9-2013

# Lessons Learned about Public Health from Online Crowd Surveillance

Shawndra Hill
*University of Pennsylvania*

Raina Merchant
*University of Pennsylvania*

Lyle Ungar
*University of Pennsylvania*



Recommended Citation
Hill, S., Merchant, R., & Ungar, L. (2013). Lessons Learned about Public Health from Online Crowd Surveillance. *Big Data, 1* (3), 160-167. http://dx.doi.org/10.1089/big.2013.0020

# Lessons Learned about Public Health from Online Crowd Surveillance

**Abstract**
The Internet has forever changed the way people access information and make decisions about their healthcare needs. Patients now share information about their health at unprecedented rates on social networking sites such as Twitter and Facebook and on medical discussion boards. In addition to explicitly shared information about health conditions through posts, patients reveal data on their inner fears and desires about health when searching for health-related keywords on search engines. Data are also generated by the use of mobile phone applications that track users' health behaviors (e.g., eating and exercise habits) as well as give medical advice. The data generated through these applications are mined and repackaged by surveillance systems developed by academics, companies, and governments alike to provide insight to patients and healthcare providers for medical decisions. Until recently, most Internet research in public health has been surveillance focused or monitoring health behaviors. Only recently have researchers used and interacted with the crowd to ask questions and collect health-related data. In the future, we expect to move from this surveillance focus to the "ideal" of Internet-based patient-level interventions where healthcare providers help patients change their health behaviors. In this article, we highlight the results of our prior research on crowd surveillance and make suggestions for the future.

**Disciplines**
Business | Communication Technology and New Media | Health Communication | Medicine and Health | Public Health | Social Media

ORIGINAL ARTICLE


# LESSONS LEARNED ABOUT PUBLIC HEALTH FROM ONLINE CROWD SURVEILLANCE

*Shawndra Hill,[1] Raina Merchant,[2] and Lyle Ungar[3]*

## Abstract

*The Internet has forever changed the way people access information and make decisions about their healthcare needs. Patients now share information about their health at unprecedented rates on social networking sites such as Twitter and Facebook and on medical discussion boards. In addition to explicitly shared information about health conditions through posts, patients reveal data on their inner fears and desires about health when searching for health-related keywords on search engines. Data are also generated by the use of mobile phone applications that track users' health behaviors (e.g., eating and exercise habits) as well as give medical advice. The data generated through these applications are mined and repackaged by surveillance systems developed by academics, companies, and governments alike to provide insight to patients and healthcare providers for medical decisions. Until recently, most Internet research in public health has been surveillance focused or monitoring health behaviors. Only recently have researchers used and interacted with the crowd to ask questions and collect health-related data. In the future, we expect to move from this surveillance focus to the "ideal" of Internet-based patient-level interventions where healthcare providers help patients change their health behaviors. In this article, we highlight the results of our prior research on crowd surveillance and make suggestions for the future.*

## Introduction

Widespread Internet usage and social networking have permanently changed the way people access information and make decisions about their healthcare needs. Patients search for health and medical information online, use mobile phone applications to track their health behaviors (e.g., eating, sleep, and exercise habits), and now have an unprecedented ability to share personal health information on medical discussion boards, as well as on social networking sites such as Twitter and Facebook, revealing their inner fears and hopes by sharing explicit information about their health in social media posts and searching for health-related keywords on search engines. These data, generated by keyword searches, social media posts, and mobile applications, are mined and repackaged by health surveillance systems that have been designed through collaboration among academics, private companies, and government agencies to provide insight into the medical decisions of both patients and healthcare providers.

Collecting data through these means and mining the data for insights is called *online crowd surveillance*. Most Internet research in the field of public health has until now focused on monitoring health behaviors; however, researchers have recently begun to interact with users to collect a wider variety of

[1]*Operations and Information Management Department, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania.*
*Departments of [2]Emergency Medicine and [3]Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania.*



BIG DATA SEPTEMBER 2013 • DOI: 10.1089/big.2013.0020

health-related data. In the near future, we expect to move from a largely surveillance focus to the "ideal" of Internet-based patient-level interventions, where healthcare providers actually help patients to change their health behaviors, for example, by helping them eat more healthfully or stop smoking. In this article, we highlight the results of our prior research on online crowd surveillance, using a unique dataset to illustrate one of its limitations and provide suggestions for how "big data" might be utilized in the public health field in the future.

## Surveillance

The Centers for Disease Control[1] referred to surveillance as, "The systematic, ongoing, collection, management and interpretation of these data to public health programs to stimulate public health action." The attractiveness of the Internet as a research tool to health policy researchers for online crowd surveillance lies in its population-level scale and its ability to access the uncensored thoughts of patients, all for minimal cost. In essence, Internet users comprise a larger focus "crowd" group than other traditional methods make practicable, where the "voices of millions" can be heard. With the massive amounts of data this makes available, it is no surprise that researchers have used the Internet for surveillance.[2]

**"COLLECTING DATA THROUGH THESE MEANS AND MINING THE DATA FOR INSIGHTS IS CALLED *ONLINE CROWD SURVEILLANCE*."**

Indeed, through surveillance, researchers have access to surprisingly rich public health-related data, generated when patients congregate, seek information, and discuss their concerns and outcomes.[3] Twitter especially has proven to be an abundant source of such information. For example, although many postings on Twitter communicate seemingly mundane accounts of everyday life and experiences, this chatter often also includes disclosure of emotional and physical well-being.[4–10] Recent studies have suggested that 8.5% of English-language tweets relate to disease of some type, and 16.6–25.1% relate to health.[11] This information can be downloaded, geocoded, and characterized by researchers for content and demographics.[12]

Twitter has served as a source of health-related data in numerous novel ways. In particular, Twitter's immediacy has permitted real-time assistance in the case of natural disasters (hurricanes and earthquakes, for instance) by allowing for the widescale broadcast of available resource, enabling people in need of medical assistance to locate help.[10,13,14] This immediacy also allows for much quicker surveillance for targeting infection "hot spots" in pandemic situations, as was done by companies such as Google in the H1N1 crisis.[9,15,16] However, the potential application is much broader than simply emergency situations or healthcare: linguists and sociologists, among others, have mined tweets for their research, among other things, succeeding in distinguishing local dialects and forecasting the moods and opinions of populations in specific geographic regions.[17,18]

In terms of nonemergency healthcare, many studies offer important public health insights about linking the origin of sadness and depression to a number of serious medical conditions, and new methods of identifying them are always welcome. For example, researchers have recently been able to link changes in tweeting behavior to postpartum depression.[19] Others have used Twitter to quantify medical misconceptions (e.g., sequelae of concussions) and the spread of poor medical compliance (e.g., antibiotic use).[8,20] In our recent work,[21] we have used Twitter to understand how people communicate online about cardiovascular health. Specifically, we sought to characterize how Twitter users seek and share information related to cardiac arrest, which is a time-sensitive cardiovascular condition where initial treatment is often reliant on public knowledge and response. This project demonstrated that tweets about cardiovascular health could be identified, sorted, and characterized relative to content and the person generating the content. Twitter offers promise as a research tool not only because of its immense scale, but also because the content of messages can be systematically searched.[22] The immediacy of Twitter offers another great advantage as a research tool. For example, emergency departments in Boston learned about the 2013 marathon bombings through Twitter *before* announcements from conventional sources such as the media or established emergency service communication channels.[23] While terrorist attacks are an extreme case, the general principle holds.

Surveillance opportunities extend far beyond Twitter, however, with the Internet offering significant opportunities for researchers and public health officials alike. Patients discuss their health with others on medical discussion boards and review sites, which provide a test-bed for public health surveillance. In our work,[24–27] for instance, we used medical discussion board data to successfully link drugs and homeopathic remedies to relevant side effects.[27] We developed a methodology for establishing a corpus of medical message board posts, anonymizing the corpus and successfully extracting information on potential adverse drug effects discussed by users. In addition, we used these data to determine the extent to which patients use social media to discuss side effects related to medications. In addition to linking drug use to side effects, we also focused our research more specifically on discussions by breast cancer patients related to using aromatase inhibitors (AIs), with particular




emphasis on AI-related arthralgia, and sought to understand the frequency and content of side effects and associated *adherence* behaviors. We found that online discussions of AI-related side effects are common and often relate to drug switching and discontinuation.[24] Obviously, physicians would benefit from awareness of the implications of these discussions and should promote optimal adherence by guiding patients in managing side effects effectively. It is this type of awareness—of what the "person in the street" is saying—that research such as ours can provide to an unparalleled extent.

In addition to posting information about their health, patients search for solutions on the Internet and often click on links to health-related websites. When collected, these link data are useful indicators of public health. Data resulting from search queries have been found to be highly predictive of a wide range of population-level health behaviors. For example, trends in Google and Yahoo search queries can be used to predict epidemics of illnesses such as flu and dengue fever,[28] the seasonality of mental health, depression and suicide,[29,30] the prevalence of Lyme disease,[31] incidence of kidney stone,[31] and the prevalence of smoking and electronic cigarette use.[32] Web logs, which serve as histories of data about where people click, are predictive of individual characteristics such as mental health and dietary preferences.[33] While the availability of vast amounts of information about health on the Web means that people will find information when they search, we have found that search keyword selection is critical for arriving at reliable curated health content.[34]

## Limitations to Surveillance

While the collection and analysis of Internet data is a promising path to better understanding of health behaviors, this strategy suffers from several limitations. First, eavesdropping on such communication involves privacy concerns that have not been fully resolved. People have an expectation of and right to privacy, particularly when they discuss health-related issues. Internet-based data gathering thus represents both logistic challenges (e.g., how to get people to opt in to share their Facebook status updates) and potential ethics dilemmas (if one predicts that someone is at risk for suicide based on his/her posts, should one intervene in some way?). Second, such data are obtained without context; it does not include a patient's health history or medical outcomes, merely a snapshot of their daily lives. (Health history is almost impossible to come by if one only collects anonymized tweets or posts.) In the absence of context, causal claims about specific behaviors and health conditions are thus difficult to substantiate. Third, Internet-based data are seldom curated; with no distinction between genuine and spurious information, it becomes increasingly important to develop methodologies for isolating "the signal from the noise." Fourth, a commonly expressed concern about data from Twitter and similar services relates to defining the sample populations. Twitter users do not represent a random sample of the population; for instance, the elderly and young children are less likely to use Twitter than people between the ages of 18 and 40. Although studies have shown that Twitter represents broad demographic segments of the population,[35–37] drawing conclusions without considering the populations can be problematic. In our current work, we seek to understand how bias in the representation of Internet users impacts the conclusions drawn at the population level.

**"TWITTER OFFERS PROMISE AS A RESEARCH TOOL NOT ONLY BECAUSE OF ITS IMMENSE SCALE, BUT ALSO BECAUSE THE CONTENT OF MESSAGES CAN BE SYSTEMATICALLY SEARCHED."**

To illustrate the severity of the problem of relying on tweet data to draw population-level conclusions, we present below results from a large-scale survey of U.S. households, the Simmons National Consumer Study, annually issued to over 12,000 adults over the age of 18. The survey asks respondents questions on all aspects of their daily lives, including product purchases, news consumption, Internet usage, opinions, and health. To demonstrate the problems that may exist when generalizing to the entire population if special care is not taken to poststratify the information to match the general population, in Table 1, we combine answers from the survey about Internet usage and health from the Simmons survey. Table 1 presents the number of people in the U.S. population over age 18 who have diseases or conditions queried about in the Simmons survey in 2011 and 2012. For each year, we present the estimated counts of people in the population with the disease and people on Twitter with the disease. These data come directly from the Simmons survey. Survey respondents were asked about both their health conditions and whether they used Twitter. Therefore, we can cross-tabulate users by both of these characteristics. When we rank the conditions by their prevalence, some obvious differences appear. First, conditions more prevalent in the elderly, such as hypertension, arthritis, and high cholesterol, show up in the top five in the population, but not for Twitter users. On the other hand, conditions that skew young, like acne and anxiety, rank higher in prevalence on Twitter.

Much more serious problems than the differences in Twitter versus population demographics, however, arise from the facts that words are ambiguous (e.g., "heart attack" or "MI" mostly do not refer to heart attacks) and that people mention diseases without necessarily experiencing them.

Table 1. Ranking of 47 Health Symptoms and Diseases by Prevalence in the US Population* and Prevalence of Twitter Users for 2011–2012

| | 2012 | | | | 2011 | | | |
|---|---|---|---|---|---|---|---|---|
| | *US* | *Rank* | *Twitter* | *Rank* | *US* | *Rank* | *Twitter* | *Rank* |
| Total | 230124 | | 15631 | | 227008 | | 11629 | |
| Hypertension/High blood pressure | 43459 | 1 | 1480 | 16 | 43464 | 2 | 1158 | 8 |
| Backache | 42043 | 2 | 2605 | 1 | 47488 | 1 | 2151 | 1 |
| High cholesterol | 37861 | 3 | 1668 | 12 | 39707 | 3 | 585 | 16 |
| Any arthritis | 34412 | 4 | 1293 | 21 | 32043 | 5 | 365 | 22 |
| Acid reflux disease (gerd) | 32383 | 5 | 2445 | 3 | 35293 | 4 | 1161 | 7 |
| Overweight (30 lbs or more) | 27051 | 6 | 2137 | 6 | 30133 | 6 | 1613 | 4 |
| Heartburn | 26799 | 7 | 2218 | 4 | 26029 | 7 | 1387 | 6 |
| Arthritis (osteoarthritis) | 26688 | 8 | 936 | 28 | 24133 | 8 | 264 | 28 |
| Anxiety | 18824 | 9 | 2465 | 2 | 18773 | 11 | 2071 | 2 |
| Depression | 18693 | 10 | 2173 | 5 | 18783 | 10 | 1530 | 5 |
| Gas | 18481 | 11 | 1990 | 7 | 16233 | 13 | 881 | 14 |
| Nasal allergies/Hay fever | 18232 | 12 | 1316 | 19 | 22045 | 9 | 921 | 13 |
| Flu | 17167 | 13 | 1786 | 11 | 17465 | 12 | 1671 | 3 |
| Diabetes type 2 | 16487 | 14 | 746 | 32 | 16061 | 15 | 338 | 23 |
| Migraine headache | 16422 | 15 | 1803 | 10 | 14630 | 18 | 1090 | 10 |
| Sensitive teeth | 16341 | 16 | 1527 | 14 | 16168 | 14 | 805 | 15 |
| Snoring/Sleep apnea | 16056 | 17 | 1414 | 17 | 14462 | 19 | 573 | 17 |
| Insomnia/Sleep disorder | 13671 | 18 | 1853 | 9 | 15752 | 16 | 923 | 12 |
| Cold sores | 13461 | 19 | 1593 | 13 | 12229 | 22 | 933 | 11 |
| Asthma | 12423 | 20 | 1000 | 24 | 15007 | 17 | 1091 | 9 |
| Indigestion | 12192 | 21 | 671 | 33 | 12343 | 21 | 391 | 20 |
| Acne | 11220 | 22 | 1985 | 8 | | | | |
| Hemorrhoids | 11076 | 23 | 1512 | 15 | 10540 | 24 | 436 | 19 |
| Arthritis (rheumatoid arthritis) | 11071 | 24 | 509 | 37 | 11021 | 23 | 151 | 33 |
| Chronic pain | 10438 | 25 | 978 | 25 | 12575 | 20 | 276 | 26 |
| Urinary tract infection (uti) | 9992 | 26 | 1025 | 23 | 8528 | 26 | 472 | 18 |
| Nail fungus | 9386 | 27 | 1348 | 18 | 10365 | 25 | 383 | 21 |
| Athlete's foot | 8679 | 28 | 1306 | 20 | 8256 | 27 | 272 | 27 |
| Overactive bladder | 7426 | 29 | 940 | 27 | 7490 | 31 | 109 | 36 |
| Irritable bowel syndrome | 7363 | 30 | 943 | 26 | 7910 | 30 | 288 | 25 |
| Constipation (chronic) | 6651 | 31 | 204 | 42 | 7258 | 33 | 169 | 32 |
| Eczema/Psoriasis | 6531 | 32 | 1192 | 22 | 7321 | 32 | 187 | 31 |
| Osteoporosis | 6040 | 33 | 131 | 43 | 7925 | 29 | 142 | 34 |
| Heart disease/Congestive heart failure | 5876 | 34 | 460 | 38 | 8164 | 28 | 42 | 43 |
| Hiatal hernia | 5580 | 35 | 647 | 35 | 4382 | 37 | 97 | 38 |
| COPD (Chronic obstructive pulmonary dis) | 5451 | 36 | 862 | 30 | 5226 | 35 | 56 | 41 |
| Cancer | 5031 | 37 | 460 | 39 | 4202 | 39 | 27 | 46 |
| Add/Adhd | 4860 | 38 | 879 | 29 | 4944 | 36 | 230 | 29 |
| Diabetes Type 1 | 4328 | 39 | 450 | 40 | 4260 | 38 | 98 | 37 |
| Chronic Bronchitis | 4077 | 40 | 781 | 31 | 5980 | 34 | 60 | 40 |
| Impotence/Loss of Libido | 4069 | 41 | 652 | 34 | 3861 | 41 | 128 | 35 |
| Stomach Ulcers | 3298 | 42 | 31 | 47 | 3574 | 42 | 227 | 30 |
| Heart attack/Stroke | 2997 | 43 | 109 | 45 | 3945 | 40 | 33 | 44 |
| Emphysema | 2592 | 44 | 636 | 36 | 2424 | 43 | 32 | 45 |
| Genital Herpes | 1808 | 45 | 333 | 41 | 1692 | 46 | 48 | 42 |
| Chronic Kidney Disease | 1773 | 46 | 64 | 46 | | | | |
| Human Papilloma Virus | 1456 | 47 | 119 | 44 | 2114 | 45 | 299 | 24 |

*18 and over.

Thus, keywords searched for on Twitter do not necessarily accurately represent the incidence of specific medical problems. For example, Table 2 shows the number of tweets on Twitter about the 10 most prevalent diseases as well as the rank of the disease in the US population. We collected the tweets during the week August 7–13, 2013. We simply searched Twitter for the listed keywords and counted the resulting tweets. We see again that the Twitter ranking by keywords differs greatly from the incidence rate. For example, the most tweeted-about terms related to names of the top 10 symptoms and conditions were anxiety and depression, whereas these are at the bottom of the top 10 list in terms of prevalence. It is important to also note that the proportion of individuals tweeting about certain conditions is very low. For example, very few people tweet about arthritis or the word "obese." Instead, most of the tweets containing these words


Table 2. Ranking of the Top 10 Health Symptoms and Diseases in US Poulation* Compared to Number of Tweets Collected During the Week August 7 to 13, 2013

| | *US* | *Rank* | *Keywords* | *Tweets* | *Proportion of tweets about having the "disease"* | *Proportion of tweets from individuals (not organizations)* |
|---|---|---|---|---|---|---|
| Hypertension/high blood pressure | 43459 | 1 | hypertension/high blood pressure | 63 | 0.03 | 0.44 |
| Backache | 42043 | 2 | backache | 61 | 0.70 | 0.95 |
| High cholesterol | 37861 | 3 | cholesterol | 55 | 0.00 | 0.35 |
| Any arthritis | 34412 | 4 | arthritis | 50 | 0.00 | 0.14 |
| Acid reflux disease (gerd) | 32383 | 5 | acid reflux | 22 | 0.14 | 0.41 |
| Overweight (30 lbs or more) | 27051 | 6 | obese | 89 | 0.00 | 0.19 |
| Heartburn | 26799 | 7 | heartburn | 26 | 0.31 | 0.42 |
| Arthritis (osteoarthritis) | 26688 | 8 | arthritis | 50 | 0.00 | 0.14 |
| Anxiety | 18824 | 9 | anxiety | 305 | 0.02 | 0.27 |
| Depression | 18693 | 10 | depressed | 405 | 0.02 | 0.40 |

*US Population 18 years and older.

are from health organizations. Finally, with the exception of backache, very few people are tweeting about having the condition themselves. Instead, they are sharing news and using the related terms to mean something other than the health condition. It is likely that no one factor accounts for this; a variety of reasons, including word ambiguity, omission of synonyms, stigma about the disease, the geographic location and demographics of Tweeters, and the different government and NGO involvement in disease all affect the tweet rate. In ongoing work, we are studying how to correct for biases introduced by these and other factors.

## Calling the Crowd to Action

While much of our work has been focused on mining social media data, there are other ways to employ Internet users to help solve public health–related challenges, for example, through crowd-sourcing. The Internet provides access to millions of users who can potentially answer a call for action, as has been demonstrated by the success of crowd-sourcing projects in many areas, including health challenges. As mentioned above, we see the opportunity for public health officials to move from simple surveillance to using the power of crowd-sourcing to collect public health data.[38–58] During a recent literature review, we found that in addition to surveillance, crowd-sourcing was frequently used for problem solving, data processing, and surveying.[59]

"RESEARCHERS WILL BETTER UNDERSTAND PATIENTS AND PATIENTS WILL BETTER UNDERSTAND THEMSELVES AS THEY BECOME MORE PROACTIVE ABOUT THEIR HEALTH."

Crowd-sourcing has been used to provide data processing relating to a wide range of health-related tasks, including classifying polyps in computer tomography colonography images,[54] and then providing feedback to help optimize presentation of the polyps[53]; annotating public webcam images to determine how the addition of a bike lane changed the mode of transportation observed in the images[57]; and examining red blood cells for the presence of infection[51,52] or thick blood smears containing[50] malaria parasites (*Plasmodium falciparum*). In a survey of workers on Amazon.com's Mechanical Turk, the crowd workforce was surveyed for malarial symptoms as part of a study to assess the prevalence of malaria in India.[46] Another survey provided a mobile phone application that allowed users to report potential flulike symptoms along with GPS coordinates and other details. Response data from the survey enabled researchers to chart the incidence of flu symptoms that matched relatively well with Centers for Disease Control data.[40]

Crowd-sourcing can be used both as a way of gathering public health data and as a way of getting "crowd-sourced workers" (e.g., Mechanical Turk) to sift through and locate health data. In our work, we sought to determine the feasibility of using mobile workforce technology to validate locations of automated external defibrillators (AEDs), which are an emergency public health resource. We developed a crowd-sourcing application, the MyHeartMap Challenge, to organize the public reporting of AED locations throughout a major U.S. metropolitan area. This study had three purposes. First, we wanted to investigate the capacity of crowd-sourcing and social media for collecting meaningful public health data regarding an underutilized health-related technology. Second, we wanted to determine the locations of existing AEDs and build a serviceable inventory of AEDs within a defined region for use by laypeople and municipal service providers during life-threatening emergencies. The study provided a baseline snapshot of AED locations at a particular point in time. This will serve as the foundation for updating and maintaining a

database of the devices over time. The third purpose was to evaluate the survey process of data collection itself, including the demographics and motivations of participants who submitted the crowd-sourced information, as well as the validity of the data submitted. Although we used the crowd, we noted that as with other Internet studies, participants were demographically limited. A major challenge when calling a crowd to action is incentivizing participation for a survey population with certain health conditions from across all walks of life. Nevertheless, despite its problems, the crowd-sourcing of health information presents tremendous opportunities, since the available survey population is still much larger than the traditional focus groups that were employed for health-related studies in the past.

## The Future Is Intervention

What should we expect in the near future? Certainly, there will be further advances in healthcare surveillance methodology that integrates information from disparate sources such as Tweets, Facebook posts, medical records, purchases, and cell phone data. The forms in which data are available are also diversifying as patients increasingly gather health information from sources such as YouTube videos and their personal electronic medical records, and self-monitor their health behaviors using devices such as Nike wristbands or other medical measuring devices that are linked to smart phones. Additionally, we expect crowd-sourcing to play a major role in gathering health information. The data generated will be useful to both researchers and individuals. Researchers will better understand patients and patients will better understand themselves as they become more proactive about their health.

The biggest change, however, will be the shift from merely monitoring people's activities to actually using this information to induce behavioral changes that can impact individual health-related practices. Many of the most actionable health issues involve individual behaviors that can be modulated by feedback and social influence; these include exercise, obesity, smoking, drunk driving, lack of medication compliance, and seeking treatment for problems such as depression. Having access to a wealth of personal health information available, and the ability to develop interventions via cell phones or social networking sites open up a multitude of ways to improve the general health of the population-related behaviors.

Over the last decade, the doctor–patient relationship has shifted. Patients now routinely use the Internet to obtain medical information as well as a second—or sometimes first—opinion on their healthcare options. For example, upon receiving a diagnosis that a relative has cancer, or that one's mother does, a common first response is to Google the illness in order to understand the treatment options and potential outcomes. Patients then bring this knowledge—factual or not—to their next meeting with their doctor. While patients generally perceive physicians and other clinicians as highly credible and influential sources for health-related information, it is believed that people are also highly influenced by the opinions of friends and by information obtained from the Internet, whether or not these can be verified. The effect of these often nonprofessional opinions can be misinformation. This observation becomes even more significant when considering the amount of time the average person spends in a clinical setting in direct communication with a health professional compared with the amount of time s/he spend communicating with other people. Most individuals spend less than 2 hours a year with a physician, compared with the annual 5,000 hours spent in communication with others. Given that because of the spacing effect, repetition and convenience of access to information offer a greater likelihood of its retention, it is clear that nonclinical methods of imparting health information are likelier to have an effect than visits to a clinician, despite the latter's greater authority. Therefore, it is critical to provide reliable health information on the Web for patients.

This use of the Internet for health information goes beyond the management of one's health that has typically been the doctor's purview: people want to know not only how to best treat illnesses, but also, increasingly, how to be healthier and happier in general. For example, research has overwhelmingly shown that exercise has significant health benefits, as do being happy and having good relationships. This being the case, it is evident that attaining positive health outcomes involves a host of small daily decisions, many of which can be supported through mechanisms such as phone and social network reminders and support groups. The move from healthcare surveillance to actually helping people take control of their health presents healthcare professionals with a plethora of exciting opportunities. Data mining will play a crucial role in this effort by helping to determine which interventions are effective, at which times, and for which people. Further refinement of data mining abilities will doubtless increase the possibilities, and it will then be possible, thanks to these data, not only to see which interventions work, but also to plan new ones with a higher likelihood of success.

## Acknowledgments

We would like to thank our many collaborators and research assistants on our prior work discussed in this article. The prior work was supported by the National Library of Medicine (RC1LM010342) and K23 grant 10714038. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine or the National Institutes of Health. The funding source did not play any role in the study design, in the collection, analysis and interpretation of data; in the writing of the manuscript; or in the decision to submit the manuscript for publication.


## Disclosure Statement

No competing financial interests exist.

## References

1. Thacker SB, Qualters JR, Lee LM. Public health surveillance in the United States: Evolution and challenges. 2012. Available online at www.cdc.gov/mmwr/preview/mmwrhtml/su6103a2.htm (Last accessed on Aug. 13, 2013).
2. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection—Harnessing the Web for public health surveillance. N Engl J Med 2009; 360:2153–2157.
3. Lagu T, Lindenauer PK. Putting the public back in public reporting of health care quality. JAMA 2010; 304:1711–1712.
4. Twitter. 2013. Available online at twitter.com (Last accessed on Jun. 1, 2013).
5. Collier N, Son NT, Nguyen NM. OMG U got flu? Analysis of shared health messages for bio-surveillance. J Biomed Semantics 2011; 2 Suppl 5:S9.
6. Bosley JC, et al. Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation 2013; 84:206–212.
7. Lyles CR, et al. "5 mins of uncomfyness is better than dealing with cancer 4 a lifetime": An exploratory qualitative analysis of cervical and breast cancer screening dialogue on Twitter. J Cancer Educ 2013; 28:127–133.
8. Sullivan SJ, et al. "What's happening?" A content analysis of concussion-related traffic on Twitter. Br J Sports Med 2012; 46:258–263.
9. Chew C, Eysenbach G. Pandemics in the age of Twitter: Content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 2010; 5:e14118.
10. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. Am J Trop Med Hyg 2012; 86:39–45.
11. Paul MJ, Dredze M. A Model for Mining Public Health Topics from Twitter. Baltimore: Johns Hopkins University, 2011, pp. 16–26.
12. Hill S, Benton A, Xu J. Talkographics: Using what viewers say online to calculate audience affinity networks for social TV-based recommendations. 2012. Available online at http://ssrn.com/abstract=2273381 (Last accessed on August 26, 2013).
13. Merchant RM, Elmer S, Lurie N. Integrating social media into emergency-preparedness efforts. N Engl J Med 2011; 365:289–291.
14. Keim ME, Noji E. Emergent use of social media: A new age of opportunity for disaster resilience. Am J Disaster Med 2011; 6:47–54.
15. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One 2011; 6:e19467.
16. St. Louis C, Zorlu G. Can Twitter predict disease outbreaks? BMJ 2012; 344:e2353.
17. Mocanu D, et al. The Twitter of babel: Mapping world languages through microblogging platforms. PLoS One 2013; 8:e61981.
18. Dodds PS, et al. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. PLoS One 2011; 6:e26752.
19. De Choudhury M, Counts S, Horvitz E. Predicting postpartum changes in emotion and behavior via social media. Proceedings of the 2013 ACM Annual Conference on Human Factors in Computing Systems, ACM, 2013; pp. 3267–3276.
20. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: Twitter and antibiotics. Am J Infect Control 2010; 38:182–188.
21. Bosley JC, et al. Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. Resuscitation 2013; 84:206–212.
22. Gawande A. Why Boston's hospitals were ready. The New Yorker, April 17, 2013.
23. Cassa CA, et al. Twitter as a sentinel in emergency situations: Lessons from the Boston marathon explosions. PLoS Curr 2013; 5.
24. Mao JJ, et al. Online discussion of drug side effects and discontinuation among breast cancer survivors. Pharmacoepidemiol drug Saf 2013; 22:256–262.
25. Benton A, et al. A system for de-identifying medical message board text. BMC Bioinform 2011; 12(Suppl 3):S2.
26. Benton A, et al. Medpie: An information extraction package for medical message board posts. Bioinformatics 2012; 28:743–744.
27. Benton A, et al. Identifying potential adverse effects using the Web: A new approach to medical hypothesis generation. J Biomed Inform 2011; 44:989–996.
28. Chan EH, et al. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. PLoS Negl Trop Dis 2011; 5:e1206.
29. Ayers JW, et al. Seasonality in seeking mental health information on Google. Am J Prev Med 2013; 44:520–525.
30. McCarthy MJ. Internet monitoring of suicide risk in the population. J Affect Disord 2010; 122:277–279.
31. Seifter A, et al. The utility of "Google Trends" for epidemiological research: Lyme disease as an example. Geospatial Health 2010; 4:135–137.
32. Ayers JW, Ribisl KM, Brownstein JS. Tracking the rise in popularity of electronic nicotine delivery systems (electronic cigarettes) using search query surveillance. Am J Prev Med 2011; 40:448–453.
33. West R, White RW, Horvitz E. From cookies to cooks: Insights on dietary patterns via analysis of web usage logs. Proceedings of the 22nd International Conference on World Wide Web, International World Wide Web Conferences Steering Committee, 2013; pp. 1399–1410.
34. Hill S, et al. Natural supplements for H1N1 influenza: Retrospective observational infodemiology study of in-

formation and search activity on the Internet. J Med Internet Res 2011; 13:e36.
35. Nielsen. State of the media: The social media report—Q3. 2011. Available online at http://blog.nielsen.com/nielsenwire/social/ (Last accessed on Jun. 1, 2013).
36. Duggan M, Brenner J. The demographics of social media users—2012. The Pew Internet and American Life Project. 2013. Available online at www.pewinternet.org/Reports/2013/social-media-users.aspx (Last accessed on Jun. 1, 2013).
37. Epstein JOB, Smith NA, Xing EP. A latent model for geographic lexical variation. EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, 2010, pp. 1277–1287.
38. Cooper S, et al. The challenge of designing scientific discovery games. Proceedings of the Fifth International Conference on the Foundations of Digital Games. Monterey, CA: ACM, 2010, pp. 40–47.
39. Cooper S, et al. Predicting protein structures with a multiplayer online game. Nature 2010; 466:756–760.
40. Freifeld CC, et al. Participatory epidemiology: Use of mobile phones for community-based health reporting. PLoS Med 2010; 7:e1000376.
41. Behrend TS, et al. The viability of crowdsourcing for survey research. Behav Res Methods 2011; 43:800–813.
42. Bender J, et al. Collaborative authoring: A case study of the use of a wiki as a tool to keep systematic reviews up to date. Open Med 2011; 5:e201–e208.
43. Cooper S, et al. Analysis of social gameplay macros in the Foldit cookbook. Proceedings of the 6th International Conference on Foundations of Digital Games. Bordeaux, France: ACM, 2011, pp. 9–14.
44. Khatib F, et al. Crystal structure of a monomeric retroviral protease solved by protein folding game players. Nat Struct Mol Biol 2011; 18:1175–1177.
45. Khatib F, et al. Algorithm discovery by protein folding game players. Proc Natl Acad Sci USA 2011; 108:18949–18953.
46. Chunara R, et al. Online reporting for malaria surveillance using micro-monetary incentives, in urban India 2010–2011. Malar J 2012; 11:43.
47. Eiben CB, et al. Increased Diels-Alderase activity through backbone remodeling guided by Foldit players. Nat Biotechnol 2012; 30:190–192.
48. Jarmolowicz DP, et al. Using crowdsourcing to examine relations between delay and probability discounting. Behav Processes 2012; 91:308–312.
49. Kawrykow A, et al. Phylo: A citizen science approach for improving multiple sequence alignment. PLoS One 2012; 7:e31362.
50. Luengo-Oroz MA, Arranz A, Frean J. Crowdsourcing malaria parasite quantification: An online game for analyzing images of infected thick blood smears. J Med Internet Res 2012; 14:e167.
51. Mavandadi S, et al. Distributed medical image analysis and diagnosis through crowd-sourced games: A malaria case study. PLoS One 2012; 7:e37245.
52. Mavandadi S, et al. Crowd-sourced BioGames: Managing the big data problem for next-generation lab-on-a-chip platforms. Lab Chip 2012; 12:4102–4106.
53. McKenna MT, et al. Strategies for improved interpretation of computer-aided detections for CT colonography utilizing distributed human intelligence. Med. Image Anal. 2012; 16:1280–1292.
54. Nguyen TB, et al. Distributed human intelligence for colonic polyp classification in computer-aided detection for CT colonography. Radiology 2012; 262:824–833.
55. Turner AM, Kirchhoff K, Capurro D. Using crowdsourcing technology for testing multilingual public health promotion materials. J Med Internet Res 2012; 14:e79.
56. Crump MJ, McDonnell JV, Gureckis TM. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. PLoS One 2013; 8:e57410.
57. Hipp JA, et al. Emerging technologies: Webcams and crowd-sourcing to identify active transportation. Am J Prev Med 2013; 44:96–97.
58. Merchant RM, et al. A crowdsourcing innovation challenge to locate and map automated external defibrillators. Circ Cardiovasc Qual Outcomes 2013; 6:229–236.
59. Ranard BL, Ha YP, Meisel ZF, et al. Crowdsourcing—harnessing the masses to advance health and medicine, a systematic review. J Gen Intern Med 2013; pp. 1–17.

Address correspondence to:

*Shawndra Hill*
*Operations and Information Management*
*University of Pennsylvania*
*3730 Walnut Street, Suite 500*
*Philadelphia, PA 19103*

*E-mail:* shawndra@wharton.upenn.edu