Linking social media and medical record data: a study of adults presenting to an academic, urban emergency department

Kevin A Padrez,[1,2] Lyle Ungar,[2,3,4] Hansen Andrew Schwartz,[2,3,4] Robert J Smith,[2,4,5] Shawndra Hill,[2,6] Tadas Antanavicius,[2,4] Dana M Brown,[2,5] Patrick Crutchley,[2,4] David A Asch,[2,7] Raina M Merchant[2,5]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/bmjqs-2015-004489).

For numbered affiliations see end of article.

Correspondence to
Raina M Merchant, Department of Emergency Medicine, Penn Medicine Social Media and Health Innovation Lab University of Pennsylvania, Philadelphia, PA, USA;
Raina.merchant@uphs.upenn.edu

Received 14 June 2015
Revised 14 September 2015
Accepted 16 September 2015

To cite: Padrez KA, Ungar L, Schwartz HA, *et al*. *BMJ Qual Saf* Published Online First: [*please include* Day Month Year] doi:10.1136/bmjqs-2015-004489

ABSTRACT

Background Social media may offer insight into the relationship between an individual's health and their everyday life, as well as attitudes towards health and the perceived quality of healthcare services.

Objective To determine the acceptability to patients and potential utility to researchers of a database linking patients' social media content with their electronic medical record (EMR) data.

Methods Adult Facebook/Twitter users who presented to an emergency department were queried about their willingness to share their social media data and EMR data with health researchers for the purpose of building a databank for research purposes. Shared posts were searched for select terms about health and healthcare.

Results Of the 5256 patients approached, 2717 (52%) were Facebook and/or Twitter users. 1432 (53%) of those patients agreed to participate in the study. Of these participants, 1008 (71%) consented to share their social media data for the purposes of comparing it with their EMR. Social media data consisted of 1 395 720 posts/tweets to Facebook and Twitter. Participants sharing social media data were slightly younger (29.1±9.8 vs 31.9±10.4 years old; p<0.001), more likely to post at least once a day (42% vs 29%; p=0.003) and more likely to present to the emergency room via self-arrival mode and have private insurance. Of Facebook posts, 7.5% (95% CI 4.8% to 10.2%) were related to health. Individuals with a given diagnosis in their EMR were significantly more likely to use terms related to that diagnosis on Facebook than patients without that diagnosis in their EMR (p<0.0008).

Conclusions Many patients are willing to share and link their social media data with EMR data. Sharing patients have several demographic and clinical differences compared with non-sharers. A database that merges social media with EMR data has the potential to provide insights about individuals' health and health outcomes.

INTRODUCTION

Facebook and Twitter combined have >1.5 billion users worldwide, and many of these users share information daily.[1 2] Information in the form of posts, photos, 'likes', 'pins', location-specific check-ins and other 'digital breadcrumbs' of social media can provide a detailed narrative about an individual's day-to-day activities and behaviours.[3–6] Prior work has demonstrated that data from social media can be used in a variety of contexts for health-related research such as tracking cholera outbreaks, predicting heart disease mortality rates and identifying individuals with postpartum depression.[7–11] When a person with asthma posts about wheezing or a patient tweets about receiving a flu shot, there is an assumed connection with health. Other measures of health may be revealed more subtly, not just through content but word patterns. In this context, quantitative changes related to a person's baseline profile could be related to long-term adjustments in health or health behaviour, such as a variation in word complexity suggesting cognitive decline or a change in the number of words per post or network size suggesting a depressed mental status.[11–15] Together these data may offer insight into the health behaviours of social media users and their illnesses, as well as their attitudes about

health, perceived quality of healthcare services and needs for healthcare information.

Despite the potential uses of these data, much of social media content is neither publically available for research nor linked with electronic medical records (EMRs), making it difficult to use or validate reported information at an individual level. One result is that most health-related analyses of social media are passive and may not include consenting patients about accessing their data.[16]

Little is known about individual and social acceptability of sharing social media data for purposes of individual or public health, or medical research, and yet building a 'digital health databank' demonstrates many parallels with the current construct of DNA banking. While initially met with scepticism from the general public, consent rates for donating blood for the purpose of conducting genetic research reached upward of 80%, albeit with significant demographic variation.[17–19] Given that health is determined at least as much by social endowments and exposures as genetic endowments and biomedical exposures, associations between social media information and health might offer considerably leverage for prevention and cure or at least information that is complementary to our genetic makeup.

We examined the acceptability of building a research databank that links consenting patients' social media content with their EMR data. To determine whether such a social media databank contained enough information to be useful from a research perspective, we also examined the quantity of language data available for analysis and the quality of specific health-related content.

METHODS

Study design

This was a study of patients seeking care in an urban academic adult emergency department (ED) from 28 March 2014 to 31 October 2014. During selected shifts, balanced to reflect different days of the week, two or three research assistants (RAs) approached patients as they were brought into the ED from the triage area. Patients were excluded if they were <18 years of age, in police custody, evaluated in the trauma resuscitation bay, intoxicated, in respiratory distress or with other evidence of severe illness including intubation or sedation, altered mental status or haemodynamic instability requiring acute intervention. Non-English-speaking patients were also excluded as to enable language analysis of social media posts. Patients were verbally asked if they used Facebook or Twitter. Those who indicated yes were asked if they would be willing to participate in a study about social media and health. They were informed that the study involved sharing with researchers all information from their Facebook and/or Twitter accounts and data from their EMR. Patients who indicated that they did not have a Facebook or Twitter account were not queried further. Patients who were unwilling to share their social media data ('non-sharers') and EMR data were verbally asked to consent to complete a brief survey about their demographics and reasons for not sharing.

Patients willing to share their Facebook and/or Twitter data for analysis with their EMR data ('sharers') underwent a two-step consent process. First, a written informed consent form was read and reviewed by an RA with the patient. This document explained the research study and the type of information that would be extracted from their Facebook profile and EMR. A copy of this document was given to all participants. Facebook sharers were then directed to a link using a tablet computer that allowed subjects to log in to their Facebook account and add a plug-in. Before adding this plug-in, an agreement screen appeared that again listed the precise information that was going to be shared and therefore accessible to researchers. To participate, subjects were required to click a checkbox and enter the study identification number provided by the RA. Once installed, this plug-in application appeared in the sidebar of a sharer's Facebook profile listed under 'Apps' but did not affect use of Facebook in any way. Participants received verbal and written instructions about how to remove the application from their profile at any time or unenroll from the study.

If sharing Twitter, sharers underwent the same written informed consent and shared their Twitter username or 'handle'. Participants sharing social media information were entered in to a raffle to win an iPad. This was described to participants as an odds of winning of 1 in 2500.

Patients willing to share their social media data also completed a brief survey about their demographics (age, race, gender) and frequency of social media use (number of posts per day or per week) (see online supplementary appendix for survey instrument). The survey instrument for both sharers and non-sharers was designed by the research team and piloted using an initial sample of 48 sharers and 29 non-sharers. Due to the large variation in responses from the pilot, an 'other' response was added that enabled participants to add free text regarding their reason for not sharing social media with researchers.

Data extracted from participants ED visit included weight/height, triage class, mode of arrival to ED, discharge diagnoses, insurance type, medical history, disposition, prior ED visits and prior hospital admissions. Data extracted from participants' longitudinal EMR included all International Classification of Diseases, ninth revision (ICD)-9 diagnosis codes and diagnosis dates for each patient from outpatient visits, emergency room visits, hospitalisations, as well as demographic information, lab results and medications.

Extracting data from social media platforms

The Facebook plug-in application designed by this research team and used previously was created using

Facebook's public API; it allowed for extraction and secure storage of past and future content including age, gender, language, 'News Feed' content, work history, education history, hometown, interests, relationship status, current city, religious or political views, items 'liked' and 'status updates'.[20] The plug-in was created for this research study using Facebook's public API. The application itself is not available for purchase or distribution and does not serve any purpose besides data extraction for research purposes. The purpose and intent of the app was verbally explained to participants and provided in written form. Once the app was installed by the user, all posts were retroactively extracted from January 2009 until the date of enrolment.

For Twitter sharers, data were extracted using the Twitter handles provided by participants and included tweet content, number of tweets, number of followers, number following and geographic data. The purpose of collecting both Twitter and Facebook data from our patients was to maximise the quantity of social media data gathered. Acknowledging that significant differences exist between Twitter and Facebook and the type of information users may share, we performed separate analyses using each type of platform. Both Facebook and Twitter sharers, however, were combined for comparing health and demographic characteristics of 'sharers' compared with 'non-sharers'.

Social media data were not linked or uploaded to patients' EMRs and likewise no EMR data were linked or made accessible on an individual's social media account. No changes or manipulations were made to individuals' EMR or social media account.

Assessing references to health in social media data

In order to complete meaningful content analysis of language data, a minimum threshold of data is necessary. To identify whether there were sufficient data within our sample, we quantified the total number of words and posts in the sample as well as the average number of words and posts for each participant.

To determine the frequency of health-related posts, we selected a random sample of 600 posts. Two annotators (RJS, TA) independently identified whether or not each of the posts was contextually related to health. Prior to independently coding the sample, the annotators met to develop consensus on the contextual meaning of health-related by reviewing 100 sample posts together. For example, the post "I have cancer and received chemo today" was coded as health-related; however, the post "Cancer is my astrological sign" was not. Inter-rater reliability across annotators was assessed using a kappa statistic (K=0.64).[21] Using the percentage of posts mentioning health in the random sample and the average number of Facebook posts per sharer, we then estimated the number of health-related posts for the average sharer. CIs were approximated from the SE calculated over 10 randomly selected equal-size subsets of the annotated messages based on a normal distribution.

In order to link and reflect the data present in the EMR with the language of social media, we searched for the top 15 most prevalent diagnoses present within the historical medical records of study participants, as determined by ICD-9 codes. We identified health terms associated with each of these ICD-9 codes. We determined the prevalence of these health terms in the Facebook language of our participants. We then compared the prevalence of these terms between participants with and without the corresponding ICD-9 codes. A data dictionary publicly available through the Research Data Distribution Center at the Medicare Provider Analysis and Review was used to identify common terms related to health systems.[22] We used these terms to determine the proportion of participants in our sample who use specific words related to health systems.[23 24]

Statistical analysis

Summary statistics were used to characterise demographics, social media use and self-reported health rating of sharers and non-sharers. A sample of patients who visited the same ED during the time of enrolment (N=19 184) was used to compare the sharing participants with the overall ED population. Differences between shares and non-sharers were assessed using χ^2 test for categorical variables and t tests for continuous variables. Two-tailed p values were calculated for the ORs using Fisher's exact test.

Reports of the proportion of the study sample with specific health and health systems terms in posts were based on automatic analysis of the entire Facebook database of posts using language processing software (http://www.wwbp.org/data.html).

The 15 most common ICD-9 diagnosis codes were identified within the longitudinal electronic record of our patient sample. Differences in the prevalence of health terms between patients with and without specific ICD-9 codes were assessed using two-sample test of proportions. Overall comparison of the prevalence of health terms on Facebook between patients with and without diagnoses was calculated using a Wilcoxon signed-rank t test.

All statistical analyses were performed using SAS V.9.1 (SAS Institute, Cary, North Carolina, USA) or the Python programming language.

RESULTS

In total, 5256 patients were approached and 2717 (52%) indicated that they used Facebook or Twitter. Of these social media users, 1433 (53%) agreed to participate in a study about social media and health (figure 1) and of those 1008 (71%) agreed to share their social media and EMR data for analysis. Among non-sharers, privacy concerns were reported as the primary reason

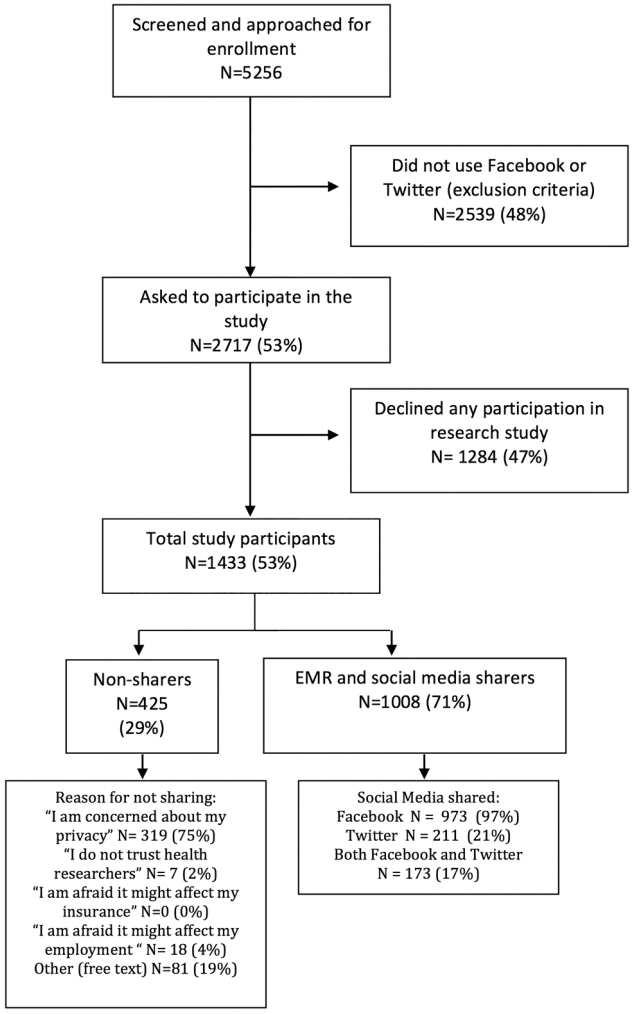

Figure 1 Enrolment flow chart.

for not sharing social media content, 319 (75%). Other reasons for electing to not share their data included fear that sharing could affect employment, 18 (4%), or 'other', 81 (19%), which included a variety of free text responses including a general lack of desire to share, belief that sharing would not be useful for research or a belief that Facebook should be reserved for 'social' or 'personal' spheres.

Characteristics of study participants

Participants sharing social media data were younger than those not sharing (29.1±9.8 vs 31.9±10.4; p<0.001). Participants across both groups were more likely to be female: sharers 719 (71%) and non-sharers, 313 (74%) (table 1). While self-reported Facebook use was similar between sharers and non-sharers, sharers were much more likely to post at least once a day, 420 (42%), compared with non-sharers, 122 (29%) (p=0.003). When compared with the sample of ED patients, the sharing participants were more likely to present to the ER via self-arrival mode, have respiratory diseases and have private insurance (table 2).

Volume of shared social media content

Among the 973 participants agreeing to share their Facebook accounts, 174 (18%) were unable to access their Facebook account during their ED visit due to either poor internet access in the ED or reported forgotten passwords. Of the 787 accessible accounts, Facebook data included 708 300 distinct posts and 99 311 058 words. The median number of Facebook posts per individual was 448 (IQR 124 to 1157), and the median number of total words was 60 400 (IQR 16 200 to 159 800). Most, 655 (83%), Facebook sharers had >1000 words, a requirement previously identified by study coauthors (HAS and LU) for automatic language content analysis of social media data.[20]

Among the 211 sharers agreeing to provide their Twitter handles, 131 (62%) of those accounts were publicly accessible for analysis. Shared Twitter social media data represented 232 200 distinct tweets and 2 363 829 words. The median number of tweets per individual was 1927 (IQR 245 to 3146) and the median number of words was 18 704 (IQR 2924 to 30 800). Most, 101 (84%), Twitter sharers had

Table 1 Characteristics and demographics of sharers and non-sharers

	ED sample (n=19 182) n (%)	Participants not sharing social media* data (n=425) n (%)	Participants sharing social media* and EMR data (n=1008) n (%)	OR†	p Value‡
Age group					
18–30	5957 (31)	232 (55)	660 (65)	1.5	<0.001
31–40	2844 (15)	102 (24)	210 (21)	0.83	0.2
41–50	2275 (12)	54 (13)	87 (9)	0.65	<0.5
51–60	2645 (14)	22 (5)	42 (4)	0.8	0.4
≥61	4394 (23)	6 (1)	8 (1)	0.56	0.4
Gender					
Female	11 003 (57)	313 (74)	719 (71)	0.86	0.3
Race/ethnicity					
White	5494 (29)	142 (33)	211 (21)	0.56	<0.0001
Black	11 752 (61)	232 (55)	691 (69)	1.9	<0.0001
Asian	484 (3)	24 (5)	21 (2)	0.39	<0.005
Hispanic	389 (2)	28 (3)	30 (3)	0.43	<0.005
Frequency of posting on social media*	Not available				
3+ times daily		56 (13)	203 (20)	1.6	<0.005
1–3 times daily		66 (16)	217 (22)	1.5	<0.01
Every few days		108 (25)	260 (26)	1	0.9
Once per week		52 (12)	121 (12)	1	0.9
Every few weeks		60 (14)	98 (10)	0.66	<0.05
Less than once a month		33 (8)	62 (6)	0.78	0.3
A few times a year		27 (6)	30 (3)	0.45	<0.01
Never		23 (5)	17 (2)	0.3	<0.001

*Social media refers to Facebook or Twitter.
†OR calculated using the group compared to non-group (eg, white vs non-white). OR <1 means this group is under-represented among sharers versus non-sharers.
‡Overall ED sample is not included in the calculation of OR or p value.
ED, emergency department; EMR, electronic medical record.

>1000 words. Given the low number of unique Twitter accounts (N=131), Twitter posts were not extracted for content or language analysis in the present study.

Use of health and healthcare terms in social media content

Based on review of status content, we estimated that 7.5% (95% CI 4.8% to 10.2%) of Facebook posts were contextually related to health. In total, 8447 Facebook posts contained one of the health terms related to common health conditions (table 3). In a random 10% sample of posts, 218 (91%) were identified as contextually pertaining to health by study team annotators.

Patients with highly prevalent diseases and symptoms, as derived from ICD-9 codes in their medical record, were more likely to use terms related to those diagnoses on Facebook than patients without those diagnoses, as shown in table 3 (p=0.0008). Among individuals diagnosed with abdominal pain (ICD=789), for example, 81 (21%) used terms such as 'stomach pain' and 'belly ache' on Facebook compared with 21 (8%) of individuals without that diagnosis who used those terms (p<0.0001). The difference in proportion of individuals containing disease-related Facebook language between patients with and without a given ICD-9 code was significant for the following diagnoses: abdominal pain, nausea/vomiting, headache, normal delivery, anaemia, asthma and throat pain. In searching for health system-related words, we also identified several commonly occurring terms (table 4). These include 'surgery' (28%), 'hospital' (55%) and 'ER' (25%).

DISCUSSION

A database linking information from an individual's social media data with validated information from their medical record has the potential to provide insights about health and healthcare. This study illustrates the potential for creating this data repository. First, we demonstrated that among social media users seeking care in an urban ED many consented to share access to their social media data (Facebook, Twitter) and EMR. Second, these participants' posts contained sufficient content to enable automated

Table 2 Health characteristics and healthcare use of subjects sharing their social media data for the study versus a sample of patients visiting same ED during study enrolment period

	ED sample (n=19 182)	Participants sharing social media and EMR data (n=1008)		
	n (%)	n (%)	OR	p Value
Triage class				
1 (most urgent)	103 (1)	2 (0)	0.36	0.2
2	5090 (27)	109 (11)	0.33	<0.0001
3	8466 (45)	551 (55)	1.47	<0.0001
4	3709 (20)	337 (33)	2.04	<0.0001
5 (least urgent)	762 (4)	0 (0)	0	<0.0001
Arrival mode to the ED				
Emergency services	3944 (20)	79 (8)	0.32	<0.0001
Self	14 708 (78)	909 (90)	2.56	<0.0001
Other	148 (1)	20 (2)	0.63	0.5
Primary insurance				
Uninsured/self-pay	4028 (21)	270 (27)	1.34	<0.0001
Private	6522 (34)	634 (63)	3.19	<0.0001
Medicaid or Medicare	7673 (40)	95 (9)	0.15	<0.0001
Other	959 (5)	5 (0)	0.09	<0.0001
ED discharge diagnoses*				
Diseases of the musculoskeletal system and connective tissue	5190 (28)	407 (40)	1.78	<0.0001
Diseases of the genitourinary system	2284 (12)	281 (28)	2.8	<0.0001
Diseases of the nervous system and sense organs	1191 (6)	266 (26)	5.3	<0.0001
Diseases of the respiratory system	1364 (7)	235 (23)	3.89	<0.0001
Injury and poisoning	4962 (26)	263 (26)	0.98	0.9
Diseases of the digestive system	1669 (9)	177 (17)	2.19	<0.0001
Diseases of the circulatory system	1273 (7)	143 (14)	2.28	<0.0001
Diseases of the skin and subcutaneous tissue	1293 (7)	104(10)	1.56	<0.0001
Complications of pregnancy, childbirth and the puerperium	301 (2)	42 (4)	2.67	<0.0001
Symptoms, signs and ill-defined conditions and factors influencing health status	13 070 (70)	502 (50)	0.44	<0.0001
Disposition				
Treated and discharged	12 994 (68)	865 (86)	2.71	<0.0001
Admitted to inpatient	3347 (18)	83 (8)	0.41	<0.0001
Left without treatment complete	190 (1)	7 (1)	0.69	0.4
Transferred	110 (1)	2 (0)	0.34	0.1
Assigned to observation	1323 (7)	45 (4)	0.62	0.001
Two or more ED visits to 1 of 3 hospitals in the EMR network†				
Yes	Not available	222 (22)		
No	Not available	770 (76)		

*Included are all listed discharge diagnoses from the ED visit, totals >100% as an individual may have up to three diagnoses across or within categories according to Clinical Classifications Software.[46]
†Administrative database for ED sample did not include ED visit history; these data were manually extracted from sample of sharers for consented to share their EMR with researchers.
ED, emergency department; EMR, electronic medical record.

characterisations of their postings using language algorithms that could support speed and scale of analysis. Third, although social media is often used for entertainment purposes and personal exchanges, this study also showed that posted content on Facebook often (1 in 13) included health terms. In evaluating the combined social media posts and EMR data of study participants, we also observed that individuals with particular health conditions identified in the EMR were more likely to post about those conditions on Facebook than individuals without those conditions identified in the EMR. Overall, these findings demonstrate the feasibility of merging large collections of data about individuals' thoughts, feelings and daily activities, found on social media, with detailed validated information in their EMR to evaluate possible predictors and insights about health and health outcomes.

Table 3 Proportion of individuals with and without specific International Classification of Diseases, ninth revision (ICD-9) diagnoses who use related health terms on Facebook

ICD-9 diagnosis code	Terms searched	n (%) of patients with diagnosis who used term	n (%) of patients without diagnosis who used term
Abdominal pain (789)	Abdominal pain, stomach pain, belly pain, tummy pain, stomach hurts, belly hurts, tummy hurts, tummyache, stomachache, bellyache	81 (21)	21 (8)*
Nausea/vomiting (787)	Nausea, vomiting, vomit, throwing up, spitting up, threw up, puke, puked, vomited	101 (29)	69 (22)*
Headache (339 and 784.0)	Headache, migraine, head hurts	141 (59)	192 (46)*
Pain in limb (729.5)	Leg hurts, arm hurts, finger hurts, toe hurts	5 (3)	5 (1)
Urinary tract infection (599.0)	UTI, urinary tract infection	1 (1)	4 (1)
Back pain (724.1, 724.2, 724.3, 724.4 and 724.5)	Back pain, backache, back hurts	29 (15)	51 (11)
Cough (786.2)	Cough, coughing, coughed	40 (26)	109 (22)
Normal delivery (V27.0)	Giving birth, gave birth	62 (33)	148 (10)*
Anaemia (280, 281, 282, 283, 284 and 285)	Anaemia	3 (2)	2 (0)*
Dizziness (780.4)	Dizzy, dizziness, vertigo	28 (22)	79 (15)
Asthma (493)	Asthma	36 (28)	36 (7)*
Acute upper respiratory tract infection (465)	Caught a cold, have a cold	7 (7)	24 (4)
Throat pain (784.1)	Sore throat, throat hurts	26 (24)	62 (11)*
Depression (311)	Depression, depressed	35 (38)	169 (30)

*p<0.05.

Social media and EMR data: consent and transparency

The large percentage of individuals using Facebook and Twitter in the present study is consistent with the widespread use of these platforms. Data from the Pew Research Center show that of adults in the USA who are online, 71% report using Facebook and 23% report using Twitter.[25] New is the observation that a substantial proportion of social media users would consent to share both their social media and health data with researchers. While both data sets contain potentially sensitive information, protections are in place for EMR data that do not exist for social media data. Social media data may, however, contain a significant amount of information that is implied or revealed about health that may not be evident to individual social media users. Prior studies of social media data have demonstrated that it can be used for identifying discussions about diseases and healthcare, tracking infectious diseases outbreaks, surveillance in public health emergencies, identifying individuals with mental illnesses, following public opinions about health policies and detecting hospital quality.[7–11 26–39] Merging these data may be helpful in determining this intersection and informing patients about what can be inferred from their digital footprint. Critically important privacy concerns must be interpreted alongside the social good that can come from this kind of health research.[40–42] Of note, this study focused on collecting data for purposes of a research database and did not involve integrating social media data into an actual EMR. The latter could involve increased responsibility for a health system to protect social media data and would require validated approaches for interpreting the merged data and a process for ensuring the security of the linked data.

Table 4 Proportion of sharers using healthcare system terms on Facebook (N=787 unique accounts)

Category	Healthcare term	n (%) of participants using the term on Facebook
Emergency care	ER	194 (25)
	Emergency room	53 (7)
Ambulatory care	Outpatient	6 (1)
	Doctor's office	81 (10)
	Clinic	45 (6)
Hospital	Hospital	430 (55)
	Surgery	219 (28)
	Intensive care	46 (6)
Nursing home	Nursing home	21 (3)
Ambulance	Ambulance	39 (5)
Dialysis	Dialysis	4 (0.5)

Marketers often rely on mining individual's data without their explicit awareness or consent for the purpose of tailoring their advertising message or managing their products.[43–45] There are also many 'Facebook-enhanced' applications or websites that connect with Facebook using a common log-in.[46] Many of these websites will also collect information

from that user's Facebook profile—information similar to this investigation, but perhaps with much less knowledge and consent of the user. Public opinion research reveals that the public fundamentally cares most about the *purpose* for which their information is used[17 47] and is much more accepting of situations in which their personal health information is used for research purposes *without* their consent than for marketing purposes *with* their consent.[42]

Contextualising the volume of social media data available for analyses

Large-scale open-vocabulary language analysis requires sufficient data for meaningful results.[20] With >11 million words extracted from social media for this study, there was sufficient volume to conduct large-scale language analysis of social media output. This project suggests that the social media and health databank could support linguistic analysis in future research. This investigation also highlights the importance of interdisciplinary collaboration for such analysis. Novel insights using this type of language database for health research will likely benefit from collaborations between patients, physicians, computer scientists, data scientists and others.

In linking validated health records with social media posts, our study also showed an increased likelihood of posting about health terms by individuals with specific diagnoses. This unique finding suggests the potential to further characterise patients' digital footprints (eg, words used, places frequented, images captured, things 'liked') that could be associated with specific diseases.

Data quality: identifying health terms and health content on social media

Our finding that many Facebook posts included health terms suggests the potential for using a social media platform to study conversations and information exchange about health. While our analysis showed that many posts using health terms were contextually related to health, next steps will require a further content analysis of the actual posts. For example, a health term may have different implications if used in reference to the individual posting (eg, "I have chest pain and worry I may be having a heart attack") or to another person ("My grandmother is in the hospital after having a heart attack") or more broadly to a population (eg, "Join me in celebrating heart attack awareness month").

Potential applications of a merged social media and EMR database

With further granularity about how health-related language is used on social media, a databank that links patient-level EMR data with social media could potentially be used to identify disease severity, onset or progression, as well as use of healthcare resources. This database could allow for predictions about changes in health status, functional status and/or health behaviours based on explicit or subtler content in postings. In this context, posted content could reveal information about adherence to prescribed medications, exercise and diets based on posted content. When compared with data from traditional sources, studying social media posts may also allow for more rapid analysis. This is clearly an area requiring further investigation, and lessons learned from using Google searches to track influenza suggest that this will require rapid data validation, replicability, transparency and precise robust algorithms for data interpretation.[34]

Much of the utility of social media data will rely on being able to accurately analyse content and understand differences in content relative to demographic variability and other factors. Interpretability will also require understanding how individuals who participate in the database differ from those who do not with regard to health status and demographics. In our study, non-sharers frequently cited privacy concerns as reasons for non-participation. Understanding these concerns and how they may be impacted by the sensitivity of the data, intended use of the data, prior experience with data sharing and data breeches and other factors will be an important focus of future work. This is an evolving area, and social media platforms and their associated features, privacy controls and safety measures will inevitably vary as individuals choose to share both more and less information about their health and day-to-day lives. Today individuals share photos, text and other data on social media. However, in the near future that type of data may change to include even more personalised information and the potential questions that can be answered from merging these data with health data may vary.

While the research database described here is being used only for retrospective analysis, there is potential for active monitoring and intervention using a similar database. This would require patient consent and an infrastructure for intervention if there is the potential for harm. Previous studies have examined the use of social media for tracking disease risk factors, such as the risk of suicide attempts in the USA.[26] At present, this type of passive monitoring of well-being is already being tested with Facebook in partnership with organisations such as the National Suicide Prevention Lifeline.[48]

Limitations

This study has several limitations. It was conducted in a single, academic, urban ED, which may limit the generalisability of the results. The acute care setting may have also affected patients' willingness or unwillingness to participate or share data due to severity of illness, patient flow through the department and other factors impacting a patient's ability to engage with an RA. Thus, participation results are unlikely to be

generalisable to other clinical settings. In addition, participants willing to share social media in the present study were younger, less acutely ill and more likely to have private insurance than the general ED population, suggesting a potential selection or response bias. We are limited in understanding the specific privacy concerns of non-sharers and whether these concerns relate to the merging of social media and EMR data, data access, intent of data usage, context for being approached about the study (ie, acute care setting), concerns about research in general or other reasons. Future work can focus on expanding to more diverse settings and expanding questions for non-sharers regarding non-participation.[49]

CONCLUSION

Over a third of adult patients using Facebook or Twitter at a single, urban ED were willing to share both their social media content and their EMR data with health researchers for the purpose of building a research databank that links digital health information with social media data. It is evident that individuals with a given medical diagnosis are more likely to use language related to that diagnosis on Facebook than individuals without that diagnosis. This finding suggests that social media is a rich and promising avenue for exploring how patients conceptualise and communicate about their specific health issues. A linked EMR-social media databank could provide valuable data about patients' day-to-day activities related to their health. Just as genetic information is banked to track potential future health, previously unobservable social media postings—made up of words, language, conversations—may also be banked from consenting individuals and evaluated for potential correlations with health and health outcomes. With patients' consent, this information can be studied for a variety of more specific clinical, public health and policy investigations.

Author affiliations
[1]Emergency Medicine, University of California, San Francisco, California, USA
[2]Penn Medicine Social Media and Health Innovation Lab, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[3]Computer and Information Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[4]Positive Psychology Center, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[5]Department of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[6]Operations and Information Management Department, University of Pennsylvania, Philadelphia, Pennsylvania, USA
[7]Center for Health Equity Research and Promotion, Philadelphia VA Medical Center, Philadelphia, Pennsylvania, USA

Twitter Raina Merchant at @RainaMerchant

Contributors KAP, RMM and RJS: design, acquisition, analysis and interpretation of the data, drafting and critical revision of the manuscript. LU, HAS and SH: design, acquisition of data, analysis and interpretation of the data and critical revision of drafted manuscript. DA: design, analysis and interpretation of data and critical revision of drafted manuscript. TA and PC: analysis and interpretation of data and critical revision of drafted manuscript. DMB: design, acquisition of data and critical revision of drafted manuscript. All authors approved the final version of the article and are accountable for all aspects of the study including the accuracy and integrity of the investigation.

Funding Leonard Davis Institute of Health Economics. RMM received funding support from NIH, R01, and (HL122457-01A1) and NIH, K23 (109083-01).

Competing interests None declared.

Ethics approval University of Pennsylvania Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Twitter company blog. 2014-09-11. https://blog.twitter.com/ (accessed 09 11 2014). (Archived by WebCite®at http://www.webcitation.org/6SVuuCTpB).
2 Facebook company info. Facebook 014-09-11. http://newsroom.fb.com/ (accessed 11 Sep 2014). (Archived by WebCite®at http://www.webcitation.org/6SVv3acn3).
3 Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet. *J Med Internet Res* 2009;11:e11.
4 Fjeldsoe BS, Marshall AL, Miller YD. Behavior change interventions delivered by mobile telephone short-message service. *Am J Prev Med* 2009;36:165–73.
5 Maher CA, Lewis LK, Ferrar K, *et al*. Are health behavior change interventions that use online social networks effective? A systematic review. *J Med Internet Res* 2014;16:e40.
6 Portnoy DB, Scott-Sheldon LA, Johnson BT, *et al*. Computer-delivered interventions for health promotion and behavioral risk reduction: a meta-analysis of 75 randomized controlled trials, 1988–2007. *Prev Med* 2008;47:3–16.
7 Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. *Am J Trop Med Hyg* 2012;86:39–45.
8 Prieto VM, Matos S, Álvarez M, *et al*. Twitter: a good place to detect health conditions. *PLoS ONE* 2014;9:e86191.
9 Park S, Lee SW, Kwak J, *et al*. Activities on Facebook reveal the depressive state of users. *J Med Internet Res* 2013;15:e217.
10 Eichstaedt JC, Schwartz HA, Kern ML, *et al*. Psychological language on Twitter predicts county-level heart disease mortality. *Psychol Sci* 2015;26:159–69.
11 De Choudhury M, Gamon M, Counts S, *et al*. Predicting depression via social media. *Paper presented at the International AAAI Conference on Weblogs and Social Media*; 27 May 2014; Ann Arbor, MI.
12 Akay A, Dragomir A, Erlandsson BE. Network-based modeling and intelligent data mining of social media for improving care. *IEEE J Biomed Health Inform* 2015;19:210–18.
13 Schwartz H, Eichstaedt J, Kern M, *et al*. Towards assessing changes in degree of depression through facebook. *Paper presented at the Conference of the Association for Computational Linguistics*; 27 June 2014; Baltimore, MD.
14 Berisha V, Wang S, LaCross A, *et al*. Tracking discourse complexity preceding Alzheimer's disease diagnosis: a case study comparing the press conferences of presidents ronald reagan and george herbert walker bush. *J Alzheimers Dis* 2015;45:959–63.

15 Snowdon DA, Kemper SJ, Mortimer JA, *et al*. Linguistic ability in early life and cognitive function and Alzheimer's disease in late life. Findings from the Nun Study. *JAMA* 1996;275:528–32.
16 Moorhead SA, Hazlett DE, Harrison L, *et al*. A new dimension of health care: Systematic review of the uses, benefits, and limitations of social media for health communication. *J Med Internet Res* 2013;15:85.
17 Grande D, Mitra N, Shah A, *et al*. Public preferences about secondary uses of electronic health information. *JAMA Intern Med* 2013;173:1798–806.
18 Sterling R, Henderson GE, Corbie-Smith G. Public willingness to participate in and public opinions about genetic variation research: a review of the literature. *Am J Public Health* 2006;96:1971–8.
19 Rice MS, Custer BS, Hindes DA, *et al*. Genetic research in the blood bank: Acceptability to northern California donors. *Transfusion* 2010;50:1951–8.
20 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:e73791.
21 Viera AJ, Garrett JM. "Understanding Interobserver Agreement: The Kappa Statistic." *Fam Med* 2005;37:360–3. Retrieved 2 February 2015. http://www.ncbi.nlm.nih.gov/pubmed/15883903
22 Medicare provider analysis and review (MEDPAR). Centers for Medicare and Medicaid Services. 09 11 2014. http://www.cms.gov/Research-Statistics-Data-and-Systems/Files-for-Order/IdentifiableDataFiles/MedicareProviderAnalysisandReviewFile.html (accessed 11 Sep 2014). (Archived by WebCite® at http://www.webcitation.org/6SVv8rJIR).
23 St Sauver JL, Warner DO, Yawn BP, *et al*. Why patients visit their doctors: assessing the most prevalent conditions in a defined American population. *Mayo Clin Proc* 2013;88:56–67.
24 Zeng QT. Consumer Health Vocabulary Initiative website. 09 11 2014. http://www.layhealthinformatics.org/ (accessed 11 Sep 2014). (Archived by WebCite® at http://www.webcitation.org/6SVvFwfKw).
25 Duggan M, Ellison NB, Lampe C, *et al*. Social Media Update 2014. 9 January 2015 (accessed 8 Aug 2015). http://www.pewinternet.org/2015/01/09/social-media-update-2014/
26 Jashinsky J, Burton SH, Hanson CL, *et al*. Tracking suicide risk factors through twitter in the US. *Crisis* 2014;35:51–9.
27 Moreno MA, Christakis DA, Egan KG, *et al*. A pilot evaluation of associations between displayed depression references on Facebook and self-reported depression using a clinical scale. *J Behav Health Serv Res* 2012;39:295–304.
28 Settanni M, Marengo D. Sharing feelings online: studying emotional well-being via automated text analysis of Facebook posts. *Front Psychol* 2015;6:1045.
29 Cutrona SL, Roblin DW, Wagner JL, *et al*. Adult willingness to use email and social media for peer-to-peer cancer screening communication: quantitative interview study. *JMIR Res Protoc* 2013;2:e52.
30 Ram S, Zhang W, Williams M, *et al*. Predicting asthma-related emergency department visits using big data. *IEEE J Biomed Health Inform* 2015;19:1316–23.
31 Merchant RM, Elmer S, Lurie N. Integrating Social Media into Emergency-Preparedness Efforts. *N Engl J Med* 2011;365:289–91.
32 Paul MJ, Dredze M. Discovering health topics in social media using topic models. *PLoS ONE* 2014;9:e103408. eCollection 2014.
33 Gesualdo F, Stilo G, Agricola E, *et al*. Influenza-like illness surveillance on Twitter through automated learning of naïve language. *PLoS ONE* 2013;8:e82489..
34 Nagel AC, Tsou MH, Spitzberg BH, *et al*. The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *J Med Internet Res* 2013;15:e237.
35 Bosley JC, Zhao NW, Hill S, *et al*. Decoding twitter: Surveillance and trends for cardiac arrest and resuscitation communication. *Resuscitation* 2013;84:206–12.
36 Merchant RM, Asch DA, Hershey JC, *et al*. A crowdsourcing innovation challenge to locate and map automated external defibrillators. *Circ Cardiovasc Qual Outcomes* 2013;6:229–36.
37 Greaves F, Laverty AA, Cano DR, *et al*. Tweets about hospital quality: a mixed methods study. *BMJ Qual Saf* 2014;23:838–46.
38 Greaves F, Ramirez-Cano D, Millett C, *et al*. Harnessing the cloud of patient experience: using social media to detect poor quality healthcare. *BMJ Qual Saf* 2013;22:251–5.
39 Rozenblum R, Bates DW. Patient-centred healthcare, social media and the internet: the perfect storm? *BMJ Qual Saf* 2013;22:183–6.
40 Kramer ADI, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. *Proc Natl Acad Sci USA* 2014;111:8788–90.
41 Goel V. As data overflows online, researchers grapple with ethics. *The New York Times*. 12 August 2014. http://www.nytimes.com/2014/08/13/technology/the-boon-of-online-data-puts-social-science-in-a-quandary.html?_r=0 (accessed 11 Sep 2014). (Archived by WebCite® at http://www.webcitation.org/6SVvgOJ3k).
42 Determann L. Social media privacy: a dozen myths and facts. *Stanford Technol Law Rev* 2012;7. https://journals.law.stanford.edu/sites/default/files/stanford-technology-law-review/online/determann-socialmediaprivacy.pdf
43 bStelzner MA. Social media marketing industry report. How Marketers Are Using Social Media to Grow Their Businesses. http://www.socialmediaexaminer.com/SocialMediaMarketingIndustryReport2012.pdf (accessed 11 Sep 2014). (Archived by WebCite® at http://www.webcitation.org/6SVvVx3pt).
44 Lovejoy K, Waters RD, Saxton GD. Engaging stakeholders through twitter: How nonprofit organizations are getting more out of 140 characters or less. *Public Relations Rev* 2012;38:313–18.
45 Kaplan AM, Haenlein M. Users of the world, unite! The challenges and opportunities of social media. *Bus Horiz* 2010;53:59–68.
46 bFacebook Privacy Policy. Facebook website. 09 11 2014. https://www.facebook.com/about/privacy/ (accessed 11 Sep 2014). (Archived by WebCite® at http://www.webcitation.org/6SVvIu7iC).
47 Grande D, Mitra N, Shah A, *et al*. The importance of purpose: moving beyond consent in the societal use of personal health information. *Ann Intern Med* 2014;161:855–62.
48 Dara K. Facebook boosts efforts on suicide prevention. CNET. 23 January 2013. http://www.cnet.com/news/facebook-boosts-efforts-on-suicide-prevention/ (accessed 11 Feb 2015).
49 HCUP CCS. Healthcare cost and utilization project (HCUP). 2014-09-11. http://www.hcup-us.ahrq.gov/toolssoftware/icd_10/ccs_icd_10.jsp (accessed 11 Sep 2014). (Archived by WebCite® at http://www.webcitation.org/6SVvNYk8R).