Published in final edited form as:
JAMA Cardiol. 2016 December 01; 1(9): 1032–1036. doi:10.1001/jamacardio.2016.3029.

Twitter as a potential data source for cardiovascular disease research

Lauren E. Sinnenberg[1,2], **Christie L. DiSilvestro**[1,2], **Christina Mancheno**[1,2], **Karl Dailey**[1,2], **Christopher Tufts, MS**[1,2], **Alison M. Buttenheim, PhD,MBA**[3,4], **Fran Barg, PhD Med**[5,6], **Lyle Ungar, PhD**[1,7,8], **H Schwartz, PhD**[7], **Dana Brown**[1,2], **David A. Asch, MD, MBA**[1,9], and **Raina M. Merchant, MD, MSHP**[1,2]

[1] Penn Medicine Social Media and Health Innovation Lab, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[2] Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3] Department of Family and Community Health, University of Pennsylvania School of Nursing, Philadelphia, Pennsylvania, USA

[4] Center for Health Incentives and Behavioral Economics, University of Pennsylvania, Philadelphia, PA, USA

[5] Department of Family Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[6] Department of Anthropology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[7] Positive Psychology Center, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[8] Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[9] Center for Health Equity Research and Promotion - Philadelphia Veterans Affairs Medical Center, Philadelphia, Pennsylvania, USA

Abstract

Importance—Twitter has recently been explored as a data source by researchers in multiple fields, though, as a whole, the research remains nascent. Even less is known about using Twitter to study cardiovascular disease.

Objective—We sought to characterize the volume and content of tweets related to cardiovascular disease, and the characteristics of Twitter users.

Corresponding Author Raina M. Merchant, MD MSHP, Department of Emergency Medicine, Penn Medicine Social Media and Health Innovation Lab, 423 Guardian Drive, University of Pennsylvania, Philadelphia, PA, USA, Ph: 215 746-8681, Fax: 215 746-1224, Raina.merchant@uphs.upenn.edu.

Disclosures
Sinnenberg, DiSilvestro, Mancheno, Dailey, Buttenheim, Barg, Schwartz, Asch: none
Ungar: research funding: Templeton Religious Trust
Merchant: research funding NIH K23 109083, NIH R01 122457

Design—We used the Twitter Application Programming Interface (API) to access a random sample of tweets from July 2009 to February 2015. We filtered approximately 10 billion tweets for keywords related to cardiovascular disease. We included only English tweets originating from US counties. We characterized each tweet relative to estimated user demographics. A random subset of 2,500 tweets was hand-coded for tweet content and modifiers.

Setting—Twitter, a social media platform

Participants—Twitter users tweeting about cardiovascular disease

Exposures—None

Main outcomes and measures—Our main outcomes included the volume of tweets about cardiovascular disease on Twitter and the content of these tweets.

Results—Diabetes (n=239,989) and myocardial infarction (269,907) terms were used more frequently than heart failure (9,414) terms. Users tweeting about cardiovascular disease were more likely to be older than the average Twitter population (mean age= 28.68 vs. 25.36, p<0.01), and less likely to be male (47% vs. 49%, p <0.01). Most tweets (94%) were health related. Common themes included tweets related to risk factors (42%), awareness (23%) and management (23%).

Conclusions and relevance—Twitter offers promise to characterize public understanding and communication about heart disease.

INTRODUCTION

Person-to-person communication is one of the most persuasive ways people deliver and receive information[1,2]ntil 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.

METHODS

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

Data source

Twitter is a social media platform which allows users to send and receive 140-character messages or “tweets.” Our tweet data, spanning July 2009 through February 2015, was made up of the "Twitter Decahose", a 10% sample (covering 52 months) and the "Twitter Spritzer", a 1% sample (covering the other 15 months).

From this group of tweets, we searched for keywords related to five cardiovascular diseases: hypertension, diabetes, myocardial infarction, heart failure, and cardiac arrest. To generate a set of search terms, we used the Consumer Health Vocabulary (CHV)[6], the Unified Medical Language System (UMLS)[7], and the consensus of the study authors. Keywords identified from these sources were as follows: diabetes (blood sugar, mellitus), heart attack (coronary attack, cardiac infarction, myocardial infarction, heart infarction, myocardial infarct, myocardial necrosis); cardiac arrest (asystolic, asystole, cardiac arrest, heart arrest, ventricular fibrillation, pulseless electrical activity), heart failure (cardiac failure), and hypertension (high blood pressure). To ensure tweets with these keywords were in the English language, 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 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 user

To characterize Twitter users, we collected information from the Twitter API including number of friends and followers. Additional data about the Twitter user can be estimated based on their Twitter behavior. Applying established language-based algorithms trained on users with known demographics to the Twitter user tweets in our sample, we imputed the age and gender of each user.[9] These gender and age data were compared to a random sample of Twitter users.

Tweet content

To describe the content of tweets, two authors (CLD, CM) used NVivo to code 500 tweets for each of the 5 cardiovascular diseases, adjudicating differences with a larger group of authors (DB, FB, RMM). After coding the set of 2500 Tweets, total agreement for each category was greater than 90% and the average kappa was 0.77.

Tweet rate

We measured the number of tweets per topic per day. To account for variability in baseline tweet count, we identified three peaks in Twitter posts for each US-based disease topic against 7-day running averages. Two authors (CT, LS) then identified the triggers for the peaks by identifying the common theme in the tweets for that day.

Statistical analysis

Chi-squared test were used to compare gender and paired t-test was used to compare the age of Twitter users tweeting about cardiovascular disease to the general Twitter population.

RESULTS

Tweet volume and rate

From an initial sample of 10 billion tweets, we identified 4.9 million tweets with cardiovascular disease related terms; 550,338 were in English and originating from a US county (Supplement 1). Diabetes and myocardial infarction represented over 200,000 tweets while the heart failure disease topic returned fewer than 10,000 tweets (Table 1). Similar findings were noted when analyzing data from a sample of tweets geocoded to the US (Supplement 2). Peaks in tweet rate were most often associated with thematically connected news events (Supplement 3).

Twitter users

Those tweeting about heart disease tended to be older than the average Twitter user (28 vs. 25, $p< 0.01$) and mean age and sex varied across the different cardiovascular conditions (Table 1). Users tweeting about heart disease were less likely to be male when compared to the average Twitter user (47% vs. 49% male, $p <0.01$).

Tweet content

Tweet content varied across and within cardiovascular disease terms (Table 2). The majority (94%) of hand-coded Tweets in our sample included health-related information. The most commonly represented theme was risk factors (42%) (Table 2). Approximately a quarter of all tweets (23%) discussed awareness, frequently in the setting of fundraising for disease. Many tweets, (22%) discussed treatment and management of disease, often focusing on topics like diet and exercise. Of tweets that discussed outcomes from cardiovascular disease (10%), the majority (78%) mentioned death.

Tweet modifiers

Tweets could be characterized by tone, style, and perspective of the tweet. Tweets related to cardiovascular disease often used metaphor (44%), emotional language with positive or negative sentiment (39%) and first person accounts (34%). Three percent of included self-identification of having cardiovascular disease (Table 3).

DISCUSSION

This study has three main findings. First, we identified a large volume of US-based tweets about heart disease on Twitter. 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 reflects real-time changes in discussion of a disease topic.

We were able to identify approximately 5 million heart disease tweets in this study. Of the hand-coded sample, 94% were related in some manner to health rather than a colloquial but non health-related 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 observed that Twitter users respond to events like 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 five specific cardiovascular conditions. Broader terms like “heart disease”, specific terms like “sudden cardiac death,” and slang terms like “DM2” or “diabeetus,” may have captured other themes related to these diagnoses. This study characterized only US-based English-language tweets. We did not characterize the impression for each tweet, nor the identities of those who received each tweet. Hand coding was used to read the content of tweets and infer content, purpose, and sentiment. The true nature or intent of the user could not be verified.

Conclusion

Twitter, as a nascent data source for research, has potential to help characterize public understanding and communication about heart disease.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Supported by R01-HL1422457 from the National Heart, Lung, and Blood Institute.

The funding organization did not play a role in design and conduct of the study; collection, management, analysis and interpretation of the data; preparation, review or approval of the manuscript; or decision to submit the manuscript for publication. Raina Merchant and Lauren 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.

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 G, Ambler T. The SAGE Handbook of Advertising. 2007
3. Company @ About.Twitter.Com. https://about.twitter.com/company.
4. Twitter Usage Statistics. http://www.internetlivestats.com/twitter-statistics/. Accessed February 22, 2016.
5. Creswell J. Mixed methods procedures. Res Des Qual Quant Mix methods approaches. 2009:203–226.
6. Zeng QT, Tse T. Exploring and developing consumer health vocabularies. J Am Med Informatics Assoc. 2006; 13(1):24–29. doi:10.1197/jamia.M1761.
7. Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993; 32(4):281–291. http://www.ncbi.nlm.nih.gov/pubmed/8412823. Accessed October 6, 2016. [PubMed: 8412823]
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. doi:10.1371/journal.pone.0073791. [PubMed: 24086296]

9. Sap, M.; Park, G.; Eichstaedt, J.; Kern, M. Developing age and gender predictive lexica over social media. 2014. 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. doi: 10.1177/0956797614557867. [PubMed: 25605707]

Table 1

Characteristics of the study sample: tweet data and user data geotagged to United States county

	All	Diabetes terms	Hypertension terms	Myocardial infarction terms	Cardiac Arrest terms	Heart Failure terms
Tweet data						
Total English tweets geotagged in US	550,338	239,989	23,459	269,907	12,238	9,414
Total retweets by topic	24%	22 %	21%	26%	28%	21%
User data						
Distinct users	364,406	121,494	18,072	233,168	10,852	7,822
Estimated gender (male)	47%	47%	44%	50%	46%	46%
Estimated mean age	28	29	30	28	27	30
Distinct users overall tweets mean+/− stdev (max)	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)
Distinct users followers mean+/− stdev (max)	2,933+/−78,780 (21,189,904)	3,633+/−73,950 (14,903,807)	3,242+/−35,625 (2,737,196)	3,000+/−86,073 (21,189,904)	4,597+/−90,876 (7,919,031)	4,575+/−35,580 (1,528,879)
Distinct user friends mean+/− stdev (max)	942+/−6,348 (1,074,133)	1,135+/−7,022 (965,509)	1,322+/−7,848 (461,204)	913+/−6,180 (1,053,316)	1,131+/−6,203 (279,019)	1,751+/−15,093 (1,074,133)
Potential reach* (Average followers × total statuses)	1,614,141,354	871,880,037	76,054,078	809,721,000	36,714,000	43,069,050

Table 2

Tweet semantic content

		DM n=500	HTN n=500	MI n=500	CA n=500	HF n=500	Examples
Risk Factor	42%	19%	37%	58%	55%	41%	• 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	23%	24%	49%	4%	12%	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/ management	22%	23%	43%	2%	13%	27%	• 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	14%	8%	36%	8%	10%	9%	• Hormone Misfires in #Obese People - #diabetes • Fat Fighters helps your body balance blood sugar 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	10%	5%	9%	6%	15%	15%	• 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	9%	9%	2%	6%	4%	4%	• 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	4%	9%	5%	2%	2%	2%	• Working out for just 30 minutes a day, five days a week may help protect your body against diabetes

		DM n=500	HTN n=500	MI n=500	CA n=500	HF n=500	Examples
							• Best Group of Foods for #Cardiovascular #Health and Why... Prevents heart attacks, heart failure and strokes • Walk for health! 3hrs of walking a week can reduce risk of heart attack by 35%
Support	2%	6%	1%	1%	2%	1%	• 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

DM, diabetes mellitus; HTN, hypertension; MI, myocardial infarction; CA, cardiac arrest; HF, heart failure.

Table 3

Tweet modifiers

Modifiers	All tweets (n=2500)	Examples
User		
First person	35%	• 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	3%	• 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	17%	• 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	5%	• 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	5%	• Just stole some candy from a baby. Because I care about preventing juvenile diabetes. • I had high blood pressure for 4 years. 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	44%	• 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	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!!!