Using Search Queries to Understand Health Information Needs in Africa

Rediet Abebe
Cornell University
red@cs.cornell.edu

Shawndra Hill
Microsoft Research
shawndra@microsoft.com

Jennifer Wortman Vaughan
Microsoft Research
jenn@microsoft.com

Peter M. Small
Stony Brook University
Peter.Small@stonybrook.edu

H. Andrew Schwartz
Stony Brook University
has@cs.stonybrook.edu

ABSTRACT

The lack of comprehensive, high-quality health data in developing nations creates a roadblock for combating the impacts of disease. One key challenge is understanding the health information needs of people in these nations. Without understanding people's everyday needs, concerns, and misconceptions, health organizations and policymakers lack the ability to effectively target education and programming efforts. In this paper, we propose a bottom-up approach that uses search data from individuals to uncover and gain insight into health information needs in Africa. We analyze Bing searches related to HIV/AIDS, malaria, and tuberculosis from all 54 African nations. For each disease, we automatically derive a set of common search themes or *topics*, revealing a wide-spread interest in various types of information, including disease symptoms, drugs, concerns about breastfeeding, as well as stigma, beliefs in natural cures, and other topics that may be hard to uncover through traditional surveys. We expose the different patterns that emerge in health information needs by demographic groups (age and sex) and country. We also uncover discrepancies in the quality of content returned by search engines to users by topic. Combined, our results suggest that search data can help illuminate health information needs in Africa and inform discussions on health policy and targeted education efforts both on- and offline.

KEYWORDS

Computational Social Science, Language, Infectious Diseases, Developing Nations, Large-scale Data Sources

ACM Reference format:
Rediet Abebe, Shawndra Hill, Jennifer Wortman Vaughan, Peter M. Small, and H. Andrew Schwartz. 2016. Using Search Queries to Understand Health Information Needs in Africa.

1 INTRODUCTION

New technologies and sources of data are constantly being leveraged to upgrade and supplement the design, monitoring, and evaluation of health policy, a phenomenon the United Nations has dubbed the *Data Revolution* [6]. There is, however, a substantial gap in the availability and quality of health data between developing and developed nations. In many developing nations, even when health-related information is collected, it is often neither comprehensive nor digitized. The 2014 regional report by the African Union highlights this issue, noting: "Unless gaps are identified early and accurately, simply providing a raft of general interventions will not meet the real health needs of the people in the Region" [5].

This lack of data can be a roadblock to identifying major public health concerns and implementing effective interventions, such as targeted education efforts. While targeted education that addresses individuals' health needs is a critical tool for combating disease, health organizations and policy makers struggle to identify what knowledge individuals in developing nations seek and whether their health information needs are being met. It is especially urgent to understand how information needs vary by region and demographic groups since the impact of a disease—its prevalence, progression, and transmission rates—as well as people's knowledge and attitudes about the disease often vary regionally and demographically. Limited understanding on the health information needs of individuals limits the efficacy of gender- and age-specific programming [2, 4, 7, 8, 16, 19, 21].

In this paper, we take a step towards narrowing the gap, focusing on the problem of identifying and measuring people's everyday health information needs, concerns, and misconceptions. We use search queries originating in all 54 African nations to explore which themes, or *topics*, related to infectious disease people are most interested in getting information about, as evidenced by their searches. We focus on HIV/AIDS, malaria, and tuberculosis because together, these three diseases account for 22% of the disease burden in sub-Saharan Africa [1].

Search engine data provide a wealth of information on people's real-time activities, experiences, concerns, and misconceptions relatively cheaply [24, 28, 33], allowing us to obtain potentially hard-to-survey information in a bottom-up manner. In contrast, most data-driven efforts aimed at mitigating the impact of disease in data-sparse regions, including the Global Burden of Disease Study and the African Health Observatory, have utilized a top-down approach [1, 10], actively collecting data with a particular goal or question in mind. Such approaches, while helpful, are often limited in their ability to provide a thorough and comprehensive overview of people's information needs, attitudes, and misconceptions. Existing bottom-up solutions to this problem, such as the West Africa Health Organization's study of health information needs in West Africa [12], primarily make use of manual surveys or interviews. These approaches can obtain a comprehensive picture of individuals' needs, but are difficult to scale, expensive, and time-consuming. Analyzing search data is a natural candidate for scaling up studies not only because it addresses some of these challenges, but also

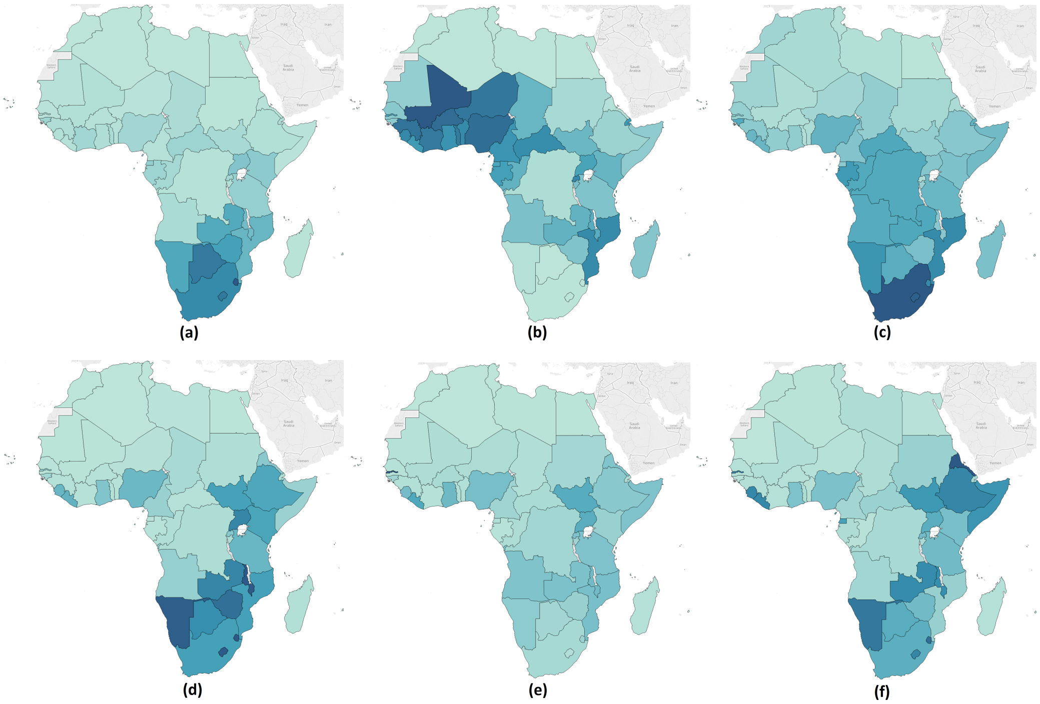

Figure 1: Top: Heat maps showing 2016 rates of (a) HIV/AIDS prevalence (ages 15–49), (b) malaria incidence, and (c) tuberculosis incidence. Bottom: Heat maps showing percentage of total search traffic containing the words (d) "HIV" or "AIDS", (e) "malaria", and (f) "tb" or "tuberculosis." The Spearman correlation is $\rho = 0.714$ $[0.689, 0.737]$ for HIV/AIDS, $\rho = 0.402$ $[0.360, 0.442]$ for malaria, and $\rho = 0.462$ $[0.422, 0.499]$ for tuberculosis.

because search logs have already been shown to contain large quantities of information related to serious and stigmatizing conditions in other contexts [17, 28]. Despite the fact that Internet penetration in Africa is growing rapidly—31% of the population is currently covered, with nearly 8,500% growth since 2000 [9]—to our knowledge, no prior work has looked specifically at search data to understand health information needs in African nations.

To uncover the themes in which individuals in Africa are interested, we use latent Dirichlet allocation (LDA), a generative statistical model that can be used to automatically extract lexical topics—sets of semantically-related words—from text [14]. The topics that emerge cover basics such as symptoms and treatment, as well as stigma and discrimination, natural cures and remedies, and concerns about breastfeeding and gender inequality. We explore the ways in which the popularity of these topics vary by gender, sex, and location. Finally, we compare the organic search results returned for different topics and ask whether the quality of information returned varies by search query topic.

We conclude with a discussion of the potential implications of our results on targeted health policy and education efforts.

2 DATA AND RESULTS

To generate the data set of HIV/AIDS queries, we first obtained all Bing search queries containing at least one of the terms "HIV" or "AIDS" that originated in any of the 54 African nations between January 2016 and June 2017. Each query record in the data consisted of the raw search query, country of origin, and date, along with age and gender of the user when available. We removed all queries with two or fewer words (at least one of which must be "HIV" or "AIDS"). We scrubbed the data to ensure that HIPAA identifiers were not included [18]. We removed names, addresses, IP addresses, phone numbers, Bing user ids, among others. The data were completely anonymized for Bing for business purposes prior to the researchers using the data making it virtually impossible for the researchers to identify users. The data sets of malaria and tuberculosis queries were generated in an analogous manner, starting with queries containing, respectively, the term "malaria", or at least one of the terms "tb" and "tuberculosis."

Figure 1 shows two heat maps for each disease. The maps at the top illustrate the 2016 disease prevalence (for HIV) or incidence (for malaria/tuberculosis) rates for each country, obtained from the World Bank Databank [3]. The maps at the bottom illustrate the fraction of total searches made in each country that contain the specific disease terms. There is a high correlation between the fraction of searches about a given disease in a particular country and the rate of the disease. In particular, the Spearman correlation is $\rho = 0.714$ $[0.689, 0.737]$ for HIV/AIDS, $\rho = 0.402$ $[0.360, 0.442]$ for malaria, and $\rho = 0.462$ $[0.422, 0.499]$ for tuberculosis. Each correlation coefficient has a p-value < 0.01. We view this as reassurance that search queries filtered in this way are pertinent to the diseases in question.

Table 1: Sample LDA Topics for HIV/AIDS, Malaria, and Tuberculosis with Representative Words and Sample Queries

Disease	Topic	25 Most Representative Words	Sample Queries from Top 100
HIV/ AIDS	Symptoms (2.28%)	pain, sign, lymph, swollen, nodes, sore, symptom, symptoms, throat, infection, body, back, positive, pains, stomach, fever, neck, headache, glands, patient, feet, symtoms, caused, cough, feel	hiv painfull jaw hiv swollen lymph nodes hiv swollen gland throat
	Natural cure (0.74%)	cure, oil, black, healing, heal, healed, seed, herbs, natural, cures, moringa, kill, cured, testimonials, coconut, traditional, god, garlic, lemon, aloe, prayer, medicine, treat, juice, people	prophet bushiri hiv miracles hiv garlic lemon honey coloidal silver hiv testimonials
	Epidemiology (0.59%)	statistics, report, 2015, global, unaids, 2016, united, epidemic, besigye, kizza, children, 2014, progress, 2010, response, nations, nigeria, prevalence, million, sa, 2013, zambia, uganda, namibia	unaids global aids report mia khalifa hiv hiv 2030
	Drugs (0.85%)	drug, treatment, patients, abuse, therapy, drugs, resistance, antiretroviral, substance, adherence, alcohol, art, failure, spread, leads, patient, relationship, transmission, effect, lead, arv, infected, hiv-1, children	stanford hiv drug resistance hiv drug therapy resistance virological failure hiv
	Breastfeeding (0.66%)	positive, baby, mother, breastfeeding, breast, mothers, child, born, feeding, babies, birth, give, breastfeed, infant, infected, feed, milk, pregnant, safe, exposed, negative, im, months, woman, unborn	hiv exclusive fomular feeding exclusive breast feeding and hiv hiv mom can breast feed baby
	Stigma (0.46%)	stigma, issues, discrimination, related, ethical, legal, prevention, safety, pdf, workplace, relating, precaution, work, dies, surrounding, universal, reduce, address, namibia, gender, social, singer, effects	hiv aids ethical dilema safty issues relating to hiv-aids aids stigma in garissa
Malaria	Symptoms (0.93%)	pregnancy, effects, symptoms, early, pathophysiology, treatment, management, effect, pdf, complications, mouth, disease, sign, sore, bitter, throat, nigeria, symptom, dar, treat, incidence, taste, symtoms, es	malaria lip sores malaria blisters on lips bitterness in mouth and malaria
	Natural cure (1.02%)	cure, natural, home, treat, treatment, remedy, remedies, fever, typhoid, treating, herbal, herbs, medicine, leaf, leaves, good, cures, naturally, lemon, water, treatments, moringa, fruits, tea	pawpaw leave malaria remedy papaya leaf malaria lipton tea for malaria
	Epidemiology (17.39%)	disease, people, year, africa, deaths, download, die, communicable, nigeria, song, number, virus, cases, died, million, caused, mp3, tropical, soty, information, occur, diseases, burden, bacteria	malaria free sri lanka lyrics malaria theme song stoy- malaria mp3
	Drugs (1.31%)	prophylaxis, treatment, quinine, dosage, pregnancy, dose, cdc, doxycycline, prevention, artesunate, children, malarone, chloroquine, severe, table, treat, guidelines, fansidar, treating, drugs, injection, artemether, tablets, country	fansidar malaria dose quinine maximum dose malaria artefan malaria dose
	Breastfeeding (1.05%)	drug, drugs, anti, baby, treat, mother, treatment, breastfeeding, medicine, pregnancy, cancer, fight, good, child, taking, months, affect, medication, babies, treating, breast, prevent, tablet, month,	malaria breast milk malaria through breast feeding interval to treat malaria in toddler
	Diagnosis (1.24%)	parasite, blood, test, parasites, film, smear, thick, stain, slide, thin, microscope, giemsa, procedure, images, staining, field, medicine, count, density, positive, pictures, picture, meaning, microscopy	dar es salaam malaria malaria swamp swollen lip and malaria
TB	Symptoms (1.80%)	symptoms, signs, early, stages, warning, list, sign, infection, pulmonary, symtoms, babies, children, infants, symptons, kids, toddlers, cough, baby, symtomps, symptomes, stage, exposure, sighns, sympoms	night sweat in tuberculosis tuberculosis dry cough tb feet and face swelling
	Natural cure (0.85%)	cure, treat, home, treatment, natural, history, remedies, medicine, mdr, group, patient, taboola, utm_source, disease, long, rememdy, traditional, utm_campaign, treatments, herbs, type, utm_medium, discovered, cures, www	tuberculosis cure discovered does moringa seed cure tb tb reducing natural remedies mdr tb natural remedy
	Epidemiology (0.93%)	africa, south, statistics, deaths, 2010, death, provinces, sa, stats, show, rate, prevalence, province, 2016, incidence, african, graph, 2015, showing, graphs, mortality, caused, west, chart	tb death toll sa tuberculosis graphs tb death provincial statistic
	Drug Side-Effects (1.44%)	drugs, effects, side, treatment, anti, medication, effect, drug, line, medications, liver, list, anti-tb, pregnancy, dosage, anti-tuberculosis, patients, adverse, induced, pills, action, mdr, combination, taking	anti-tuberculosis drugs anti tuberculosis combination 2nd line anti tb
	Diagnosis (1.28%)	diagnosis, culture, sputum, mycobacterium, gene, test, laboratory, genexpert, smear, xpert, testing, lab, microscopy, expert, negative, stain, diagnostic, procedure, pulmonary, collection, positive, methods, samples, gram, standard	tb auramine staining tb culture sensitivity mycobacterium tuberculosis acid-fast stain
	Drug Resistance (1.11%)	drug, resistant, resistance, multidrug, treatment, multi, management, therapy, drugs, multiple, pdf, multi-drug, mycobacterium, patients, antibiotic, mdr, active, rifampicin, latent, anti, extensively, mdr-tb, regimen, choice	multdrug resistant tuberculosis extensively drug resistant vs multi drug resistant tuberculosis

2.1 Disease Topics

We extracted topics from each data set using LDA, a standard generative statistical model in which each *document* (in our case, an individual search query) is associated with a distribution over *topics*, and each topic is defined as a distribution over words [14]. We used the implementation of LDA provided by the Mallet package [27] and the Differential Language Analysis ToolKit [32] as an interface to Mallet for further analysis [31]. We retained all default parameters in Mallet, with the exception of the parameter α, which controls the prior on the per-document topic distribution.[1] In LDA, the number of topics extracted is a free parameter. Based on the size of our datasets, we chose 100 topics for the HIV/AIDS data set and 50 each for malaria and tuberculosis.

Our data reveal a rich set of themes. The topics output by LDA range from those about standard health information, such as disease symptoms, testing, medication, and epidemiological inquiries, to those about hard-to-survey concerns such as stigma and discrimination, natural cures and remedies, and issues regarding healthy lifestyles. Table 4 shows six sample topics extracted from the data by LDA, which were hand-chosen to illustrate both the breadth of themes that emerged from the analysis as well as the coverage of hard-to-survey topics. The second column provides a label for each topic, which was hand-generated by the authors, along with a measure of the frequency with which the topic occurs in the data.[2] Note that some themes, such as breastfeeding, are captured by more than one topic, so the actual overall frequency with which these themes occur in the data is higher than the numbers suggest. The third column presents the 25 most representative words for the given topic. The final column shows a few randomly selected samples from the 100 queries determined by LDA to be most closely related to the topic; we opt to show a random sample of queries since the top few most highly ranked queries often differ by only one letter or word. All typos in the word lists and representative queries are intentional, stemming from common typos in the search data. An additional sample of six topics is included in the Supplementary Information.

2.2 Prevalence of Topics by Region and User Demographics

In this section, we explore whether health information needs, manifested as web search queries, vary by country and user demographics. Due to space limitations, we only present results for queries related to HIV/AIDS. The corresponding results for queries related to malaria and tuberculosis can be found in the Supplementary Information.

For each of the six HIV/AIDS topics listed in Table 4, we estimate the number of times an HIV/AIDS topic is queried relative to the overall number of queries for HIV/AIDS for the country. We call this quantity topic prevalence and denote it by prevalence(topic|country), which is a measure of the frequency with which a topic is mentioned in queries from a given country.

To estimate the prevalence, we need two values—the frequency the topic is searched for in the country and the overall number of searches for any topics in the country. We do not have the exact number of times a topic is mentioned, but we can estimate the frequency with which the topic is used by utilizing the words associated with a topic using the posterior probabilities, derived from LDA, of a topic given a word, $p(\text{topic}|\text{word})$. We then combine these estimated counts with the relative frequencies of a word given a country, frequency(word|country) to get the overall prevalence of the topic in the country:

$$\sum_{\text{word}\in\text{country}} p(\text{topic}|\text{word}) \times \text{frequency}(\text{word}|\text{country})$$

Here, frequency(word|country) is derived from maximum likelihood estimation given all words used in queries from the country or simply the count of the number of times the word appears, divided by the total count of all words in the country. [3]

To explore the association of topic prevalence with HIV prevalence rates across countries, we ran a linear regression using the prevalence values for each of the 100 topics within a given country as the explanatory variables and the 2016 HIV prevalence rate in that country as the dependent variable. Of the six topics listed in Table 4, we found a significant (multi-test corrected[4]) relationship between the *Stigma* topic and the HIV prevalence rate ($r = 0.473$; multi-test corrected $p < 0.01$), as illustrated in Figure 2.

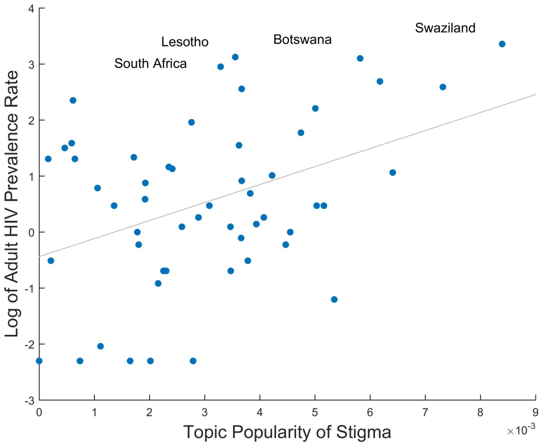

Figure 2: Comparison between topic popularity of *Stigma* in each country with the log of the 2016 adult HIV prevalence rate. Countries with higher topic popularity for *Stigma* tend to have higher HIV prevalence rates.

[1] In particular, we set $\alpha = 2$ to allow for the fact that search queries, by nature, are shorter than the documents for which LDA is typically used.

[2] Specifically, given $\theta_{\text{query}} = p(\text{topic}|\text{query})$, the values in brackets correspond to $\sum_{\text{query}} \theta_{\text{query}}$ over all queries, expressed as a percentage. This corresponds to the popularity of the given topic.

[3] Since we use an estimate of the number of times a topic is searched for, we provide additional results for estimating frequency differently, namely by considering only the top 50 words associated with a topic ranked by $p(\text{topic}|\text{word})$. We provide the additional analysis the Supplementary Information for robustness and show that we obtain qualitatively similar results regardless if we consider all words or the top words to estimate the frequency the topic is searched for in a country.

[4] Relationships are considered significant if they pass the Benjamini-Hochberg false discovery rate (multi-test correction) with $\alpha = .05$ [13].

Table 2: Popular topics by age and sex.

Ages 18–24 *(0.083)*: symptoms, signs, early, women, men, infection, stages, months, symptoms, earliest, children, major, syptoms, systoms, rare

Ages 25–34 *(0.070)*: positive, negative, partner, person, man, sex, woman, tested, im, infected, pregnant, baby, husband, wife, infect

Ages 35–49 *(0.063)*: cure, latest, news, research, treatment, discovery, today, vaccine, update, 2015, 2016, 2017, google, breakthrough, recent

Female *(0.115)*: positive, baby, mother, breastfeeding, breast, mothers, child, born, feeding, babies, birth, given, breastfeed, infant, infected

Male *(0.133)*: cure, news, 2016, latest, vaccine, breaking, 2017, www, development, today, headline, updates, breakthrough, found, feb

Although our correlation analysis alone cannot explain the reasons behind our findings, the observation that the popularity of the *Stigma* topic is correlated with HIV prevalence is consistent with findings from the public health literature. In particular, smaller scale studies have shown that that HIV-related stigma can lead to more risky behavior, lower testing rates, and decreased adherence to antiretroviral therapy, all of which increase transmission rates [15, 20]. Some of the outliers in Figure 2 are also consistent with the literature. For example, we see that the popularity of the *Stigma* topic is nearly twice as high in Botswana as it is in Lesotho, despite the two countries having comparable HIV prevalence rates (22.2 versus 22.7). Botswana has struggled with issues of discrimination and stigma, including various policies and proposals for mandatory HIV testing [29]. Further, despite high HIV prevalence rates, it has been reported that the level of stigma has been declining in South Africa, in part due to targeted programming to combat stigmatization of HIV infected individuals [26]; this, again, is consistent with the results shown in Figure 2.

Other topics also exhibit variance in popularity by country. To explore this, we ran a standard logistic regression for each country, with the topic weights (distribution) of a given search query as the explanatory variables and a binary dependent variable indicating whether or not the query originated in that country. There is notable contrast in the popularity of queries associated with the *Natural Cure* topic across the countries. For instance, the *Natural Cure* topic is popular in Malawi ($\beta = 0.005; p < 0.05$), which has a relatively high HIV prevalence rate of 10.6, and in Botswana ($\beta = 0.007; p < 0.01$), which has an HIV prevalence rate of 22.2. In Mozambique, which has an HIV prevalence rate of 11.5, the popularity of the *Natural Cure* topic is relatively low ($\beta = -0.007; p < 0.01$).

To explore topic popularity by sex, we ran a logistic regression with the topic weights of a given search query as the explanatory variables and the self-reported sex of the user as the dependent variable, limiting ourselves to those queries for which user demographic information was available. Similarly, to explore topic popularity by age group, we ran an ordinary least squares regression, again with the topic weights of a given search query as the explanatory variables, but now with users' self-reported age group as the dependent variable. To understand which topics are more popular for a given demographic group, we ordered the topics by their respective correlation coefficients. The fifteen words with the highest weight from the top ranked topic for each group are shown in Table 2.

Our analysis reveals that topics related to news on HIV/AIDS cures are more popular among men, as well as the 35–49 age group. Topics related to breastfeeding, pregnancy, and family care are more popular among women. For the 18–24 and 25–34 age groups, topics related to symptoms are more popular. Among the former group, topics related to the socioeconomic implications of HIV/AIDS, such as gender inequality, are more popular, while topics related to concerns about transmission to partner and child are more popular among the 25–34 age group. [5]

Finally, we looked at the topic popularity of the six topics of interest from Table 4. Table 3 lists the correlation coefficients, where $**$ indicates a p-value of less than 0.01 and $*$ indicates a p-value of less than 0.05.

Table 3: Topic Popularity by User Demographics

	Female	Ages 18–24	Ages 25–34	Ages 35–49
Symptoms	-0.052**	0.000	-0.019*	-0.018*
Natural Cure	-0.010	-0.050**	-0.018*	0.043**
Epidemiology	-0.052**	-0.080**	-0.019*	0.019*
Drugs	-0.016	-0.020**	-0.041**	0.030**
Breastfeeding	0.115**	-0.031**	0.061**	-0.008
Stigma	0.025**	0.032**	-0.047**	0.004

We see again that *Breastfeeding* has a higher correlation coefficient for women than for men and for the 25–34 age group compared to the other age groups. Notably, women and users aged 18–24 are more interested in *Stigma* compared to their demographic counterparts. *Natural Cure* has the highest popularity among the oldest age group (35–49), and the lowest among the youngest age group (18–24). Despite expressing higher interest in *Natural Cure*, the 35–49 age group also has more interest in *Drugs* compared to the other age groups.

2.3 Differences in User Behavior and Quality of Results

We next examine whether user behavior and the quality of search results returned vary across different topics. To answer these questions, we used an expanded version of our HIV/AIDS data set consisting of only those queries that were made during June 2017. In

[5] Some themes appear in more than one topic. When we identify a novel pattern relating a particular topic to demographics or location in the data (for example, breastfeeding is more popular among women), we confirm the same pattern exists for the most similar topics. To measure similarity between topics, we calculate the pairwise hamming distance between topics using the top 20 most representative words. If the observed pattern does not hold across similar topics, we omit the pattern from our positive findings.

addition to raw queries, country, and search date, this data set contains a list of the first 10 organic webpages returned to the user for each query. It also contains information about which webpages the user clicked on, the amount of time spent on each webpage, and the total time spent on the *results page*, the page containing the ten initial links presented to a user after entering a search query.

To compare user behavior across topics, we focused on several standard metrics from the information retrieval literature [30]. *Dwell time* measures the total amount of time a user spends looking at the results page and any links that are followed. *Click count* is the total number of links on which a user clicks.[6] Note that these metrics can be used to measure various properties related to user engagement and satisfaction. For instance, dwell time can be used to measure both interest in a webpage and ease-of-use, depending on the context and intent of the search query. In our study, we use these metrics to measure user activity. Specifically, we are interested in measuring whether there is variance in user activity by topic as measured by these metrics.

Figure 3 shows how these metrics vary by topic. Both dwell time and click count are significantly lower for the *Natural Cure* topic compared with the other topics of interest. That is, on average, users issuing queries related to the *Natural Cure* topic spend less time exploring the results page and click on fewer links. There are many reasons why this may be the case. It could simply be selection bias—perhaps different types of users search for queries related to the *Natural Cure* topic compared with *Epidemiology*, *Drugs*, or other topics. It could be that users seeking information related to the *Natural Cure* topic find the information they are seeking faster. Another possibility is that the quality of information returned could vary by query.

To examine variance in the quality of content returned, for each of the topics in Table 4, we extracted the first link returned to the user at the top of the web results page for each of the 30 queries most strongly associated with the topic. We consider only distinct user/query pairs, which means we ignored duplicate queries from the same user. Each resulting link was independently evaluated for quality (described in terms of relevance, accuracy, and objectiveness, as is standard in information retrieval [22]) and was ranked by three research assistants on a scale of 1 to 5, with higher values indicating better quality. Details of the evaluation are included in the Supplementary Information. All of the research assistants who provided rankings have graduate-level training in medicine or public health, and each website was evaluated by at least one research assistant specializing in the disease of interest. We took the average of these three ratings as the rating for a webpage.

Figure 4 shows the average rating across all links and all raters for each topic. On average, the quality of links returned for queries related to the *Natural Cure* topic is low, with an average quality rating of 1.45. In contrast, links returned for queries related to the *Stigma*, *Breastfeeding*, and *Drugs* topics have much higher average quality ratings (4.22, 4.36, and 3.99, respectively). A t-test comparison between *Natural cure* with each of these topics yields $p < 0.01$. Corresponding results for malaria and tuberculosis can be found in the Supplementary Information. Consistent with prior research on H1N1 outbreaks [23], the quality of content that is returned to users varies by topic, especially when we compare natural cures vs. drugs which are both treatment options for HIV/AIDS.

3 DISCUSSION

We have shown that search data can provide valuable insights into the health information needs, concerns, and misconceptions of individuals across Africa, which can complement existing top-down or survey-based approaches and allow us to move one step closer to the goal of narrowing the gap in available health data between developing and developed nations. We conclude with a discussion of the limitations of our techniques as well as implications and next steps for future work.

3.1 Limitations

There are several limitations to the use of search data. First, the Bing users we study—and Internet users in general—are not a representative sample of the entire population of Africa. It is challenging to extrapolate observations obtained through the analysis of search data to the wider population of countries, and the health concerns of entire communities who are not on the web could be overlooked. This concern will diminish as Internet penetration continues to grow, but the studied population will never be fully representative.

Second, the results of this study depend on proprietary data from Bing which can limit the ability for health organizations to extend the research. However, the data are readily available to researchers at companies with search engines.

Another limitation is the use of imprecise language in search queries, as well as queries in languages other than English. An initial exploration of the Bing query logs showed that many users search for HIV/AIDS, malaria, and tuberculosis by their English names, but it is likely that the filtering method we used still led us to exclude many relevant searches. Furthermore, the excluded searches are more likely to come from regions in which the use of English names is less common, further biasing the data collection. These concerns could be amplified for other illnesses, such as respiratory infections, which have multiple common names in different languages and for which users commonly search for symptoms instead of the disease name itself. A multi-language approach would be necessary to more fully extract all of the information on individual's health information needs that is captured in search data.

3.2 Practical Implications

Our methods have great potential to inform targeted education efforts in data-sparse regions. Gender and age impact an individual's chance of contracting HIV/AIDS, malaria, or tuberculosis [21], and health information needs are often specific to demographic groups and geographic locations. For these reasons, stakeholders have emphasized the need for gender-responsive and age-responsive programming in resource-constrained regions [2, 4, 7, 8, 16, 19, 21]. Efforts to understand health information needs in developing nations by demographic group have mostly used surveys and interviews, which are limited in their scale [7, 11, 25, 34]. In contrast, by

[6] We also examined *successful click count*, which measures the number of pages a user clicks on with dwell time at least 30 seconds, and *maximum dwell time*, which measures the maximum amount of time spent on a webpage. Results for these are similar to the click count and dwell time results, respectively, and are included in the Supplementary Information.

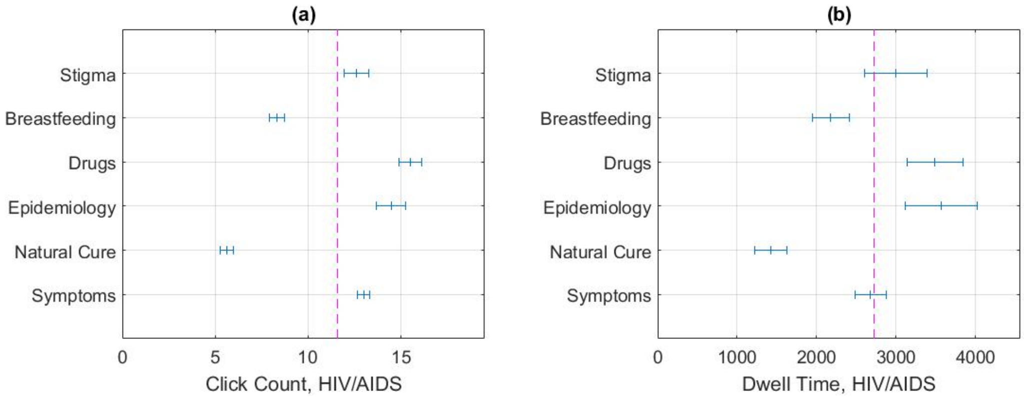

Figure 3: Average (a) click count and (b) total dwell time for queries associated with HIV/AIDS topics of interest. The vertical lines represent the mean values across topics.

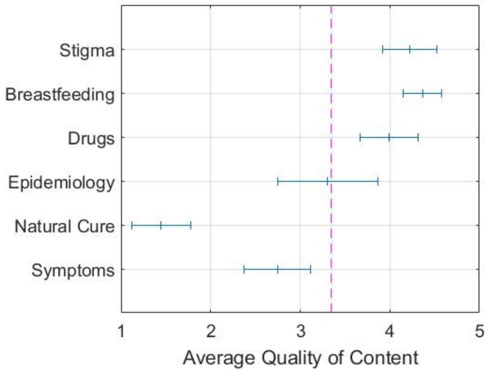

Figure 4: Average quality of content of webpages returned to users for the 30 queries most strongly associated with each HIV/AIDS topic of interest. The webpages were rated by research assistants and show variance in quality between webpages shown for *Natural Cure* compared to the other topics of interest.

analyzing search data we can study health needs at a much larger scale.

Our analysis could also be used to investigate the quality and volume of information available to individuals with different health information needs. To get a sense for how much high-quality public health information on Natural Cures for HIV/AIDS is available, we used the queries we found to be the most highly associated with the *Natural Cure* and *Drugs* topics on Bing to search authoritative websites on HIV/AIDs. The authoritative websites include the World Health Organization (WHO), the Joint United Nations Programme on HIV/AIDS (UNAIDS), the Center for Disease Control and Prevention (CDC), and the National Institutes of Health (NIH). We posed each of the top 30 queries for *Natural Cure* in turn to each of the aforementioned authoritative websites and noted the number of webpages that were returned on each website for each query. We performed the same actions for the *Drugs* topic. We then reported the average number of webpages available for the 30 queries corresponding to the two topics by website.

We found that on the CDC site, there were an average of $56,705.85$ webpages corresponding to *Natural Cure* (over the 30 queries corresponding to the topic) compared to 258,948.6 for the top 30 queries for *Drugs* ($p = 0.01$). Similarly, for the NIH site, there were $91,840.8$ and $456,982.3$ ($p = 0.00$) ; for the WHO, there were $46,600.1$ and $305,528.2$ ($p = 0.00$) ; and for UNAIDS, there were $8,926.5$ and $65,954.0$ ($p = 0.02$) webpages for *Natural Cure* and *Drugs*, respectively. These results indicate there are consistently fewer high-quality documents for natural cures than for pharmaceutical drugs on authoritative websites.

Search engines themselves could potentially be an effective platform for implementing targeted interventions to improve access to health information. For instance, gender- or age-specific targeted advertisements for health campaigns could be triggered by queries associated with specific health topics. Insights garnered from the analysis of search data could help health organizations prepare material and develop interventions aimed at regions where specific misconceptions are especially common. These interventions could take the form of highlighted high-authority links to discourage misinformation, advertisements for support groups triggered by searches related to stigma, or advertisements for testing clinics for testing-related searches.

Search data could also be used to monitor other aspects of public health, for example by providing marketing surveillance for new medications or measuring the impact of public health campaigns. In principle, search engines and health organizations could also work together on case finding, a strategy that directs resources at individuals or groups suspected to be at risk for a particular disease, which is a key strategy in communicable disease outbreak management. Of course, this pursuit would need to be handled with great care, with consideration for the risks and ethics involved.

Finally, as it is detailed and available in real time, search data could be especially valuable for monitoring the impacts of emerging health concerns in developing nations. For instance, noncommunicable diseases such as cancer, cardiovascular disease, and diabetes are of growing concern due to the expansion of the middle class in developing countries and a lack of resources and programs aimed at minimizing their impact [5]. Since the portion of the population affected by these diseases is likely to have Internet access, search data could play an instrumental role in understanding attitudes about these diseases, implementing interventions aimed at improving access to health information, and highlighting overlooked aspects of the impacts of these diseases.

4 ACKNOWLEDGEMENTS

We would like to thank Siam Amis, Yeonsoo Baik, Simret Habtezqi, Maria Ma, Hema Rao, Chelsea Shover, and Afoma Umeano for assistance with rating the quality of content of webpages. We would also like to thank Abraham Flaxman, Thorsten Joachims, Jon Kleinberg, Masha Krupenkin, Michael Macy, John Redd, David Rothschild, Siddhartha Sen, Siddharth Suri, Hanna Wallach, Duncan Watts, Kim Weeden, James Wright, as well as members of the Cornell Center for the Study for Inequality, Department of Computer Science, and Social Dynamics Lab, and Microsoft Research for their valuable input. RA was partially supported by a Google Scholarship and a Facebook scholarship.

REFERENCES

[1] Global burden of disease compare. https://vizhub.healthdata.org/gbd-compare/. Accessed: 2017-10-17.

[2] Strategic investments for adolescents in HIV, tuberculosis and malaria programs. https://www.theglobalfund.org/media/1292/core_adolescents_infonote_en.pdf. Accessed: 2018-01-10.

[3] The world bank databank: World development index database archives. http://databank.worldbank.org/data/reports.aspx?source=WDI-Archives. Accessed: 2017-10-17.

[4] Gender, health, and malaria. http://who.int/gender/documents/gender_health_malaria.pdf, June 2007. Accessed: 2018-01-10.

[5] The health of the people: What works – the African regional health report. Technical report, 2014.

[6] A world that counts: Mobilising the data revolution for sustainable development. Technical report, November 2014. Report prepared at the request of the United Nations Secretary-General by the Independent Expert Advisory Group on Data Revolution for Sustainable Development.

[7] UNDP discussion paper: Gender and malaria. http://www.undp.org/content/dam/undp/library/HIV-AIDS/Gender%20HIV%20and%20Health/Discussion%20Paper%20Gender_Malaria.pdf, December 2015. Accessed: 2018-01-10.

[8] UNDP discussion paper: Gender and TB. http://www.undp.org/content/dam/undp/library/HIV-AIDS/Gender%20HIV%20and%20Health/Gender%20and%20TB%20UNDP%20Discussion%20Paper%20(1).pdf, December 2015. Accessed: 2018-01-10.

[9] International Telecommunication Union statistics. https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx, 2017. Accessed: 2018-01-08.

[10] African Health Observatory. http://www.aho.afro.who.int/, 2018. Accessed: 2018-01-08.

[11] Julie Abimanyi-Ochom, Hasheem Mannan, Nora Ellen Groce, and Joanne McVeigh. Hiv/aids knowledge, attitudes and behaviour of persons with and without disabilities from the uganda demographic and health survey 2011: Differential access to hiv/aids information and services. *PLOS ONE*, 12(4):1–20, 04 2017.

[12] Winston J. Allen, Albert Ouedraogo, and Leela McCullough. Health information needs in West Africa: Results of a survey on the role of the West Africa Health Organization (waho). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.177.9604&rep=rep1&type=pdf. Accessed: 2018-01-08.

[13] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[14] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March 2003.

[15] Brian Chan and Alexander Tsai. Trends in hiv-related stigma in the general population during the era of antiretroviral treatment expansion: an analysis of 31 sub-Saharan African countries. In *Open Forum Infectious Diseases*, volume 2. Oxford University Press, 2015.

[16] Maria De Bruyn. Gender, adolescents and the hiv/aids epidemic: the need for comprehensive sexual and reproductive health responses. *Ipas. Available at http://www. un. org/womenwatch/daw/csw/hivaids/De% 20bruyn. htm.(Accessed 28 January 2010)*, 2000.

[17] Munmun De Choudhury, Meredith Ringel Morris, and Ryen White. Seeking and sharing health information online: Comparing search engines and social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[18] Centers for Disease Control, Prevention, et al. Hipaa privacy rule and public health. guidance from cdc and the us department of health and human services. *MMWR: Morbidity and mortality weekly report*, 52(Suppl. 1):1–17, 2003.

[19] Raoul Fransen-dos Santos. Young people, sexual and reproductive health and hiv. *Bulletin of the World Health Organization*, 87(11):877–879, 2009.

[20] Becky L Genberg, Zdenek Hlavka, Kelika A Konda, Suzanne Maman, Suwat Chariyalertsak, Alfred Chingono, Jessie Mbwambo, Precious Modiba, Heidi Van Rooyen, and David D Celentano. A comparison of hiv/aids-related stigma in four countries: Negative attitudes and perceived acts of discrimination towards people living with hiv/aids. *Social science & medicine*, 68(12):2279–2287, 2009.

[21] Adrienne Germain. Integrating gender into hiv/aids programmes in the health sector: tool to improve responsiveness to women's needs. *Bulletin of the World Health Organization*, 87(11):883–883, 2009.

[22] Layla Hasan and Emad Abuelrub. Assessing the quality of web sites. *Applied Computing and Informatics*, 9(1):11 – 29, 2011.

[23] Shawndra Hill, Jun Mao, Lyle Ungar, Sean Hennessy, Charles E Leonard, and John Holmes. Natural supplements for h1n1 influenza: retrospective observational infodemiology study of information and search activity on the internet. *Journal of medical Internet research*, 13(2), 2011.

[24] M. L. Kern, G. Park, J. C. Eichstaedt, H. A. Schwartz, M. Sap, L. K. Smith, and L. H. Ungar. Gaining insights from social media language: Methodologies and challenges. 2016.

[25] Xiaoming Li, Chongde Lin, Zuxin Gao, Bonita Stanton, Xiaoyi Fang, Qin Yin, and Ying Wu. Hiv/aids knowledge and the implications for health promotion programs among chinese college students: geographic, gender and age differences. *Health Promotion International*, 19(3):345–356, 2004.

[26] Ngozi C Mbonu, Bart van den Borne, and Nanne K De Vries. Stigma of people with hiv/aids in sub-Saharan Africa: a literature review. *Journal of tropical medicine*, 2009, 2009.

[27] Andrew Kachites McCallum. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[28] Michael J. Paul and Mark Dredze. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 9(5):1–183, 2017.

[29] Rofiah O Sarumi and Ann E Strode. New law on hiv testing in botswana: The implications for healthcare professionals. *Southern African Journal of HIV Medicine*, 16(1):1–4, 2015.

[30] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press, 2008.

[31] H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*, 8(9):1–16, 09 2013.

[32] H. Andrew Schwartz, Salvatore Giorgi, Maarten Sap, Patrick Crutchley, Lyle Ungar, and Johannes Eichstaedt. DLATK: Differential language analysis toolkit. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing System Demonstrations*, 2017.

[33] H Andrew Schwartz and Lyle H Ungar. Data-driven content analysis of social media: A systematic overview of automated methods. *The ANNALS of the American Academy of Political and Social Science*, 659:78–94, 2015.

[34] Jianming Wang, Yang Fei, Hongbing Shen, and Biao Xu. Gender difference in knowledge of tuberculosis and associated health-care seeking behaviors: a cross-sectional study in a rural area of china. *BMC Public Health*, 8(354), 2008.

A ADDITIONAL DETAILS ON THE DATA SETS

A.1 Data Cleaning

After generating the initial sets of all queries containing the disease terms ("HIV" or "AIDS," "malaria," or "tuberculosis" or "TB," respectively), removing queries with two or fewer words, and scrubbing the data to remove personal information, including all HIPAA identifiers [18], we manually examined a sample of 1,000 queries from each data set to check whether they were relevant to the disease. The scrubber we used, Tee anonymizer, extracts and replaces PII with more suitable specific placeholders. For example an email address gets replaced by the text fiemailpiifi. There are 13 different types of PII that are replaced including Name, Phone, Address, SSN, CC, and so on. Of these samples, we found that all queries in the malaria data set were related to malaria, but 4.5% of the queries in the HIV/AIDS data set and 16.7% of the queries in the tuberculosis data set were off-topic, containing phrases such as "tb dresses," "TB Joshua," or "4 TB." To improve the quality of the tuberculosis data set, we used these off-topic queries to generate a list of common phrases that we then employed to filter out irrelevant queries from the full tuberculosis data set. After this filtering step, we sampled a fresh set of 1,000 queries and found that only 7.0% were off-topic.

A.2 Coverage of Africa

Our search data covers all 54 nations of Africa. Figure 5 shows a heat map of the total search traffic during January 2016–June 2018 period by country and a heat map of the 2016 population of each country. The Spearman correlation coefficient between search traffic and country population is $\rho = 0.622$ $[0.591 - 0.651]$ with $p < 0.01$. The correlation between search traffic and Internet penetration is $\rho = 0.574[0.540, 0.606]$ with $p < 0.01$.

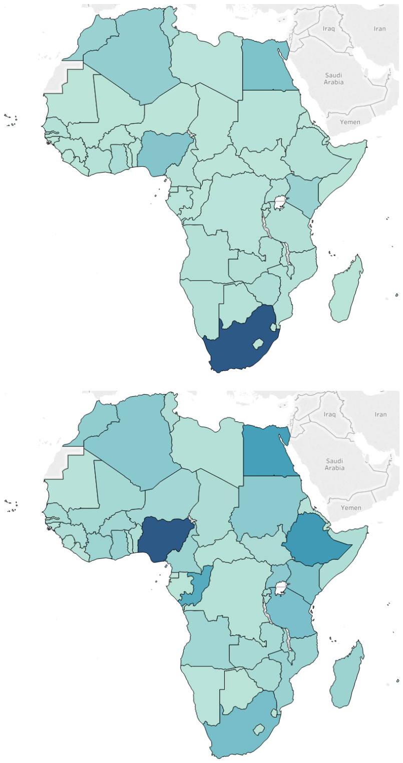

Figure 5: Top: Heat map of the total search traffic in each country during the January 2016–June 2018 period. Bottom: Heat map of the population in each country in 2016.

B ADDITIONAL DETAILS ON THE TOPICS

As described in the main paper, the number of topics output by latent Dirichlet allocation (LDA) is a parameter that can be tuned. Choosing a large number of topics leads to the discovery of highly specific topics that overlap in theme, while choosing a small number leads to very general topics that may be difficult to interpret as each topic covers several themes [33]. Before running our analyses, we ran LDA on each data set with different numbers of topics (10, 20, 50, 100, 200, 500, 1,000, and 2,000). Based on a manual inspection of the interpretability and coherence of topics, we chose 100 topics for the HIV/AIDS data set and 50 each for the malaria and tuberculosis data sets, which are smaller.

The Table 1 in the main body contains examples of representative queries for each topic. We would ideally like to define representative queries as queries with high weight for the topic. However, the existence of rare words or strings (such as obscure URLs) in a query can result in a query having an artificially high weight for a given topic. We thus excluded words that appear fewer than 10 times in the data set.

To give a better sense of the breadth of topics output by LDA, we provide an additional six example topics for each disease in Table 4. These topics were again manually selected by the authors to show the wide range of themes that emerge. As in the main text, the labels in the second column are provided by the authors, the third column displays representative words for each topic, and the fourth column contains a randomly selected sample of the 100 most representative queries. Before running LDA, we scrubbed the data to ensure that HIPAA identifiers were not included. We then manually scrubbed the output of LDA to further remove identifiers, such as celebrity names, which have been redacted here.

C ADDITIONAL ANALYSES OF THE POPULARITY OF TOPICS

C.1 Topic Popularity by Country

In the main paper, we discuss the association between the popularity of the *Stigma* topic for HIV/AIDS in a country and that country's disease prevalence. We ran similar tests on the six topics included in the table in the main body for the malaria and tuberculosis data sets. While in most cases we found no statistically significant association, we did find a significant (multi-test corrected) relationship between the popularity of the *Epidemiology* topic for tuberculosis and the tuberculosis incidence rate ($r = 0.509$ multi-test corrected $p < 0.01$). See Figure 6.[7]

C.2 Topic Popularity by User Demographics

As we did for HIV/AIDS, we looked at the topic popularity of the six topics of interest from the topic table in the main paper. Tables

[7] As in the main text, relationships are considered significant if they pass the Benjamini-Hochberg false discovery rate (multi-test correction) with $\alpha = 0.5$.

Table 4: Additional sample LDA topics for HIV/AIDS, malaria, and tuberculosis with representative words and sample queries.

Disease	Topic	25 most representative words	Sample queries from top 100
HIV/ AIDS	Transmission (0.61%)	sex, oral, penis, infected, man, woman, vagina, person, sucking, risk, contact, pussy, positive, contract, workers, chances, condom, girl, transmitted, sperm, vaginal, female, anal, inside	aids-infected penis hiv cunnilingus sucking penis transmit hiv
	Testing kits (0.69%)	test, home, kit, testing, treat, kits, buy, clicks, rapid, tests, tester, app, pharmacy, finger, phone, online, download, free, dischem, price, accurate, cost, remedies, scanner	hiv home test kit cvs download hiv fingerprint scanner hiv kit dischem
	Gender inequality (0.45%)	spread, gender, contribute, power, relations, infections, ways, inequality, unequal, infection, discuss, relation, namibia, imbalance, contributes, pdf, poverty, zimbabwe, lead, spreading, influence, leads, sexuality, contracting	hiv spread gender equality relations unequal hiv infections
	Healthy lifestyle (0.59%)	food, positive, people, person, diet, healthy, living, eat, good, patients, patient, medication, nutrition, foods, lifestyle, eating, importance, manage, tea, supplements, vitamin, benefits, drink, security	hiv healthy food diet food insecurity and hiv hiv vitamins and supplements
	Disease progression (0.48%)	cd4, count, positive, blood, cells, bebe, winans, low, story, cell, patients, patient, person, white, high, treatment, virus, negative, infection, cd, infected, normal, vision, diagnosis	hiv cd4 count 500 cd4 cd8 ratio hiv t4 count in hiv
	Celebrity gossip (0.95%)	***, ***, positive, hiv-positive, status, ***, ***, ***, TRUE, ***, ***, ***, ***, ***, ***, ***, ***, ***, ***, ***, ***, drivers, revealed, couples, info, vasculopathy, sighns, morphology, cinnamon	*** *** hiv-positive *** *** hiv status is *** hiv-positive
Malaria	Prevalence (1.07%)	map, areas, africa, countries, risk, endemic, kenya, high, area, botswana, cdc, mozambique, affected, namibia, list, african, country, found, zimbabwe, zones, tanzania, top, world, free	top malaria countries malaria endemic areas map how europe eliminated malaria
	Mortality (0.80%)	mortality, children, rate, nigeria, morbidity, death, impact, due, child, questions, malawi, africa, prevalence, effects, infection, years, maternal, poverty, anaemia, related, infant, disease, population, birth	malaria morbidity and mortality malaria premature child mortality due to malaria
	Testing kit (1.32%)	test, rapid, diagnostic, kit, testing, rdt, kits, tests, pf, antigen, sd, positive, results, result, negative, bioline, diagnosis, pan, urine, rdts, reporting, hiv, strip, carestart	buy binaxnow malaria combo kit rapid malaria test kit false negative rapid malaria test
	Parasite (1.54%)	cycle, life, parasite, 10, icd, plasmodium, diagram, code, stages, history, cdc, mosquito, pdf, parasites, lifecycle, host, describe, explain, transmission, drawing, human, circle, annotated, video	life cyle malaria cdc malaria life cycle diagram malaria icd
	Pregnancy (1.11%)	pregnant, drugs, pregnancy, woman, drug, women, anti, treat, treatment, safe, trimester, weeks, nigeria, list, good, medication, early, medicine, treating, effect, month, tablet, taking, anti-malaria	36 weeks with malaria malaria in pregnant woman malaria prophylaxes
	Prevention (0.91%)	prevent, ways, control, prevention, measures, preventing, spread, methods, preventive, method, controlling, prevented, pdf, ddt, reduce, avoid, areas, awareness, community, campaign, eradicate, government, effective, state	ways of controling malaria malaria preventative measures anti-malaria campaigns
TB	Patient care (1.11%)	person, people, patients, patient, pictures, food, eat, images, infected, spread, lungs, diet, prevent, picture, hiv, healthy, foods, good, avoid, suffering, living, kissing, stomach, dangerous	eating healthy with tuberculosis foods to avoid tb food supplements for tb patients
	Testing (1.23%)	test, positive, skin, pictures, negative, results, blood, gold, quantiferon, sputum, tests, testing, lam, result, urine, diagnosis, diagnostic, pcr, FALSE, rapid, mantoux, bovine, tested, reaction	quantiferon gold tb capilla tb test cdc tb skin test reading
	Prevention (0.95%)	stop, strategy, end, dots, partnership, life, reach, meaning, quality, program, cycle, order, logo, treatment, price, application, key, wave, strategies, nigeria, control, populations, www, kenya	stop tb partnership tuberculosis life cycle tb reach wave 6
	HIV co-infection (0.94%)	hiv, coinfection, research, nigeria, co-infection, job, description, 2017, patients, positive, jobs, aids, related, tvoroyri, solution, tanzania, project, children, eradication, field, officer, society, person, model	tb-hiv co-infection tuberculosis description tb coordinator job description
	National programs (1.23%)	national, control, hiv, plan, program, leprosy, health, strategic, programme, infection, policy, guidelines, prevention, kenya, sti, ministry, aids, manual, training, nigeria, uganda, ghana, malawi, zambia	tb crisis plan tanzania strategic plan tb tb helpline ghana
	Global programs (0.92%)	hiv, aids, malaria, global, fund, fight, health, project, program, services, nigeria, funding, diseases, impact, integration, challenges, lesotho, grant, call, cholera, integrated, indicators, evaluation, funds	global fund fight aids tuberculosis malaria tb malaria cholera meningitis

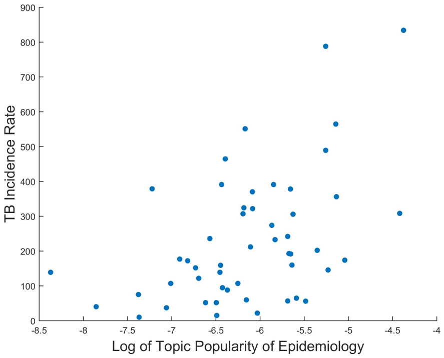

Figure 6: Popularity of the *Epidemiology* topic vs. tuberculosis incidence.

5 and 6 list the correlation coefficients, where ∗∗ indicates a p-value of less than 0.01 and ∗ indicates a p-value of less than 0.05. As we saw in the HIV/AIDS data set, female users and users in the 25–34 age group expressed relatively more interest in topics related to pregnancy, breastfeeding, and family care compared to their male counterparts in the malaria data set. Additionally, for both the malaria and tuberculosis data sets, users in the 25–34 age group were relatively more interested in symptom-related topics. While it might not explain the whole story, this is consistent with the literature that nearly half of all new HIV infections occur among the 15–24 age group [19]. The 25–34 age group corresponds to the time-frame where these infections may progress significantly, and even develop to AIDS.

Table 5: Relative topic popularity by user demographics for malaria.

	Female	Ages 18–24	Ages 25–34	Ages 35–49
Drug	0.006	-0.068**	-0.008	0.031
Natural Cure	0.027	-0.051**	-0.014	0.051**
Breastfeeding	0.073**	-0.089**	0.060**	0.031
Epidemiology	-0.084**	-0.002	-0.042*	-0.005
Diagnosis	-0.065**	0.032	0.071**	-0.058**
Symptoms	0.055**	0.001	0.062**	-0.019

D ADDITIONAL DETAILS ON USER BEHAVIOR AND QUALITY OF RESULTS

D.1 User Behavior

We examined whether user behavior varies across different topics for the HIV/AIDS, malaria, and tuberculosis data sets. We used four popular metrics in the information retrieval literature: dwell time, maximum dwell time, click count, and successful click count. Dwell time and click count are discussed in the main text. *Maximum dwell time* measures the maximum amount of time that a user spends on any link that is followed. *Successful click count* is the total number of links the user clicks that have a dwell time of at least 30 seconds.

Table 6: Relative topic popularity by user demographics for TB data set.

	Female	Ages 18–24	Ages 25–34	Ages 35–49
Epidemiology	0.006	0.037*	-0.078**	0.007
Drug Resist.	0.003	0.007	-0.027	0.002
Diagnosis	-0.043*	0.004*	-0.011*	0.001
Symptoms	0.090**	0.014	0.074**	-0.042
Drug Side-effect	0.027	0.028	-0.004	-0.016
Natural Cure	0.018	-0.020	-0.016	0.022

We use the same methodology described in the main text and the regression coefficients to report the average values. Results are shown in Figure 7. Note that for the HIV/AIDS and malaria data sets, users issuing queries associated with *Natural cure* exhibited relatively low activity (by all four metrics) compared to many of the other topics of interest; this is not true for the tuberculosis data set.

D.2 Quality of Content

To measure the quality of the links presented by topic, we examined the set of links returned in the first position for the thirty most representative queries for each of the six topics from the malaria and tuberculosis data sets that appear in the table in the main text, in the same way described in the main text for the HIV/AIDS data set. Each link was evaluated by three research assistants, each of whom has graduate-level training in medicine or public health, and at least one of whom specializes in the corresponding disease. In all three cases, the research assistants were asked to assess the relevance, accuracy, and objectiveness of the links returned. In particular, the research assistants were presented with the following questions.

- *Relevance:* How comprehensive and complete is the information provided on the website and does it appear to provide the right level of detail? Is the URL related to the disease?
- *Accuracy:* Are sources of information properly identified, and are there any glaring omissions or misinformation?
- *Objectiveness:* Does the information presented appear in an objective manner without political, cultural, religious, or institutional bias?

The research assistants were asked to assign a single rating for each link on a scale from 1 to 5, with 1 equal to bad quality and 5 equal to high quality. Values were defined as:

1. Bad quality: several serious issues concerning all three of relevance, accuracy, and objectiveness
2. Subpar quality: several serious issues covering at least two of relevance, accuracy, or objectiveness
3. Mediocre quality: several issues concerning relevance, accuracy, or objectiveness
4. Good quality: mostly relevant, accurate, and objective, with a few small issues
5. High quality: very relevant, accurate, and objective

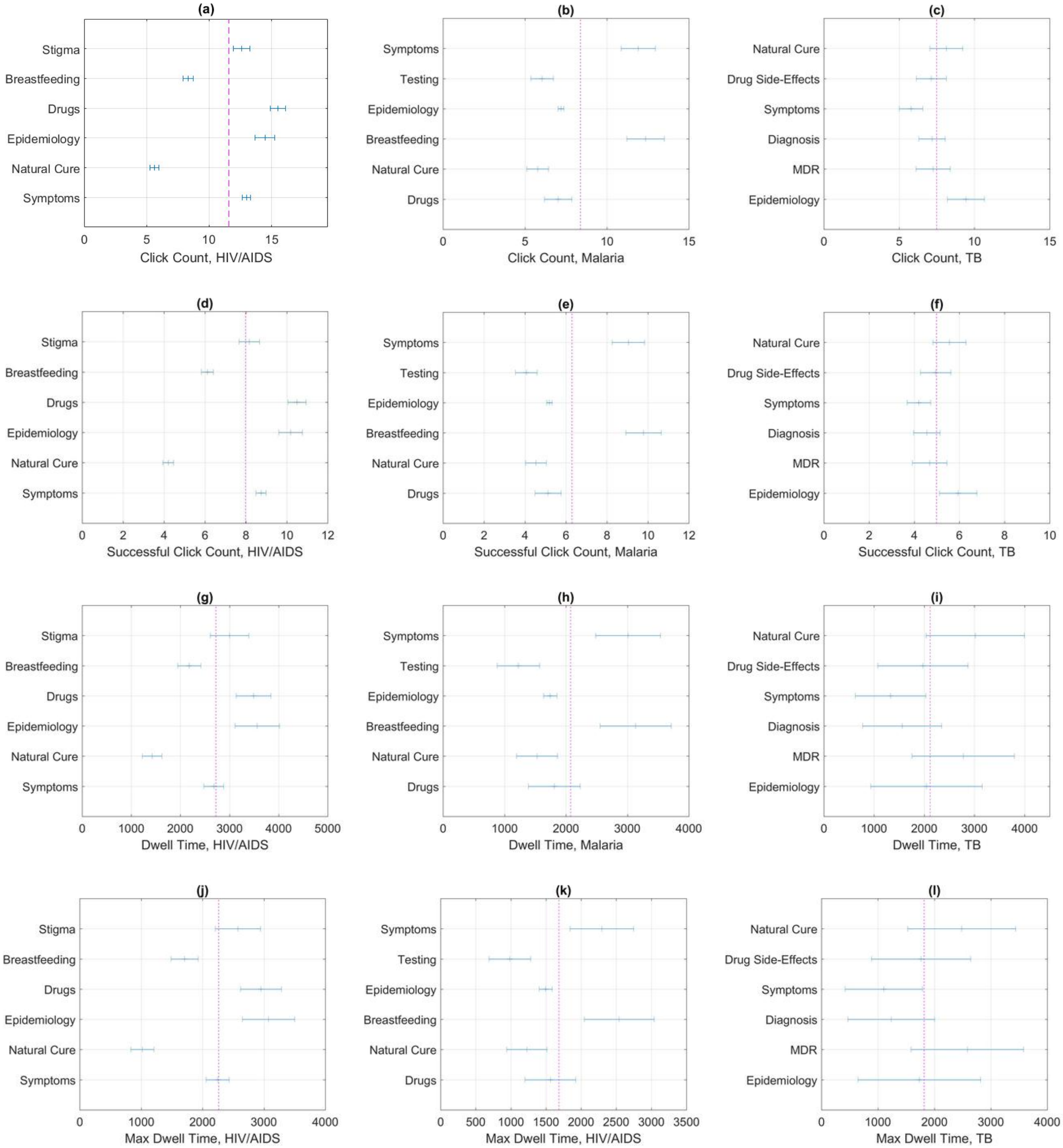

Figure 7: Rows, from top to bottom: average click count, successful click count, dwell time, and maximum dwell time for queries associated with different topics. Columns, from left to right: HIV/AIDS, malaria, and tuberculosis data sets. The vertical lines represent the mean values across topic.

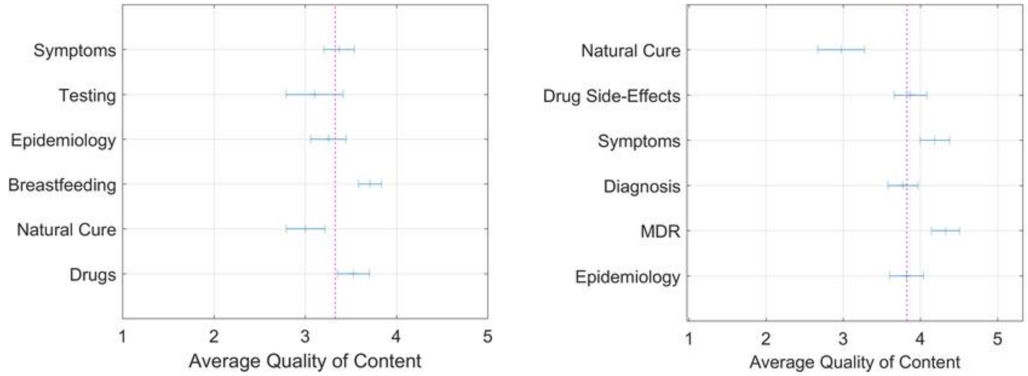

Figure 8: Average quality of content of webpages returned to users for the 30 queries most strongly associated with the six malaria (left) and tuberculosis (right) topics.

Note that the research assistants were not asked to take into account the website design, interface, usability, or other metrics unrelated to content.

Consistent with the observations for the HIV/AIDS data set, for the tuberculosis data set, the quality of content returned to users was, on average, rated lower for queries related to the *Natural cure* topic than for other topics of interest. In contrast, for the malaria data set, the quality of links returned was indistinguishable among queries related to the *Natural cure*, *Epidemiology*, *Testing*, and *Symptoms* topics.