Methodological Considerations in Analyzing Twitter Data

Article *in* JNCI Monographs · December 2013

DOI: 10.1093/jncimonographs/lgt026 · Source: PubMed

6 authors, including:

Annice E Kim
RTI International
47 PUBLICATIONS **1,792** CITATIONS

SEE PROFILE

DOI:10.1093/jncimonographs/lgt026

Methodological Considerations in Analyzing Twitter Data

Annice E. Kim, Heather M. Hansen, Joe Murphy, Ashley K. Richards, Jennifer Duke, Jane A. Allen

Correspondence to: Annice Kim, PhD, MPH, RTI International, 3040 Cornwallis Rd, PO Box 12194, Research Triangle Park, NC 27709 (e-mail: akim@rti.org).

Twitter is an online microblogging tool that disseminates more than 400 million messages per day, including vast amounts of health information. Twitter represents an important data source for the cancer prevention and control community. This paper introduces investigators in cancer research to the logistics of Twitter analysis. It explores methodological challenges in extracting and analyzing Twitter data, including characteristics and representativeness of data; data sources, access, and cost; sampling approaches; data management and cleaning; standardizing metrics; and analysis. We briefly describe the key issues and provide examples from the literature and our studies using Twitter data to understand public health issues. For investigators considering Twitter-based cancer research, we recommend assessing whether research questions can be answered appropriately using Twitter, choosing search terms carefully to optimize precision and recall, using respected vendors that can provide access to the full Twitter data stream if possible, standardizing metrics to account for growth in the Twitter population over time, considering crowdsourcing for analysis of Twitter content, and documenting and publishing all methodological decisions to further the evidence base.

J Natl Cancer Inst Monogr 2013;47:140–146

Twitter is a microblogging tool that is growing in popularity worldwide. Registered users can publish an unlimited number of 140-character posts, or "tweets," which are by default visible to the public, including nonregistered users. As of March 2013, more than 200 million active users are creating more than 400 million tweets each day (1). The sheer volume of tweets being produced is staggering. The US Library of Congress—which began archiving new and historical tweets in 2011—has acquired approximately 170 billion tweets and 133 terabytes of data as of January 2013 (2,3).

Twitter is emerging as an important channel for communicating about cancer. Studies show that Twitter is used by cancer patients to receive and give psychological support (4,5) and share personal narratives of cancer, from diagnosis through survivorship (6). Individuals share personal experiences related to cancer screenings, such as Pap smears and mammograms (7). Public health organizations use Twitter to promote smoking prevention (8) and cessation (9,10), and oncologists use Twitter to share research findings and discuss treatment options (11). Because this body of tweets is publicly available, Twitter data have become a rich source of cancer information, which can be used to identify needs of patients and survivors, deliver and evaluate cancer prevention campaigns and interventions, and advance cancer research and treatment.

Although analysis of Twitter data represents an important opportunity for the cancer prevention and control community, it presents investigators with a number of unique challenges. Because this area of research is so new, there are few standardized metrics or research methodologies (12). Tools used to conduct Twitter research are evolving, but they vary in cost and functionality, and commercial vendors provide limited information about proprietary products. The size and character of the Twitter population is fluid. Finally, the massive volume of user-generated content can be overwhelming in terms of data management and analysis.

This paper introduces cancer prevention and control investigators to key methodological challenges in conducting Twitter research. We begin with a brief overview of the population of Twitter users, then explore the following issues: data sources, attributes, and cost; sampling approaches; data management and cleaning; and analysis. For each topic, we discuss the key issues and provide examples from the literature and our Twitter work related to cancer prevention and other public health topics, described elsewhere (13–17). We conclude with recommendations for cancer prevention and control investigators conducting Twitter research. Supplemental files on the characteristics of Twitter data and the time frame of analysis are provided (Supplementary Material, available online).

Population of Twitter Users

The Pew Research Center regularly publishes demographic characteristics of Twitter users in the United States (18). In 2012, among adult Internet users (80% of the US population), 15% used Twitter and 8% used Twitter on a typical day. Twitter use is fairly well distributed across gender, income, and education levels (18). Twitter users tend to be younger (26% of Internet users aged 18–29 are on Twitter, compared with 14% of those aged 30–49) and more racially diverse (28% of African American Internet users use Twitter, compared with 14% of Hispanic and 12% of white Internet users) than the overall population of Internet users (18). Given the known and unknown systematic differences between those who do and do not use Twitter, investigators must be cautious when generalizing results to any particular population.

Data Sources, Attributes, and Cost

Twitter data can be obtained from a variety of sources, which influence the representativeness of the study sample and the data processing

and/or cleaning needs and costs (Table 1). Twitter Search and Twitter application programming interface (API) are provided directly via Twitter. "Twitter Search" refers to the search function available on the Twitter Web site. "Twitter API" refers to a more automated approach of data retrieval through the API. However, these sources provide only a small sample of Twitter data. Twitter licenses its full data stream to commercial vendors that provide clean Twitter data and basic automated analysis in real time via automated dashboards (19). The full data stream is also provided in raw format by Twitter data resellers, allowing users to customize data retrieval, storage, and analysis. Benefits and drawbacks to these data sources are discussed below.

The fastest and least expensive approach is to conduct manual searches using keywords through Twitter Search (twitter.com/search). Twitter Search can retrieve up to 7 days of historical data or 1500 tweets. This method requires investigators to manually copy and paste search results into a database, which may be cumbersome for a study examining a long time period.

Twitter API can be used to automate the data retrieval process. However, the number of tweets retrieved through Twitter API is capped at approximately 1% of all tweets, with no assurance of a random or representative sample. Twitter API is free, but there are costs associated with monitoring the API for issues such as being disconnected from the stream due to slow processing. Additionally, data from Twitter API are in JavaScript Object Notation format and may contain unexpected or missing fields or duplicate records, adding costs for data cleaning. Data are collected prospectively, so historical data cannot be obtained using this approach.

Automated dashboards, available through commercial vendors, can make data retrieval, preparation, and basic analysis somewhat easier. Investigators can purchase cleaned data from Twitter's full data stream (also called "firehose" data), including historical data. Vendors provide easy-to-use dashboards with built-in visualization tools (eg, word cloud, figures) that analyze data in real time. Along with tweet content and metadata, dashboards may include tweet "sentiment" (content coded as positive, negative, or neutral), the Klout score of the individual who produced the tweet (a measure of social media activity), and other social media data (eg, public Facebook posts). Costs are driven by volume and type of data retrieved. One major drawback is that, in most cases, users cannot customize the algorithms used to generate metrics such as sentiment.

Twitter data resellers are the most expensive option in terms of Twitter data retrieval and preparation. These companies provide the full raw Twitter data stream so investigators can customize the programming infrastructure and computational algorithms needed for analysis. Although this approach gives investigators complete control of how to retrieve, store, and analyze the full sample of Twitter data, it is infrequently used given the extensive cost and resources needed to build such a system.

Sampling Approaches

Essentially all Twitter studies begin the sampling process by using search terms and/or hashtags to identify relevant data. Investigators should give careful thought to their search terms to avoid under- or overestimating the volume of discussion or obscuring patterns of interest. Qualitative review of tweet content is helpful in determining whether search terms are identifying relevant content. Some

Table 1. Characteristics of Twitter data obtained from common sources*

Characteristic	Data sources			
	Twitter Search	Twitter API	Automated dashboard vendors	Twitter data reseller
Data provider and/or examples of vendors	Twitter	Twitter	Radian6 Salesforce, Crimson Hexagon	Gnip, DataSift
Study population and/or base sample	Up to 1500 of past week tweets	Up to 1% of the population of tweets	100% population of tweets	100% population of tweets
Availability of historical data	No	No	Yes (availability varies by vendor)	Yes
Setup and data retrieval	Manual	Semiautomated	Automated	Manual
Data cleaned	Not clean	Not clean	Clean	Not clean
Data storage	By user	By user	Vendor	By user
Cost	Free†	Free†	Cost varies by volume of data	Costly

* API = application programming interface.

† There are no costs involved in acquiring the data from Twitter Search or Twitter API, but there are resource costs to setting up API, retrieving data, and processing data for analysis. Disclaimer: The vendors highlighted in this paper are for case study purposes only and should not be considered an endorsement.

terms may generate a high number of false positives, thereby overestimating conversation volume. Bosley et al. (20) found that only 25% of the tweets generated by their initial keyword search were related to their topic. In our analysis of the #TFF hashtag for the Tobacco Free Florida media campaign, we found that only 1% of the 3104 tweets with #TFF hashtags were related to the campaign. Omitting a particular search term can bias the sample, especially if a term was more prevalent during a subperiod of analysis. In our study of tweets about health-care reform from 2009 to 2011, had we not included the search term "Obamacare," a substantial portion of tweets would have been missed (Figure 1).

Ultimately, compiling the list of search terms must be an iterative process. Investigators who are evaluating a campaign or intervention can use campaign-generated messages as a starting point. Investigators who are examining trends in public discourse can begin by identifying a specific event or news story related to their topic, to assess how it is being discussed publicly. However, because of the dynamic nature of user-generated content, investigators should consider what slang, abbreviations, and hashtags may be related to the topic of interest. The search query can generate a substantial volume of data, ranging from hundreds of thousands to more than 1 million tweets. Researchers cannot analyze all data retrieved. To date, investigators have primarily used random sampling to generate study samples (15,20–22), but they have also tailored the data reduction to their research questions and study designs. For example, Chew et al. (22) created a study sample of flu-related tweets by taking a random sample of 25 posts from every hour of every fourth Monday in the study period. We sampled the tweets with the most followers to examine the key influencers who may have shaped public opinion about mammography guidelines (14). This is akin to sampling newspapers with the largest circulation. Powell et al. (23) tested four sampling methods: a simple random sample of tweets, simple random sample among those in the top quartile in terms of followers, probability proportionate to size based on number of followers, and simple random sample after removal of retweets. The simple random and probability proportionate to size samples had ratios of positive to negative tweets that most closely represented the full coded data. Ultimately, the sampling strategy used should be driven by the investigators' research question (eg, Are everyone's opinions equally important? Should more weight be given to those with larger networks?) and available resources. Researchers should test different approaches and document the sampling process used.

Data Management and Cleaning

A major challenge associated with Twitter research is managing the volume of data. Twitter data sets can include millions of cases; many traditional software packages used by public health researchers cannot handle this volume. As a result, it may be necessary to export Twitter data in batches and painstakingly assemble a working data set. For example, in our analysis of health-care reform tweets, we identified a total sample of approximately 1.5 million tweets. Radian6 export files were limited to 5000 cases, resulting in 300 csv export files, which were imported into SPSS Text Analytics software as 78 separate files. Other software is better equipped to manage large volumes of data (eg, statistical software R, open source Apache Hadoop for processing, text analysis with Python), but the associated cost and learning curve may be a deterrent for investigators conducting a one-time study. Building a quality Twitter data set requires time and/or financial resources for data cleaning. One key source of noise in Twitter data is spam, which has been increasing over time—accounting for as much as 10% of all tweets in 2009 (24). Twitter has implemented strategies to combat spam (25), but researchers have also begun developing algorithms to automatically detect spam accounts or tweets. For example, Choudhury and Breslin (26) removed as duplicates or spam any messages with only

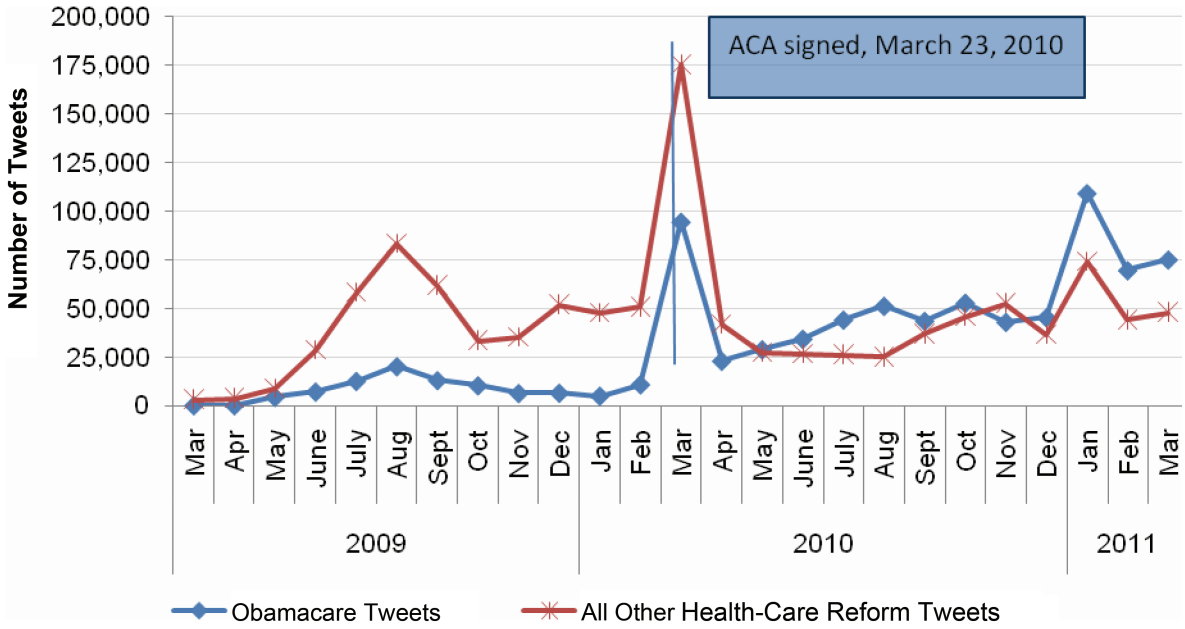

Figure 1. Comparison of health-care reform (HCR) tweets using the search term "Obamacare" vs all other HCR search terms, March 2009–March 2011. Examples of other health-care reform search terms include "health-care reform" and "Patient Protection and Affordable Care Act." ACA = Affordable Care Act.

hashtags as content, tweets with very similar content and identical time stamps, and multiple tweets with identical content from a single author. Further research is needed to determine the rules for validating the authenticity of tweets.

Analysis

Analyzing Trends and Peaks in Twitter Activity

The most common Twitter analysis consists of examining trends and peaks in discussion activity about a given topic over time. This analytic approach is particularly suited to understanding trends in public discourse, although it may be possible to detect a spike in activity around a campaign or intervention. Although researchers often point out that increasing trends may be due to the growth of both tweet volume and Twitter users over time, most do not adjust for this in their analysis. This is problematic for data interpretation.

We address this issue by adjusting tweet and Twitter author counts to reflect growth over time (13,15,27). Based on Twitter's published statistics, we estimated the number of daily and monthly Twitter users and tweet volume assuming a linear increase over time. We then calculated the volume of relevant tweets and tweet authors as a proportion of total estimated tweet volume and Twitter users monthly (Figure 2). The unadjusted figure suggests that discussions about health-care reform peaked in March 2010, whereas the adjusted figure shows that discussion peaked in August 2009. In another analysis (27), when we adjusted for the volume of tweets about flu over time, the trends in tweet volume more accurately matched patterns in Google searches about flu (Figure 3). Adjusting for Twitter growth can help account for the inherent bias in reporting raw counts over time and may provide more comparable data when comparing Twitter with other data sources.

Analyzing Twitter Content

Tweet content can be analyzed to understand conversation topics, characteristics of individuals or organizations tweeting, and public beliefs and opinions about a specific topic. Analysis of tweet content is hampered by the brevity of tweets and by the use of slang, sarcasm, and unconventional forms of written expression, including hashtags, emoticons, and acronyms. For this reason, manual coding of tweets by trained data collectors with high interrater reliability is ideal. However, the large volume of Twitter data makes manual coding cost prohibitive.

An alternative is to use computer software to automate this process. Researchers in fields such as computational linguistics use natural language processing, machine learning, and predictive analytics to mine large volumes of text data for patterns and meaning. However, the steep learning curve associated with this approach may deter some investigators.

A third approach is to purchase coded data from the growing number of commercial vendors that use computer-based algorithms to code tweet content in real time. A wide range of commercial vendors (eg, Clarabridge, Lexalytics) and social media–monitoring tools (eg, Radian6) offer automated content analysis of opinion expressed in a tweet (ie, positive, negative, and neutral), but most systems use proprietary algorithms that are not customizable. Most automated solutions have been optimized for coding opinions about brands and products, which likely do not reflect how people talk about health.

In our study of health-care reform tweets, we compared the results of manual coding of tweets by two trained coders ("gold standard") in relation to automated content analysis provided by a social media–monitoring vendor and found that the accuracy of the automated methods was poor (Table 2). Rates of agreement were 17.7% for positive content and 3.3% for negative content. Agreement for neutral content was high (91.6%), largely because the majority of the tweets were classified as neutral. Others have found that the accuracy of automated coding was lower (ranging from 7% to 48% accuracy) when neutral units were removed (28).

An alternative strategy that addresses cost and quality concerns for tweet coding is crowdsourcing, whereby human workers are leveraged to manage large-scale data analysis (29,30). Crowdsourcing involves outsourcing discrete tasks to a large network of people who work at their own convenience and are paid for the work they complete (31). In the study mentioned above, we had the same set of tweets coded via crowdsourcing using the CrowdFlower platform. Workers were given our codebook of definitions used during the manual coding process but were not formally trained. We found that the overall level of agreement between the trained coders and the crowdsourced coders was high (89.7%) and substantially better for positive (82.4%) and negative (100%) content than using the automated approach. CrowdFlower may have been able to achieve high accuracy rates as a result of their methods: all tweets were coded by three coders; the scores were averaged; and coders who performed poorly on test tweets administered throughout the coding process were dropped. These results suggest that crowdsourcing may be a good alternative to automated content analysis of tweets because humans are able to process linguistically nuanced text more efficiently than computer-based systems that need extensive training. It also suggests that crowdsourcing could be used to analyze tweet content beyond positive and negative tone, such as topics discussed (eg, prevention vs treatment) or source type (eg, health organization or news media). We summarize these three approaches in terms of cost, speed, and accuracy in Table 3.

Conclusion and Recommendations

This paper presents key issues for investigators considering Twitter research. The fact that Twitter research is now being published by peer-reviewed health and social science journals, apart from proceedings from conferences that intersect computer science, linguistics, and technology (32), indicates that Twitter has emerged as a respected data source. Our recommendations for investigators who are considering research with Twitter data are as follows:

1. Determine whether your research question can be appropriately answered using Twitter data, given the benefits and limitations of Twitter data.
2. Convene an investigative team with an understanding of Twitter data; textual analysis; natural language processing and computational linguistics; predictive analytics and data mining; and, if using API, programming and database management.
3. Consider your search terms carefully, and use other sources to assess whether they best represent the concept you wish to study.
4. Consider the time frame that will best enable you to answer your research question before selecting your sample (Supplementary Material, available online).

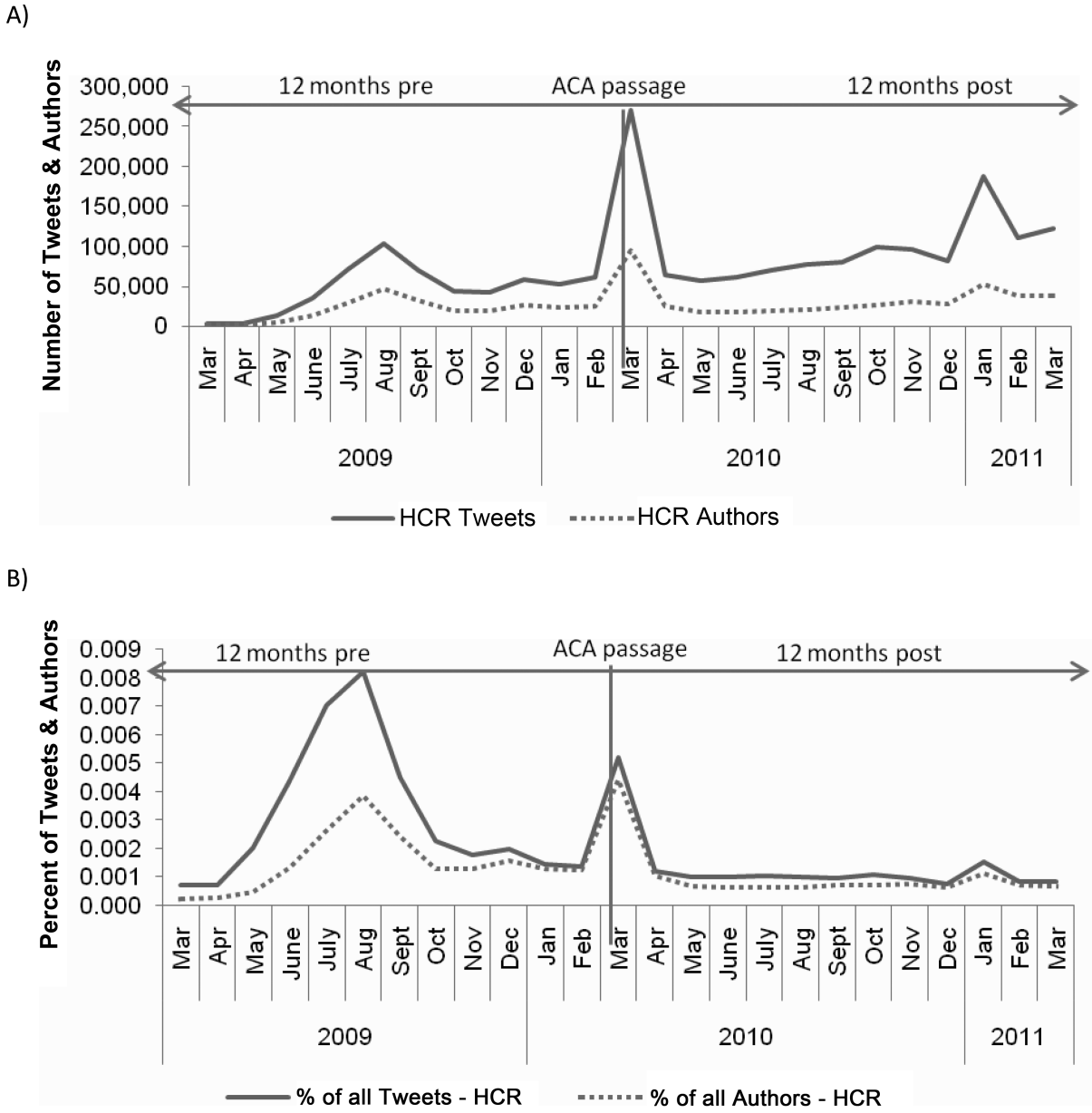

Figure 2. Volume of health-care reform (HCR) tweets and Twitter users tweeting about HCR, (**A**) unadjusted vs (**B**) adjusted, March 2009–March 2011. ACA = Affordable Care Act. Figure reproduced from Kim et al. (15) with permission of John Wiley & Sons, Inc.

5. To the degree that funding allows, use respected vendors that can provide access to the full Twitter data stream. Budget calculations should include costs associated with cleaning freely obtained data. Prepare to spend more time, energy, and money on data monitoring, management, and cleaning than you would on survey data.
6. Standardize your metrics to account for the changes in the Twitter population, to improve the precision of your estimates.
7. If working with a large volume of data that are challenging to code because of nuances in language, rare occurrences, or topic complexity, consider crowdsourcing.
8. Document and publish all of your decisions regarding methodologies—including information on vendors and tools used; the process of identifying search terms and the terms used; data sampling and characteristics; data management, cleaning, and adjustment; and coding used in content analysis—so other investigators can learn from your experience, and we can develop standard methods for Twitter analysis.

Analysis of Twitter data represents an important opportunity for the cancer prevention and control community. Each time a cancer patient, oncologist, public health professional, investigator, or other individual uses Twitter to share cancer information, he or she contributes to a growing body of publicly available cancer data. We hope this work will accelerate the pace of Twitter-based cancer research and help advance our understanding of public knowledge and perceptions about cancer, in addition to increasing our awareness of how to improve cancer treatment and support for cancer patients and survivors.

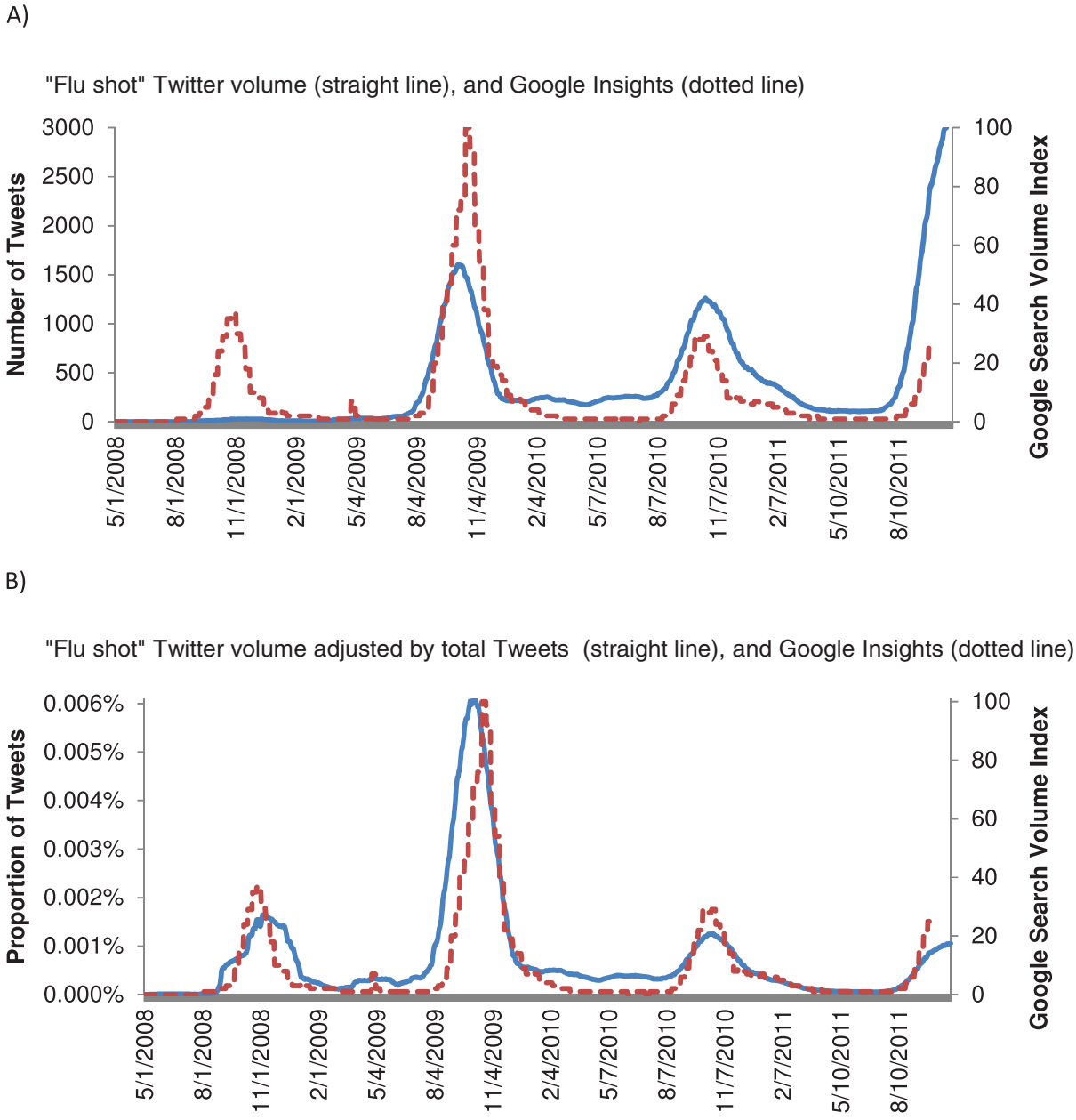

Figure 3. Volume of tweets and Google searches about flu shot. **A**) Twitter unadjusted, **B**) Twitter adjusted, April 2008–August 2011.

Table 2. Agreement of automated coding and crowdsourced coding with manual coding

		RTI (manual)		
		Positive (n = 34)	**Neutral (n = 24)**	**Negative (n = 30)**
Radian6 (automated)	Positive	17.7%	8.3%	3.3%
	Neutral	76.5%	91.7%	93.3%
	Negative	5.9%	0.00%	3.3%
	Total	100.0%	100.0%	100.0%
CrowdFlower (crowdsourced)	Positive	82.4%	4.2%	0.0%
	Neutral	17.7%	87.5%	0.0%
	Negative	0.0%	8.3%	100.0%
	Total	100.0%	100.0%	100.0%

Table 3. Characteristics of automated, crowdsourced, and manual coding of Twitter content*

	Characteristics		
	Automated	**Crowdsourced**	**Manual**
Cost	Lowest cost: included in price of obtaining data from social media–monitoring tools	Minimal cost	Highest cost: cost of professional staff labor
Speed	Fastest: near instant and independent of data set size	Midrange: dependent on data set size	Slowest: dependent on data set size and staff availability
Accuracy	Least accurate	Moderately accurate	Most accurate

* Social media–monitoring tools offer limited coding of tweet content (eg, opinion expressed as positive, negative, and neutral), whereas crowdsourced and manual coding options give researchers more control over what to code. Machine learning and text mining approaches are not characterized here.

References

1. Wickre K. Celebrating #Twitter7. Twitter Blog Web site. https://blog.twitter.com/2013/celebrating-twitter7. Accessed July 20, 2013.
2. Kessler S. Inside the Library of Congress's mission to organize 170 billion tweets. Fast Company Web site. http://www.fastcompany.com/3004480/inside-library-congresss-mission-organize-170-billion-tweets. Accessed April 25, 2013.
3. Allen E. Update on the Twitter archive at the Library of Congress. Library of Congress Web site. http://blogs.loc.gov/loc/2013/01/update-on-the-twitter-archive-at-the-library-of-congress/. Accessed July 20, 2013.
4. De la Torre-Diez I, Diaz-Pernas FJ, Anton-Rodriguez M. A content analysis of chronic diseases social groups on Facebook and Twitter. *Telemed J E Health*. 2012;18(6):404–408.
5. Sugawara Y, Narimatsu H, Hozawa A, Shao L, Otani K, Fukao A. Cancer patients on Twitter: a novel patient community on social media. *BMC Res Notes*. 2012;5:699.
6. Keim-Malpass J, Steeves RH. Talking with death at a diner: young women's online narratives of cancer. *Oncol Nurs Forum*. 2012;39(4):373–8, 406.
7. Lyles CR, Lopez A, Pasick R, 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(1):127–133.
8. Truth campaign Web site. http://www.thetruth.com/. Accessed April 25, 2013.
9. Tobacco Free Florida Web site. http://www.tobaccofreeflorida.com/. Accessed April 25, 2013.
10. Tips from former smokers. Centers for Disease Control and Prevention Web site. http://www.cdc.gov/tobacco/campaign/tips/. Accessed April 25, 2013.
11. Chaudhry A, Glodé LM, Gillman M, Miller RS. Trends in twitter use by physicians at the American Society of Clinical Oncology annual meeting, 2010 and 2011. *J Oncol Pract*. 2012;8(3):173–178.
12. Bruns A, Stieglitz S. Towards more systematic Twitter analysis: metrics for tweeting activities. *Int J Soc Res Methodol*. 2013;16(2):91–108.
13. Murphy JJ, Kim A, Hansen HM, et al. Twitter feeds and Google search query surveillance: can they supplement survey data collection. Paper presented at Association for Survey Computing Sixth International Conference; September 22, 2011; Bristol, IL.
14. Squiers LB, Holden DJ, Dolina SE, Kim AE, Bann CM, Renaud JM. The public's response to the U.S. Preventive Services Task Force's 2009 recommendations on mammography screening. *Am J Prev Med*. 2011;40(5):497–504.
15. Kim AE, Murphy J, Richards A, et al. Can tweets replace polls? Case study comparing tweets about U.S. healthcare reform to public opinion poll data. In: Hill C, Murphy J, Dean E, eds. *Social Media, Sociality, and Survey Research*. Hoboken, NJ: Wiley; 2013.
16. Kim A, Hansen HM, Murphy JJ. Methodological considerations in analyzing Twitter data. Paper presented at American Association for Public Opinion Research Annual Conference; May 18, 2012; Orlando, FL.
17. Kim A, Richards AK, Murphy JJ, et al. Can automated sentiment analysis of Twitter data replace human coding? Paper presented at American Association for Public Opinion Research Annual Conference; May 18, 2012; Orlando, FL.
18. Smith A, Brenner J. Twitter use 2012. Pew Internet & American Life Project Web site. http://pewinternet.org/Reports/2012/Twitter-Use-2012.aspx. Published May 31, 2012. Accessed December 10, 2013.
19. Williams D. Twitter certified products: tools for businesses. Twitter Blog Web site. https://blog.twitter.com/2012/twitter-certified-products-tools-businesses. Accessed October 16, 2013.
20. Bosley JC, Zhao NW, Hill S, et al. Decoding Twitter: surveillance and trends for cardiac arrest and resuscitation communication. *Resuscitation*. 2013;84(2):206–212.
21. 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.
22. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of tweets during the 2009 H1N1 outbreak. *PLoS One*. 2010;5(11):e14118.
23. Powell RJ, Kim AE, Richards A, et al. Quantity versus quality: the impact of sampling tweets on health care reform opinions. Paper presented at Conference of the Midwest Association for Public Opinion Research; November 16–17, 2012; Chicago, IL.
24. Chowdhury A. State of Twitter spam. Twitter Blog Web site. http://blog.twitter.com/2010/03/state-of-twitter-spam.html. Accessed November 1, 2012.
25. Twitter. Shutting down spammers. Twitter Blog Web site. http://blog.twitter.com/2012/04/shutting-down-spammers.html. Accessed November 1, 2012.
26. Sovinova H, Sadilek P, Csemy L. Development of Smoking Prevalence in the Adult Population of the Czech Republic 1997–2011 [in Czech]. Prague, Czech Republic: Institute of Public Health; 2012.
27. Murphy J. Relatively normal? adjusting tweet volume to account for Twitter's rise in popularity. RTI International Web site. https://blogs.rti.org/surveypost/2012/04/27/relatively-normal-adjusting-tweet-volume-to-account-for-twitters-rise-in-popularity/. Accessed June 1, 2012.
28. Rhodes M. The problem with automated sentiment. Fresh Minds Web site. http://www.freshnetworks.com/blog/2010/05/the-problem-with-automated-sentiment-analysis/. Accessed November 15, 2012.
29. Kamel Boulos MN, Resch B, Crowley DN, et al. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: trends, OGC standards and application examples. *Int J Health Geogr*. 2011;10:67.
30. Behrend TS, Sharek DJ, Meade AW, Wiebe EN. The viability of crowdsourcing for survey research. *Behav Res Methods*. 2011;43(3):800–813.
31. Parvanta C, Roth Y, Keller H. Crowdsourcing 101: a few basics to make you the leader of the pack. *Health Promot Pract*. 2013;14(2):163–167.
32. AAAI Publications. AAAI Web site. http://www.aaai.org/ocs/index.php/index/index/index/index. Accessed October 16, 2013.

Funding

This study was supported by internal research and development funds at RTI International.

Note

The authors are grateful to Susan Murchie, for editing this manuscript, and two anonymous reviewers, whose suggestions greatly strengthened this manuscript.

Affiliation of authors: Public Health Policy Research Program, Survey Research Division Program on Digital Technology and Society RTI International, Research Triangle Park, NC (AEK, HMH, AKR, JD, JAA) and Chicago, IL (JM).