Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Current Opinion in Psychology

Research opportunities at the intersection of social media and survey data

Emma S Spiro

This article reviews literature related to the use of social media, specifically Twitter, to study health behavior. It will discuss the potential for studies that link social media data with survey data. After detailing study design considerations and outlining guidelines for work in this area, it will propose opportunities for novel contributions to health research.

Address
Information School, Mary Gates Hall Ste. 370, University of Washington, Seattle, WA 98195, USA

Corresponding author: Spiro, Emma S (espiro@uw.edu)

Current Opinion in Psychology 2016, **9**:67–71

This review comes from a themed issue on **Social media and applications to health behavior**

Edited by **Clayton Neighbors** and **Melissa Lewis**

http://dx.doi.org/10.1016/j.copsyc.2015.10.023

Introduction

As social media continue to become integrated into daily life, the extensive records these systems archive as part of normal operation promise to change the available avenues of inquiry in the social and behavioral sciences [1]. Researchers from a variety of disciplines are using social media data to explore human behavior [2,3•,4–6,7••,8]. A growing number of studies explore questions within the health domain [9••], where social media have been used to explore the effects of peer influence on health behaviors [10•,11], to aid in the detection of mental health concerns [12,13••], and to quantify deviant behaviors [14–16]. Other work has looked at the use of social media as a tool for health communication [17], particularly within the domains of public health [18–20,21•,22], health information seeking [23] and general health status or well-being [24,25••,26]. While the majority of these studies are observational and descriptive, recent work has also employed large-scale, randomized experimental trials to infer causal relationships [27].

The focus of this review is Twitter, a platform designed to allow users to find out what other people are doing [28,29]. Within the site, 140-character messages, known as *tweets*, are posted by individual users and then delivered to that person's content subscribers, known as *followers* [30,31]. The larger stream of public tweets is searchable through the website interface. As of 2015, almost 25% of online adults use Twitter [32•], and the platform has particularly high penetration among Internet users 18–29 years old [33]. Research utilizing Twitter data specifically has dramatically increased in recent years as scholars recognize the wealth of observational data available [34,35,29]. However, while social media provide large-scale data in terms of the number of observations or cases, data are often quite poor in terms of the user features or characteristics contained. Very limited data is available about user socio-demographic characteristics, for example, driving researchers to infer such attributes from the data that is available. These techniques have become more prevalent within the literature [36,37].

For social scientists, the lack of individual-level attributes limits the scope of questions that can be addressed using social media data. Cavazos-Rehg *et al.* [38••], for example, offer a detailed look at tweets about drinking behaviors for prominent Twitter users, but they are unable to determine whether pro-drinking or anti-drinking content is more prevalent in certain socio-demographic groups, compared to others. While prior work exploring 'big social data' makes important contributions to understanding how these data provide insight into health behaviors, many studies analyze tweets in isolation (see e.g. [39•]) limiting the generalizability of the work and its applications to specific domains (e.g. youth and young adult populations) [40].

The limitations of Twitter and other social media data, however, are precisely the benefits of more traditional forms of data collection—surveys and questionnaires—where detailed individual-level attributes are self-reported by participants. Likewise, the limitations of survey data (e.g. prohibitive resource cost for large-scale efforts) are the strengths of big social data. In combining both methodological approaches, researchers can curate rich, complex observations of health behavior at scale. As such, this paper offers an alternative framework for utilizing social media data in studies of health behavior—a user-centric approach that builds from survey data. It considers the advantages and opportunities of research at this intersection, offering new directions for scientific inquiry.

Linking Twitter data to survey data

Survey research is widespread in the social and behavioral sciences [41]; both the advantages and disadvantages of survey research are well documented. Yet limited work lies at the intersection of surveys and social media. Social media have been used to aid participant recruitment [42,43], but few studies [44••,13••] have collected data about survey participants' activity on social media platforms. Survey and big social data methods complement each other by alleviating weaknesses in the other approach. Moreover, linking social media to survey data affords the ability to associate self-reported health behaviors with those expressed (directly or indirectly) online [45].

If health behaviors be detected, measured and inferred from observations on social media platforms, large-scale, early intervention and behavior change strategies may be viable [46,9••]. This is the core aim of many studies of social media and health behavior. However, without validation these techniques—which require 'ground truth' data to which to compare inferred health behaviors—these claims fall short. Survey data can provide comparison data; and while self-reported health behaviors have notable constraints (e.g. social desirability and participant recall concerns) the potential value of this work is substantial.

In practice, combining Twitter data with survey is not without challenges. Study participants can be asked to volunteer social media identities as part of contact information on a survey or questionnaire. Participants may or may not choose to volunteer this information. However, in the case of Twitter, this information (account usernames) is all that is currently needed to collect large amounts of data on the online social behaviors of study participants.

User-centric data collection on Twitter

Opportunities for novel research contributions at the intersection of social media and survey data abound. In order for researchers to capitalize on these opportunities it is necessary to design data collection systems that augment survey and questionnaire data with social media data. This section briefly discusses some of the challenges that arise in this endeavor, offering design guidelines and other considerations.

Defining the study population. Studies that utilize big social data, such as tweets, face notable issues related to the dual problems of defining the study population and sampling that population. There are multiple approaches to sampling data from Twitter which are in part determined by the restrictions of the Twitter application programming interface (API)—the standard access point for data collection most tools utilize (e.g. see [47,48•]). The Twitter API offers programmatic access at the level of the tweet or the user. While prior work tends to use data collected from the public stream of tweets, we in contrast propose methods that utilize user-centric entry points.

One of the advantages this framework is that the population of interest is well-defined—defined by the survey component of the study—avoiding issues that arise when attempting statistical sampling strategies in social media environments (which are in many cases near impossible due to unknown factors, such as the set of possible cases). Bringing to bear all the standard techniques for sampling populations already employed in the social and behavioral sciences, one obtains a sample of participants, each of whom is asked for their social media identifiers (e.g. account username of Twitter). Not only does this alleviate sampling bias towards highly active users (as is present in tweet-based samples), but it also grants the possibility of obtaining a representative sample of the population of interest.

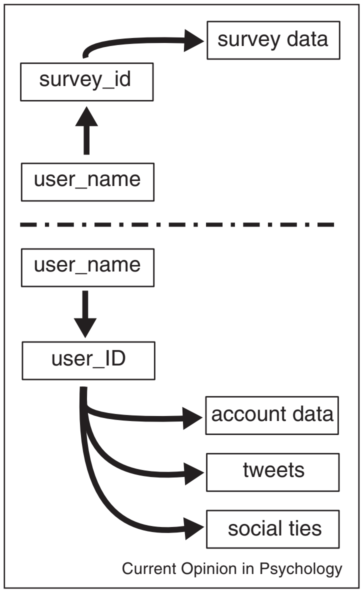
Linking survey and social media data is entirely contingent on the willingness of survey participants to share their social media identify. Whether non-response bias and/or response rates are a severe limitation remains an open question. During a pilot study of this approach, our project found that approximately 20% of study participants were willing to share social media handles as part of their contact information before taking part in a study of risky alcohol and sex-related behaviors (PI:Lewis, R01AA021379). As work in this area continues, these limitations will be further explored.

Longitudinal monitoring of behavior on Twitter. While detailed description of data collection tools is outside the scope of this paper, this section provides sufficient detail for researchers to understand the overall aims of this approach. Data collection begins with a well-defined study population, with accompanying social media account information collected during survey administration. Twitter account usernames (`user_name`) allow for an initial access point to collect participants' social media data. Twitter usernames are not static and can be changed by the user while still maintaining the same account. Therefore, it is recommended that researchers obtain a unique account identifier for each participant soon after survey data has been collected; this unique identified will be referred to as a `user_id`. IDs can be used to systematically sample Twitter data.

Three types of observational data are available via the Twitter API: Firstly, information about the user account; secondly, tweets; and finally, social tie information. Each component must be queried in turn. One can obtain a user's most recent tweets at the time of query.[1] Data collection can subsequently occur at regular intervals to continue longitudinal observation. If data collection proceeds over time the researcher can ensure (almost) complete data on each participant during that period; every tweet posted during the observation period can be

[1] At the time of writing the limit was 3200 posts.

Figure 1



User-centric data collection system architecture.

collected. The one exception is when the user deletes messages before collection. Owing to Twitter API limits, the number of data requests that can be made per time interval is finite; requests must spaced out to collect data on large study populations (e.g. thousands of participants). The overall collection framework is depicted in Figure 1.

Data processing and storage. Working with big social data comes with new practical challenges. Social media data are often available in specialized formats due to the unstructured nature of the content available. Twitter, for example, provides data in JavaScript Object Notation (JSON) which is a common format for data consisting of attribute-value pairs. After obtained, data must be processed, cleaned, and stored in a manner that facilitates subsequent analysis; more often than not this means adding data to and querying data from a database. Simply manipulating data involves technical know-how beyond the training of many social scientists, though a growing number of social scientists are obtaining these skills. As such, the capacity to work in this area will likely require interdisciplinary collaborations and training.

Privacy and ethical concerns. Users share a significant amount of personal information online [35]. Researchers must be aware of privacy and ethical concerns in this domain [49,40]. For work at the intersection of survey and social media data these issues are paramount. While many choices regarding privacy and ethics are domain and study specific, general guidelines are important. Careful consideration of privacy expectations should be discussed among the research team; while most social media data is public, users may have expectations about the availability and visibility of content. Researchers should consider not publishing specific tweets, especially when that content is archived and searchable. As research continues, scholars should assist in educating social media users and the public about what behavior online can reveal.

Current and future research opportunities

Studies that employ the approach detailed here, augmenting survey data with social media data, can contribute to a number of areas within the health domain. First, a number of methodological problems remain open. Questions about the association between self-reported behaviors and those observed (or inferred) from social media activity are many. Theses issues are vital to capitalizing on the value of social media for early detection and intervention or prevention campaigns.

Linking social media to survey data provides rich attributes about social media users. In this context, researchers can explore the social media expressions and behaviors of particular socio-demographic groups. One promising direction for future work considers how self-disclosure choices themselves are structured along demographic lines [50]. This work has consequences for methods that infer individual characteristics from social media activity.

Social media may also afford opportunities to explore health behaviors difficult to obtain reliable data on via survey techniques; behaviors that have strong social desirability bias could be estimated from social media data, where individuals may be less likely to self-censor their expressions (e.g. see [51]).

As social media continue to change, new platforms will enter the landscape providing additional opportunities for research on health behaviors. Activity tracking data, photos, videos, and more will provide rich data for study [52,53].

Conclusion

Social media data have many applications to studies of health behavior. Research utilizing data from Twitter has seen rapid growth in recent years. However, while many studies collect data at the level of tweets, alternative approaches are underutilized. The potential for studies that link social media data with more traditional forms of data—survey and questionnaire data—is substantial. A number of opportunities for novel contributions to health research under this paradigm exist.

Conflict of interest statement

Nothing declared.

References

1. Lazer D, Pentland A, Adamic LA, Aral S, Barabási A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy MW, Roy D, Van Alstyne M: **Computational social science**. *Science* 2009, **323**:721-723.

2. Salganik MJ, Watts J: **Web-based experiments for the study of collective social dynamics in cultural markets**. *Topics Cogn Sci* 2009, **1**:439-468.

3. • Golder SA, Macy MW: **Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures**. *Science* 2011, **333**:1878-1881.
This article presents an analysis of individual-level diurnal and seasonal mood rhythms across the world. The study uses a dataset of public messages on Twitter and text analysis methods to estimate mood across the day, finding that moods deteriorate as the day progresses. The study also found that baseline positive affect varies with changes in daylength.

4. Aral S: **Poked to vote**. *Nature* 2012, **489**:8-10.

5. Monroy-Hernández A, Boyd D, Kiciman E, Counts St: **Narcotweets: social media in wartime**. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media Narcotweets*; *AAAI: 2012:515-518*.

6. Spiro ES, Sutton J, Greczek M, Fitzhugh S, Pierski N, Butts CT: **Rumoring During extreme events: a case study of deepwater horizon 2010**. In *Proceedings of the 3rd Annual ACM Web Science Conference*. ACM; 2012:275-283.

7. •• Golder SA, Macy MW: **Digital footprints: opportunities and challenges for online social research**. *Ann Rev Sociol* 2014, **40**:129-152.
This article reviews current advances in online social research, focusing specifically on presenting a critical analysis of both theoretical and methodological opportunities and limitations. The article discusses online environments for both observational and experimental research. It also reviews specific substantive areas where current work has led to novel insight, such as information diffusion and collective action.

8. Choe EK, Lee NB, Lee B, Pratt W, Kientz JA: **Understanding quantified-selfer' practices in collecting and exploring personal data**. *CHI*. Toronto, Ontario, Canada: ACM; 2014, .

9. •• Centola D: **Social media and the science of health behavior**. *Circulation* 2013, **127**:2135-2144.
This article reviews how social media can be used to study health behaviors. It specifically consider the methodological challenges related to studies of social influence and health behaviors, and notes specific capabilities of social media data and platforms to address these challenges. This article focuses on how social media can be used to experimentally evaluate the effects of social influence on behavior change.

10. • Teodoro R, Naaman M: **Fitter with Twitter: understanding personal health and fitness activity in social media**. *ICWSM*. AAAI; 2013:611-620.
This study presents a qualitative analysis of Twitter posts, as well as an extensive set of interviews with experienced users who post messages on Twitter about exercise, diet, and weight loss activities. Study findings show that most participants did not seek out fitness communities, but rather found them by accident. The authors also highlight feedback and accountability as important factors for participation.

11. Fowler J, Christakis N: **Quitting in droves: collective dynamics of smoking behavior in a large social network**. *N Engl J Med* 2010, **358**:2249-2258.

12. De Choudhury M, Counts S, Horvitz E: **Social media as a measurement tool of depression in populations**. *WebScience*; *Paris, France: 2013*.

13. •• De Choudhury M, Gamon M, Counts S, Horvitz E: **Predicting depression via social media**. *ICWSM, vol 2*. AAAI; 2013:128-137.
This study is one of the few that combined survey and social media data. The authors explore the use of social media data to detect and diagnose major depressive disorder in individuals. Results demonstrate that social media contains useful signals for characterizing the onset of depression, including decreased activity and raised negative affect.

14. Heber West J: **Temporal variability of problem drinking on Twitter**. *Open J Prev Med* 2012, **2**:43-48.

15. De Choudhury M, Counts S, Horvitz E: **Major life changes and behavioral markers in social media: case of childbirth**. *CSCW*. San Antonio, Texas: ACM; 2013, .

16. Thompson L, Rivara FP, Whitehill JM: **Prevalence of marijuana-related traffic on Twitter, 2012–2013: a content analysis**. *Cyberpsychol Behav Soc Netw* 2015, **18**:311-319.

17. Sylvia Chou W-y, Hunt YM, Burke BE, Moser RP, Hesse BW: **Social media use in the United States: implications for health communication**. *J Med Internet Res* 2009, **11**:e48.

18. Scanfeld D, Scanfeld V, Larson EL: **Dissemination of health information through social networks: Twitter and antibiotics**. *Am J Infect Control* 2010, **38**:182-188.

19. Salathé M, Khandelwal S: **Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control**. *PLoS Comput Biol* 2011, **7**.

20. Krieck M, Dreesman J, Otrusina L, Denecke K: **A new age of public health: Identifying disease outbreaks by analyzing tweets**. In *Proceedings of Health Web-Science Workshop, ACM Web Science Conference*. 2011.

21. • Paul MJ, Dredze M: **You are what you Tweet: analyzing Twitter for public health**. *ICWSM*. AAAI; 2011:265-272.
This study considers the problem of extracting public heath data from Twitter. The study uses a large-scale dataset of public Twitter data to estimate cases of various health-related conditions such as allergies or influenza. The study results suggest that Twitter could be a cost effective means for estimating some public health concerns, but not all. Results are limited to population level metrics because data on individuals is sparse.

22. Lazard AJ, Scheinfeld E, Bernhardt JM, Wilcox GB, Suran M: **Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat**. *Am J Infect Control* 2015:1-3.

23. Fox S: *The Social Life of Health Information, 2011. Technical Report*. Washington DC: Pew Research Center; 2011.

24. Hawn C: **Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care**. *Health Aff* 2009, **28**:361-368.

25. •• Korda H, Itani Z: **Harnessing social media for health promotion and behavior change**. *Health Prom Pract* 2011, **14**:15-23.
This article summarizes current empirical evidence and understanding of using social media for health promotion. The article reviews not only what is known about how social media is effective at reaching public audiences, but also how social media can be used in the context of enhancing health knowledge, changing behavior, and monitoring health practices.

26. Sadilek A, Kautz H: **Modeling the impact of lifestyle on health at scale**. *WSDM*. ACM; 2013:637-646.

27. Centola D, van de Rijt A: **Choosing your network: social preferences in an online health community**. *Soc Sci Med* 2015, **125**:19-31.

28. Murthy D: **Twitter: microphone for the masses?** *Media Cult Soc* 2011, **33**:779-789.

29. Liu Y, Kliman-Silver C, Mislove A: **The Tweets They are A-Changin: evolution of twitter users and behavior**. *ICWSM*. . 2014:5-314.

30. Boyd D, Golder SA, Lotan G: **Tweet, tweet, retweet: conversational aspects of retweeting on twitter**. *HICSS-43*. IEEE; 2010:1-10.

31. Grinberg N, Naaman M, Shaw Blake, Lotan G: **Extracting diurnal patterns of real world activity from social media**. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*. AAAI; 2013:205-214.

32. • Duggan M, Ellison NB, Lampe C, Lenhart A, Madden M: *Social Media Update 2014*. 2015.
This report details findings from a 2014 Pew survey on social media adoption and use. It comments on social and demographics differences in social media site usage.

33. Duggan M, Brenner J: *The Demographics of Social Media Users—2012. Technical Report*. Pew Research Center; 2013.

34. Kiciman E: **OMG, I Have to Tweet That! A study of factors that influence tweet rates**. In In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media OMG.* Edited by AAAI. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media OMG* 2012:170-177.

35. Humphreys L, Gill P, Krishnamurthy B: **Twitter: a content analysis of personal information**. *Inform Commun Soc* 2013:1-15.

36. Mislove A, Lehmann S, Ahn Y-y, Onnela J-p, Niels Rosenquist J: **Understanding the demographics of Twitter users**. *ICWSM*. 2012:554-557.

37. Sloan L, Morgan J, Burnap P, Williams M: **Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data**. *PLOS ONE* 2015, **10**:e0115545.

38. Cavazos-Rehg PA, Krauss MJ, Sowles SJ, Bierut LJ: **"Hey •• Everyone, I'm Drunk." An evaluation of drinking-related Twitter chatter**. *J Stud Alcohol Drugs* 2014, **76**:635-639.
This study explores alcohol and drinking-related content on Twitter. By examining a large sample of public tweets posted by users with high Klout scores (a highly debated measure of social influence in social media) the authors consider the sentiment and theme of alcohol and drinking-related content within public content. Results indicate high frequency of pro-alcohol content, produced mainly by non-commercial sources.

39. Cavazos-Rehg PA, Krauss M, Fisher SL, Salyer P, Grucza Ra, Jean • Bierut L: **Twitter chatter about marijuana**. *J Adolesc Health* 2015, **56**:139-145.
In this study, similar in approach to the article cited above, the authors explore sentiment and themes in public marijuana-related content on Twitter. The study is restricted to public tweets posted by users with high Klout scores. The study finds that the majority of content expressed positive sentiment towards marijuana use. Using demographics inferred by a third party the authors suggest Tweeters of marijuana-related content were younger and a greater proportion were African-American compared with the Twitter average.

40. Tufekci Z: **Big questions for social media big data: representativeness, validity and other methodological pitfalls**. *ICWSM*. 2014:10.

41. Sudman S, Bradburn NM: *Asking questions: a practical guide to questionnaire design*. 1982.

42. Ramo DE, Prochaska JJ: **Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use**. *J Med Internet Res* 2012, **14**:e28.

43. Fenner Y, Garland SM, Moore EE, Jayasinghe Y, Fletcher A, Tabrizi SN, Gunasekaran B, Wark JD: **Web-based recruiting for health research using a social networking site: an exploratory study**. *J Med Internet Res* 2012, **14**:e20.

44. Moreno MA, Christakis DA, Egan KG, Brockman LN, Becker T: •• **Associations between displayed alcohol references on Facebook and problem drinking among college students**. *Arch Pediatr Adolesc Med* 2012, **166**:157-163.
This study examines associations between displayed alcohol use and intoxication/problem drinking (I/PD) references on Facebook and self-reported problem drinking as measured by the AUDIT scale. Study participants were undergraduate students at two universities in the United States. Facebook profiles of participants were evaluated by trained coders to determine the presence or absence of alcohol references. I/PD references were positively related to being categorized as at risk for problem drinking in survey data.

45. McCormick T, Lee H, Cesare N, Shojaie A, Spiro E: **Using Twitter for demographic and social science research: tools for data collection and processing**. *Sociol Methods Res* 2015:1-7.

46. De Choudhury M, Counts S, Horvitz E: **Predicting postpartum changes in emotion and behavior via social media**. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems—CHI'13*. 2013:3267.

47. Williams JA, Spiro ES: **Tweeting the Japanese general election of 2014—a first look**. *Workshop on Voting, Elections and Electoral Systems*. Seattle, WA: University of Washington; 2015, 1-19.

48. Burgess J, Bruns A: **Twitter archives and the challenges of "big •• social data" for media and communication research**. *M/C J* 2014, **15**:1-5.
This article describes in detail aspects of tweet data, messages on Twitter, gathered through the Twitter API. The authors discuss methodological and practical challenges related to use of these data for social research, highlighting coding literacy as one of the primary barriers for effective research.

49. Boyd D, Crawford K: **Six provocations for big data**. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*; *Oxford Internet Institute: 2011:1-17*.

50. Cesare N, Lee H, McCormick T, Spiro: **Self-presentation and information disclosure on twitter: understanding patterns along demographic lines**. *Population Association of America Annual Meeting*. San Diego, CA: University of Washington; 2015, 1-19.

51. Yardi S, Boyd D: **Dynamic debates: an analysis of group polarization over time on Twitter**. *Bull Sci Technol Soc* September 2010, **30**:316-327.

52. Hochman N, Manovich L: **Zooming into an instagram city: reading the local through social media**. *First Monday* 2013, **18**.

53. Syed-Abdul S, Fernandez-Luque L, Jian W-S, Li Y-C, Crain S, Hsu M-H, Wang Y-C, Khandregzen D, Chuluunbaatar E, Anh Nguyen P *et al.*: **Misleading health-related information promoted through video-based social media: anorexia on youtube**. *J Med Internet Res* 2013, **15**:e30.