JAMA Cardiology | Brief Report

Twitter as a Potential Data Source for Cardiovascular Disease Research

Lauren Sinnenberg, BA; Christie L. DiSilvestro, BA; Christina Mancheno, BA; Karl Dailey, BA; Christopher Tufts, MS; Alison M. Buttenheim, PhD, MBA; Fran Barg, PhD, MEd; Lyle Ungar, PhD; H. Schwartz, PhD; Dana Brown, BA; David A. Asch, MD, MBA; Raina M. Merchant, MD, MSHP

IMPORTANCE As society is increasingly becoming more networked, researchers are beginning to explore how social media can be used to study person-to-person communication about health and health care use. Twitter is an online messaging platform used by more than 300 million people who have generated several billion Tweets, yet little work has focused on the potential applications of these data for studying public attitudes and behaviors associated with cardiovascular health.

OBJECTIVE To describe the volume and content of Tweets associated with cardiovascular disease as well as the characteristics of Twitter users.

DESIGN, SETTING, AND PARTICIPANTS We used Twitter to access a random sample of approximately 10 billion English-language Tweets originating from US counties from July 23, 2009, to February 5, 2015, associated with cardiovascular disease. We characterized each Tweet relative to estimated user demographics. A random subset of 2500 Tweets was hand-coded for content and modifiers.

MAIN OUTCOMES AND MEASURES The volume of Tweets about cardiovascular disease and the content of these Tweets.

RESULTS Of 550 338 Tweets associated with cardiovascular disease, the terms *diabetes* (n = 239 989) and *myocardial infarction* (n = 269 907) were used more frequently than *heart failure* (n = 9414). Users who Tweeted about cardiovascular disease were more likely to be older than the general population of Twitter users (mean age, 28.7 vs 25.4 years; $P < .01$) and less likely to be male (59 082 of 124 896 [47.3%] vs 8433 of 17 270 [48.8%]; $P < .01$). Most Tweets (2338 of 2500 [93.5%]) were associated with a health topic; common themes of Tweets included risk factors (1048 of 2500 [41.9%]), awareness (585 of 2500 [23.4%]), and management (541 of 2500 [21.6%]) of cardiovascular disease.

CONCLUSIONS AND RELEVANCE Twitter offers promise for studying public communication about cardiovascular disease.

Editor's Note page 1036

Supplemental content at jamacardiology.com

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Raina M. Merchant, MD, MSHP, Penn Medicine Social Media and Health Innovation Lab, University of Pennsylvania, 423 Guardian Dr, Philadelphia, PA 19104 (raina.merchant@uphs.upenn.edu).

JAMA Cardiol. 2016;1(9):1032-1036. doi:10.1001/jamacardio.2016.3029
Published online September 28, 2016.

Person-to-person communication is one of the most persuasive ways people deliver and receive information.[1,2] Until recently, this communication was impossible to collect and study. Now, social media networks allow researchers to systematically witness public communication about health, including cardiovascular disease. Twitter, one such network, is used by more than 300 million people who have generated several billion Tweets.[3,4]

There are several unknowns when using social media for research on cardiovascular disease. Is it possible to separate signal from noise? Can the data be analyzed to characterize features associated with the person posting and the Tweet itself? Does the Twitter data set reflect real-time changes in conversation? We explored these questions by characterizing a sample of Tweets about cardiovascular disease from the United States.

Key Points

Question Can Twitter, a social media platform for person-to-person communication, be used as a data source to study cardiovascular disease?

Findings In this descriptive study, we identified 4.9 million Tweets about cardiovascular disease posted on Twitter. User demographics, as well as content (eg, risk factors, awareness, and treatment) and volume of Tweets, varied across cardiovascular diseases.

Meaning Twitter has potential as a data source for studying public communication about cardiovascular health.

Methods

This was an exploratory mixed-methods study of Twitter data associated with cardiovascular disease.[5] This study was approved by the University of Pennsylvania Institutional Review Board.

Data Source

Twitter is a social media platform that allows users to send and receive 140-character messages known as Tweets. Our data from July 23, 2009, through February 5, 2015, was made up of the "Twitter decahose," a 10% sample of Tweets (covering 52 months of posts), and the "Twitter spritzer," a 1% sample of Tweets (covering the other 15 months).

From this group of Tweets, we searched for keywords associated with the following 5 cardiovascular diseases: hypertension, diabetes, myocardial infarction, heart failure, and cardiac arrest. To generate a set of search terms, we used the Consumer Health Vocabulary,[6] the Unified Medical Language System,[7] and the consensus of the study authors. The following keywords were identified from these sources: *diabetes* (*blood glucose* and *mellitus*), *heart attack* (*coronary attack*, *cardiac infarction*, *myocardial infarction*, *heart infarction*, *myocardial infarct*, and *myocardial necrosis*), *cardiac arrest* (*asystolic*, *asystole*, *cardiac arrest*, *heart arrest*, *ventricular fibrillation*, and *pulseless electrical activity*), *heart failure* (*cardiac failure*), and *hypertension* (*high blood pressure*). To ensure that Tweets with these keywords were in English, we applied an English-language classifier to the sample.

Tweet Location

Reported coordinates were used to identify Tweets that could be mapped to a county in the United States.[8] For Tweets without coordinates but with location information, locations reflecting city or county plus state were mapped. Tweets that could not be mapped to a US county by this process were eliminated.

Twitter Users

To characterize Twitter users, we collected information from each user's account, including the number of friends and followers. Additional data about Twitter users can be estimated based on their behavior on the platform. By applying established language-based algorithms based on users with known demographics to the Tweets in our sample, we imputed the age and sex of each user[9]; these data were compared with a random sample of Twitter users.

Tweet Content

To describe the content of Tweets, 2 of us (C.L.D. and C.M.) used NVivo (QSR International) to code 500 Tweets for each of the 5 cardiovascular diseases, adjudicating differences with a larger group of authors (F.B., D.B., and R.M.M.). After coding the set of 2500 Tweets, total agreement for each category was greater than 90% and the mean κ was 0.77.

Tweet Rate

We measured the number of Tweets per topic per day. To account for variability in the baseline Tweet count, we identified 3 peaks in Twitter posts for each US-based disease topic against 7-day running means. Two of us (L.E.S. and C.T.) then identified the triggers for the peaks by identifying the common theme in the Tweets for that day.

Statistical Analysis

We used χ^2 test to compare the sex of individuals Tweeting about cardiovascular disease with the sex of the general population of Twitter users. Paired *t* tests were used to compare the ages of individuals Tweeting about cardiovascular disease with those of the general population of Twitter users.

Results

Tweet Volume and Rate

From an initial sample of 10 billion Tweets, we identified 4.9 million with terms associated with cardiovascular disease; 550 338 were in English and originated from a US county (eFigure 1 in the Supplement). Diabetes and myocardial infarction represented more than 200 000 Tweets each, while the topic of heart failure returned fewer than 10 000 Tweets (**Table 1**). Similar findings were noted when analyzing data from a sample

Table 1. Tweet Data and User Data Geotagged to US County

Characteristic	Tweets by Terms Used, Value[a]					
	All Tweets	Diabetes	Hypertension	Myocardial Infarction	Cardiac Arrest	Heart Failure
Tweet data						
Total English-language Tweets geotagged in the United States, No.	550 338	239 989	23 459	269 907	12 238	9414
Total retweets by topic, No. (%)	132 721 (24.1)	53 302 (22.2)	4865 (20.7)	69 204 (25.6)	3376 (27.6)	1974 (21)
User data						
Distinct users, No.	364 406	121 494	18 072	233 168	10 852	7822
Estimated						
Male, No. (%)	59 082/124 986 (47.3)	28 480/60 961 (46.7)	5873/13 245 (44.3)	17 473/34 863 (50.1)	4198/9196 (45.7)	3057/6631 (46.1)
Mean age, y	28.7	29.1	29.5	28.2	26.8	30.4
Distinct users, No., mean (SD) [maximum]						
Overall Tweets	17 776 (31 432) [1 656 982]	16 501 (35 996) [1 656 982]	22 598 (44 921) [994 549]	19 947 (32 421) [1 399 982]	26 515 (48 583) [1 128 894]	25 022 (57 728) [1 656 982]
Followers	2933 (78 780) [21 189 904]	3633 (73 950) [14 903 807]	3242 (35 625) [2 737 196]	3000 (86 073) [21 189 904]	4597 (90 876) [7 919 031]	4575 (35 580) [1 528 879]
User friends	942 (6348) [1 074 133]	1135 (7022) [965 509]	1322 (7848) [461 204]	913 (6180) [1 053 316]	1131 (6203) [279 019]	1751 (15 093) [1 074 133]
Potential reach (mean followers × total statuses), No.	1 614 141 354	871 880 037	76 054 078	809 721 000	36 714 000	43 069 050

[a] Data are presented as number (percentage) unless otherwise indicated.

Table 2. Semantic Content of Tweets

Characteristic	No. (%)						Examples
	Total (n = 2500)	Diabetes Type 1 and Type 2 (n = 500)	HTN (n = 500)	MI (n = 500)	CA (n = 500)	HF (n = 500)	
Risk factor	1048 (41.9)	97 (19.4)	184 (36.8)	290 (58)	273 (54.6)	204 (40.8)	"High Blood Pressure Common Among Overweight Kids" "Air #pollution tied to high blood pressure in #pregnancy" "Chronic Heart Failure: Iron deficiency was found to be associated with a 58% increased risk"
Awareness	585 (23.4)	122 (24.4)	245 (49)	22 (4.4)	61 (12.2)	135 (27)	"October is Sudden Cardiac Arrest Month. How can you protect yourself and your loved ones?" "Next week I ride 100km to raise funds for @diabetesql. Pls help by donating." "Walking for a cure for diabetes!"
Treatment and management	541 (21.6)	116 (23.2)	215 (43)	11 (2.2)	63 (12.6)	136 (27.2)	"The DASH diet can help lower high blood pressure" "Big Breakfast May Be Best for Diabetes Patients." "Exercise 'just as good as drugs' for treating heart failure and stroke."
Mechanism	357 (14.3)	42 (8.4)	182 (36.4)	39 (7.8)	50 (10)	44 (8.8)	"Hormone Misfires in #Obese People - #diabetes" "Fat Fighters helps your body balance blood glucose levels. Great for diabetics." "Sudden cardiac arrest occurs when electrical impulses in the heart become rapid or chaotic, causing the heart to suddenly stop beating."
Outcomes	247 (9.9)	25 (5)	44 (8.8)	28 (5.6)	73 (14.6)	77 (15.4)	"High Blood Pressure. Silent Killer." "What You Don't Know Will Kill You...The Shocking Truth About Sudden Cardiac Arrest" "High Blood Pressure Linked to Declining Brain Function"
Symptoms	121 (4.8)	43 (8.6)	11 (2.2)	31 (6.2)	18 (3.6)	18 (3.6)	"One of the symptoms of a heart attack is left arm hurting" "Take a urine test if you have a frequent urge to go. This could be a sign of diabetes." "Very pale nails can be suggestive of anemia, congestive heart failure, liver disease or malnutrition."
Prevention	103 (4.1)	47 (9.4)	27 (5.4)	10 (2)	9 (1.8)	10 (2)	"Working out for just 30 min a day, 5 days a week may help protect your body against diabetes" "Best Group of Foods for #Cardiovascular #Health and Why... Prevents heart attacks, heart failure and strokes" "Walk for health! 3 hrs of walking a week can reduce risk of heart attack by 35%."
Support	50 (2)	30 (6)	4 (0.8)	3 (0.6)	9 (1.8)	4 (0.8)	"Thx @tomhanks for coming out about your type 2 #diabetes. Need MORE courageous celebrities like you." "Explaining Tweetchats - Our Diabetes #ourD" "Tell us how you win over diabetes every day"

Abbreviations: CA, cardiac arrest; DASH, dietary approaches to stop hypertension; HF, heart failure; HTN, hypertension; MI, myocardial infarction; Thx, thanks.

of Tweets geocoded to the United States (eTable in the Supplement). Peaks in Tweet rate were associated most often with thematically connected events reported in the news (eFigure 2 in the Supplement).

Twitter Users

Those tweeting about cardiovascular disease tended to be older than the general population of Twitter users (mean age, 28.7 vs 25.4 years; $P < .01$); mean age and sex varied across the different cardiovascular conditions (Table 1). Users tweeting about cardiovascular disease were less likely to be male compared with the general population of Twitter users (59 082 of 124 896 [47.3%] vs 8433 of 17 270 [48.8%]; $P < .01$).

Tweet Content

Tweet content varied across and within cardiovascular disease terms (**Table 2**). Most of the hand-coded Tweets in our sample (2338 of 2500 [93.5%]) included health-related information. The most commonly represented theme was risk factors for cardiovascular disease (1048 of 2500 [41.9%]) (Table 2). Approximately one-fourth of all Tweets (585 of 2500 [23.4%]) discussed awareness, frequently in the setting of fundraising for disease. Many Tweets (541 of 2500 [21.6%]) discussed the treatment and management of cardiovascular disease, often focusing on topics such as diet and exercise. Of Tweets that discussed outcomes of cardiovascular disease (247 of 2500 [9.9%]), most (193 [78.1%]) mentioned death.

Tweet Modifiers

Tweets could be characterized by tone, style, and perspective. Tweets associated with cardiovascular disease often used metaphor (1106 of 2500 [44.2%]), emotional language with positive or negative sentiment (974 of 2500 [39%]), and first-person accounts (872 of 2500 [34.9%]) (**Table 3**). Three percent of Tweets included a statement that the individual posting the Tweet identified as having cardiovascular disease.

Table 3. Tweet Modifiers

Modifier	All Tweets, No. (%) (n = 2500)	Examples
User		
First person	872 (34.9)	"I'm fine. It's my dad. Cardiac arrest." "The nurse practitioner told me rice will give me diabetes #rude" "I might have a heart attack this is too intense"
Self-reported diagnosis	80 (3.2)	"I wanna wake up one day and say I used to have diabetes" "I went from battling heart failure in 2010 to being able to jog" "So I was diagnosed with stroke-level high blood pressure and tomorrow I have to go do a treadmill test"
Intent		
News	418 (16.7)	"Meta-Analysis Finds Potassium to Prevent Strokes, Heart Attacks, and High Blood Pressure" "Denmark cardiac arrest survival triples after teaching the nation CPR" "NIH funds trio to build tools that predict heart failure"
Advertisement	131 (5.2)	"Nitroxyl (HNO): a Novel Approach for the Acute Treatment of Heart Failure" "Two-Med Combo to Prevent Diabetes" "Effective High Blood Pressure Home Remedy in East Kingston"
Humor	126 (5)	"Just stole some candy from a baby. Because I care about preventing juvenile diabetes." "I had high blood pressure for 4 y. My doctor said it was the result of being a Wizards fan." "The worst time to have a heart attack is during a game of charades"
Rhetoric		
Metaphor	1106 (44.2)	"Yikes, a heart attack waiting to happen. How to Make Your Own Cheesy Mac Attack Burger" "Stop the silent killer, lower your high blood pressure naturally" "Simple tips on how to battle diabetes"
Sentiment	974 (39)	"NOTHING EVER WORKS FOR PPL WITH DIABETES" "This guy..Trying to cause me a heart attack I swear. I'm over here going crazy." "High blood pressure. Paralyzed. Ah!!!"

Abbreviations: CPR, cardiopulmonary resuscitation; NIH, National Institutes of Health; PPL, people.

Discussion

This study has 3 main findings. First, we identified a large volume of US-based Tweets about cardiovascular disease. Second, we were able to characterize the volume, content, style, and sender of these Tweets, demonstrating the ability to identify signal from noise. Third, we found that the data available on Twitter reflect real-time changes in discussion of a disease topic.

We were able to identify 4.9 million Tweets associated with cardiovascular disease. Of the hand-coded sample, 94% included some form of information associated with health rather than a colloquial but non-health-associated use of the term. Prior work has suggested, however, that the language of Tweets, regardless of whether they arise from patients or other members of the community, can provide insight into the health behaviors of communities that are known to influence risk of disease.[10] We also observed that Twitter users respond to events, such as World Diabetes Day or celebrity deaths, within minutes to hours and that these peaks in discussion are easily identifiable in the Twitter data set.

This study has several limitations. Our study focused on Tweets relevant to 5 specific cardiovascular conditions. Broader terms such as *heart disease*, specific terms such as *sudden cardiac death*, and slang terms such as *DM2* or *diabeetus* may have captured other themes associated with these diagnoses. This study characterized only US-based English-language Tweets. We did not characterize the impression for each Tweet or the identities of those who received each Tweet. Hand coding was performed to read the text of Tweets and infer content, purpose, and sentiment. The true nature or intent of the user could not be verified.

Conclusions

Twitter may be useful for studying public communication about cardiovascular disease. The use of Twitter for clinical research is still in its infancy. Its value and direct applications remain to be seen and warrant further exploration.

ARTICLE INFORMATION

Accepted for Publication: July 13, 2016.

Published Online: September 28, 2016. doi:10.1001/jamacardio.2016.3029

Author Affiliations: Penn Medicine Social Media and Health Innovation Lab, University of Pennsylvania, Philadelphia (Sinnenberg, DiSilvestro, Mancheno, Dailey, Tufts, Ungar, Brown, Asch, Merchant); Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia (Sinnenberg, DiSilvestro, Mancheno, Dailey, Tufts, Brown, Merchant); Department of Family and Community Health, University of Pennsylvania School of Nursing, Philadelphia (Buttenheim); Center for Health Incentives and Behavioral Economics, University of Pennsylvania, Philadelphia (Buttenheim); Department of Family Medicine, University of Pennsylvania, Philadelphia (Barg); Department of Anthropology, University of Pennsylvania, Philadelphia (Barg); Positive Psychology Center, University of Pennsylvania, Philadelphia (Ungar, Schwartz); Computer and Information Science, University of Pennsylvania, Philadelphia (Ungar); Center for Health Equity Research and Promotion, Philadelphia Veterans Affairs Medical Center, Philadelphia, Pennsylvania (Asch).

Author Contributions: Dr Merchant and Ms Sinnenberg had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.
Study concept and design: Sinnenberg, DiSilvestro, Dailey, Buttenheim, Barg, Ungar, Schwartz, Brown, Asch, Merchant.
Acquisition, analysis, or interpretation of data: Sinnenberg, DiSilvestro, Mancheno, Dailey, Tufts, Ungar, Schwartz, Brown, Merchant.
Drafting of the manuscript: Sinnenberg, DiSilvestro, Mancheno, Dailey, Tufts, Barg, Merchant.
Critical revision of the manuscript for important intellectual content: Sinnenberg, DiSilvestro, Dailey, Tufts, Buttenheim, Ungar, Schwartz, Brown, Asch, Merchant.
Statistical analysis: Sinnenberg, DiSilvestro, Mancheno, Dailey, Tufts, Ungar, Schwartz, Merchant.
Obtaining funding: Ungar, Merchant.
Administrative, technical, or material support: Dailey, Barg, Schwartz, Brown, Merchant.
Study supervision: Mancheno, Buttenheim, Barg, Brown, Merchant.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Drs Buttenheim, Barg, Schwartz, and Asch; Mss Sinnenberg, DiSilvestro, and Mancheno; and Mr Dailey are employed by the US Government. No other conflicts were reported.

Funding/Support: This study was supported by grant R01-HL1422457 from the National Heart, Lung, and Blood Institute, Templeton Religious Trust (Dr Ungar), and grants K23 109083 and R01 122457 from the National Institutes of Health (Dr Merchant).

Role of Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Cialdini RB, Demaine LJ, Sagarin BJ, Barrett DW, Rhoads K, Winter PL. Managing social norms for persuasive impact. *Soc Influ*. 2006;1(1):3-15. doi:10.1080/15534510500181459

2. Tellis GJ, Ambler T, eds. *The SAGE Handbook of Advertising*. Thousand Oaks, CA: SAGE Publications; 2007.

3. Company @ About.Twitter.com. https://about.twitter.com/company. Accessed August 22, 2016.

4. Twitter usage statistics. http://www.internetlivestats.com/twitter-statistics/. Accessed February 22, 2016.

5. Creswell JW. Mixed methods procedures. In: *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*. 3rd ed. Thousand Oaks, CA: SAGE Publications; 2009:203-226.

6. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. *J Am Med Inform Assoc*. 2006;13(1):24-29.

7. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med*. 1993;32(4):281-291.

8. Schwartz HA, Eichstaedt JC, Kern ML, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. *PLoS One*. 2013;8(9):e73791.

9. Sap M, Park G, Eichstaedt J, et al. Developing age and gender predictive lexica over social media. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.672.9851. Accessed February 19, 2016.

10. Eichstaedt JC, Schwartz HA, Kern ML, et al. Psychological language on Twitter predicts county-level heart disease mortality. *Psychol Sci*. 2015;26(2):159-169.

Editor's Note

Twitter and Cardiovascular Disease
Useful Chirps or Noisy Chatter?

Mintu P. Turakhia, MD, MAS; Robert A. Harrington, MD

As modern society continues to organize around a digital, connected way of life, information from our daily interactions and exposures are now measured, recorded, and memorialized in ways previously unimaginable. This tapestry of information includes data from social media or electronic tools, such as websites and applications, that enable users to create, share, and exchange content.[1] Twitter is one such social networking service whose 310 million active users post short public messages known as Tweets.

Related article page 1032

In this issue of *JAMA Cardiology*, Sinnenberg and colleagues[2] explore the characteristics of Twitter users and Tweets associated with cardiovascular disease. They found a large volume of Tweets (4.9 million) on cardiovascular disease and were able to characterize tone, style, and perspective of these Tweets, as well as some basic demographics of the users posting them. Most notably, Sinnenberg and colleagues found that Tweet volume and content were temporally associated with news events that were thematically connected with cardiovascular disease.

This brief report differs from much of the original investigation in *JAMA Cardiology*. We accepted it because it highlights the potential for using these emerging data sources such as Twitter for cardiovascular research, in this case to evaluate public communication about cardiovascular medicine in a manner not previously possible on such a scale. Furthermore, application programming interfaces allow persons with basic coding skills to mine these data as well as data from other social media platforms, which are often publicly accessible, thereby adding to the mix of open data and potentially engaging investigators and data scientists outside the traditional venues of cardiovascular research. Other uses of social media in areas related to clinical care or research are rapidly being ex-