHL decile. Two categories showed strong correlations: diet and fitness ($R^2 = .59$, slope = 0.002) and social media ($R^2 = .30$, slope = −0.00005). These correlations indicate that people with limited health literacy browsed more diet and fitness pages than people with adequate health literacy but fewer social media pages.

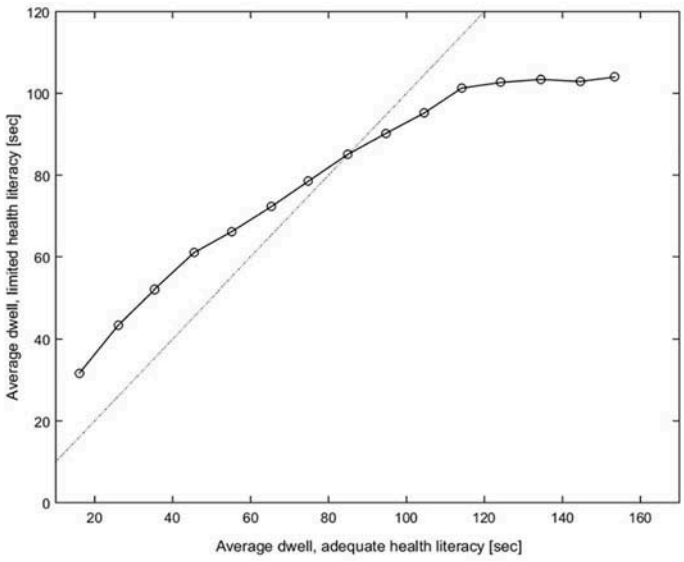

Fig. 1. Dwell time by literacy for matched pages. The dotted line shows identical median dwell times. sec = seconds.

Table 2. Percentage of pages and clicks on pages from each category

Category	Percentage of pages	Percentage of clicks
Medical information	4	80
General search engines	15	<0.5
Social media	1	<0.5
Diet and fitness	69	16
E-magazines for diabetics	7	1
Nonspecific e-magazines	1	<0.5
Other	2	3

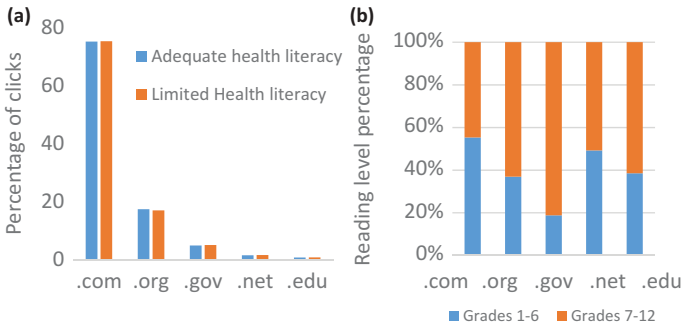

Fig. 2. Percentage of (a) page clicks by top-level domains and (b) reading level scores by top-level domains.

People tend to use the source of a page as an indicator of its authoritativeness (Walther, Wang, & Loh, 2004). Indeed, pages vary in their information content, style, and readability according to their source, and this may affect people's ability to understand page content. Therefore, we analyzed pages according to their source, as evident in their address, and measured the correlation between readability and health literacy, stratified by page source.

Figure 2a shows the percentage of clicks to pages according to their top-level domain (i.e., .com, .gov, etc.), stratified by whether users had adequate or limited health literacy. There were little to no differences between the percentages of clicks. Figure 2b shows the reading level scores of these domains; most diabetes-related pages were skewed toward higher reading level scores. Consumers with limited health literacy were more likely to visit pages with a higher reading difficulty.

Modelling the Transitions Between Hidden States

As explained in the Materials and Methods section, we examined the order in which users browsed pages in different categories using an HMM. The best number of hidden states was found using holdout, as follows: We trained the HMM using browsing data from 50% of the users and tested the model accuracy by comparing the predicted sequences of searches for the remaining 50% of users with their actual behavior. This

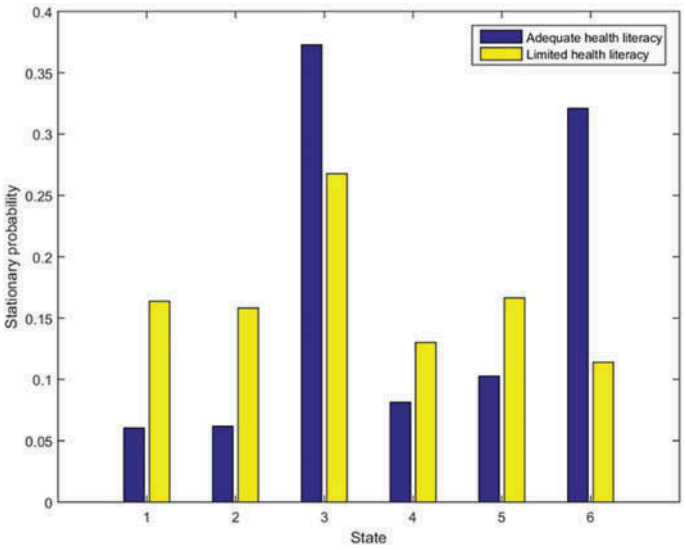

Fig. 3. Stationary probabilities of the hidden states of the Hidden Markov Model.

procedure was repeated 10 times with a random selection of the initial model parameter to reduce the chance of convergence to a local minimum. We conducted this procedure separately for users with limited and adequate health literacy and attempted to fit a model with two to 20 states. The HMM that reached the lowest holdout error in both cases had six hidden states. We therefore report results using this number of states.

Figure 3 shows that the stationary probabilities of the HMMs—that is, the probability that a user will end his or her search process in each of the hidden states—differ by health literacy level. Note that if searches ended at random, each state would be equally likely, that is, a uniform distribution of the stationary probabilities. People with adequate health literacy had two dominant states (3 and 6) compared to people with limited health literacy, where a slight tendency toward State 3 was observed, but overall the stationary probabilities were similar. In addition, the average ratio between the probability of staying within a state and that of transitioning to another state was 22.2 for people with limited health literacy compared to 9.2 for people with adequate health literacy. Taken together, these findings suggest that people with limited health literacy tend to explore less (move less within the HMM) and also end their process almost randomly, perhaps because they are unable to complete it to their satisfaction.

Table 3 shows the most likely output states for each of the hidden states for users with both limited and adequate health literacy. As can be expected, medical information appeared prominently, and the diet and fitness category was a likely output state for users with limited health literacy. It is interesting that output states for people with adequate health literacy were from low reading level scores in four of the six states compared to three in people with limited health literacy. This means that people with limited health literacy have a slight preference for pages that require a high reading level, even though they may not possess it and, as shown previously, may be less likely to be able to understand them.

Discussion

The Pew Internet Research Survey shows that in the United States, the vast majority of Internet users who have a medical concern turn to the Internet to learn about it. In this article we provide a large-scale analysis of how people obtain information online about one of the most prevalent chronic conditions, Type 2 diabetes. This analysis was based on observational data collected through the normal process of Internet search and is therefore unique in its scale and setting. To enable stratification of users by health literacy we used a community scoring–based approach that used location as a proxy for health literacy rates. This approach has been utilized in the past to approximate voting patterns and vaccination rates (Yom-Tov et al., 2015; Yom-Tov & Fernandez-Luque, 2014), but to the best of our knowledge this is the first time it has been used to impute health literacy on a large scale and for the purpose of examining health information–seeking behavior.

Our investigation suggests that there are important differences between the online learning processes of people with limited health literacy and those with adequate health literacy, which may cause the former to benefit less from the Internet as an informational source for learning about one's condition and management. This is of concern, as prior studies have found that less literate patients may be less likely to seek out health information beyond their physician and that they also ask fewer questions of their doctor during medical encounters compared to patients with adequate literacy (Katz et al., 2007; Serper et al., 2013; Wolf et al., 2004).

Table 3. Most likely output category

Hidden state number	Adequate health literacy	Limited health literacy
1	Medical information (high read)	Uncategorized (low read)
2	Uncategorized (high read)	Medical information (high read)
3	Uncategorized (low read)	Medical information (low read)
4	Medical information (low read)	Diet and fitness (high read)
5	Uncategorized (low read)	Uncategorized (high read)
6	Uncategorized (low read)	Medical information (low read)

Note. High/low read refers to high/low reading level scores.

The differences in the learning processes of people with limited health literacy compared to people with adequate health literacy are evident in how users access individual pages as well as the learning process itself. As shown in Figure 1, people with limited health literacy typically spend less time on pages that people with adequate health literacy spend a long time reading and vice versa. We attribute this to the difficulty people with limited health literacy have in parsing certain pages. When pages are not too difficult this can be overcome by spending a longer amount of time on each page. However, beyond a certain point these pages become too difficult, and people with limited health literacy abandon their effort.

Though there are also differences in the topics accessed by these two populations, these are relatively small. Specifically, diet and fitness is positively correlated with health literacy, whereas social media is negatively correlated. The former may be explained in a few ways. It may be that as the more common topics associated with the management of diabetes, the preponderance of searches by individuals with lower health literacy revolve around these issues. These individuals may be less knowledgeable of other relevant issues to apply toward their search, or they require more background information about these topics than individuals with adequate health literacy whose baseline understanding of diet and fitness might be higher. Yet another explanation might pertain to the strong correlation between health literacy levels and the availability of comprehensive medical insurance (Kutner, Greenberg, Jin, & Paulsen, 2006). Thus, if one has less access to medical services, nonmedical diabetes self-management solutions may be sought out. The latter, inverse health literacy association with social media may indicate a stronger tendency among people with adequate health literacy to engage with others to find a solution to their condition. In addition, there were little to no differences in the percentages of clicks to different domains made by people of different health literacy levels (see Figure 2), though most diabetes-related pages were skewed toward higher reading level scores. Consumers with limited health literacy were more likely to visit pages with a higher reading difficulty. This is also evident in Table 3, where people with limited health literacy abandon their search in states associated with pages of a higher reading level, lending additional support to the understanding that these people abandon their search because they are frustrated with their inability to gain the relevant information from the pages they read. We conclude that people with limited health literacy are reading information that may be above their ability to process.

Taken together, our findings show that people with limited health literacy spend more time than others on pages that can be read relatively quickly but less time on pages that require a longer time to understand. They also read pages that are typically written in language that is less interpretable to them, including, for example, government pages and those from educational facilities (e.g., universities). Therefore, to some extent, although the authoritativeness of pages probably draws people to them, people are hindering their learning process through their use of inappropriate selection of language.

The learning process itself also differs by health literacy. As the HMM demonstrated, people with limited health literacy tend to end their search at a random state. This again suggests an unsatisfactory end to the search process compared to the process experienced by people with adequate health literacy.

Limitations

Our study has several limitations. Although we had an unprecedented opportunity to examine consumers' diabetes-related Internet searches on a massive scale, our study data were from one search engine, and the nature of the data did not allow us to have a wide list of individual-level covariates. Similarly, we had to impute health literacy levels using community-level and demographic data. Although this approach has been validated in other studies, it is very possible that there was misclassification error. However, our general trends have considerable face validity and are supported by previous related findings pertaining to health information seeking among less literate adults. Our outcome also was Internet information-seeking behavior and not a reflection of actual retrieval and comprehension of diabetes material. Differences by health literacy level were also very small, and even though statistical significance was attained this was more than likely the result of our sample size and statistical power. Finally, our sample included consumers who had access to the Internet, and therefore we may have had an underrepresentation of individuals of lower socioeconomic status, those who are less educated, and ultimately those who are less computer literate (likely to overlap with those who are also less health literate). Future studies should remedy these many concerns by prospectively engaging a diverse sample of adults (especially those with lesser proficiency) and observing their Internet search proficiencies and subsequent comprehension and use of acquired health information.

In summary, individuals with lower health literacy and access to the Internet may disproportionately struggle with Internet searches and fail to get an equivalent benefit from this supplementary health information resource compared to others who are more health literate. Without proper orientation and technical support, this access barrier may explain documented disparities in Internet use for health care purposes and further well-established health care inequities if these individuals are less informed about their disease and treatment. Future research should seek to confirm these findings and link Internet health information–seeking behavior to comprehension and health behavior outcomes. Once confirmed, future interventions should target the quality and ease of navigation of health care websites and find ways to leverage care coordinators, health coaches, librarians, and other relevant professionals to be available resources to encourage and promote successful information access on the Web (Bailey et al., 2015).

Our investigation was observational in nature and used proxies to infer health literacy. We are currently attempting to replicate our results in a laboratory setting, which, albeit smaller in scale, allows for direct measurement of the experimental parameters. This will provide direct validation of our results and allow us to test possible solutions to providing appropriate content to individuals with varying levels of health literacy.

References

Bailey, S. C., O'Conor, R., Bojarski, E. A., Mullen, R., Patzer, R. E., Vicencio, D., … Wolf, M. S. (2015). Literacy disparities in patient

access and health-related use of Internet and mobile technologies. *Health Expectations*, *18*(6), 3079–3087. doi:10.1111/hex.12294

Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., & Crotty, K. (2011). Low health literacy and health outcomes: An updated systematic review. *Annals of Internal Medicine*, *155*(2), 97–107. doi:10.7326/0003-4819-155-2-201107190-00005

Borra, E., & Weber, I. (2012). Political insights: Exploring partisanship in Web search queries. *First Monday*, *17*(7). doi:10.5210/fm

Butow, P. N., Maclean, M., Dunn, S. M., Tattersall, M. H. N., & Boyer, M. J. (1997). The dynamics of change: Cancer patients' preferences for information, involvement and support. *Annals of Oncology*, *8*(9), 857–863. doi:10.1023/A:1008284006045

Cartright, M. A., White, R. W., & Horvitz, E. (2011, July). Intentions and attention in exploratory health search. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 65–74). ACM Digital Library. Retrieved from http://dl.acm.org/citation.cfm?id=2009929

Centers for Disease Control and Prevention. (n.d.). *United States Diabetes Surveillance System: Diagnosed diabetes*. Retrieved from http://gis.cdc.gov/grasp/diabetes/DiabetesAtlas.html

Collins-Thompson, K., Bennett, P. N., White, R. W., De La Chica, S., & Sontag, D. (2011, October). Personalizing Web search results by reading level. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management* (pp. 403–412), ACM Digital Library. Retrieved from http://dl.acm.org/citation.cfm?id=2063639&CFID=826865161&CFTOKEN=37290699

Eddy, S. R. (2004). What is a hidden Markov model? *Nature Biotechnology*, *22*(10), 1315–1316. doi:10.1038/nbt1004-1315

Fox, S., & Rainie, L. (2014). *The Web at 25 in the US*. Retrieved from the Pew Research Center website: http://www.pewinternet.org/2014/02/27/the-web-at-25-in-the-us

Hanchate, A. D., Ash, A. S., Gazmararian, J. A., Wolf, M. S., & Paasche-Orlow, M. K. (2008). The Demographic Assessment for Health Literacy (DAHL): A new tool for estimating associations between health literacy and outcomes in national surveys. *Journal of General Internal Medicine*, *23*(10), 1561–1566. doi:10.1007/s11606-008-0699-5

Katz, M. G., Jacobson, T. A., Veledar, E., & Kripalani, S. (2007). Patient literacy and question-asking behavior during the medical encounter: A mixed-methods analysis. *Journal of General Internal Medicine*, *22*(6), 782–786. doi:10.1007/s11606-007-0184-6

Kutner, M., Greenberg, E., Jin, Y., & Paulsen, C. (2006). *The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy* (NCES 2006-483). Retrieved from the National Center for Education Statistics website: https://nces.ed.gov/pubs2006/2006483.pdf

Martin, L. T., Ruder, T., Escarce, J. J., Ghosh-Dastidar, B., Sherman, D., Elliott, M., … Lurie, N. (2009). Developing predictive models of health literacy. *Journal of General Internal Medicine*, *24*(11), 1211–1216. doi:10.1007/s11606-009-1105-7

Ofran, Y., Paltiel, O., Pelleg, D., Rowe, J. M., & Yom-Tov, E. (2012). Patterns of information-seeking for cancer on the Internet: An analysis of real world data. *PLoS ONE*, *7*(9), e45921. doi:10.1371/journal.pone.0045921

Parker, R. M., Ratzan, S. C., & Lurie, N. (2003). Health literacy: A policy challenge for advancing high-quality health care. *Health Affairs*, *22*(4), 147–153. doi:10.1377/hlthaff.22.4.147

Serper, M., McCarthy, D. M., Patzer, R. E., King, J. P., Bailey, S. C., Smith, S. G., … Wolf, M. S. (2013). What patients think doctors know: Beliefs about provider knowledge as barriers to safe medication use. *Patient Education and Counseling*, *93*(2), 306–311. doi:10.1016/j.pec.2013.06.030

Walther, J. B., Wang, Z., & Loh, T. (2004). The effect of top-level domains and advertisements on health Web site credibility. *Journal of Medical Internet Research*, *6*(3), e24. doi:10.2196/jmir.6.3.e24

Weber, I., & Castillo, C. (2010). The demographics of Web search. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 523–530), ACM Digital Library. Retrieved from http://dl.acm.org/citation.cfm?id=1835537

Williams, M. V., Parker, R. M., Baker, D. W., Parikh, N. S., Pitkin, K., Coates, W. C., & Nurss, J. R. (1995). Inadequate functional health literacy among patients at two public hospitals. *Journal of the American Medical Association*, *274*(21), 1677–1682. doi:10.1001/jama.1995.03530210031026

Wolf, M. S., Davis, T. C., Cross, J. T., Marin, E., Green, K., & Bennett, C. L. (2004). Health literacy and patient knowledge in a southern US HIV clinic. *International Journal of STD & AIDS*, *15*(11), 747–752. doi:10.1258/0956462042395131

Yom-Tov, E., Borsa, D., Hayward, A. C., McKendry, R. A., & Cox, I. J. (2015). Automatic identification of Web-based risk markers for health events. *Journal of Medical Internet Research*, *17*(1), e29. doi:10.2196/jmir.4082

Yom-Tov, E., Dumais, S., & Guo, Q. (2013). Promoting civil discourse through search engine diversity. *Social Science Computer Review*, *32*(2), 145–154.

Yom-Tov, E., & Fernandez-Luque, L. (2014). Information is in the eye of the beholder: Seeking information on the MMR vaccine through an Internet search engine. *AMIA annual Symposium Proceedings*, *2014*, 1238.

Appendix A: List of Terms Used for Filtering Queries

Actoplus	glumetza	precose	hgba1c
Actos	glyburide	rosiglitazone	high AND a1c
Amaryl	glynase	saxagliptin	high sugar
Avandamet	insulin glargine	tradjenta	diabetic
Avandia	invokana	victoza	insulin
Bydureon	janumet	aodm	type 1 AND diabetes
Byetta	januvia	dmii	hypoglycemia
canagliflozin	kombiglyze	niddm	low AND blood AND sugar
Diabeta	lantus	t2dm	sugar AND urine
Exenatide	linagliptin	adult AND diabetes	metformin
Fortamet	liraglutide	type 2 AND diabetes	insulin AND pumps
glimepiride	metaglip	type ii AND diabetes	glucometer
Glipizide	metformin	sugar AND diabetes	sugar AND meter
glucophage	micronase	diabetes	antidiabetic drug
Glucotrol	onglyza	blood AND sugar	
glucovance	pioglitazone	a1c	

Appendix B: Can Health Literacy, Imputed From Zip Codes, Be Used to Identify the Health Literacy of Individuals?

We asked 30 people from the crowdsourcing website CrowdFlower to complete the Short Assessment of Health Literacy–English (available at http://www.ahrq.gov/professionals/quality-patient-safety/quality-resources/tools/literacy/index.html) and provide us with their current 5-digit zip code. We then measured the correlation between the number of correct responses and the community-based health literacy score for their county.

The correlation was –.436 (p = .016), indicating a good match between imputed health literacy and health literacy as measured by the Short Assessment of Health Literacy–English.