Good starting point for ACA coverage gap causes analysis

Using Search Queries to Understand Health Information Needs in Africa

Rediet Abebe
Cornell University

Shawndra Hill
Microsoft Research

Jennifer Wortman Vaughan
Microsoft Research

Peter M. Small
Rockefeller Foundation

H. Andrew Schwartz
Stony Brook University

arXiv:1806.05740v2 [cs.CY] 17 Apr 2019

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 health information needs of people. Without understanding people's everyday concerns, health organizations and policymakers are less able to effectively target education and programming efforts. In this paper, we propose a bottom-up approach that uses search data to uncover and gain insight into health information needs of individuals 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 topics, revealing a widespread 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 gender) and country. Using finer-grained data, we also uncover discrepancies in the quality of content returned by search engines to users by topic and highlight differences in user behavior and satisfaction. 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 off-line.

Introduction

New technologies and data-sources are constantly being leveraged to upgrade and supplement the design, monitoring, and evaluation of health policy in the developed world. 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. A 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" (Sambo 2014).

This lack of data can be a roadblock to identifying major public health concerns and implementing effective interventions. While targeted education addressing 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 such needs vary by region and demographic groups since the impact of diseases—their prevalence, progression, and transmission rates—as well as people's disease knowledge and attitudes vary regionally and demographically. Limited understanding on the health information needs of individuals hinders the efficacy of gender- and age-specific programming (Germain 2009; De Bruyn 2000; Global Fund 2016; UNDP 2015b; UNDP 2015a).

In this paper, we take a step towards narrowing this gap, focusing on the problem of identifying and measuring people's everyday health information needs, concerns, and misconceptions. We use Bing search queries originating in all 54 African nations to explore which themes 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 (IHME 2016).

Search data provide a wealth of information on people's real-time activities, experiences, concerns, and misconceptions relatively cheaply (Paul and Dredze 2017; Kern et al. 2016), 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 used a top-down approach, actively collecting data with a particular goal in mind (IHME 2016; AHO 2010). 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, primarily make use of manual interviews (Allen, Ouedraogo, and McCullough 2010). These approaches can obtain a comprehensive picture of individuals' needs, but are

Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

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 because search logs have already been shown to contain large quantities of information related to serious conditions in other contexts (De Choudhury, Morris, and White 2014; Paul and Dredze 2017). 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 (ITU 2017)—to our knowledge, no prior work has looked specifically at search data to understand health information needs in all African nations.

The Present Work. We analyze Bing search data related to HIV/AIDS, malaria, and tuberculosis from all 54 African nations. We uncover themes in which individuals are interested using latent Dirichlet allocation (LDA), a standard generative model for automatically extracting topics from text (Blei, Ng, and Jordan 2003). The topics that emerge cover basics such as symptoms, testing, and treatment, as well as hard-to-survey topics such as stigma and discrimination, beliefs in natural cures and remedies, and concerns about the impact of gender inequality in HIV transmission. We explore the ways in which the popularity of these topics vary by age, gender, and location. We expose patterns including that searches related to pregnancy and breastfeeding and relatively more popular among women while searches related to cure news are relatively more popular among men.

Delving into the content returned to users, we compare the organic search results returned for different topics and quantify the discrepancies in the quality of information returned to users. These results highlight unmet health information needs, concentrated misinformation related to specific health topics, and differences in user satisfaction by topic. We discuss the limitations of our approach, including the difficulty of extrapolating our observations to the wider population of Africa, and the danger of overlooking the health concerns of communities who are not on the web. Finally, we highlight potential implications of these analyses on health policy and education efforts both on- and off-line.

Related Work

Health Information Needs. Health information seeking behavior plays a key role in combating the burden of diseases. Online behavior can provide an important lens, especially for stigmatizing conditions (such as STIs), where off-line behavior may be harder to collect (Fox and Duggan 2013). Health information is central to disease control; for instance, HIV management requires extensive informational support to maintain the well-being of those affected and their caretakers. However, there is inadequate understanding of individuals' health information seeking behavior, disease knowledge, and perceptions for the three diseases of interest in this paper (Chan and Tsai 2018; Hogan and Palmer 2005).

This lack of knowledge is especially prominent for individuals living in developing nations. There are relatively few studies, and existing studies are often limited to specific subpopulations. For instance, Abimanyi-Ochom et al. (2017) explore the HIV/AIDS knowledge, attitudes, and practices of Ugandan individuals with disabilities. Similarly, Gombachika et al. (2013) set out to understand how couples living with HIV in Malawi obtain sources of information and reproductive decisions. Studies covering all 54 nations in the continent have often focused on aggregate health outcomes, such as quantifying the burden of diseases by country, rather than health information consumption.

Search Data for Health. Search and other large-scale Web data have emerged as key for understanding health patterns and health information consumption (De Choudhury and De 2014; Fox and Duggan 2013; Sillence et al. 2007; Liu et al. 2013; Eysenbach and Köhler 2002; Spink et al. 2004; De Choudhury, Morris, and White 2014). Online health information seeking behavior is known to be connected to off-line behavior and can inform health policy (Fiksdal et al. 2014; Ling and Lee 2016; Zheluk et al. 2013; Ocampo, Chunara, and Brownstein 2013; O'Grady 2008). Search data is especially valuable as it is real-time, detailed, relevant, and gives less-filtered insights into individuals' health information needs (Eysenbach and Köhler 2004).

Despite these findings, studies focusing on the use of search engines as a medium for obtaining health information related to the three diseases have remained small scale, often limited to surveys and focused on developed nations (O'Grady 2008; Hogan and Palmer 2005; Shuyler and Knight 2003). There is need to understand individuals' search behavior before attempting to target relevant information to individuals whether on- or off-line (Fiksdal et al. 2014).

A closely related line of work to ours is surveillance and case finding, where there is extensive work related to HIV/AIDS, malaria, and tuberculosis (Zhou and Shen 2010; Ocampo, Chunara, and Brownstein 2013). This work shows promising results connecting on- and off-line behavior and suggests that search data can be valuable as an information source for health. Similar work related to other diseases include using search data for influenza outbreaks (Ginsberg et al. 2009; Polgreen et al. 2008; Santillana et al. 2015; Yang, Santillana, and Kou 2015; Yuan et al. 2013), dengue fever (Althouse, Ng, and Cummings 2011; Chan et al. 2011), norovirus outbreaks (Desai et al. 2012), bacterial infections (CDC 1998), and many other outbreaks (Brownstein, Freifeld, and Madoff 2009; Carneiro and Mylonakis 2009; Hay et al. 2013; Rothman et al. 2008).

Impact of Demography on Health. Health outcomes can vary drastically by demographics, especially in developing nations. Gender and age impact likelihood of being infected and ability to obtain care and treatment. For instance, it is known that 61% of all sub-Saharan individuals living with HIV are women, and women in the 15–24 age group are three times more likely than men in the same age group to acquire HIV (WHO 2009). Men are more likely to develop and die from tuberculosis (UNDP 2015b). Pregnant women are disproportionally impacted by malaria (UNDP 2015a).

This variance extends to individuals' knowledge about the diseases of interest (Fransen-dos Santos 2009). Young indi-

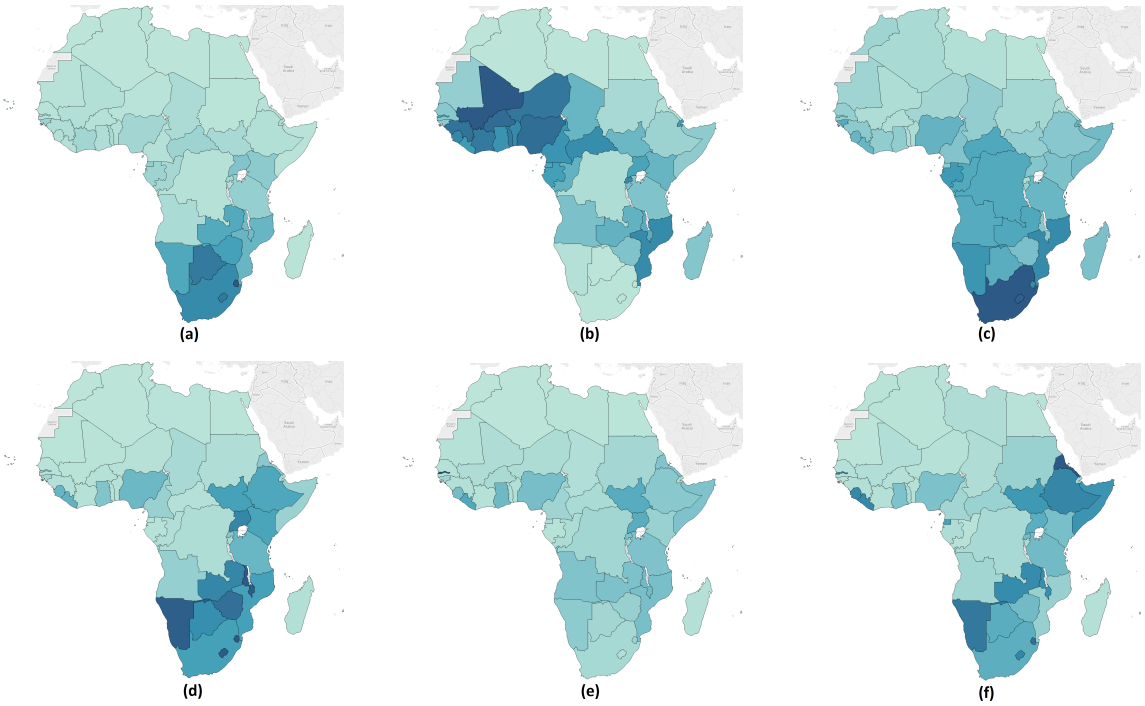

Figure 1: Top: Heat maps showing 2016 rates of (a) HIV/AIDS prevalence (ages 15–49), (b) malaria incidence, and (c) tuberculosis incidence. Bottom: Heat maps showing percentage of total search traffic containing the words (d) "HIV" or "AIDS", (e) "malaria", and (f) "tb" or "tuberculosis."

viduals are evaluated to have incomprehensive knowledge about these diseases, and young women especially so (Stonbraker et al. 2017; WHO 2009; Li et al. 2004; Kumar and Mmari 2004; Wang et al. 2008). At the same time, it is difficult to find large age and gender-disaggregated data as they are not routinely collected or reported (UNDP 2015a; UNDP 2015b). This highlights a research gap and potential for computationally-informed policy contributions for effective prevention, coverage, and treatment of these diseases. In resource-constrained settings where funds must be allocated strategically, it is especially prudent to take the varied needs of these groups into account.

Data and Methodology

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. We consider this time-period since data at the level of granularity described here was available for this time period. Both mobile and desktop searches were retrieved. Each query record in the data consisted of the raw search query, country of origin, and date, along with self-reported age and gender of the user when available. We scrubbed the data to remove HIPAA identifiers: names, addresses, IP addresses, phone numbers, Bing user ids, among others. The data were anonymized for Bing for business purposes prior to the researchers' access. The data sets of malaria and tuberculosis queries were generated in an analogous manner, except with keywords "malaria" as well as "tb" or "tuberculosis."

Figure 1 shows two heat maps for each disease. The top maps illustrate the 2016 disease prevalence (for HIV) or incidence (for malaria/tuberculosis) rates for each country, obtained from the World Bank Databank (WB 2018). The bottom maps 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 disease rate. 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 $p < 0.01$. We view this as reassurance that search queries filtered in this way are pertinent to the diseases in question.

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 a distribution over *topics*, and each topic is a distribution over words (Blei, Ng, and Jordan 2003). We used the implementation of LDA provided by the Mallet package (McCallum 2002) and the Differential Language Analysis ToolKit as an interface to Mallet for further analysis (Schwartz et al. 2017; Schwartz et al. 2013). We retained all default parameters, with the exception of α, the prior on the per-document topic distribution, which we set to 2 since search queries are shorter than the documents for which LDA is typically used.

The number of topics extracted is a free parameter that can

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

Disease	Topic	20 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	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	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	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	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	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	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	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	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	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	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	malaria breast milk can a breast feeding mother take malaria drugs
	Diagnosis (1.24%)	parasite, blood, test, parasites, film, smear, thick, stain, slide, thin, microscope, giemsa, procedure, images, staining, field, medicine, count, density	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	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	tuberculosis cure discovered does moringa seed cure tb tb reducing natural remedies
	Epidemiology (0.93%)	africa, south, statistics, deaths, 2010, death, provinces, sa, stats, show, rate, prevalence, province, 2016, incidence, african, graph, 2015, showing	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	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	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	multdrug resistant tuberculosis extensively drug resistant vs multidrug resistant tuberculosis

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 general, multi-theme, and difficult to interpret topics (Schwartz and Ungar 2015). 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 by a health expert, we chose 100 topics for the HIV/AIDS data set and 50 each for the malaria and tuberculosis data sets, which are smaller than the HIV/AIDS data set. These topics were then labeled by a health expert and manually inspected by the authors and checked for any overlooked HIPAA identifiers.

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 (abnormally high probability of belonging to a single topic). We thus excluded words that appear fewer than 10 times in the data set. For the same reason, we removed all queries with two or fewer words since these often contained similar issues. (Note that at least one of these must be the name of the disease.)

Additional methods are described below alongside the corresponding results. All statistical significance tests were conducted correcting for false discovery rate with $\alpha = .05$ of testing either 50 or 100 topics using the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995).

Analysis and Results

Our analyses reveal a rich set of themes. The topics output by LDA range from those about standard health information, such as *Symptoms*, *Drugs*, and *Epidemiology*, to those about hard-to-survey concerns, such as *Stigma* and *Natural Cure*. Table 4 shows six sample topics for each disease extracted from the data by LDA, which were hand-chosen and labeled by a health expert to illustrate both the breadth of themes that emerged from the analysis as well as the coverage of hard-to-survey topics. For the HIV/AIDS data set, the full list of topics additionally includes themes such as *Transmission*, *Testing Kits*, *Testing Clinics*, *Gender Inequality*, *Healthy Lifestyle*, *Disease Progression*, and *Celebrity Gossip*. Likewise, the malaria and tuberculosis data sets display a rich set of themes ranging from *Patient Care*, and *Pregnancy*, to *National Programs*.

The second column in Table 4 provides a label for each topic along with a measure of the frequency with which the topic occurs in the data. Specifically, given $\theta_{\text{query}} = p(\text{topic}|\text{query})$, the values in parentheses correspond to $\sum_{q \in \text{query}} \theta_q \times p(q)$ over all queries, expressed as a percentage. This corresponds to the popularity of the given topic. Some themes, such as breastfeeding, are captured by more than one topic, so the overall frequency with which these themes occur in the data is higher than the numbers suggest. The third column presents the 20 most representative words, according to posterior probabilities of word given topic. The final column shows randomly selected queries from among the 100 most closely related to the topic. We 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 queries are unaltered.

Topic Prevalence by Region and Demographics

We explore whether health information needs, manifested as 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 malaria and tuberculosis can be found in the appendix.

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$(\text{topic}|\text{country})$, 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$(\text{word}|\text{country})$ 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$(\text{word}|\text{country})$ is derived from maximum likelihood estimation given all words used in queries from the country: the ratio of the number of times a word appears and the total count of all words for the the country.[1]

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 relationship between the *Stigma* topic and the HIV prevalence rate ($r = 0.473$; multi-test corrected $p < 0.01$), as illustrated in Figure 2.

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 (often based on survey data), have shown 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 (Chan and Tsai 2015; Genberg et al. 2009). An outlier in Figure 2, *Stigma* is nearly twice as frequent in Botswana as in Lesotho, despite the

[1] Since we use an estimate of the number of times a topic is searched for, we provide results for estimating frequency differently, namely by considering only the top 50 words associated with a topic ranked by $p(\text{topic}|\text{word})$, in the appendix. We obtain qualitatively similar results when we consider all words or the top words to estimate the frequency the topic is searched for in a country.

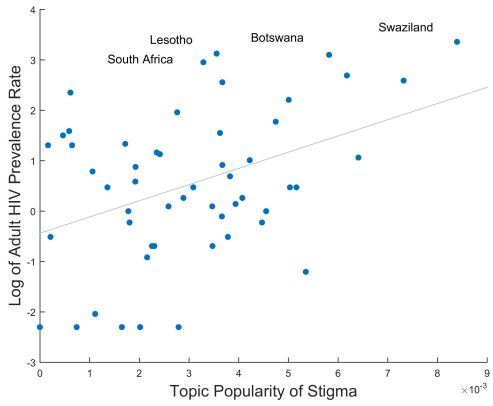

Figure 2: Comparison between topic popularity of *Stigma* in each country with the log of the 2016 adult HIV prevalence rate. Countries with higher topic popularity for *Stigma* tend to have higher HIV prevalence rates.

two countries having comparable HIV prevalence rates (22.2 versus 22.7). This is consistent with literature indicating that Botswana has struggled with discrimination issues including proposals for mandatory HIV testing (Sarumi and Strode 2015). Similarly, Swaziland is also known to have one of the highest rates of HIV prevalence, as well as stigma surrounding the disease (Root 2010). Further, despite high HIV prevalence rates, the level of stigma has been declining in South Africa, in part due to targeted HIV destigmatization efforts (Mbonu, van den Borne, and De Vries 2009).

Other topics also exhibit variance in popularity by country. To explore this, we ran a logistic regression for each country, with the standardized 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 ($\beta = 0.05; p < 0.05$), which has a relatively high HIV prevalence rate of 10.6, and in Botswana ($\beta = 0.07; p < 0.01$), 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 ($\beta = -0.07; p < 0.01$).

To explore topic popularity by gender, we ran a logistic regression with the topic weights of a given search query as the explanatory variables and the self-reported gender 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 linear 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. We then 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.

Table 2: Popular topics by age and gender.

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
Women *(0.115)*: positive, baby, mother, breastfeeding, breast, mothers, child, born, feeding, babies, birth, given, breastfeed, infant, infected
Men *(0.133)*: cure, news, 2016, latest, vaccine, breaking, 2017, www, development, today, headline, updates, breakthrough, found, feb

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.[2]

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

	Women	Ages 18–24	Ages 25–34	Ages 35–49
Sympt.	-0.052**	0.000	-0.019*	-0.018*
Natl Cure	-0.010	-0.050**	-0.018*	0.043**
Epid.	-0.052**	-0.080**	-0.019*	0.019*
Drugs	-0.016	-0.020**	-0.041**	0.030**
Breastf.	0.115**	-0.031**	0.061**	-0.008
Stigma	0.025**	0.032**	-0.047**	0.004

We again confirm *Breastfeeding* has a higher correlation coefficient for women than for men and for the 25–34 age group compared to the other age groups. Less expected,

[2]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 (e.g., "breastfeeding is more popular among women"), we confirmed 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.

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.

User Behavior and Quality of Results

We examine whether user behavior and the quality of search results returned vary across different topics. Differences here would highlight unmet health information needs, concentration of misinformation related to specific topics, and differences in user satisfaction by topic.

We used an expanded version of our HIV/AIDS data set consisting of only those queries that were made during June 2017. In addition to raw queries, country, and search date, this data set contains a list of the first 10 organic web pages returned to the user for each query. It also contains information about which web pages the user clicked on, the amount of time spent on each web page, 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 (Manning, Raghavan, and Schütze 2010). *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.[3] 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 web page 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.

Plots (a) and (b) in Figure 3 show 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 (Hasan and Abuelrub 2011)) and was ranked by three research assistants on a scale of 1 to 5, with higher values indicating better quality. 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 web page.

Plot (c) in Figure 3 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 $p < 0.01$. Results for malaria and tuberculosis are similar. Consistent with prior research on H1N1 outbreaks (Hill et al. 2011), the quality of content that is returned to users varies by topic, especially when we compare *Natural Cure* vs. *Drugs*.

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 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 web pages 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 web pages available for the 30 queries corresponding to the two topics by website.

We found that on the CDC site, there were an average of $56,705.85$ web pages 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 $65,954.0$ ($p = 0.02$) web pages for *Natural Cure* and *Drugs*, respectively. By this measure, there are consistently fewer high-quality documents for natural cures than for pharmaceutical drugs on authoritative websites.

Discussion and Conclusion

We have shown that search data, with well-chosen analyses, can provide valuable insights into the health infor-

[3] 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 web page. Results for these are similar to the click count and dwell time results, respectively.

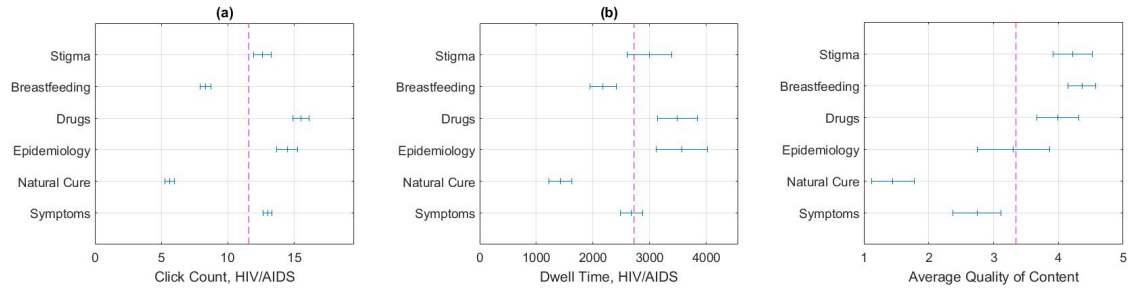

Figure 3: Average (a) click count and (b) total dwell time for queries associated with HIV/AIDS topics of interest. The vertical lines represent the mean values across topics. (c) denotes the average quality of content of web pages returned to users for the 30 queries most strongly associated with each HIV/AIDS topic of interest.

mation needs, concerns, and misconceptions of individuals across Africa. Such analyses can complement existing top-down approaches and allow us to narrow 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.

Limitations. There are several limitations to using search data. First, the Bing users we study—and Internet users in general—are not a representative sample of the entire population of Africa. Throughout this work, we have included analyses which give evidence that the data is associated with the diseases at hand. However, since ground-truth data, and especially disaggregated by demographics, is hard to come by due to the health data gap, we are further limited in our ability to measure the representativeness of this data. It is therefore 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. Still, many findings corroborated previous smaller scale studies using representative samples. A second limitation is that the results of this study depend on proprietary data from Bing which can limit the ability for health organizations to extend the research.

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 individuals' health information needs that is captured in search data.

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 (Germain 2009), 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. Efforts to understand health information needs in developing nations by demographic group have mostly used surveys and interviews, which are limited in their scale (Li et al. 2004; Wang et al. 2008; Abimanyi-Ochom et al. 2017). Search data allows us to study health needs at a much larger scale.

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. Recent work has studied the potential to use computational techniques to combat health misinformation on social media (Ghenai and Mejova 2018).

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

More data typically the better, but limitations and biases need to be discussed and cannot be used to harm others

Something I've noticed -- I wish we talked more about these "risks and ethics" when ideas are presented in papers, rather than punting it for later

to the expansion of the middle class in developing countries and a lack of resources and programs aimed at minimizing their impact (Sambo 2014). 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 to improve access to health information, and highlighting overlooked aspects of the impacts of these diseases.

References

[Abimanyi-Ochom et al. 2017] Abimanyi-Ochom, J.; Mannan, H.; Groce, N. E.; and McVeigh, J. 2017. 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*.

[AHO 2010] AHO. 2010. African health observatory.

[Allen, Ouedraogo, and McCullough 2010] Allen, W. J.; Ouedraogo, A.; and McCullough, L. 2010. Health information needs in west africa: Results of a survey on the role of the west africa health organization (waho).

[Althouse, Ng, and Cummings 2011] Althouse, B. M.; Ng, Y. Y.; and Cummings, D. A. 2011. Prediction of dengue incidence using search query surveillance. *PLoS neglected tropical diseases*.

[Benjamini and Hochberg 1995] Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*.

[Blei, Ng, and Jordan 2003] Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent Dirichlet allocation. *JMLR* 3.

[Brownstein, Freifeld, and Madoff 2009] Brownstein, J. S.; Freifeld, C. C.; and Madoff, L. C. 2009. Digital disease detectionharnessing the web for public health surveillance. *New England Journal of Medicine*.

[Carneiro and Mylonakis 2009] Carneiro, H. A., and Mylonakis, E. 2009. Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical infectious diseases* 1557–1564.

[CDC 1998] CDC. 1998. Active bacterial core surveillance report, emerging infections program network, streptococcus pneumoniae.

[CDC 2003] CDC. 2003. 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.

[Chan and Tsai 2015] Chan, B., and Tsai, A. 2015. Trends in HIV-related stigma in the general population during the era of antiretroviral treatment expansion: an analysis of 31 sub-Saharan African countries.

[Chan and Tsai 2018] Chan, B. T., and Tsai, A. C. 2018. HIV knowledge trends during an era of rapid antiretroviral therapy scale-up: an analysis of 33 sub-Saharan African countries. *Journal of the International AIDS Society*.

[Chan et al. 2011] Chan, E. H.; Sahai, V.; Conrad, C.; and Brownstein, J. S. 2011. Using web search query data to monitor dengue epidemics: a new model for neglected tropical disease surveillance. *PLoS neglected tropical diseases*.

[De Bruyn 2000] De Bruyn, M. 2000. Gender, adolescents and the HIV/AIDS epidemic: the need for comprehensive sexual and reproductive health responses.

[De Choudhury and De 2014] De Choudhury, M., and De, S. 2014. Mental health discourse on reddit: Self-disclosure, social support, and anonymity.

[De Choudhury, Morris, and White 2014] De Choudhury, M.; Morris, M. R.; and White, R. 2014. Seeking and sharing health information online: Comparing search engines and social media.

[DESA 2017] DESA. 2017. World Population Prospects: The 2017 Revision. United Nations: Department of Economic and Social Affairs.

[Desai et al. 2012] Desai, R.; Hall, A. J.; Lopman, B. A.; Shimshoni, Y.; Rennick, M.; Efron, N.; Matias, Y.; Patel, M. M.; and Parashar, U. D. 2012. Norovirus disease surveillance using google internet query share data. *Clinical Infectious Diseases* e75–e78.

[Eysenbach and Köhler 2002] Eysenbach, G., and Köhler, C. 2002. How do consumers search for and appraise health information on the world wide web? qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ*.

[Eysenbach and Köhler 2004] Eysenbach, G., and Köhler, C. 2004. Health-related searches on the internet. *JAMA*.

[Fiksdal et al. 2014] Fiksdal, A. S.; Kumbamu, A.; Jadhav, A. S.; Cocos, C.; Nelsen, L. A.; Pathak, J.; and McCormick, J. B. 2014. Evaluating the process of online health information searching: a qualitative approach to exploring consumer perspectives. *Journal of Medical Internet research*.

[Fox and Duggan 2013] Fox, S., and Duggan, M. 2013. Health online 2013. *Washington, DC: Pew Internet & American Life Project* 1.

[Fransen-dos Santos 2009] Fransen-dos Santos, R. 2009. Young people, sexual and reproductive health and hiv. *Bulletin of the WHO* 87.

[Genberg et al. 2009] Genberg, B. L.; Hlavka, Z.; Konda, K. A.; Maman, S.; Chariyalertsak, S.; Chingono, A.; Mbwambo, J.; Modiba, P.; Van Rooyen, H.; and Celentano, D. D. 2009. 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*.

[Germain 2009] Germain, A. 2009. Integrating gender into HIV/AIDS programmes in the health sector: tool to improve responsiveness to women's needs. *Bulletin of the WHO*.

[Ghenai and Mejova 2018] Ghenai, A., and Mejova, Y. 2018. Fake cures: User-centric modeling of health misinformation in social media. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):58.

[Ginsberg et al. 2009] Ginsberg, J.; Mohebbi, M. H.; Patel, R. S.; Brammer, L.; Smolinski, M. S.; and Brilliant, L. 2009. Detecting influenza epidemics using search engine query data. *Nature*.

[Global Fund 2016] Global Fund. 2016. Strategic investments for adolescents in HIV, tuberculosis and malaria programs.

[Gombachika et al. 2013] Gombachika, B. C.; Chirwa, E.; Maluwa, A.; et al. 2013. Sources of information on HIV and sexual and reproductive health for couples living with HIV in rural southern malawi. *AIDS Research and Treatment* 2013.

[Hasan and Abuelrub 2011] Hasan, L., and Abuelrub, E. 2011. Assessing the quality of web sites. *Applied Computing and Informatics*.

[Hay et al. 2013] Hay, S. I.; George, D. B.; Moyes, C. L.; and Brownstein, J. S. 2013. Big data opportunities for global infectious disease surveillance. *PLoS medicine* e1001413.

[Hill et al. 2011] Hill, S.; Mao, J.; Ungar, L.; Hennessy, S.; Leonard, C. E.; and Holmes, J. 2011. Natural supplements for H1N1 influenza: retrospective observational infodemiology study of information and search activity on the internet. *JMIR*.

[Hogan and Palmer 2005] Hogan, T. P., and Palmer, C. L. 2005. Information preferences and practices among people living with HIV/AIDS: results from a nationwide survey. *Journal of the Medical Library Association*.

[IHME 2016] IHME. 2016. GBD compare data visualization.

[ITU 2017] ITU. 2017. International Telecommunication Union statistics.

[Kern et al. 2016] Kern, M. L.; Park, G.; Eichstaedt, J. C.; Schwartz, H. A.; Sap, M.; Smith, L. K.; and Ungar, L. H. 2016. Gaining insights from social media language: Methodologies and challenges. *Psychological methods* 21(4):507.

[Kumar and Mmari 2004] Kumar, S., and Mmari, K. 2004. Programming considerations for youth-friendly HIV care and treatment services. *From the Ground Up*.

[Li et al. 2004] Li, X.; Lin, C.; Gao, Z.; Stanton, B.; Fang, X.; Yin, Q.; and Wu, Y. 2004. HIV/AIDS knowledge and the implications for health promotion programs among chinese college students: geographic, gender and age differences. *HPI*.

[Ling and Lee 2016] Ling, R., and Lee, J. 2016. Disease monitoring and health campaign evaluation using google search activities for HIV and AIDS, stroke, colorectal cancer, and marijuana use in canada: a retrospective observational study. *JMIR*.

[Liu et al. 2013] Liu, L. S.; Huh, J.; Neogi, T.; Inkpen, K.; and Pratt, W. 2013. Health vlogger-viewer interaction in chronic illness management.

[Manning, Raghavan, and Schütze 2010] Manning, C.; Raghavan, P.; and Schütze, H. 2010. *Introduction to information retrieval*, volume 16. Cambridge university press.

[Mbonu, van den Borne, and De Vries 2009] Mbonu, N. C.; van den Borne, B.; and De Vries, N. K. 2009. Stigma of people with HIV/AIDS in sub-Saharan Africa: a literature review. *Journal of Tropical Medicine*.

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

[Ocampo, Chunara, and Brownstein 2013] Ocampo, A. J.; Chunara, R.; and Brownstein, J. S. 2013. Using search queries for malaria surveillance, thailand. *Malaria journal*.

[O'Grady 2008] O'Grady, L. 2008. Meeting health information needs of people with HIV/AIDS: sources and means of collaboration. *Health Information & Libraries Journal*.

[Paul and Dredze 2017] Paul, M. J., and Dredze, M. 2017. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services*.

[Polgreen et al. 2008] Polgreen, P. M.; Chen, Y.; Pennock, D. M.; Nelson, F. D.; and Weinstein, R. A. 2008. Using internet searches for influenza surveillance. *Clinical infectious diseases*.

[Root 2010] Root, R. 2010. Situating experiences of hiv-related stigma in swaziland. *Global public health*.

[Rothman et al. 2008] Rothman, K. J.; Greenland, S.; Lash, T. L.; et al. 2008. Modern epidemiology. 3.

[Sambo 2014] Sambo, L. G. 2014. *The health of the people: what works: the African Regional Health Report 2014*. World Health Organization.

[Santillana et al. 2015] Santillana, M.; Nguyen, A. T.; Dredze, M.; Paul, M. J.; Nsoesie, E. O.; and Brownstein, J. S. 2015. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS computational biology*.

[Sarumi and Strode 2015] Sarumi, R. O., and Strode, A. E. 2015. New law on HIV testing in botswana: The implications for healthcare professionals. *Southern African Journal of HIV Medicine*.

[Schwartz and Ungar 2015] Schwartz, H. A., and Ungar, L. H. 2015. 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.

[Schwartz et al. 2013] Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E. P.; and Ungar, L. H. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLOS ONE*.

[Schwartz et al. 2017] Schwartz, H. A.; Giorgi, S.; Sap, M.; Crutchley, P.; Ungar, L.; and Eichstaedt, J. 2017. DLATK: Differential language analysis toolkit.

[Shuyler and Knight 2003] Shuyler, K. S., and Knight, K. M. 2003. What are patients seeking when they turn to the internet? qualitative content analysis of questions asked by visitors to an orthopaedics web site. *Journal of Medical Internet research*.

[Sillence et al. 2007] Sillence, E.; Briggs, P.; Harris, P. R.; and Fishwick, L. 2007. How do patients evaluate and make use of online health information? *Social science & medicine*.

[Spink et al. 2004] Spink, A.; Yang, Y.; Jansen, J.; Nykanen, P.; Lorence, D. P.; Ozmutlu, S.; and Ozmutlu, H. C. 2004. A

study of medical and health queries to web search engines. *Health Information & Libraries Journal*.

[Stonbraker et al. 2017] Stonbraker, S.; Befus, M.; Nadal, L. L.; Halpern, M.; and Larson, E. 2017. Factors associated with health information seeking, processing, and use among HIV positive adults in the dominican republic. *AIDS and Behavior*.

[UNDP 2015a] UNDP. 2015a. UNDP discussion paper: Gender and malaria.

[UNDP 2015b] UNDP. 2015b. UNDP discussion paper: Gender and TB.

[Wang et al. 2008] Wang, J.; Fei, Y.; Shen, H.; and Xu, B. 2008. 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*.

[WB 2018] 2018. The world bank databank: World development index database archives.

[WHO 2009] WHO. 2009. Integrating gender into HIV/AIDS programmes in the health sector: tool to improve responsiveness to women's needs. Technical report.

[Yang, Santillana, and Kou 2015] Yang, S.; Santillana, M.; and Kou, S. C. 2015. Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*.

[Yuan et al. 2013] Yuan, Q.; Nsoesie, E. O.; Lv, B.; Peng, G.; Chunara, R.; and Brownstein, J. S. 2013. Monitoring influenza epidemics in china with search query from baidu. *PloS one*.

[Zheluk et al. 2013] Zheluk, A.; Quinn, C.; Hercz, D.; and Gillespie, J. A. 2013. Internet search patterns of human immunodeficiency virus and the digital divide in the russian federation: infoveillance study. *Journal of medical Internet research*.

[Zhou and Shen 2010] Zhou, X.-c., and Shen, H.-b. 2010. Notifiable infectious disease surveillance with data collected by search engine. *Journal of Zhejiang University Science C*.

Additional Details on Data and Methodology

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 (CDC 2003), 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 emailpii. 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.

Languages Used. One potential source of bias in our data is that the way in which the data were filtered may have caused us to miss relevant queries made in languages other than English. To understand the extent of this potential problem, we examined the set of all search queries made on Bing anywhere in Africa during February 5–11, 2016. We sampled 1,000 random queries from this set and manually determined how many of these queries appeared to be in English. Overall, 22.5% of the queries were in languages other than English. Of the remaining 77.5%, all were either in English or could potentially have been in English (i.e., the name of a celebrity or the name of a country). Most non-English queries were in either Arabic, French, or Portuguese, and most non-English searches were concentrated in a few countries. Morocco, Algeria, and Egypt had especially high concentrations of non-English searchers; 49% of queries from Morocco, 55% of queries from Algeria, and 61% of queries from Egypt were not in English.

User Demographic Distributions. As reported in the main text, a portion of the queries in our data sets were accompanied by self-reported age and/or gender of the user. Users are identified by an anonymous ID. In this section, we present some analyses to indicate the distribution of these values and associations with population demographic statistics.

We first took all anonymous IDs that are associated with a query in any one of the HIV/AIDS, malaria, or tuberculosis data sets. We look at all users who reported ages between 18 and 50. Recall that this age range corresponds to that considered for our analysis for age and topic usage. (We did not consider individuals who report an age below 18 due to ethical considerations and above 50 due to data sparsity concerns.) We found that this age range account for 90%, 86%, and 87% of the queries made by users with age available for the HIV/AIDS, malaria, and tuberculosis data sets, respectively.

We further analyzed association of the reported age and gender of the users with population statistics. Specifically, we use the statistics presented by (DESA 2017), which contains estimated population statistics for different age groups given in 5 year age groups. We took the age ranges between 20 and 49 and consider the fraction of the population in the age ranges 20–24, 25–29, . . . , 45–49. Likewise, we consider the fraction of users who reported ages in these ranges. We find a correlation coefficient of $\rho = 0.9564\ [0.6478, 0.9954]$ with $p < 0.01$. Note, however, that this correlation coefficient drops significantly when adding older ages due to the sparsity concerns present in our data.

Furthermore, 54.19% percent of anonymous IDs which report their gender are men while 45.81% were women. On the other hand, 50.13 % of the African population is female while 49.87% is male (DESA 2017). This digital gap observed by gender is consistent with findings that women (as well as older age-groups, families in rural regions, low-literacy individuals, and other under-served communities)

have less access to Internet (ITU 2017). Note, however, accurate statistics disaggregated by the above groups is challenging to obtain in many African nations.

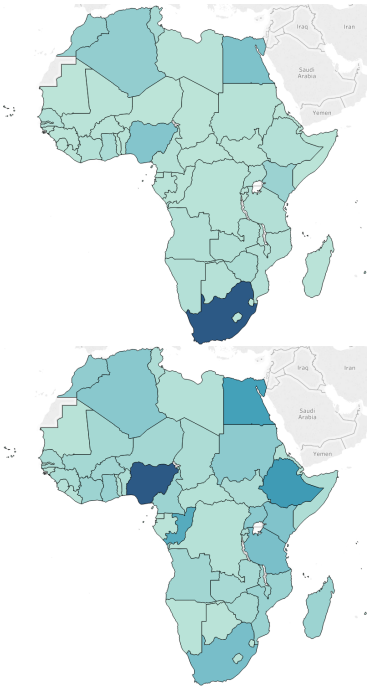

Figure 4: Top: Heat map of the total search traffic in each country during the January 2016–June 2018 period. Bottom: Heat map of the population in each country in 2016.

Coverage of Africa. Our search data covers all 54 nations of Africa. Figure 4 shows a heat map of the total search traffic during January 2016–June 2017 period by country and a heat map of the 2016 population of each country. The Spearman correlation coefficient between total search traffic during the January 2016–June 2017 period by country and the 2016 population of each country 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 $p < 0.01$.

Building on the results from Figure 1, we run a multiple linear regression with the HIV prevalence rate as the dependent variable and the fraction of searches associated with the disease, Internet penetration, percent of population in urban settings, population, and GDP as the explanatory variables. We find that the fraction of searches containing the disease name explain some of the variance in disease prevalence even after controlling for the other explanatory variables.

Additional Details on the Topics

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.

Analyses of the Popularity of Topics

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. We 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 5.[4]

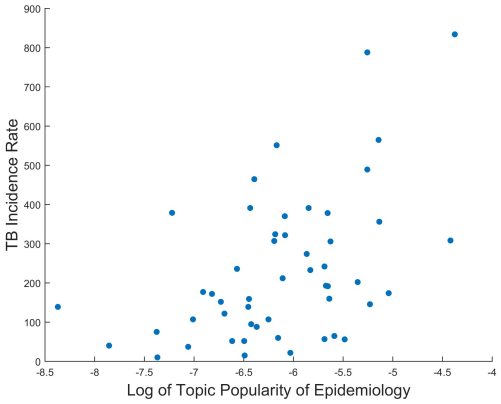

Figure 5: Popularity of the *Epidemiology* topic vs. tuberculosis incidence.

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 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, women and individuals in the 25–34 age group expressed relatively more interest in topics related to pregnancy, breastfeeding, and family care compared to men 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. This is consistent with the literature that nearly half of all new HIV infections occur among the 15–24 age group. The 25–34 age group corresponds to the time-frame where these infections may progress significantly, and even develop to AIDS.

[4] As in the main text, relationships are considered significant if they pass the Benjamini-Hochberg false discovery rate (multi-test

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

Disease	Topic	20 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	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	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	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	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	hiv cd4 count 500 cd4 cd8 ratio hiv t4 count in hiv
	Celebrity gossip (0.95%)	***, ***, positive, hiv-positive, status, ***, ***, ***, TRUE, ***, ***, ***, ***, ***, ***, ***, ***, ***, ***, ***, ***	*** *** 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	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	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	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	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	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	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	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	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	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	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	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	global fund fight aids tuberculosis malaria tb malaria cholera meningitis

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

	Women	Ages 18–24	Ages 25–34	Ages 35–49
Drug	0.006	-0.068**	-0.008	0.031
Natl Cure	0.027	-0.051**	-0.014	0.051**
Breast.	0.073**	-0.089**	0.060**	0.031
Epidem.	-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

Table 6: Relative topic popularity by user demo. for TB.

	Women	Ages 18–24	Ages 25–34	Ages 35–49
Epidm.	0.006	0.037*	-0.078**	0.007
Drug Res.	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
Side-effe.	0.027	0.028	-0.004	-0.016
Natl Cure	0.018	-0.020	-0.016	0.022

We looked at the topic popularity of the six topics of interest from the topic table in the main paper. Women and individuals in the 25–34 age group expressed relatively more interest in topics related to pregnancy, breastfeeding, and family care compared to men 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.

Additional Details on User Behavior and Quality of Results

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. We use the same methodology described in the main text and the regression coefficients to report the average values. Results are shown in Figure 6. 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.

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

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.

correction) with $\alpha = 0.5$.

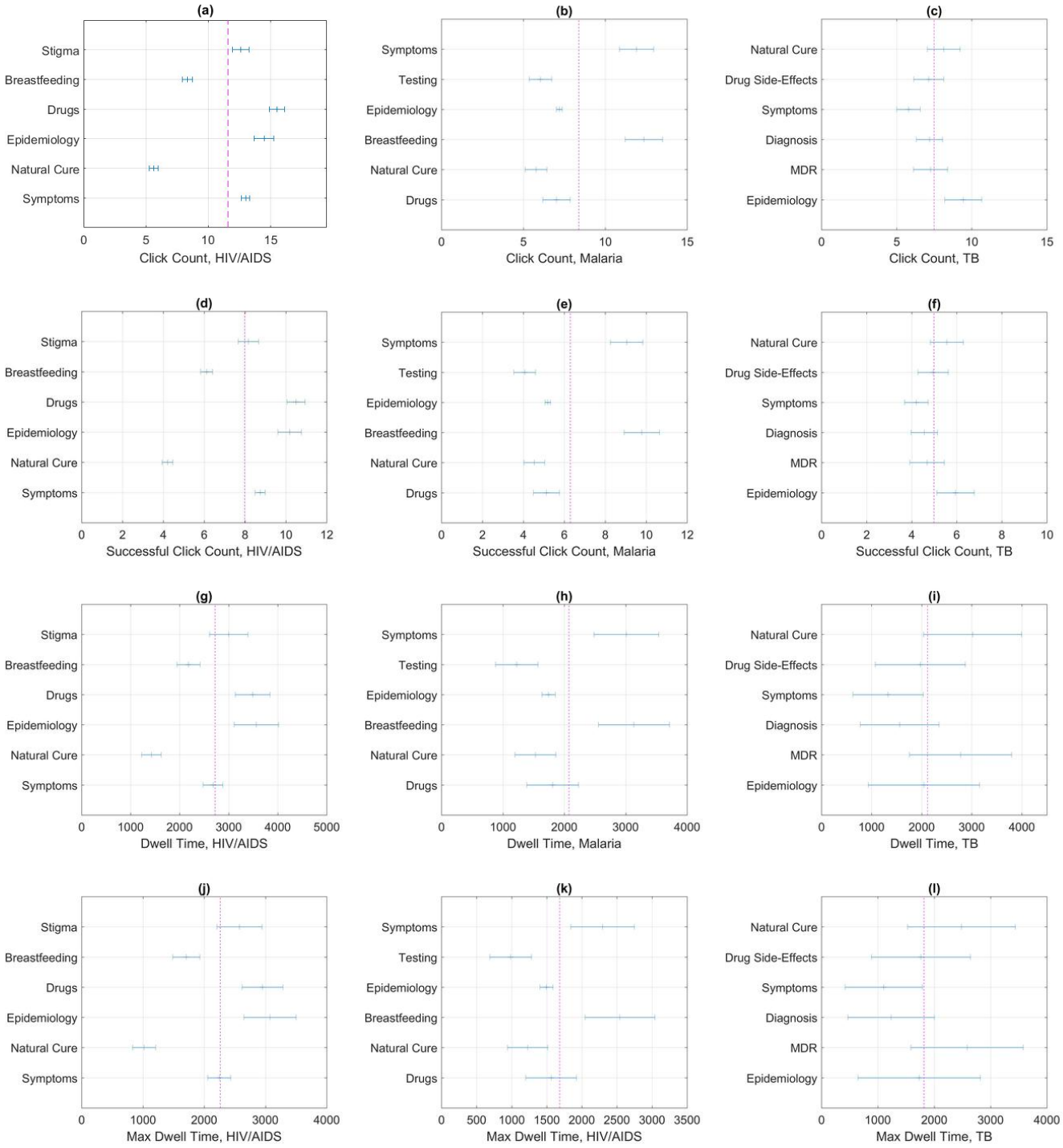

Figure 6: Rows, from top to bottom: average click count, successful click count, dwell time, and maximum dwell time for queries associated with different topics. Columns, from left to right: HIV/AIDS, malaria, and tuberculosis data sets. The vertical lines represent the mean values across topic.

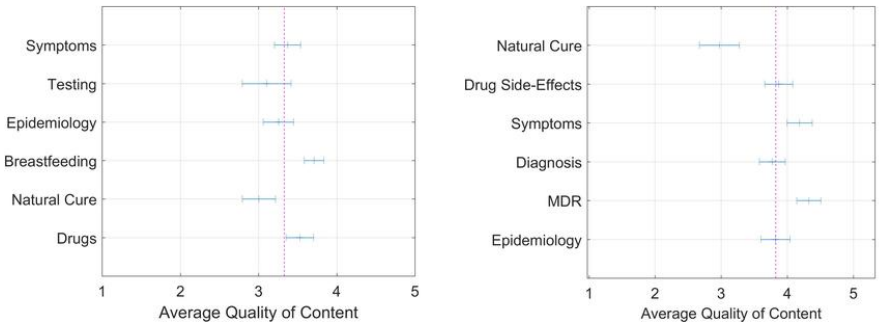

Figure 7: Average quality of content of webpages returned to users for the 30 queries most strongly associated with the six malaria (left) and tuberculosis (right) topics.