Could Behavioral Medicine Lead the Web Data Revolution?

John W. Ayers, Ph.D., M.A.[1], **Benjamin M. Althouse, Ph.D., Sc.M.**[2], and **Mark Dredze, Ph.D.**[3]

[1]Graduate School of Public Health, San Diego State University, San Diego, CA, USA

[2]Santa Fe Institute, Santa Fe, NM, USA

[3]Human Language Technology Center of Excellence, Johns Hopkins University Baltimore, MD, USA

Digital footprints left on search engines, social media, and social networking sites can be aggregated and analyzed as health proxies, yielding anonymous and instantaneous insights. On the one hand, nearly all the existing work has focused on acute diseases. This means the value-added from web surveillance is reduced, because the effectiveness of even high profile systems, such as *Google Flu Trends,* have been found inferior to already strong traditional surveillance.[1] On the other hand, the future of web surveillance is promising in an area where traditional surveillance is largely incomplete: behavioral medicine, a multidisciplinary field incorporating medicine, social science, and public health and focusing on health behaviors and mental health.

The proportion of illness (or death) attributable to health behaviors or psychological well-being has steadily increased over the last half century, while surveillance of these outcomes has remained largely unchanged. Investigators simply ask people about their health on surveys. However, surveys have well-known limitations, such as respondents' reluctance to participate, social desirability biases, difficulty in accurately reporting behaviors, long lags between data collection and availability, and provisions (sometimes legal) curtailing the inclusion of politically sensitive topics like gun violence. Most importantly, the expense of surveys means many topics are either not covered or covered restrictively (*e.g.*, clinical depression screeners are included in the *Behavioral Risk Factor Surveillance System* just every other year). Given the current budget climate, survey capacity will likely worsen before it improves. To overcome these limits, behavioral medicine should now embrace web data.

First, behavioral medicine requires observing behavior or the manifestation of mental health problems. Doing so online is easier, more comprehensive, and more effective than with surveys, because many outcomes are passively exhibited there. For example, one study showed how precise health concerns changed during the United States recession of December 2008 through 2011, by systematically selecting Google search queries and using the content of each query to describe the concern and the change in volume to describe concern prevalence. "Stomach ulcer symptoms," for example, were 228% (95%CI, 35–363)

Address Correspondents: JWA, Graduate School of Public Health, San Diego State University, 9245 Sky Park Court, San Diego, CA 92123. Fax: 858.505.4770. Phone: 619.371.1846.

higher than expected during the recession, with queries thematically related to arrhythmia, congestion, pain (including many foci like head, tooth and back) also elevated.[2] This approach highlights how web data can reveal largely assumption-free insights, via systematic data generation of hundreds of possible outcomes rather than arbitrary *a priori* selection of a few outcomes by investigators.

Second, web data reflects more than the individual, because social context can also be captured online. Online networks can reveal how mechanistic drivers such as social norms spread and influence population health. For example, social patterns in obesity promotion and suppression have been described by pooling *Facebook* posts that encourage television watching or going outdoors, which ultimately explained variability in neighborhood obesity rates.[3] Moreover, social support concepts are often expressed in web data, like observing specific instances of caregiving and confidence on Twitter. As a result, online behavioral medicine can move away from understanding aggregation based purely on location and towards understanding health in the context of our human interconnectedness.

Third, web data are potentially the only source for real-time insights into behavioral medicine, where web data can be available almost immediately compared to a 365-day lag time between annual surveys. By harnessing these data around social events or interventions, programs can be evaluated as they are implemented, hypothetically generating real-time feedback to maximize their effectiveness. Web data in this vein also hold promise for guiding investigator resources. In 2011, when tobacco journals were debating snus (a smokeless tobacco product), and funders were soliciting proposals to understand the snus pandemic, electronic cigarettes already attracted more searches on Google than any other smoking alternative, snus included.[4]

In this same way, web data can guide traditional surveillance, like vetting the inclusion of questions on surveys using online proxies.

Fourth, given all hypotheses are based on some data, web data can be an important source for identifying new hypotheses. Many hypotheses in behavioral medicine can be traced directly to data availability and can appear *ad hoc* to lay audiences. Many studies have explored birthdate seasonality in mental health problems. Why? Birthdates are routinely found in traditional surveillance, while some mental health problems are too rare to assess incidence or increased severity seasonality. As a result, obvious questions are never explored, until now. Is schizophrenia seasonal? Online interest in schizophrenia and its symptoms – as well as 8 other outcomes - peak in the winter.[5] What is the healthiest day? Online interest in quitting smoking across the globe is highest on Monday.[6] Behavioral medicine needs to escape the confines of limited data to more fully specify the next frontier of research questions, and going online is one such escape.

Fifth, it is beyond present scientific limits for a hypothetical arm to reach out of the screen to inoculate against infection. In behavioral medicine, however, substantial resources have been used to develop online interventions that treat or prevent illness with effectiveness equivalent to their offline counterparts. For example, as early as the mid-1990s, investigators implemented online programs to promote behavioral health. A meta-analysis found these

programs relatively increased quitting smoking 44%,[7] yet a research agenda for harnessing the surveillance potential of the web has not been articulated. Improving the online surveillance capacity means online interventions can be better disseminated via online screening or linking subjects to existing online treatments (i.e., what advertisements for an online program are most effective?).

Sixth, some of the most effective interventions in behavioral medicine involve changes in public policy. Web data can identify alerts for policy changes and pathways for health advocacy. For instance, by archiving online media, places considering policy changes can be identified, and this information can then be passed onto advocacy groups. Case in point, Brazilian President Lula's laryngeal cancer prompted broad changes in media coverage of tobacco control, and soon after, Brazil became the largest smoke-free nation to date.[8] By prospectively analyzing news media content, advocacy resources may be more cost-effectively spent during opportunistic times, including events like Lula's diagnosis, will be possible.

A major criticism is that web data have sampling biases. However, such biases are increasingly eroding at the population level as more people go online. In addition, several studies have demonstrated that valid trends reflecting the entire population, and even subsets of the population, can be extracted from online data. For example, computer science has already developed approaches for identifying the gender, ethnicity or education associated with a Twitter account using the content of a user's Tweets. Going forward, the research community may mimic these studies and validate methods for obtaining high quality, actionable information in behavioral medicine, then further realizing the comparative value of web data to traditional data.

Billions of digital footprints from nearly all parts of the United States and from countries around the world provide a powerful opportunity to expand the evidence-base across medicine. However, for the above reasons and more related reasons yet to be expressed, behavioral medicine potentially has the most to gain from web data and could be essential to the broader web data revolution.

Acknowledgments

References

1. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol. 2013; 9:e1003256. [PubMed: 24146603]

2. Althouse BM, Allem J-P, Childers MA, Dredze M, Ayers JW. Population Health Concerns During the United States' Great Recession. Am J Prev Med. 2014; 46:166–170. [PubMed: 24439350]

3. Chunara R, Bouton L, Ayers JW, Brownstein JS. Assessing the online social environment for surveillance of obesity prevalence. PLoS One. 2013; 8:e61373. [PubMed: 23637820]

4. 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. [PubMed: 21406279]

5. Ayers JW, Althouse BM, Johnson M, Cohen JE. Circaseptan (weekly) rhythms in smoking cessation considerations. JAMA Intern Med. 2014; 174:146–148. [PubMed: 24166181]

6. Ayers JW, Althouse BM, Allem JP, Rosenquist JN, Ford DE. Seasonality in seeking mental health information on google. Am J Prev Med. 2013; 44:520–525. [PubMed: 23597817]

7. Myung SK, McDonnell DD, Kazinets G, Seo HG, Moskowitz JM. Effects of Web- and computer-based smoking cessation programs: meta-analysis of randomized controlled trials. Arch Intern Med. 2009; 169:929–937. [PubMed: 19468084]

8. Ayers JW, Althouse BM, Noar SM, Cohen JE. Do celebrity cancer diagnoses promote primary cancer prevention? Prev Med. 2014; 58:81–84. [PubMed: 24252489]