BMJ Open

Characterizing and Measuring Expressions of Loneliness in Individuals using Twitter

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-030355
Article Type:	Original research
Date Submitted by the Author:	10-Mar-2019
Complete List of Authors:	Guntuku, Sharath Chandra; University of Pennsylvania, Department of Emergency Medicine Schneider, Rachelle; University of Pennsylvania, Department of Emergency Medicine Pelullo, Arthur; University of Pennsylvania, Department of Emergency Medicine Young, Jami; University of Pennsylvania, Department of Emergency Medicine Wong, Vivien; University of Pennsylvania, Department of Emergency Medicine Ungar, Lyle ; University of Pennsylvania, Department of Emergency Medicine Polsky, Daniel ; Leonard Davis Institute of Health Economics, Volpp, Kevin ; University of Pennsylvania Perelman School of Medicine, Medical Ethics and Policy Merchant, Raina; University of Pennsylvania, Department of Emergency Medicine
Keywords:	loneliness, social media, natural language processing, MENTAL HEALTH, STATISTICS & RESEARCH METHODS

Characterizing and Measuring Expressions of Loneliness in Individuals using Twitter

Sharath Chandra Guntuku, PhD[1,4,5], Rachelle C. Schneider, BS [1,5], Arthur Pelullo, MS[1,4,5], Jami F. Young, PhD[5,7], Vivien Wong, BS[1,5], Lyle H. Ungar, PhD[3,4], Daniel Polsky, PhD[5,6], Kevin Volpp, MD, PhD[5,6], Raina M. Merchant, MD, MSHP[1,2,5]

[1]Penn Medicine Center for Digital Health, Philadelphia, PA 19104

[2]Penn Medicine Center for Healthcare Innovation, Philadelphia, PA 19104

[3]Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104

[4]Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

[5]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

[6]The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

[7]Children's Hospital of Philadelphia, Philadelphia, PA 19146

Corresponding author and request for reprints

Sharath Guntuku

3400 Civic Center Blvd

Philadelphia PA 19104

Email: sharathg@sas.upenn.edu

Word count: 2899, 26 pages, 3 tables, 2 figures

Keywords: loneliness expressions; social media; twitter; natural language processing; mental health

Abstract

Objectives: Loneliness affects approximately 30% of individuals in the United States and is associated with high morbidity. We sought to characterize the (online) lives of people who express being lonely and correlate their posts with predictors of mental health.

Setting and design: A leading social media platform (Twitter) was the main focus of the study. We collected approximately 400 million tweets from in Pennsylvania, USA, between 2012-2016. We identified users whose posts contained the words 'lonely' or 'alone' and compared them to a control group matched by age, gender, and period of posting. Using natural-language processing, we characterized what and when users post, their association with linguistic markers of mental health, and if language can predict manifestations of loneliness. The statistical analysis, data synthesis, and model creation was conducted in 2018-2019.

Primary outcome measures: We evaluated counts of language features in the lonely group compared to the control group. These language features were measured by (1) open-vocabulary topics and (2) linguistic markers of anger, depression, and anxiety. We also evaluated the prediction of expressions of loneliness compared to the control group, measured by Area Under Curve.

Results: Users in the lonely group (N=6202) posted more about difficult interpersonal relationships, psychosomatic symptoms, substance use, wanting change, unhealthy eating, and having troubles with sleep. Their posts were also associated with linguistic markers of anger, depression, and anxiety. A random forest model predicted expressions of loneliness online with an accuracy of 77%.

Conclusions: Posts with the words lonely or alone often include psychosocial features and can provide insight about how individuals express and experience loneliness. This can inform online surveillance for high risk individuals experiencing loneliness and interventions focused on addressing morbidity in this condition.

Strengths and Limitations of this study

- Study's novel focus on timelines of social media users to study expressions of loneliness and correlation with predictors of mental health.
- The study sample consists of social media users and is not representative of the general population.
- Though we manually annotated a subset of posts mentioning loneliness, some may have been metaphorical or non sequiturs.

Introduction

Loneliness is a major public health problem affecting 1 in 3 adults in the United States (US).[1, 2] It has been described as "the psychological embodiment of social isolation, reflecting the individual's experienced dissatisfaction with the frequency and closeness of their social contacts or the discrepancy between the relationships they have and the relationships they would like to have." [1, 3, 4] Loneliness is also one of the primary underlying causes and correlates for chronic mental health conditions and physician visits in some populations.[1, 5-9] Prior research has found several risk factors associated with loneliness in specific subgroups. Risk factors for older adults include reduction in the quality of social connections, as well as institutionalization.[10] Risk factors for young adults include drug use and low self-esteem.[11-12] Prior work has evaluated the effect of social relationships on the health of individuals and social support was found to reduce morbidity and mortality.[1, 13-15] Despite high morbidity associated with loneliness,[1, 16, 18-19] few reports have focused on quantifying the experience of loneliness expressed online.

Online data on social networks is growing exponentially. More than 2.3 billion individuals use social media regularly (e.g. Facebook 1.71 billion, Twitter 320 million, Instagram 400 million).[20] Increasingly, individuals are using social media as a platform to post about their thoughts, feelings, perceptions, and experiences.[21-22] The regular production of data on online platforms also allows for tracking of health in real-time. These data offer promise as they provide different insights than data from traditional surveys. Another opportunity is in the ability of digital platforms to not only provide markers of health but also serve as platforms that can be used for direct intervention.[23-24] Users on social media often post about how they are coping (or not) with life stressors and their support networks. Specifically, expressions of loneliness have been associated with feeling unloved, depressed, bored, and not having friends.[21-22] Prior research has also demonstrated that users' mental health conditions, such as depression and anxiety, can be predicted from their social media language.[25-26]

Social media seeks to ‘connect’ people, yet several studies have reported an association between social media use and increased perceived social isolation.[27] As loneliness can impact health outcomes, identifying ways to track prevalence and manifestations of loneliness online would be useful for developing approaches for identifying and offering support for these individuals.

We sought to identify data from a widely used publicly available social network, Twitter, to characterize what and when individuals post about loneliness, association of posts with mental health, and how manifestations of loneliness can be predicted across individuals.

Methods

This was a retrospective analysis of publicly available data on users posting about loneliness on Twitter in Pennsylvania. This study was approved by the University of Pennsylvania Institutional Review Board.

Twitter Data

Twitter is a popular social media platform which allows users to send and receive short 140 character messages, or ‘tweets’ (at the time of this study; the character limit was later increased to 280). First, the Twitter Streaming API was used to collect a random 1% sample of public tweets from 2012-2016. This initial dataset was then filtered to contain only geolocated tweets or tweets originating from users with nonempty location fields in their profile. The county of origin of each tweet user was determined, and the dataset was filtered to obtain only tweets for users in Pennsylvania. To increase the sample size of tweets from the state, all unique user IDs were recorded, and the Twitter search API was used to extract timelines (each user's prior 3200 tweets) filtered by timestamps ranging from 2012-2016.

Study Sample

We identified users who posted the word "alone" or "lonely" at least once in their timeline (25,966 users). Of these, 6,202 users posted messages with "alone" or "lonely" at least 5 times. As social media includes colloquial, metaphorical, and light-hearted language (eg. "If I see Justin Bieber, I will have a heart attack")[28] we sought to identify the proportion of tweets in which lonely seemed to refer to the public health meaning rather than other uses of the term (e.g. metaphor, joke). Two co-authors independently coded a random set of 100 tweets to identify them as presumed to be associated with the feeling of loneliness or other. The Kappa was 0.70 and we identified that 76% of users' tweets indicate presumably feeling lonely. A few examples are as follows: "i'm feelin real depressed, confused, & lonely", "im always the only up around this time, feeling a lil lonely" and "I'm so Lonely in life :-(I just wish I can have love again it feels so go to be in love with someone whom loves you.". This research was done without patient involvement.

Control group

We then identified a control group of users by matching each user in the above dataset to another user by age, gender and period of activity (dates of first and last posting on twitter). We obtained the age and gender estimates by using lexica developed previously.[17] Then, we selected users with a minimum of 500 words across all their posts to have sufficient language for linguistic analyses.[29] We excluded non-English, non-US tweets, retweets, and tweets that were used to identify users in the lonely group in all analyses. Hereafter, we use ‘lonely’ group to indicate users who had more than 5 posts with the words ‘lonely’ or ‘alone’, and ‘control’ group to represent the matched set of users who had no such posts.

Deriving language features to characterize individuals expressing loneliness

We used three sets of language features: a) open-vocabulary topics,[30] b) dictionary-based psycholinguistic features,[31] c) mental well-being attributes such as anxiety, depression by applying previously developed statistical models,[32] d) number of drug words and time of posts as past research [11] has shown an association between loneliness and substance use. These language features have been shown to be predictive of several health outcomes, such as depression, schizophrenia, attention deficit hyperactivity disorder (ADHD), and general well-being.[33; 26]

Open-vocabulary: As closed-vocabulary approaches like LIWC include only a small subset of the entire language used on social media, we use an open-vocabulary approach to improve the coverage and find topics that people who express being lonely talk about. Two hundred topics (groups of co-occurring words) were generated using tweets across all users in the dataset of lonely and control users using the Mallet implementation of Latent Dirichlet Allocation (LDA).[34] The topic distribution of each user aggregated across all the messages was then calculated.

Dictionary-based: From each post, we extracted the relative frequency of single words and phrases (consisting of two or three consecutive words). Then, all words used by less than 1% of users were removed from analysis so as to remove uncommonly used words (outliers). Additionally, all messages used to identify our study group were removed prior to further analysis. The distribution of Linguistic Inquiry Word Count (LIWC) dictionary features are also extracted for each post. For each user, we measure the proportion of word tokens that fall into a given LIWC category. Then, we compare it against the word tokens from the control data using an empirical distribution of the proportion of language attributable to each LIWC category.

Mental well-being attributes: We used automatic text-regression methods to assign to each user scores on the Depression, Anxiety and Anger facets for users.[32] This model was trained on a sample of over 28,749 users who had taken the International Personality Item Pool Neuroticism-Extraversion-Openness Personality Inventory Revised (IPIP NEO-PI-R) survey that contains the Depression, Anxiety and Anger Facets of the Neuroticism Factor.[32] The text model was trained using tokens and topics extracted from status updates as features. In the original validation, the model achieved a Pearson correlation of r = .32 predictive performance, which is considered high in psychology, especially when measuring internal states.[35]

Use of Drug-words: We also extracted the frequency (aggregated to every user) of most common drug words as used on social media.[36]

Temporal patterns: We determined the frequency of posts across different hours of the day by users in both the lonely and control groups to understand the diurnal patterns in posting.

Identifying differentially expressed language features in the lonely group

We isolated the patterns in users' loneliness expressions using the linguistic attributes and user traits by correlating them with the lonely and control groups. We used Benjamini-Hochberg p-correction and use $p<0.001$ for indicating meaningful correlations and the effect size was measured using Cohen's d. The statistical analysis, data synthesis, and model creation was conducted in 2018-2019.

Predicting the likelihood of posting about loneliness online

We then looked at the feasibility of predicting whether a user is likely to express that they are lonely or not based on their social media language. Automated analysis of social media is accomplished by building predictive models, which use 'features,' or variables that have been extracted from social media data. For this analysis we used LIWC and topics as features. Features are then treated as independent variables in an algorithm (Random Forests) to predict the dependent variable of an outcome of interest (e.g., users' saying that they are lonely or not). For cross validation, the predictive model was trained, using Random Forests, on the training set and then evaluated on a test set to avoid overfitting. The prediction performances are reported as one of several possible metrics on an out-of-sample 5-fold cross validation setting.

Results

Of the 408,296,620 tweets posted by users geo-located in Pennsylvania, USA, 25,966 users with 46,160,774 posts in their timelines, had at least one post with the words 'lonely' or 'alone', and 6,202 users (referred to as 'lonely' group hereafter) with 17,995,084 posts in their timelines, had more than five such posts (Table 1). The lonely group had 1.9 times more posts in the study time period as the control (Table 1). The median estimated age of this cohort was 21 years, and 69% female.

Table 1: Descriptive statistics for the lonely group about loneliness and the control group

Descriptive Statistics of the Dataset		
	Lonely Group (n= 6,202)	Control group (n= 6,202)
Median Age	21	21
# Messages in timelines	17,995,084	9,219,677
# Females	4,400	4,400
# Males	1,802	1,802

*the lonely group is defined as any user posting at least 5 times about loneliness and the control group is defined as any user who does not have any posts about loneliness

Identifying differentially expressed language features in the lonely group

Open vocabulary approach: Analyzing differences in individual words and phrases used across both groups, we observed (Figure 1a) that users in the lonely group referred to themselves (‘myself’ (d=.18), ‘I’ (d=.16)) in their Twitter posts significantly more than the control group. They also posted about relationship issues (‘want_somebody’ (d=.08), ‘no_one_to’ (d=.1), needs and feelings (‘i_just_wanna (d=.12), ‘in_my_feelings’ (d=.1), ‘i_need’ (d=.12), ‘i_cant’ (d=.1)), and included more expletives. Users in the control group (Figure 1b) engaged in a lot more conversations as indicated by ‘<user>’ (d=-.2) (we anonymize ‘@’ mentions in users tweets as ‘<user>’) compared to the lonely group. The control group also posted more about games

(’season’ (d=-.09) ,‘coach’ (d=-.07), ‘team’ (d=-.1)) and positivity (‘!’ (d=-.13), ‘awesome’ (d=-.09), ‘:)’ (d=-.08)). Figure 1 illustrates the words and phrases most prominently associated with the lonely and control groups.

Using topics generated from LDA, we identified the themes which occur more frequently in posts in the lonely group. Posts were about interpersonal relationships (d=.28) (and associated issues (d=.22)), self-reflection (d=.21) (accompanied with wondering about the future (d=.12)), drug/alcohol use (d=.29) (considering them to be the ‘only friend’), insomnia (d=.27), uncontrolled emotions (d=.28) (accompanied by confusion (d=.11)), and psychosomatic symptoms (d=.29). Table 2 shows the effect sizes between most prominent topic distributions and the users who have more than 5 posts with the words lonely or alone.

Dictionary-based: Association of LIWC categories with the posts by users in the lonely group are shown in Table 3. Individuals who posted about being alone or lonely used increased self-references (first person pronouns, d=.18), words indicating cognitive processes (including certainty, d=.15, discrepancies, d=.14, differentiation, d=.13 and tentativeness, d=.13), and negative emotions (anger, d=.12 and swearing, d=.11).

Mental well-being*:* Users in the lonely group were more likely to have posts associated with anger (d=.95), depression (d=.81) and anxiety (d=.75) when compared to the control group.

Use of Drug Words: We also identified the distribution of words pertaining to drugs in the posts of users in the lonely group, and these were more likely to reference a blunt (d=.16), smoke (d=.13), and heroin (d=.1), and included prescribed medications for treatment, recreational drug use, and recreational drugs.

Temporal patterns*:* Users in the lonely group were found to post more during the night (d=.1). We also see themes associated with night-time posting and having difficulty sleeping (d=.27) in the open-vocabulary analysis.

Table 2: Highly correlated topics with expressions of loneliness.

* Effect size is measured using Cohen's d. Only significant topics after Benjamini-Hochberg p-correction and use p<0.001 are shown.

Predictive Analysis: A random forest model predicted language associated with lonely expressions with an AUC of .86 using a combination of LIWC and LDA topics as linguistic features.

Table 3: LIWC categories with expressions of loneliness

Category	**Cohen's d***
Pronouns	
1st Person Pronouns	0.18
Cognitive Processes	
Certainty	0.15
Discrepancies	0.15
Differentiation	0.14
Tentativeness	0.13
Negative Emotions	
Anger	0.12

Swearing	0.11

*Only significant categories after Benjamini-Hochberg p-correction and p<0.001 are shown.

Discussion

This paper has three main findings. First, we identified themes and contexts associated with users posting about loneliness on Twitter. Second, we observed that users posting about loneliness used language associated with linguistic models for anger, depression, and anxiety. Third, posts about loneliness were more likely to occur in the evening or night.

We identified themes and contexts of users posting about loneliness on Twitter. Themes associated with people expressing loneliness on Twitter were about interpersonal relationships, self-reflection, substance use, insomnia, uncontrolled emotions, food/hunger, and psychosomatic symptoms. Some of these themes are consistent with prior literature about substance use, emotional dysregulation, and troubles with relationships. For example, in one study, a high positive correlation was found between alcoholism and groups of lonely people, and lonely people were also found to express negative feelings towards relationships.[37] Lonely individuals were also reported to focus on overcoming past events as well as showing feelings of helplessness.[37]

Association of the lonely group with linguistic estimates of anger, depression, and anxiety corroborate prior research.[5-6] Specifically, anxiety, anger, and negative mood were reported as higher in lonely young adults.[38] Tweets by users in the lonely group were more self-focused compared to the control group. Prior researchers have found that “first person singular pronouns are a modest linguistic marker of depression."[39] This presents the potential for early identification and assessment to intervene on loneliness as well as mental health conditions for this group.

Trends in temporal variation in posting may reflect difficulties in terms of engaging in online activity and doing so during hours typically devoted to sleep. Prior work has shown that sleep deprivation can contribute to social withdrawal and loneliness.[40] A better understanding of the temporality of posting could inform timing of interventions designed to address loneliness, as well as provide insight for other researchers to test the inter-relationships between loneliness and the motivations for using social media during nighttime.

Loneliness is known to be one of the primary underlying causes and correlates for chronic mental health conditions.[5-6] As loneliness is becoming increasingly recognized as a public health issue associated with chronic mental and physical health problems, several groups have taken action to address it. For example, the United Kingdom appointed a Minister for Loneliness who is responsible for addressing loneliness within communities.[41] CareMore, a health plan and delivery system providing care for enrollees in Medicare Advantage and Medicaid health plans in seven states across the U.S., launched the "Togetherness Program" in a clinical setting to address loneliness in elderly patients.[42] Through this work, CareMore reported that participation in exercise programs increased by 56.6%, emergency room utilization decreased by 3.3%, and hospital admissions among participants were 20.8% lower per thousand compared to the "intent to treat population."[43] Additionally, social network interventions targeting loneliness have been found to be effective in reducing social isolation among individuals with severe mental health conditions but these interventions are not included in the treatment plans for individuals with a mental illness.[44] Using natural language processing and machine learning to automatically identify a person expressing loneliness on Twitter could inform interventions targeted at early identification and support for affected and at risk individuals.

Future work that builds off this study could be to validate whether the characteristics of people who are using the words 'lonely' or 'alone' on Twitter can be used to track community health risks, particularly, the risk of social isolation. Our methods can potentially be used to identify problematic loneliness for community public health monitoring.

Limitations and Ethics

The study sample consists of social media users and is not representative of the general population. An estimated 40% of US adults using Twitter are between the ages of 18 and 29, so our analysis is skewed towards younger people.[45] Posts mentioning loneliness may have been metaphorical or non sequiturs.

The feasibility of social media-based assessments of loneliness expressions (and mental health more broadly) needs further assessment. Privacy of individuals is an ongoing concern, especially with social media users not fully realizing the amount of health insights that can be gleaned by their online posts. Employers and insurance companies, for example, may be motivated to derive these assessments, but could use these insights against those suffering from mental illness. As mental illnesses carry social stigma and may engender discrimination, data protection and ownership frameworks are needed to make sure the data is not used against the users' interest.[46]

Further, transparency about which indicators are derived by whom for what purpose should be part of ethical and policy discourse.

There are also open questions around the impact of misclassifications, and how derived mental health indicators can be responsibly integrated into systems of care.[47]

Conclusions

In this study we characterized expressions of loneliness on Twitter at the individual level. Furthermore, we identified specific contexts, themes, and traits in the posts of individuals expressing loneliness on Twitter. As loneliness is a public health challenge, a better understanding of how loneliness is described online can inform tracking of loneliness and interventions targeted at addressing this important public health problem in regards to the behavior of lonely individuals that may be at risk of developing a severe mental health condition.[42]

Acknowledgements

Funding: This project is funded, in part, under a grant with the Pennsylvania Department of Health (ORG-FUND-PROG-CREF is 4290-567862-2446-2049). No sponsor of funding source played a role in: "study design and the collection, analysis, and interpretation of data and the writing of the article and the decision to submit it for publication." All researchers are independent from funders

Data Sharing Statement: Because of our IRB requirements, data will be shared upon request from the corresponding author.

Contributors: S.C. Guntuku and R. Merchant originated the study. S.C. Guntuku, R. Schneider, A. Pelullo, L.H. Ungar, and R. Merchant developed methods, interpreted analysis, and contributed to the writing of the article. J.F. Young, V. Wong, D. Polsky, and K. Volpp assisted with the interpretation of the findings and contributed to the writing of the article.

Disclosures: None

References:

(1) Gerst-Emerson K, Jayawardhana J. Loneliness as a Public Health Issue: The Impact of Loneliness on Health Care Utilization Among Older Adults. *American Journal of Public Health*. 2015;105(5):1013-1019. doi:10.2105/ajph.2014.302427.

(2) New Cigna Study Reveals Loneliness at Epidemic Levels in America. Cigna, a Global Health Insurance and Health Service Company. https://www.cigna.com/newsroom/news-releases/2018/new-cigna-study-reveals-loneliness-at-epidemic-levels-in-america. Accessed February 18, 2019.

(3) Jong-Gierveld JD. Developing and testing a model of loneliness. *Journal of Personality and Social Psychology*. 1987;53(1):119-128. doi:10.1037//0022-3514.53.1.119.

(4) Peplau LA, Perlman D. *Loneliness: a Sourcebook of Current Theory, Research, and Therapy*. New York: Wiley; 1982.

(5) Stravynski A, Boyer R. Loneliness in Relation to Suicide Ideation and Parasuicide: A Population-Wide Study. *Suicide and Life-Threatening Behavior*. 2001;31(1):32-40. doi:10.1521/suli.31.1.32.21312.

(6) Blai B. Health Consequences of Loneliness: A Review of the Literature. *Journal of American College Health*. 1989;37(4):162-167. doi:10.1080/07448481.1989.9938410.

(7) Heinrich LM, Gullone E. The clinical significance of loneliness: A literature review. *Clinical Psychology Review*. 2006;26(6):695-718. doi:10.1016/j.cpr.2006.04.002.

(8) Richard A, Rohrmann S, Vandeleur CL, Schmid M, Barth J, Eichholzer M. Loneliness is adversely associated with physical and mental health and lifestyle factors: Results from a Swiss national survey. *Plos One*. 2017;12(7). doi:10.1371/journal.pone.0181442.

(9) Rico-Uribe LA, Caballero FF, Olaya B, et al. Loneliness, Social Networks, and Health: A Cross-Sectional Study in Three Countries. *Plos One*. 2016;11(1). doi:10.1371/journal.pone.0145264.

(10) Pinquart M, Sorensen S. Influences on Loneliness in Older Adults: A Meta-Analysis. *Basic and Applied Social Psychology*. 2001;23(4):245-266. doi:10.1207/S15324834BASP2304_2.

(11) Rokach A. Determinants of Loneliness of Young Adult Drug Users. *The Journal of Psychology*. 2002;136(6):613-630. doi:10.1080/00223980209604823.

(12) Vanhalst J, Luyckx K, Scholte RHJ, Engels RCME, Goossens L. Low Self-Esteem as a Risk Factor for Loneliness in Adolescence: Perceived - but not Actual - Social Acceptance as an Underlying Mechanism. *Journal of Abnormal Child Psychology*. 2013;41(7):1067-1081. doi:10.1007/s10802-013-9751-y.

(13) Seeman T. How Do Others Get under Our Skin? *Emotion, Social Relationships, and Health*. 2001:189-220. doi:10.1093/acprof:oso/9780195145410.003.0006.

(14) Steptoe A, Shankar A, Demakakos P, Wardle J. Social isolation, loneliness, and all-cause mortality in older men and women. *Proceedings of the National Academy of Sciences*. 2013;110(15):5797-5801. doi:10.1073/pnas.1219686110.

(15) Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. *Social*

Science & Medicine. 2000;51(6):843-857. doi:10.1016/s0277-9536(00)00065-4.

(16) Hawkley LC, Thisted RA, Masi CM, Cacioppo JT. Loneliness predicts increased blood pressure: 5-year cross-lagged analyses in middle-aged and older adults. *Psychology and Aging*. 2010;25(1):132-141. doi:10.1037/a0017805.

(17) Sap M, Park G, Eichstaedt J, et al. Developing Age and Gender Predictive Lexica over Social Media. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014. doi:10.3115/v1/d14-1121.

(18) Knox S, Uvnäs-Moberg K. Social isolation and cardiovascular disease: an atherosclerotic pathway? *Psychoneuroendocrinology*. 1998;23(8):877-890.

(19) Cacioppo JT, Hughes ME, Waite LJ, Hawkley LC, Thisted RA. Loneliness as a specific risk factor for depressive symptoms: Cross-sectional and longitudinal analyses. *Psychology and Aging*. 2006;21(1):140-151. doi:10.1037/0882-7974.21.1.140.

(20) U.S. population with a social media profile 2018. Statista. https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/. Accessed February 18, 2019.

(21) Kivran-Swaine F, Ting J, Naaman M. “Understanding Loneliness in Social Awareness Streams: Expressions and Responses. *ICWSM 2014*. 2014. https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef.

(22) Kamvar S, Kamvar S, Harris JJ. *We Feel Fine: an Almanac of Human Emotion*. New York: Scribner; 2009.

(23) Coppersmith G, Ngo K, Leary R, Wood A. Exploratory Analysis of Social Media Prior to a Suicide Attempt. *Proceedings of the Third Workshop on Computational*

Lingusitics and Clinical Psychology. 2016. doi:10.18653/v1/w16-0311.

(24) Qntfy, Inc. OurDataHelps. OurDataHelps. https://ourdatahelps.org/. Accessed February 18, 2019.

(25) Guntuku SC, Buffone A, Jaidka K, Eichstaedt J, Ungar L. Understanding and Measuring Psychological Stress using Social Media. *ICWSM 2019*. 2018.

(26) Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*. 2017;18:43-49. doi:10.1016/j.cobeha.2017.07.005.

(27) Primack BA, Shensa A, Sidani JE, et al. Social Media Use and Perceived Social Isolation Among Young Adults in the U.S. *American Journal of Preventive Medicine*. 2017;53(1):1-8. doi:10.1016/j.amepre.2017.01.010.

(28) Sinnenberg L, DiSilvestro CL, Mancheno C, et al. Twitter as a Potential Data Source for Cardiovascular Disease Research. *JAMA Cardiology*. December 2016. https://jamanetwork.com/journals/jamacardiology/fullarticle/2556216.

(29) Jaidka K, Guntuku SC, Buffone A, Schwartz HA, Ungar L. Facebook versus Twitter: Differences in Self-Disclosure and Trait Prediction. *ICWSM*. 2018.

(30) 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). doi:10.1371/journal.pone.0073791.

(31) Pennebaker Jw, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works*. 2014:1-22.

(32) Schwartz HA, Eichstaedt J, Kern ML, et al. Towards Assessing Changes in Degree of Depression through Facebook. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014. doi:10.3115/v1/w14-3214.

(33) Guntuku SC, Ramsay JR, Merchant RM, Ungar LH. Language of ADHD in Adults on Social Media. *Journal of Attention Disorders*. 2017:108705471773808. doi:10.1177/1087054717738083.

(34) Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *The Journal of Machine Learning Research*. 2003;3.

(35) Meyer GJ, Finn SE, Eyde LD, et al. Psychological testing and psychological assessment. A review of evidence and issues. *The American psychologist*. 2001;56(2):128-165.

(36) Jones R. Drug Slang Dictionary - Words Starting With G. http://www.noslang.com/drugs/dictionary.php. Accessed February 19, 2019.

(37) Booth R. Toward an Understanding of Loneliness. *Social Work*. 1983;28(2):116-119. doi:10.1093/sw/28.2.116.

(38) Cacioppo JT, Hawkley LC, Ernst JM, et al. Loneliness within a nomological net: An evolutionary perspective. *Journal of Research in Personality*. 2006;40(6):1054-1085. doi:10.1016/j.jrp.2005.11.007.

(39) Edwards T, Holtzman NS. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*. 2017;68:63-68. doi:10.1016/j.jrp.2017.02.005.

(40) Simon EB, Walker MP. Sleep loss causes social withdrawal and loneliness. *Nature Communications*. 2018;9(1). doi:10.1038/s41467-018-05377-0.

(41) Yeginsu C. U.K. Appoints a Minister for Loneliness. *The New York Times*. https://www.nytimes.com/2018/01/17/world/europe/uk-britain-loneliness.html. Published January 17, 2018.

(42) Rubin R. Loneliness Might Be a Killer, but What's the Best Way to Protect Against It? *Jama*. 2017;318(19):1853. doi:10.1001/jama.2017.14591.

(43) CareMore Health Announces New Outcomes Data from First-of-its-Kind Togetherness Program. Business Wire A Berkshire Hathaway Company . https://www.businesswire.com/news/home/20181218005059/en/. Published December 18, 2018. Accessed February 18, 2019.

(44) Perese EF, Wolf M. Combating Loneliness Among Persons With Severe Mental Illness: Social Network Interventions Characteristics, Effectiveness, And Applicability. *Issues in Mental Health Nursing*. 2005;26(6):591-609. doi:10.1080/01612840590959425.

(45) U.S. Twitter reach by age group 2018 | Statistic. Statista. https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/. Accessed February 18, 2019.

(46) Mckee R. Ethical issues in using social media for health and health care research. *Health Policy*. 2013;110(2-3):298-301. doi:10.1016/j.healthpol.2013.02.006.

(47) Inkster B, Stillwell D, Kosinski M, Jones P. A decade into Facebook: where is psychiatry in the digital age? *The Lancet Psychiatry*. 2016;3(11):1087-1090. doi:10.1016/s2215-0366(16)30041-4.

Figure legends

Figure 1: Words/Phrases more likely to be posted by Twitter users with a) self-reported loneliness (Individuals with at least 5 posts with the words ‘lonely’ or ‘alone’ group compared to the b) control group.

Word size indicates the strength of the correlation and word color indicates relative word frequency. ($p<0.01$, Bonferroni p-corrected)

Figure 2: Temporal variation showing diurnal patterns of post frequency of both the ‘lonely’ and ‘control’ groups.

The dotted line indicates the percentage of posts at different hours of the day by the group of users with at least 5 posts containing the word ‘lonely’ or ‘alone’ and the solid line indicates users who do not have any posts about loneliness. The x-axis represents the hour of the day each post occurs and the y-axis indicates the number of posts for each group.

Figure 1:

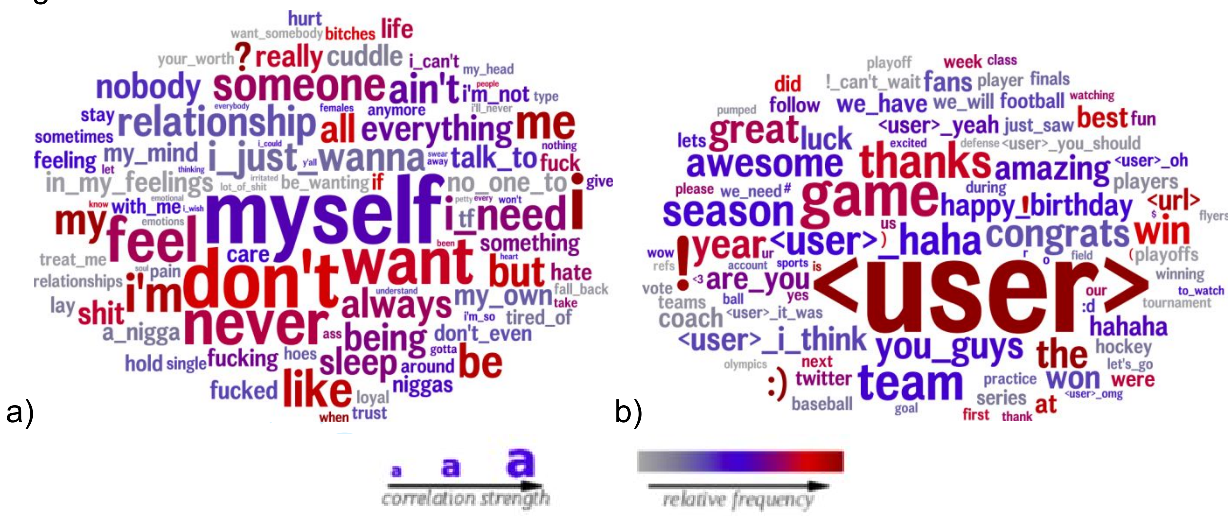
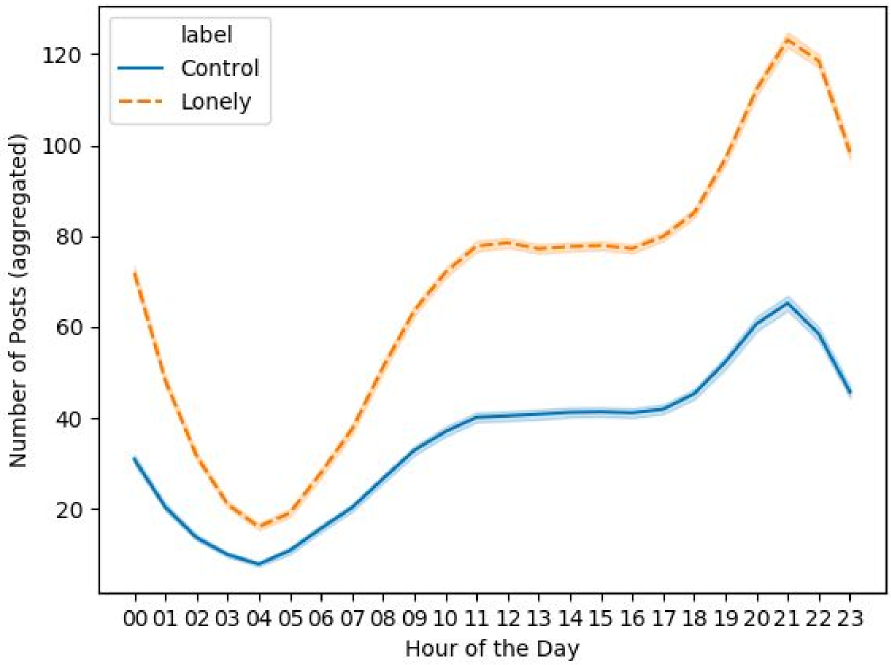

a)
hurt
life
bitches
really
cuddle
nobody
someone
ain't
i'm_not
relationship
all
everything
me
my_mind
i_just_wanna
talk_to
fuck
feeling
sometimes
in_my_feelings
no_one_to
myself
i_need
my
feel
want
something
don't
i'm
but
hate
always
my_own
shit
never
being
don't_even
tired_of
a_nigga
sleep
be
hold
single
fucking
fucked
like
niggas
loyal
trust
b)
week
fans
player
finals
did
we_have
we_will
football
follow
<user>_yeah
just_saw
best
fun
great
luck
thanks
awesome
amazing
game
season
happy_birthday
<url>
win
congrats
year
<user>_haha
!
are_you
<user>
coach
<user>_i_think
you_guys
the
hahaha
team
won
were
twitter
:)
baseball
practice
series
at
correlation strength
relative frequency

Figure 2


label
Control
Lonely
120
100
80
60
40
20
Number of Posts (aggregated)
00 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Hour of the Day

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **(pg.2)**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **(pg.2)**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **(pg.4)**
Objectives	3	State specific objectives, including any prespecified hypotheses **(pg. 4)**
Methods		
Study design	4	Present key elements of study design early in the paper **(pg.5)**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **(pg. 5)**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up **(pg. 6)**
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **(pg. 6)**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **(pg. 6)**
Bias	9	Describe any efforts to address potential sources of bias **(pg. 6)**
Study size	10	Explain how the study size was arrived at **(pg. 6)**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **(pg. 7)**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **(pg. 9)**
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) If applicable, explain how loss to follow-up was addressed
		(*e*) Describe any sensitivity analyses
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **(pg. 9)**
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **(pg. 9)**
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarise follow-up time (eg, average and total amount)
Outcome data	15*	Report numbers of outcome events or summary measures over time **(pgs 10,11)**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **(pgs 10,11)**

		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **(pgs 10,11)**
Discussion		
Key results	18	Summarise key results with reference to study objectives **(pgs. 12, 13)**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **(pgs. 14)**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **(pg. 13, 14)**
Generalisability	21	Discuss the generalisability (external validity) of the study results **(pgs. 13, 14)**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **(pg. 15)**

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

BMJ Open

Studying Expressions of Loneliness in Individuals using Twitter: An Observational Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-030355.R1
Article Type:	Original research
Date Submitted by the Author:	28-May-2019

Complete List of Authors:	Guntuku, Sharath Chandra; University of Pennsylvania, Department of Emergency Medicine Schneider, Rachelle; University of Pennsylvania, Department of Emergency Medicine Pelullo, Arthur; University of Pennsylvania, Department of Emergency Medicine Young, Jami; University of Pennsylvania, Department of Emergency Medicine Wong, Vivien; University of Pennsylvania, Department of Emergency Medicine Ungar, Lyle ; University of Pennsylvania, Department of Emergency Medicine Polsky, Daniel ; Leonard Davis Institute of Health Economics, Volpp, Kevin ; University of Pennsylvania Perelman School of Medicine, Medical Ethics and Policy Merchant, Raina; University of Pennsylvania, Department of Emergency Medicine

Primary Subject Heading:	Public health
Secondary Subject Heading:	Communication, Mental health
Keywords:	loneliness, social media, natural language processing, MENTAL HEALTH, STATISTICS & RESEARCH METHODS

Studying Expressions of Loneliness in Individuals using Twitter: An Observational Study

Sharath Chandra Guntuku, PhD[1,4,5], Rachelle C. Schneider, BS [1,5], Arthur Pelullo, MS[1,4,5], Jami F. Young, PhD[5,7], Vivien Wong, BS[1,5], Lyle H. Ungar, PhD[3,4], Daniel Polsky, PhD[5,6], Kevin Volpp, MD, PhD[5,6], Raina M. Merchant, MD, MSHP[1,2,5]

[1]Penn Medicine Center for Digital Health, Philadelphia, PA 19104

[2]Penn Medicine Center for Healthcare Innovation, Philadelphia, PA 19104

[3]Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104

[4]Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

[5]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

[6]The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

[7]Children's Hospital of Philadelphia, Philadelphia, PA 19146

Corresponding author and request for reprints

Sharath Guntuku

3400 Civic Center Blvd

Philadelphia PA 19104

Email: sharathg@sas.upenn.edu

Word count: 3723, 31 pages, 4 tables, 2 figures

Keywords: loneliness expressions; social media; twitter; natural language processing; mental health

Abstract

Objectives: Loneliness is a major public health problem affecting 1 in 3 older adults in the United States (U.S.). While less is known about the prevalence of loneliness in other age groups, around half of adults in the U.S. report sometimes or always feeling alone (46%). We sought to characterize the (online) lives of people who mention the words 'lonely' or 'alone' in their Twitter timeline and correlate their posts with predictors of mental health.

Setting and design: A leading social media platform (Twitter) was the main focus of the study. We collected approximately 400 million tweets from in Pennsylvania, USA, between 2012-2016. We identified users whose posts contained the words 'lonely' or 'alone' (referred to as the lonely group hereafter) and compared them to a control group matched by age, gender, and period of posting. Using natural-language processing, we characterized what and when users post, their association with linguistic markers of mental health, and if language can predict manifestations of loneliness. The statistical analysis, data synthesis, and model creation was conducted in 2018-2019.

Primary outcome measures: We evaluated counts of language features in the lonely group compared to the control group. These language features were measured by (1) open-vocabulary topics and (2) linguistic markers of anger, depression, and anxiety. We also evaluated the prediction of expressions of loneliness compared to the control group, measured by Area Under Curve.

Results: Users in the lonely group (N=6202) posted more about difficult interpersonal relationships, psychosomatic symptoms, substance use, wanting change, unhealthy eating, and having troubles with sleep. Their posts were also associated with linguistic markers of anger, depression, and anxiety. A random forest model predicted expressions of loneliness online with an accuracy of 77%.

Conclusions: Posts with the words lonely or alone often include psychosocial features and can provide insight about how individuals express and experience loneliness. This can inform online surveillance for high risk individuals experiencing loneliness and interventions focused on addressing morbidity in this condition.

Strengths and Limitations of this study

- Novel focus on timelines of social media users to study expressions of loneliness and correlation with predictors of mental health.
- The study sample consists of social media users and is not representative of the general population.
- Though we manually annotated a subset of posts mentioning loneliness, some may have been metaphorical or non sequiturs.

Introduction

Loneliness is a major public health problem affecting 1 in 3 older adults in the United States (U.S.).[1] While less is known about the prevalence of loneliness in other age groups, around half of adults in the U.S. report sometimes or always feeling alone (46%).[2] Loneliness has been described as "the psychological embodiment of social isolation, reflecting the individual's experienced dissatisfaction with the frequency and closeness of their social contacts or the discrepancy between the relationships they have and the relationships they would like to have." [1, 3, 4] Loneliness is also one of the primary underlying causes and correlates for chronic mental health conditions and physician visits in some populations.[2, 5-9]

Prior research has found several risk factors associated with loneliness in specific subgroups -- reduction in the quality of social connections and institutionalization in older population while drug use and low self-esteem in young adults.[10-12] Studies have also looked at the co-occurrence of substance use and loneliness as a risk factor in adolescent.[13] These risk factors are important to inform future targeted interventions addressing loneliness in individuals.

Online data on social networks is growing exponentially. More than 2.3 billion individuals use social media regularly (e.g. Facebook 1.71 billion, Twitter 320 million, Instagram 400 million).[14] A recent study showed that about 89% of 1060 teens between the ages 13 and 17 years-old who were interviewed used social media, with 71% of them having accounts on more than one platform.[15] With people increasingly using social media platforms to inform others about their mental states, solicit social support, as well as to keep records of their daily activities, preferences, and interests, social media has emerged as a powerful tool to passively measure behaviors of people[16-17].

Moreover, social media is being increasingly used for communicating about mental health.[18-19], opening an avenue to uncover insights that might be different from data using traditional surveys considering the passive data collection on social media. For example, stressed and depressed individuals use more first-person singular pronouns suggesting higher self-focus and communities with heart disease discuss hate more frequently.[18-20] Natural language processing and machine learning automate the analysis of posts that would have been too hard to evaluate without that automation, and have revealed their value in using social media posts to predict mental health. For example, individual's Facebook posts 6 months immediately preceding the first documented diagnosis of depression yielded a prediction AUC of 0.72.[21] Further, preliminary work studying expressions of loneliness on social media have found associations with feeling unloved, depressed, bored, and not having friends.[16-17] Another opportunity is in the ability of digital platforms to not only provide markers of health but also serve as platforms that can be used for direct intervention.[22-23]

While social media use has also been associated with increased perceived social isolation[24], in this study, we are interested to understand expressions of loneliness as they manifest on social media. Specifically, we sought to characterize individuals' posts about loneliness on Twitter. Studying the language of users who express being lonely or alone, we analyze the correlations between loneliness and users' mental health attributes, and several psycholinguistic attributes inspired from prior work at the intersection of mental health and natural language processing on social media. Privacy of individuals has to be at the forefront of this research to shield unintended use of this data, specifically with the amount of health insights that can be gleaned from social media.

We hypothesize that language usage patterns would both confirm existing understanding of loneliness and give new insights into the daily lives of those who express being lonely. As loneliness can impact health outcomes, identifying ways to track prevalence and manifestations of loneliness online would be useful for developing approaches for identifying and offering support for these individuals.

Methods

This was a retrospective analysis of publicly available data on users posting about loneliness on Twitter. This study was exempt by the University of Pennsylvania Institutional Review Board.

Twitter Data

Twitter is a popular social media platform which allows users to send and receive short 140 character messages, or 'tweets' (at the time of this study; the character limit was later increased to 280). First, from the Twitter Streaming API, we collected tweets from the 1% sample using a bounding box of location coordinates around Pennsylvania. The county of origin of each tweet user was determined. To increase the sample size of tweets from the state, all unique user IDs were recorded, and the Twitter search API was used to extract timelines (each user's prior 3200 tweets) filtered by timestamps ranging from 2012-2016 geolocated in Pennsylvania.

Patient and Public Involvement

Patients and public were not involved in the development of the research question and outcome measures.

Study Sample

We identified users who posted the word "alone" or "lonely" at least once in their timeline (25,966 users). Of these, 6,202 users posted messages with "alone" or "lonely" at least 5 times. As social media includes colloquial, metaphorical, and light-hearted language (eg. "If I see Justin Bieber, I will have a heart attack") we sought to identify the proportion of tweets in which lonely seemed to refer to the public health meaning rather than other uses of the term (e.g. metaphor, joke).[25] Two co-authors independently coded a random set of 100 tweets from individuals who used the words lonely/alone at least 5 times in their timeline to identify them as presumed to be associated with the feeling of loneliness or other. The Kappa was 0.70 and we identified that 76% of users' tweets indicate presumably feeling lonely. A few examples are as follows: "i'm feelin real depressed, confused, & lonely", "im always the only up around this time, feeling a lil lonely" and "I'm so Lonely in life :-(I just wish I can have love again it feels so go to be in love with someone whom loves you." Distribution of users with different number of lonely/alone words in their Twitter timeline and the temporal distribution of tweets containing these words is shown in supplementary file.

Control group

We then identified a control group of users by matching each user in the above dataset to another user by age, gender and period of activity (dates of first and last posting on twitter). We obtained the age and gender estimates by using lexica developed previously.[26] Then, we selected users with a minimum of 500 words across all their posts to have sufficient language for linguistic analyses.[27] We excluded non-English, non-US tweets, retweets, and tweets containing 'alone' and/or 'lonely' that were used to identify users in the lonely group in all analyses to identify linguistic features that are actually characteristics of lonelier people -- looking at their entire timeline of tweets. Hereafter, we use 'lonely' group to indicate users who had more than 5 posts with the words 'lonely' or 'alone', and 'control' group to represent the matched set of users who had no such posts.

Deriving language features to characterize individuals expressing loneliness

We used four sets of language features: a) open-vocabulary topics,[28] b) dictionary-based psycholinguistic features,[29] c) mental well-being attributes such as anxiety, depression by applying previously developed statistical models,[30] d) number of drug words and time of posts as past research has shown an association between loneliness and substance use.[11; 13] These language features have been shown to be predictive of several health outcomes, such as depression, schizophrenia, attention deficit hyperactivity disorder (ADHD), and general well-being.[31; 19]

Open-vocabulary: As closed-vocabulary approaches like LIWC include only a small subset of the entire language used on social media, we use an open-vocabulary approach to improve the coverage and find topics that people who express being lonely talk about. Topics consist of clusters of co-occurring words created using Latent Dirichlet Allocation (LDA).[32] The LDA generative model assumes that tweets contain a combination of topics, and that topics are a distribution of words. Since the words in a tweet are known, topics, which are latent variables, can be estimated through Gibbs sampling.[33] We use the Mallet implementation of the LDA algorithm, adjusting one parameter (alpha=5) to favor fewer topics per tweet.[34] All other parameters were kept at their default. An example of such a model is the following sets of words ('tuesday', 'monday', 'wednesday', ...) which clusters together days of the week by exploiting their similar distributional properties across tweets. In our study, two hundred topics were generated using tweets across all users in the dataset of lonely and control users.

Dictionary-based: From each post, we extracted the relative frequency of single words and phrases (consisting of two or three consecutive words). Then, all words used by less than 1% of users were removed from analysis so as to remove uncommonly used words (outliers). Additionally, all messages used to identify our study group were removed prior to further analysis. The Linguistic Inquiry Word Count (LIWC) dictionary is a language-specific, many-to-many mapping of tokens (including words and word stems) and psychologically validate categories. Each category (a curated list of words) is found to be correlated with and also predictive of several psychological traits and outcomes. For each user, we measure the proportion of word tokens that fall into a given LIWC category.

Mental well-being attributes: We used automatic text-regression methods to assign to each user scores on the depression, anxiety and anger facets for users.[30] This model was trained on a sample of over 28,749 users who had taken the International Personality Item Pool Neuroticism-Extraversion-Openness Personality Inventory Revised (IPIP NEO-PI-R) survey that contains the depression, anxiety and anger Facets of the Neuroticism Factor.[30] The machine learning model trained on words and phrases from Facebook posts to predict survey measure of depression, anger and anxiety resulted in a performance of $r = .32$, which is considered high in psychology, especially when measuring internal states.[35] The model was trained using status updates of users from another study[30], and has been shown to generalize to Twitter users.[36]

Use of Drug-words: We also extracted the frequency (aggregated to every user) of most common drug words as used on social media.[37]

Temporal patterns: We determined the frequency of posts across different hours of the day by users in both the lonely and control groups to understand the diurnal patterns in posting.

Identifying differentially expressed language features in the lonely group

We isolated the patterns in users' loneliness expressions using the linguistic attributes and user traits by correlating them with the lonely and control groups. We use logistic regression to distinguish open-vocabulary words, phrases, LIWC categories and topics associated with lonely and control groups and measure the effect size using Cohen's D. Details of the method are described in a previous work[28]. We used Benjamini-Hochberg p-correction and use $p<0.001$ for indicating meaningful correlations and the effect size was measured using Cohen's D. The statistical analysis, data synthesis, and model creation was conducted in 2018-2019.

Predicting the likelihood of posting about loneliness online

We then looked at the feasibility of predicting whether a user is likely to express that they are lonely or not based on their social media language. Automated analysis of social media is accomplished by building predictive models, which use 'features', or variables that have been extracted from social media data. For this analysis we used LIWC and topics as features. Features are then treated as independent variables in an algorithm (Random Forests) to predict the dependent variable of an outcome of interest (e.g., users' saying that they are lonely or not). For cross validation, the predictive model was trained, using Random Forests, on the training set and then evaluated on a test set to avoid overfitting. The prediction performances are reported as Area Under the Receiver Operating Curves (AUC) and several performance metrics on an out-of-sample 5-fold cross validation setting.

Results

Of the 408,296,620 tweets posted by users geo-located in Pennsylvania, USA, 25,966 users with 46,160,774 posts in their timelines, had at least one post with the words 'lonely' or 'alone', and 6,202 users (referred to as 'lonely' group hereafter) with 17,995,084 posts in their timelines, had more than five such posts (Table 1). The lonely group had 1.9 times more posts in the study time period as the control (Table 1). The median estimated age of this cohort was 21 years, and 69% female.

Table 1: Descriptive statistics for the lonely group about loneliness and the control group

Descriptive Statistics of the Dataset		
	Lonely Group (n= 6,202)	Control group (n= 6,202)
Median Age	21 ± 3 yrs	21 ± 3 yrs
# Messages in timelines	17,995,084	9,219,677
# Females	4,400	4,400
# Males	1,802	1,802

*the lonely group is defined as any user posting at least 5 times about loneliness and the control group is defined as any user who does not have any posts about loneliness

Identifying differentially expressed language features in the lonely group

Open vocabulary approach: Analyzing differences in individual words and phrases used across both groups, we observed (Figure 1a) that users in the lonely group referred to themselves (‘myself’ (d=.18), ‘I’ (d=.16)) in their Twitter posts significantly more than the control group. They also posted about relationship issues (‘want_somebody’ (d=.08), ‘no_one_to’ (d=.1), needs and feelings (‘i_just_wanna (d=.12), ‘in_my_feelings’ (d=.1), ‘i_need’ (d=.12), ‘i_cant’ (d=.1)), and included more expletives. Users in the control group (Figure 1b) engaged in a lot more conversations as indicated by ‘<user>’ (d=-.2) (we anonymize ‘@’ mentions in users tweets as ‘<user>’) compared to the lonely group. The control group also posted more about games (’season’ (d=-.09) ,‘coach’ (d=-.07), ‘team’ (d=-.1)) and positivity (‘!’ (d=-.13), ‘awesome’ (d=-

.09), ‘:)’ (d=-.08)). Figure 1 illustrates the words and phrases most prominently associated with the lonely and control groups.

Using topics generated from LDA, we identified the themes which occur more frequently in posts in the lonely group. Posts were about interpersonal relationships (d=.28) (and associated issues (d=.22)), self-reflection (d=.21) (accompanied with wondering about the future (d=.12)), drug/alcohol use (d=.29) (considering them to be the ‘only friend’), insomnia (d=.27), uncontrolled emotions (d=.28) (accompanied by confusion (d=.11)), and psychosomatic symptoms (d=.29). Table 2 shows the effect sizes between most prominent topic distributions and the users who have more than 5 posts with the words lonely or alone.

Table 2: Highly correlated topics with expressions of loneliness.

Topic Theme	**Highly Correlated Words in Topic**	**Effect size (Cohen’s D)**
Interpersonal Relationships	relationships, matter, perfect hurt, feelings, trust, forget	0.281 0.222
Self Reflection	times, changed, lost, i’ve	0.210
Drug/Alcohol Use	smoke, weed, blunt, drugs, drunk	0.298
Psychosomatic Symptoms	bad, stomach, hurt, head, sick	0.296

Insomnia	sleep, awake, tired, bed	0.274
Emotional Dysregulation	people, f***ing, hate, stupid	0.285
Food/Hunger	food, breakfast, eat, pizza, hungry	0.261

* Effect size is measured using Cohen's d. Only significant topics after Benjamini-Hochberg p-correction and use p<0.001 are shown. All these effect sizes are small.

Dictionary-based: Association of LIWC categories with the posts by users in the lonely group are shown in Table 3. Individuals who posted about being alone or lonely used increased self-references (first person pronouns, d=.18), words indicating cognitive processes (including certainty, d=.15, discrepancies, d=.14, differentiation, d=.13 and tentativeness, d=.13), and negative emotions (swearing, d=.11).

Table 3: Association of LIWC categories, mental health attributes, and drug words with expressions of loneliness

Category	Cohen's d*
Pronouns	
1st Person Pronouns	0.18
Cognitive Processes	
Certainty	0.15
Discrepancies	0.15
Differentiation	0.14
Tentativeness	0.13
Negative Emotions	

Swearing	0.11
Mental Well-being	
Depression	0.81
Anger	0.95
Anxiety	0.75
Drug words	
Blunt	0.16
Smoke	0.13
Heroin	0.1

*Only significant categories after Benjamini-Hochberg p-correction and $p<0.001$ are shown.

Mental well-being: Users in the lonely group were more likely to have posts associated with anger (d=.95), depression (d=.81) and anxiety (d=.75) when compared to the control group.

Use of Drug Words: We also identified the distribution of words pertaining to drugs in the posts of users in the lonely group, and these were more likely to reference a blunt (d=.16), smoke (d=.13), and heroin (d=.1), and included prescribed medications for treatment, recreational drug use, and recreational drugs.

Temporal patterns: Users in the lonely group were found to post more during the night (d=.1), shown in Figure 2. We also see themes associated with night-time posting and having difficulty sleeping (d=.27) in the open-vocabulary analysis (Table 2).

Predictive Analysis: Results from the predictive analysis are shown in Table 4. A random forest model using Topics as input features predicted expressions of loneliness in users with an AUC of

.854 (F1 score = 0.778) and LIWC features resulted in AUC of 0.859 (F1 score = 0.777). A combination of LIWC and Topics resulted in the best AUC of 0.863 (F1 score = 0.782).

Table 4: Performance of different features at predicting expressions of loneliness, reported on an out-of-sample 5-fold cross validation setting.

Feature	AUC	F1 Score	Accuracy	Precision	Recall
Topics	0.854	0.778	0.778	0.780	0.778
LIWC	0.859	0.777	0.777	0.778	0.777
LIWC + Topics	0.863	0.782	0.783	0.785	0.783

Discussion

We sought to mine data from a widely used publicly available social network, Twitter, to characterize what and when individuals post about loneliness, association of posts with mental health, and how manifestations of loneliness can be predicted across individuals. This paper has three main findings. First, we identified themes and contexts associated with users posting about loneliness on Twitter. Second, we observed that users posting about loneliness used language associated with linguistic models for anger, depression, and anxiety. Third, posts about loneliness were more likely to occur in the evening or night.

Themes associated with people expressing loneliness on Twitter are consistent with prior literature about substance use, emotional dysregulation, and troubles with relationships. For example, in one study, a high positive correlation was found between alcoholism and groups of lonely people, and lonely people were also found to express negative feelings towards relationships.[38] This expression of negativity related to relationships is likely related to a hypervigilance to social threat, associated with loneliness.[39] Lonely individuals were also reported to focus on overcoming past events as well as showing feelings of helplessness.[38]

Association of the lonely group with linguistic estimates of anger, depression, and anxiety corroborate prior research, showing that loneliness and social isolation influence psychological functioning , specifically the ability to self-regulate emotion.[5-6; 40] Specifically, anxiety, anger, and negative mood were reported as higher in lonely young adults.[41] Tweets by users in the lonely group were more self-focused compared to the control group. Prior researchers have found that “first person singular pronouns are a modest linguistic marker of depression.”[42] Also, previous research has shown that loneliness has been associated with greater self-disclosure in Facebook posts.[43] This presents the potential for early identification and assessment to intervene on loneliness as well as mental health conditions for this group.

Trends in temporal variation in posting may reflect that sleep deprivation can contribute to social withdrawal and loneliness.[44] This finding corroborates prior research associating loneliness with diminished sleep quality.[40] A better understanding of the temporality of posting could inform timing of interventions designed to address loneliness, as well as provide insight for other researchers to test the inter-relationships between loneliness and the motivations for using social media during nighttime.

Loneliness is known to be one of the primary underlying causes and correlates for chronic mental health conditions.[5-6; 45] As loneliness is becoming increasingly recognized as a public health, several groups have taken action to address it. For example, the United Kingdom appointed a Minister for Loneliness who is responsible for addressing loneliness within communities.[46] CareMore, a health plan and delivery system providing care for enrollees in Medicare Advantage and Medicaid health plans in seven states across the U.S., launched the "Togetherness Program" in a clinical setting to address loneliness in elderly patients.[47] Through this work, CareMore reported that participation in exercise programs increased by 56.6%, emergency room utilization decreased by 3.3%, and hospital admissions among participants were 20.8% lower per thousand compared to the "intent to treat population." [48] Additionally, social network interventions targeting loneliness have been found to be effective in reducing social isolation among individuals with severe mental health conditions but these interventions are not included in the treatment plans for individuals with a mental illness.[49-50]

Considering the advantage of large sample sizes and also the association between increased social media usage and individuals expressions of loneliness, it is promising to use natural language processing and machine learning to automatically identify a person expressing loneliness on Twitter to inform interventions targeted at early identification and support for affected and at risk individuals with the caveat that social media users are not representative of a random sample of individuals. To address loneliness will require being able to identify it passively, remotely, and over time. Many people rarely visit a healthcare provider so would miss the opportunity for screening. Approaches for treatment will also need to harness the tools and technologies that are accessible and integrated with the things people use every day (e.g. mobile phones). Future interventions would have to potentially rely on digital phenotyping of loneliness and using digital platforms (e.g. text messaging) to complement human-to-human interaction strategies to treat loneliness.

In this first study, our aim was to characterize loneliness expressions based on users' entire timelines. Future studies could perform a time-series analysis of the temporal variations associated with loneliness expressions. Further, works should also validate whether the characteristics of people who are using the words 'lonely' or 'alone' on Twitter can be used to track community health risks, particularly, the risk of social isolation. Other studies should replicate the findings in this study using more formal ground truth such as surveys and extend this work to investigate if Twitter can potentially map regional hotspots of loneliness to identify problematic loneliness for community public health monitoring.

Limitations and Ethics

The study sample consists of social media users and is not representative of the general population. An estimated 40% of US adults using Twitter are between the ages of 18 and 29, so our analysis is skewed towards younger people.[51] Considering we identified that 76% of users' tweets indicated presumably feeling lonely in the sample we hand coded, posts mentioning the words alone or lonely may have been metaphorical or non sequiturs. Also, considering the inclusion criteria based on number of tweets mentioning alone or lonely, we are potentially selecting users with more posts than the average twitter user. Additionally, Twitter is far from perfect to be used as a diagnostic tool. However, an automated machine learning tool could be a low-cost method to potentially detect elevated loneliness levels in a person who could then be referred to more formal screening methods. Further, the effects presented in this dataset may not be specific to loneliness considering the potential comorbidity with mental health conditions such as depression in this dataset.

The feasibility of social media-based assessments of loneliness expressions (and mental health more broadly) needs further assessment. Privacy of individuals is an ongoing concern, especially with social media users not fully realizing the amount of health insights that can be gleaned by their online posts. Employers and insurance companies, for example, may be motivated to derive these assessments, but could use these insights against those suffering from mental illness. As mental illnesses carry social stigma and may engender discrimination, data protection and ownership frameworks are needed to make sure the data is not used against the users' interest.[52] Further, transparency about which indicators are derived by whom for what purpose should be part of ethical and policy discourse.

There are also open questions around the impact of misclassifications, and how derived mental health indicators can be responsibly integrated into systems of care.[53]

Conclusions

In this study we characterized expressions of loneliness on Twitter at the individual level. Furthermore, we identified specific contexts, themes, and traits in the posts of individuals expressing loneliness on Twitter. As loneliness is a public health challenge, a better understanding of how loneliness is described online can inform tracking of loneliness and interventions targeted at addressing this important public health problem in regards to the behavior of lonely individuals that may be at risk of developing a severe mental health condition.[47]

Acknowledgements

Funding: This project is funded, in part, under a grant with the Pennsylvania Department of Health (ORG-FUND-PROG-CREF is 4290-567862-2446-2049). No sponsor of funding source played a role in: "study design and the collection, analysis, and interpretation of data and the writing of the article and the decision to submit it for publication." All researchers are independent from funders

Data Sharing Statement: Because of our IRB requirements, data will be shared upon request from the corresponding author.

Contributors: S.C. Guntuku and R. Merchant originated the study. S.C. Guntuku, R. Schneider, A. Pelullo, L.H. Ungar, and R. Merchant developed methods, interpreted analysis, and contributed to the writing of the article. J.F. Young, V. Wong, D. Polsky, and K. Volpp assisted with the interpretation of the findings and contributed to the writing of the article.

Disclosures: None

References:

(1) Gerst-Emerson K, Jayawardhana J. Loneliness as a Public Health Issue: The Impact of Loneliness on Health Care Utilization Among Older Adults. *American Journal of Public Health*. 2015;105(5):1013-1019. doi:10.2105/ajph.2014.302427.

(2) New Cigna Study Reveals Loneliness at Epidemic Levels in America. Cigna, a Global Health Insurance and Health Service Company. https://www.cigna.com/newsroom/news-releases/2018/pdf/new-cigna-study-reveals-loneliness-at-epidemic-levels-in-america.pdf Accessed February 18, 2019.

(3) Jong-Gierveld JD. Developing and testing a model of loneliness. *Journal of Personality and Social Psychology*. 1987;53(1):119-128. doi:10.1037//0022-3514.53.1.119.

(4) Peplau LA, Perlman D. *Loneliness: a Sourcebook of Current Theory, Research, and Therapy*. New York: Wiley; 1982.

(5) Stravynski A, Boyer R. Loneliness in Relation to Suicide Ideation and Parasuicide: A Population-Wide Study. *Suicide and Life-Threatening Behavior*. 2001;31(1):32-40. doi:10.1521/suli.31.1.32.21312.

(6) Blai B. Health Consequences of Loneliness: A Review of the Literature. *Journal of American College Health*. 1989;37(4):162-167. doi:10.1080/07448481.1989.9938410.

(7) Heinrich LM, Gullone E. The clinical significance of loneliness: A literature review. *Clinical Psychology Review*. 2006;26(6):695-718. doi:10.1016/j.cpr.2006.04.002.

(8) Richard A, Rohrmann S, Vandeleur CL, Schmid M, Barth J, Eichholzer M. Loneliness is adversely associated with physical and mental health and lifestyle factors: Results from a Swiss national survey. *Plos One*. 2017;12(7). doi:10.1371/journal.pone.0181442.

(9) Rico-Uribe LA, Caballero FF, Olaya B, et al. Loneliness, Social Networks, and Health: A Cross-Sectional Study in Three Countries. *Plos One*. 2016;11(1). doi:10.1371/journal.pone.0145264.

(10) Pinquart M, Sorensen S. Influences on Loneliness in Older Adults: A Meta-Analysis. *Basic and Applied Social Psychology*. 2001;23(4):245-266. doi:10.1207/S15324834BASP2304_2.

(11) Rokach A. Determinants of Loneliness of Young Adult Drug Users. *The Journal of Psychology*. 2002;136(6):613-630. doi:10.1080/00223980209604823.

(12) Vanhalst J, Luyckx K, Scholte RHJ, Engels RCME, Goossens L. Low Self-Esteem as a Risk Factor for Loneliness in Adolescence: Perceived - but not Actual - Social Acceptance as an Underlying Mechanism. *Journal of Abnormal Child Psychology*. 2013;41(7):1067-1081. doi:10.1007/s10802-013-9751-y.

(13) Page RM, Allen O, Moore L, *et al.* Co-Occurrence of Substance Use and Loneliness as a Risk Factor for Adolescent Hopelessness. *Journal of School Health* 1993;**63**:104–8. doi:10.1111/j.1746-1561.1993.tb06090.x

(14) U.S. population with a social media profile 2018. Statista. https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/. Accessed February 18, 2019.

(15) Lenhart A. Mobile Access Shifts Social Media Use and Other Online Activities. Pew Research Center: Internet, Science & Tech. 2015.http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-other-online-activities/ (accessed 9 May2019).

(16) Kivran-Swaine F, Ting J, Naaman M. Understanding Loneliness in Social Awareness Streams: Expressions and Responses. *ICWSM 2014*. 2014. https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef.

(17) Kamvar S, Kamvar S, Harris JJ. *We Feel Fine: an Almanac of Human Emotion*. New York: Scribner; 2009.

(18) Guntuku SC, Buffone A, Jaidka K, Eichstaedt J, Ungar L. Understanding and Measuring Psychological Stress using Social Media. *ICWSM 2019*. 2018.

(19) Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*. 2017;18:43-49. doi:10.1016/j.cobeha.2017.07.005.

(20) Eichstaedt JC, Schwartz HA, Kern ML, *et al.* Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science* 2015;**26**:159–69. doi:10.1177/0956797614557867

(21) Eichstaedt JC, Smith RJ, Merchant RM, *et al.* Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences* 2018;**115**:11203–8. doi:10.1073/pnas.1802331115

(22) Coppersmith G, Ngo K, Leary R, Wood A. Exploratory Analysis of Social Media Prior to a Suicide Attempt. *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. 2016. doi:10.18653/v1/w16-0311.

(23) Qntfy, Inc. OurDataHelps. OurDataHelps. https://ourdatahelps.org/. Accessed February 18, 2019.

(24) Primack BA, Shensa A, Sidani JE, et al. Social Media Use and Perceived Social Isolation Among Young Adults in the U.S. *American Journal of Preventive Medicine*. 2017;53(1):1-8. doi:10.1016/j.amepre.2017.01.010.

(25) Sinnenberg L, DiSilvestro CL, Mancheno C, et al. Twitter as a Potential Data Source for Cardiovascular Disease Research. *JAMA Cardiology*. December 2016. https://jamanetwork.com/journals/jamacardiology/fullarticle/2556216.

(26) Sap M, Park G, Eichstaedt J, et al. Developing Age and Gender Predictive Lexica over Social Media. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014. doi:10.3115/v1/d14-1121.

(27) Jaidka K, Guntuku SC, Buffone A, Schwartz HA, Ungar L. Facebook versus Twitter: Differences in Self-Disclosure and Trait Prediction. *ICWSM*. 2018.

(28) 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). doi:10.1371/journal.pone.0073791.

(29) Pennebaker Jw, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works*. 2014:1-22.

(30) Schwartz HA, Eichstaedt J, Kern ML, et al. Towards Assessing Changes in Degree of Depression through Facebook. *Proceedings of the Workshop on Computational*

Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. 2014. doi:10.3115/v1/w14-3214.

(31) Guntuku SC, Ramsay JR, Merchant RM, Ungar LH. Language of ADHD in Adults on Social Media. *Journal of Attention Disorders*. 2017:108705471773808. doi:10.1177/1087054717738083.

(32) Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *The Journal of Machine Learning Research*. 2003;3.

(33) Gelfand AE, Smith AF. Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 1990;**85**:398–409. doi:10.21236/ada208388

(34) McCallum A. Mallet: A machine learning for language toolkit. MALLET Machine Learning for Language Toolkit. 2002.http://mallet.cs.umass.edu/ (accessed 9 May2019).

(35) Meyer GJ, Finn SE, Eyde LD, et al. Psychological testing and psychological assessment. A review of evidence and issues. *The American psychologist*. 2001;56(2):128-165.

(36) Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, *et al.* What Twitter Profile and Posted Images Reveal About Depression and Anxiety. *ICWSM* Published Online First: 2019.https://arxiv.org/pdf/1904.02670.pdf

(37) Jones R. Drug Slang Dictionary - Words Starting With G. http://www.noslang.com/drugs/dictionary.php. Accessed February 19, 2019.

(38) Booth R. Toward an Understanding of Loneliness. *Social Work*. 1983;28(2):116-119. doi:10.1093/sw/28.2.116.

(39) Qualter P, Vanalst J, Harris R, *et al.* Loneliness across the life span. *Perspectives on psychological science : a journal of the Association for Psychological Science* 2015;**10**:250–64.https://journals.sagepub.com/doi/full/10.1177/1745691615568999?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub=pubmed

(40) Hawkley LC, Cacioppo JT. Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms. *Annals of Behavioral Medicine* 2010;**40**:218–27. doi:10.1007/s12160-010-9210-8

(41) Cacioppo JT, Hawkley LC, Ernst JM, et al. Loneliness within a nomological net: An evolutionary perspective. *Journal of Research in Personality*. 2006;40(6):1054-1085. doi:10.1016/j.jrp.2005.11.007.

(42) Edwards T, Holtzman NS. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research*

(43) Al-Saggaf Y, Nielsen S. Self-disclosure on Facebook among female users and its relationship to feelings of loneliness. *Computers in Human Behavior* 2014;**36**:460–8. doi:10.1016/j.chb.2014.04.014

(44) Simon EB, Walker MP. Sleep loss causes social withdrawal and loneliness. *Nature Communications*. 2018;9(1). doi:10.1038/s41467-018-05377-0.

(45) Cacioppo JT, Hawkley LC, Thisted RA. Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago Health, Aging, and Social Relations Study. *Psychology and Aging* 2010;**25**:453–63. doi:10.1037/a0017216

(46) Yeginsu C. U.K. Appoints a Minister for Loneliness. *The New York Times*. https://www.nytimes.com/2018/01/17/world/europe/uk-britain-loneliness.html. Published January 17, 2018.

(47) Rubin R. Loneliness Might Be a Killer, but What's the Best Way to Protect Against It? *Jama*. 2017;318(19):1853. doi:10.1001/jama.2017.14591.

(48) CareMore Health Announces New Outcomes Data from First-of-its-Kind Togetherness Program. Business Wire A Berkshire Hathaway Company . https://www.businesswire.com/news/home/20181218005059/en/. Published December 18, 2018. Accessed February 18, 2019.

(49) Perese EF, Wolf M. Combating Loneliness Among Persons With Severe Mental Illness: Social Network

Interventions Characteristics, Effectiveness, And Applicability. *Issues in Mental Health Nursing*. 2005;26(6):591-609. doi:10.1080/01612840590959425.

(50) Webber M, Fendt-Newlin M. A review of social participation interventions for people with mental health problems. *Social Psychiatry and Psychiatric Epidemiology* 2017;**52**:369–80. doi:10.1007/s00127-017-1372-2

(51) U.S. Twitter reach by age group 2018 | Statistic. Statista. https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/. Accessed February 18, 2019.

(52) Mckee R. Ethical issues in using social media for health and health care research. *Health Policy*. 2013;110(2-3):298-301. doi:10.1016/j.healthpol.2013.02.006.

(53) Inkster B, Stillwell D, Kosinski M, Jones P. A decade into Facebook: where is psychiatry in the digital age? *The Lancet Psychiatry*. 2016;3(11):1087-1090. doi:10.1016/s2215-0366(16)30041-4.

Figure legends

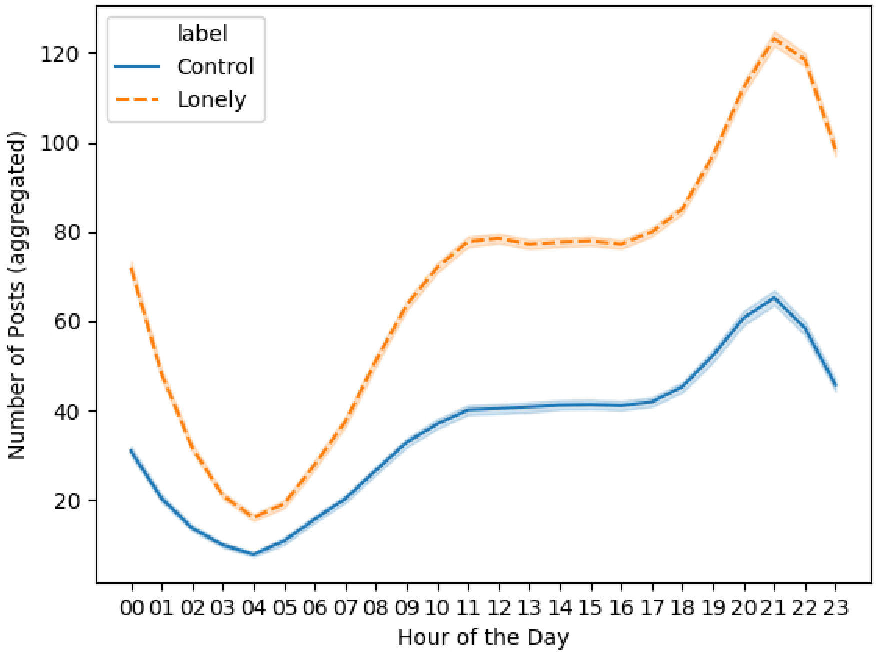
Figure 1: Words/Phrases more likely to be posted by Twitter users with a) self-reported loneliness (Individuals with at least 5 posts with the words 'lonely' or 'alone' group compared to the b) control group.

Word size indicates the strength of the correlation and word color indicates relative word frequency. ($p<0.01$, Bonferroni p-corrected)

Figure 2: Temporal variation showing diurnal patterns of post frequency of both the 'lonely' and 'control' groups.

The dotted line indicates the percentage of posts at different hours of the day by the group of users with at least 5 posts containing the word 'lonely' or 'alone' and the solid line indicates users who do not have any posts about loneliness. The x-axis represents the hour of the day each post occurs and the y-axis indicates the number of posts for each group.

Words/Phrases more likely to be posted by Twitter users with a) self-reported loneliness (Individuals with at least 5 posts with the words ‘lonely’ or ‘alone’ group compared to the b) control group.



Temporal variation showing diurnal patterns of post frequency of both the ‘lonely’ and ‘control’ groups.

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **(pg.2)**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **(pg.2)**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **(pg.4)**
Objectives	3	State specific objectives, including any prespecified hypotheses **(pg. 4)**
Methods		
Study design	4	Present key elements of study design early in the paper **(pg.5)**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **(pg. 5)**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up **(pg. 6)**
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **(pg. 6)**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **(pg. 6)**
Bias	9	Describe any efforts to address potential sources of bias **(pg. 6)**
Study size	10	Explain how the study size was arrived at **(pg. 6)**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **(pg. 7)**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **(pg. 9)**
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) If applicable, explain how loss to follow-up was addressed
		(*e*) Describe any sensitivity analyses
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **(pg. 9)**
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **(pg. 9)**
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarise follow-up time (eg, average and total amount)
Outcome data	15*	Report numbers of outcome events or summary measures over time **(pgs 10,11)**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **(pgs 10,11)**

		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **(pgs 10,11)**
Discussion		
Key results	18	Summarise key results with reference to study objectives **(pgs. 12, 13)**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **(pgs. 14)**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **(pg. 13, 14)**
Generalisability	21	Discuss the generalisability (external validity) of the study results **(pgs. 13, 14)**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **(pg. 15)**

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

BMJ Open

Studying Expressions of Loneliness in Individuals using Twitter: An Observational Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-030355.R2
Article Type:	Original research
Date Submitted by the Author:	14-Aug-2019

Complete List of Authors:	Guntuku, Sharath Chandra; University of Pennsylvania, Department of Emergency Medicine Schneider, Rachelle; University of Pennsylvania, Department of Emergency Medicine Pelullo, Arthur; University of Pennsylvania, Department of Emergency Medicine Young, Jami; University of Pennsylvania, Department of Emergency Medicine Wong, Vivien; University of Pennsylvania, Department of Emergency Medicine Ungar, Lyle ; University of Pennsylvania, Department of Emergency Medicine Polsky, Daniel ; Leonard Davis Institute of Health Economics, Volpp, Kevin ; University of Pennsylvania Perelman School of Medicine, Medical Ethics and Policy Merchant, Raina; University of Pennsylvania, Department of Emergency Medicine

Primary Subject Heading:	Public health
Secondary Subject Heading:	Communication, Mental health
Keywords:	loneliness, social media, natural language processing, MENTAL HEALTH, STATISTICS & RESEARCH METHODS

Studying Expressions of Loneliness in Individuals using Twitter: An Observational Study

Sharath Chandra Guntuku, PhD[1,4,5], Rachelle C. Schneider, BS [1,5], Arthur Pelullo, MS[1,4,5], Jami F. Young, PhD[5,7], Vivien Wong, BS[1,5], Lyle H. Ungar, PhD[3,4], Daniel Polsky, PhD[5,6], Kevin Volpp, MD, PhD[5,6], Raina M. Merchant, MD, MSHP[1,2,5]

[1]Penn Medicine Center for Digital Health, Philadelphia, PA 19104

[2]Penn Medicine Center for Healthcare Innovation, Philadelphia, PA 19104

[3]Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104

[4]Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

[5]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

[6]The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

[7]Children's Hospital of Philadelphia, Philadelphia, PA 19146

Corresponding author and request for reprints

Sharath Chandra Guntuku

3400 Civic Center Blvd

Philadelphia PA 19104

Email: sharathg@sas.upenn.edu

Word count: 3723, 31 pages, 4 tables, 2 figures

Keywords: loneliness mentions; social media; twitter; natural language processing; mental health

Abstract

Objectives: Loneliness is a major public health problem and an estimated 17% of adults aged 18-70 in the United States are classified as lonely. We sought to characterize the (online) lives of people who mention the words ‘lonely’ or ‘alone’ in their Twitter timeline and correlate their posts with predictors of mental health.

Setting and design: A leading social media platform (Twitter) was the main focus of the study. We collected approximately 400 million tweets from in Pennsylvania, USA, between 2012-2016. We identified users whose posts contained the words ‘lonely’ or ‘alone’ and compared them to a control group matched by age, gender, and period of posting. Using natural-language processing, we characterized what and when users post, their association with linguistic markers of mental health, and if language can predict manifestations of loneliness. The statistical analysis, data synthesis, and model creation was conducted in 2018-2019.

Primary outcome measures: We evaluated counts of language features in the users with posts including the words lonely or alone compared to the control group. These language features were measured by (1) open-vocabulary topics and (2) linguistic markers of anger, depression, and anxiety. We also evaluated the prediction of mentions of loneliness compared to the control group, measured by Area Under Curve.

Results: Twitter timelines of users with posts including the words lonely or alone (N=6202) were found to include themes about difficult interpersonal relationships, psychosomatic symptoms, substance use, wanting change, unhealthy eating, and having troubles with sleep. Their posts were also associated with linguistic markers of anger, depression, and anxiety. A random forest model predicted mentions of loneliness online with an accuracy of 77%.

Conclusions: Posts with the words lonely or alone often include psychosocial features and can potentially have associations with how individuals presumably express and experience loneliness. This can inform online surveillance for high risk individuals experiencing loneliness and interventions focused on addressing morbidity in this condition.

Strengths and Limitations of this study

- Novel focus on timelines of social media users to study mentions of loneliness and correlation with predictors of mental health.
- The study sample consists of social media users and is not representative of the general population.
- Though we manually annotated a subset of posts mentioning loneliness, some may have been metaphorical or non sequiturs.

Introduction

Loneliness is a major public health problem and an estimated 17% of adults aged 18-70 in the United States are classified as lonely.[1] Loneliness is defined as the discrepancy between a person's desired and actual social relationships and has been linked with an increased risk of heart disease, stroke, dementia, depression, and anxiety.[1-5] Loneliness is also one of the primary underlying causes and correlates for chronic mental health conditions and physician visits in some populations.[1, 5-9]

Reducing morbidity from loneliness requires identifying who experiences it. Traditionally this has occurred through surveys but this approach is limited by the ability to access broad populations initially and over time.[10] Social media has emerged as a tool that individuals use to share information about their mental states, solicit social support, record daily activities, and report preferences, and interests.[11-12] Social media use seeks to connect people but it also has been associated with increased perceived social isolation.[13] It is unclear if social media use causes perceived social isolation or if perceived social isolation causes social media use.

With people increasingly using social media platforms to inform others about their mental states, solicit social support, as well as to keep records of their daily activities, preferences, and interests, social media has emerged as a potentially relevant tool to passively measure health states and behaviors of people.[14-15] For example, individuals who are stressed and depressed use more first-person singular pronouns suggesting higher self-focus and communities with heart disease discuss hate more frequently.[11-12; 16] Natural language processing and machine learning have revealed their value in using social media posts to predict first documented diagnosis of depression using posts 6 months prior yielding an AUC of 0.72.[17]

While the use of social media is increasingly common, less is known about how often individuals use the platform to explicitly share about feelings of loneliness or being alone.[13] In this study, we sought to characterize Twitter timelines of individuals' whose posts include the words lonely or alone. Studying the language of users who use these terms, we analyzed the correlations between posting about loneliness and users' mental health and psycholinguistic attributes (e.g. anger and depression). This has the potential to further our understanding of how social media platforms are used for mentions of loneliness and if there is an opportunity to use these platforms for surveillance of an important but hard to track and measure condition that impacts public health. However, privacy of individuals has to be at the forefront of this research to shield unintended use of this data, specifically with the amount of health insights that can be gleaned from social media.

We hypothesize that language usage patterns would both confirm existing understanding of loneliness and give new insights into the daily lives of those who express being lonely. As loneliness can impact health outcomes, identifying ways to track prevalence and manifestations of loneliness online would be useful for developing approaches for identifying and offering support for these individuals. This presents the opportunity of digital platforms to not only provide markers of health but also potentially serve as platforms that can be used for developing interventions.[18-19]

Methods

This was a retrospective analysis of publicly available data on users posting about loneliness on Twitter. This study was exempt by the University of Pennsylvania Institutional Review Board.

Twitter Data

Twitter is a popular social media platform which allows users to send and receive short 140 character messages, or 'tweets' (at the time of this study; the character limit was later increased to 280). First, from the Twitter Streaming API, we collected tweets from the 1% sample using a bounding box of location coordinates around Pennsylvania. The county of origin of each tweet user was determined. To increase the sample size of tweets from the state, all unique user IDs were recorded, and the Twitter search API was used to extract timelines (each user's prior 3200 tweets) filtered by timestamps ranging from 2012-2016 geolocated in Pennsylvania.

Patient and Public Involvement

Patients and public were not involved in the development of the research question and outcome measures.

Study Sample

We identified users who posted the word "alone" or "lonely" at least once in their timeline (25,966 users). Of these, 6,202 users posted messages with "alone" or "lonely" at least 5 times. As social media includes colloquial, metaphorical, and light-hearted language (eg. "If I see Justin Bieber, I will have a heart attack") we sought to identify the proportion of tweets in which lonely seemed to refer to the public health meaning rather than other uses of the term (e.g. metaphor, joke).[20] Two co-authors independently coded a random set of 100 tweets from individuals who used the words lonely/alone at least 5 times in their timeline to identify them as presumed to be associated with the feeling of loneliness or other. The Kappa was 0.70 and we identified that 76% of users' tweets indicate presumably feeling lonely. A few examples are as follows: "i'm feelin real depressed, confused, & lonely", "im always the only up around this time, feeling a lil lonely" and "I'm so Lonely in life :-(I just wish I can have love again it feels so go to be in love with someone whom loves you."

Control group

We then identified a control group of users by matching each user in the above dataset to another user by age, gender and period of activity (dates of first and last posting on twitter). We obtained the age and gender estimates by using lexica developed previously.[21] Then, we selected users with a minimum of 500 words across all their posts to have sufficient language for linguistic analyses.[22] We excluded non-English, non-US tweets, retweets, and tweets containing 'alone' and/or 'lonely' that were used to identify users who had more than 5 posts with the words 'lonely' or 'alone in all analyses to identify linguistic features that are actually characteristics of lonelier people -- looking at their entire timeline of tweets. Hereafter, we indicate users who had more than 5 posts with the words 'lonely' or 'alone' as 'users with posts including the words lonely or alone', and 'control' group to represent the matched set of users who had no such posts.

Deriving language features to characterize individuals expressing loneliness

We used four sets of language features: a) open-vocabulary topics,[23] b) dictionary-based psycholinguistic features,[24] c) mental well-being attributes such as anxiety, depression by applying previously developed statistical models,[25] d) number of drug words and time of posts as past research has shown an association between loneliness and substance use.[26; 12] These language features have been shown to be predictive of several health outcomes, such as depression, schizophrenia, attention deficit hyperactivity disorder (ADHD), and general well-being.[27; 28]

Open-vocabulary: As closed-vocabulary approaches like LIWC include only a small subset of the entire language used on social media, we use an open-vocabulary approach to improve the coverage and find topics that people who mention loneliness. Topics consist of clusters of co-occurring words created using Latent Dirichlet Allocation (LDA).[29] The LDA generative model assumes that tweets contain a combination of topics, and that topics are a distribution of words. Since the words in a tweet are known, topics, which are latent variables, can be estimated through Gibbs sampling.[30] We use the Mallet implementation of the LDA algorithm, adjusting one parameter (alpha=5) to favor fewer topics per tweet.[31] All other parameters were kept at their default. An example of such a model is the following sets of words ('tuesday', 'monday', 'wednesday', ...) which clusters together days of the week by exploiting their similar distributional properties across tweets. In our study, two hundred topics were generated using tweets across all users in the dataset of users with posts including the words lonely or alone and control users.

Dictionary-based: From each post, we extracted the relative frequency of single words and phrases (consisting of two or three consecutive words). Then, all words used by less than 1% of users were removed from analysis so as to remove uncommonly used words (outliers).

Additionally, all messages used to identify our study group were removed prior to further analysis. The Linguistic Inquiry Word Count (LIWC) dictionary is a language-specific, many-to-many mapping of tokens (including words and word stems) and psychologically validate categories. Each category (a curated list of words) is found to be correlated with and also predictive of several psychological traits and outcomes. For each user, we measure the proportion of word tokens that fall into a given LIWC category.

Mental well-being attributes: We used automatic text-regression methods to assign to each user scores on the depression, anxiety and anger facets for users.[25] This model was trained on a sample of over 28,749 users who had taken the International Personality Item Pool Neuroticism-Extraversion-Openness Personality Inventory Revised (IPIP NEO-PI-R) survey that contains the depression, anxiety and anger Facets of the Neuroticism Factor.[25] The machine learning model trained on words and phrases from Facebook posts to predict survey measure of depression, anger and anxiety resulted in a performance of $r = .32$, which is consistent with other reports of mental health states identified via social media.[32] The model was trained using status updates of users from another study[25], and has been shown to generalize to Twitter users.[33]

Use of Drug-words: We also extracted the frequency (aggregated to every user) of most common drug words as used on social media.[34]

Temporal patterns: We determined the frequency of posts across different hours of the day by users in both users with posts including the words lonely or alone and control groups to understand the diurnal patterns in posting.

Identifying differentially expressed language features in users with posts including the words lonely or alone

We isolated the patterns in users' loneliness mentions using the linguistic attributes and user traits by correlating them with users with posts including the words lonely or alone and control groups. We used logistic regression to distinguish open-vocabulary words, phrases, LIWC categories and topics associated with lonely and control groups and measure the effect size using Cohen's D. The models were set up to predict the group of users with posts including the words lonely or alone against the control group (e.g., group was the dependent variable). Details of the method are described in a previous work[23]. For identifying themes from topics, researchers looked at 20 messages each with the highest topic prevalence to identify themes. We used Benjamini-Hochberg p-correction and use $p<0.001$ for indicating meaningful correlations and the effect size was measured using Cohen's D. The statistical analysis, data synthesis, and model creation was conducted in 2018-2019.

Predicting the likelihood of posting about loneliness online

We then looked at the feasibility of predicting whether a user is likely to mention loneliness or not based on their social media language. Automated analysis of social media is accomplished by building predictive models, which use 'features', or variables that have been extracted from social media data. For this analysis we used LIWC and topics as features. Features are then treated as independent variables in an algorithm (Random Forests) to predict the dependent variable of an outcome of interest (e.g., users' saying that they are lonely or not). For cross validation, the predictive model was trained, using Random Forests, on the training set and then evaluated on a test set to avoid overfitting. The prediction performances are reported as Area Under the Receiver Operating Curves (AUC) on an out-of-sample 5-fold cross validation setting.

Results

Of the 408,296,620 tweets posted by users geo-located in Pennsylvania, USA, 25,966 users with 46,160,774 posts in their timelines, had at least one post with the words 'lonely' or 'alone', and 6,202 users with 17,995,084 posts in their timelines, had more than five such posts (Table 1). Users with posts including the words lonely or alone had 1.9 times more posts in the study time period as the control (Table 1). The median estimated age of this cohort was 21 years, and 69% female.

Table 1: Descriptive statistics for users with posts including the words lonely or alone and the control group

Descriptive Statistics of the Dataset		
	Users with posts including the words lonely or alone (n= 6,202)	Control group (n= 6,202)
Median Age	21 ± 3 yrs	21 ± 3 yrs
# Messages in timelines	17,995,084	9,219,677
# Females	4,400	4,400
# Males	1,802	1,802

*users with posts including the words lonely or alone is defined as any user posting at least 5 times about loneliness and the control group is defined as any user who does not have any posts about loneliness

Identifying differentially expressed language features in users with posts including the words lonely or alone

Open vocabulary approach: Analyzing differences in individual words and phrases used across both groups, we observed (Figure 1a) that users with posts including the words lonely or alone referred to themselves ('myself' (d=.18), 'I' (d=.16)) in their Twitter posts significantly more than the control group. They also posted about relationship issues ('want_somebody' (d=.08), 'no_one_to' (d=.1), needs and feelings ('i_just_wanna (d=.12), 'in_my_feelings' (d=.1), 'i_need' (d=.12), 'i_cant' (d=.1)), and included more expletives. Users in the control group (Figure 1b) engaged in a lot more conversations as indicated by '<user>' (d=-.2) (we anonymize '@' mentions in users tweets as '<user>') compared to users with posts including the words lonely or alone. The control group also posted more about games ('season' (d=-.09) ,'coach' (d=-.07), 'team' (d=-.1)) and positivity ('!' (d=-.13), 'awesome' (d=-.09), ':)' (d=-.08)). Figure 1 illustrates the words and phrases most prominently associated with the group of users with posts including the words lonely or alone and the control group.

Using topics generated from LDA, we identified the themes which occur more frequently in posts of users with posts including the words lonely or alone. Posts were about interpersonal relationships (d=.28) (and associated issues (d=.22)), self-reflection (d=.21) (accompanied with wondering about the future (d=.12)), drug/alcohol use (d=.29) (considering them to be the 'only friend'), insomnia (d=.27), uncontrolled emotions (d=.28) (accompanied by confusion (d=.11)), and psychosomatic symptoms (d=.29). Table 2 shows the effect sizes between most prominent topic distributions and the users who have more than 5 posts with the words lonely or alone.

Table 2: Highly correlated topics with mentions of loneliness.

Topic Theme	**Highly Correlated Words in Topic**	**Effect size (Cohen's D)**
Interpersonal Relationships	relationships, matter, perfect hurt, feelings, trust, forget	0.28 0.22
Self Reflection	times, changed, lost, i've	0.21
Drug/Alcohol Use	smoke, weed, blunt, drugs, drunk	0.29
Psychosomatic Symptoms	bad, stomach, hurt, head, sick	0.29
Insomnia	sleep, awake, tired, bed	0.27
Emotional Dysregulation	people, f***ing, hate, stupid	0.28
Food/Hunger	food, breakfast, eat, pizza, hungry	0.26

* Effect size is measured using Cohen's d. Only significant topics after Benjamini-Hochberg p-correction and use p<0.001 are shown.

Dictionary-based: Association of LIWC categories of users with posts including the words lonely or alone are shown in Table 3. Individuals who had posts including the word lonely or alone used increased self-references (first person pronouns, d=.18), words indicating cognitive processes (including certainty, d=.15, discrepancies, d=.14, differentiation, d=.13 and tentativeness, d=.13), and negative emotions (swearing, d=.11).

Table 3: Association of LIWC categories, mental health attributes, and drug words with mentions of loneliness

Category	**Cohen's D***
Pronouns	
1st Person Pronouns	0.18
Cognitive Processes	
Certainty	0.15
Discrepancies	0.15
Differentiation	0.14
Tentativeness	0.13
Negative Emotions	
Swearing	0.11
Mental Well-being	
Depression	0.81
Anger	0.95
Anxiety	0.75
Drug words	
Blunt	0.16
Smoke	0.13
Heroin	0.1

*Only significant categories after Benjamini-Hochberg p-correction and p<0.001 are shown.

Mental well-being: Users with posts including the words lonely or alone were more likely to have posts associated with anger (d=.95), depression (d=.81) and anxiety (d=.75) when compared to the control group.

Use of Drug Words: We also identified the distribution of words pertaining to drugs in the posts of users with posts including the words lonely or alone, and these were more likely to reference a blunt (d=.16), smoke (d=.13), and heroin (d=.1), and included prescribed medications for treatment, recreational drug use, and recreational drugs.

Temporal patterns: Users with posts including the words lonely or alone were found to post more during the night (d=.1), shown in Figure 2. We also see themes associated with night-time posting and having difficulty sleeping (d=.27) in the open-vocabulary analysis (Table 2).

Predictive Analysis: Table 4 shows that random forest model using Topics as input features predicted mentions of loneliness in users with an AUC of .854 (F1 score = 0.778) and LIWC features resulted in AUC of 0.859 (F1 score = 0.777). A combination of LIWC and Topics resulted in the best AUC of 0.863 (F1 score = 0.782).

Table 4: Performance of different features at predicting mentions of loneliness, reported on an out-of-sample 5-fold cross validation setting.

Feature	AUC	F1 Score	Accuracy	Precision	Recall
Topics	0.854	0.778	0.778	0.780	0.778
LIWC	0.859	0.777	0.777	0.778	0.777
LIWC + Topics	0.863	0.782	0.783	0.785	0.783

Discussion

We sought to mine data from a widely used publicly available social network, Twitter, to characterize what and when individuals post about loneliness, association of posts with mental health, and how manifestations of loneliness can be predicted across individuals. Our fundamental hypothesis was that the language of users with posts including the words lonely or alone would be significantly different from matched controls, that this language would reveal differences in characteristics such as mental health attributes between both groups, and that the language usage patterns would both confirm existing understanding of loneliness and give new insights into the daily lives of those who post the words alone or lonely. Towards this goal, we took an inductive approach of computationally analysing the large volumes of social media data with the aim of better understanding the varying manifestations of loneliness. This paper has three main findings. First, we identified themes and contexts associated with users posting about loneliness on Twitter. Second, we observed that users posting about loneliness used language associated with linguistic models for anger, depression, and anxiety. Third, posts about loneliness were more likely to occur in the evening or night.

Themes associated with people mentioning loneliness on Twitter are consistent with prior literature about substance use, emotional dysregulation, and troubles with relationships. For example, in one study, a high positive correlation was found between alcoholism and groups of lonely people, and lonely people were also found to express negative feelings towards relationships.[35] This expression of negativity related to relationships is likely related to a hypervigilance to social threat, associated with loneliness.[36] Lonely individuals were also reported to focus on overcoming past events as well as showing feelings of helplessness.[35]

Researchers who coded the topics were attempting to identify these associations by looking at 20 messages each with the highest topic prevalence to identify themes, and we acknowledge that this can be subjective.

Association of users with posts including the words lonely or alone with linguistic estimates of anger, depression, and anxiety corroborate prior research, showing that loneliness and social isolation influence psychological functioning , specifically the ability to self-regulate emotion.[5-6; 37] Specifically, anxiety, anger, and negative mood were reported as higher in lonely young adults.[38] Tweets by users with posts including the words lonely or alone were more self-focused compared to the control group. Prior researchers have found that "first person singular pronouns are a modest linguistic marker of depression."[39] Also, previous research has shown that loneliness has been associated with greater self-disclosure in Facebook posts.[40] This presents the potential for early identification and assessment to intervene on loneliness as well as mental health conditions for this group.

Trends in temporal variation in posting may reflect that sleep deprivation can contribute to social withdrawal and loneliness.[41] This finding corroborates prior research associating loneliness with diminished sleep quality.[37] A better understanding of the temporality of posting could inform timing of interventions designed to address loneliness, as well as provide insight for other researchers to test the inter-relationships between loneliness and the motivations for using social media during nighttime.

Loneliness is known to be one of the primary underlying causes and correlates for chronic mental health conditions.[5-6; 42] As loneliness is becoming increasingly recognized as a public health, several groups have taken action to address it. For example, the United Kingdom appointed a Minister for Loneliness who is responsible for addressing loneliness within communities.[43] CareMore, a health plan and delivery system providing care for enrollees in Medicare Advantage and Medicaid health plans in seven states across the U.S., launched the “Togetherness Program” in a clinical setting to address loneliness in elderly patients.[44] Through this work, CareMore reported that participation in exercise programs increased by 56.6%, emergency room utilization decreased by 3.3%, and hospital admissions among participants were 20.8% lower per thousand compared to the “intent to treat population.” [45] Additionally, social network interventions targeting loneliness have been found to be effective in reducing social isolation among individuals with severe mental health conditions but these interventions are not included in the treatment plans for individuals with a mental illness.[46-47]

Considering the advantage of large sample sizes and also the association between increased social media usage and individuals mentions of loneliness, it is promising to use natural language processing and machine learning to automatically identify a person mentions the words alone or lonely on Twitter to inform interventions targeted at early identification and support for affected and at risk individuals with the caveat that social media users are not representative of a random sample of individuals. To address loneliness will require being able to identify it passively, remotely, and over time. Many people rarely visit a healthcare provider so would miss the opportunity for screening. Approaches for treatment will also need to harness the tools and technologies that are accessible and integrated with the things people use every day (e.g. mobile phones). Future interventions would have to potentially rely on digital phenotyping of loneliness and using digital platforms (e.g. text messaging) to complement human-to-human interaction strategies to treat loneliness.

In this first study, our aim was to characterize loneliness mentions based on users' entire timelines. Future studies could perform a time-series analysis of the temporal variations associated with loneliness mentions. Further, works should also validate whether the characteristics of people who are using the words 'lonely' or 'alone' on Twitter can be used to track community health risks, particularly, the risk of social isolation. Other studies should replicate the findings in this study using more formal ground truth such as surveys and extend this work to investigate if Twitter can potentially map regional hotspots of loneliness to identify problematic loneliness for community public health monitoring.

Limitations and Ethics

The study sample consists of social media users and is not representative of the general population. An estimated 40% of US adults using Twitter are between the ages of 18 and 29, so our analysis is skewed towards younger people.[48] An automated machine learning tool could be a low-cost method to potentially detect posts about loneliness or being alone that may occur with other concerning signals from digital sensors (e.g. changes in sleep, activity, purchases). These signals could trigger could then be referred to more formal screening methods or support resources.[49]

Considering we identified that 76% of users' tweets indicated presumably feeling lonely in the sample we hand coded, posts mentioning the words alone or lonely may have been metaphorical or non sequiturs. Also, considering the inclusion criteria based on number of tweets mentioning alone or lonely, we are potentially selecting users with more posts than the average twitter user. Additionally, Twitter is far from perfect to be used as a diagnostic tool. However, an automated machine learning tool could be a low-cost method to potentially detect elevated loneliness levels in a person who could then be referred to more formal screening methods. Further, the effects presented in this dataset may not be specific to loneliness considering the potential comorbidity with mental health conditions such as depression in this dataset.

The feasibility of social media-based assessments of loneliness mentions (and mental health more broadly) needs further assessment. Privacy of individuals is an ongoing concern, especially with social media users not fully realizing the amount of health insights that can be gleaned by their online posts. Employers and insurance companies, for example, may be motivated to derive these assessments, but could use these insights against those suffering from mental illness. As mental illnesses carry social stigma and may engender discrimination, data protection and ownership frameworks are needed to make sure the data is not used against the users' interest.[50] Further, transparency about which indicators are derived by whom for what purpose should be part of ethical and policy discourse.

There are also open questions around the impact of misclassifications, and how derived mental health indicators can be responsibly integrated into systems of care.[51]

Conclusions

In this study we characterized mentions of loneliness on Twitter at the individual level. Furthermore, we identified specific contexts, themes, and traits in the posts of individuals mentioning loneliness on Twitter. As loneliness is a public health challenge, a better understanding of how loneliness is described online can inform tracking of loneliness and interventions targeted at addressing this important public health problem in regards to the behavior of lonely individuals that may be at risk of developing a severe mental health condition.[44]

Acknowledgements

Funding: This project is funded, in part, under a grant with the Pennsylvania Department of Health (ORG-FUND-PROG-CREF is 4290-567862-2446-2049). No sponsor of funding source played a role in: "study design and the collection, analysis, and interpretation of data and the writing of the article and the decision to submit it for publication." All researchers are independent from funders

Data Sharing Statement: Because of our IRB requirements, data will be shared upon request from the corresponding author.

Contributors: S.C. Guntuku and R. Merchant originated the study. S.C. Guntuku, R. Schneider, A. Pelullo, L.H. Ungar, and R. Merchant developed methods, interpreted analysis, and contributed to the writing of the article. J.F. Young, V. Wong, D. Polsky, and K. Volpp assisted with the interpretation of the findings and contributed to the writing of the article.

Disclosures: None

References:

(1) Hyland P, Shevlin M, Cloitre M, *et al.* Quality not quantity: loneliness subtypes, psychological trauma, and mental health in the US adult population. *Social Psychiatry and Psychiatric Epidemiology* Published Online First: 2018. doi:10.1007/s00127-018-1597-8

(2) Gerst-Emerson K, Jayawardhana J. Loneliness as a Public Health Issue: The Impact of Loneliness on Health Care Utilization Among Older Adults. *American Journal of Public Health*. 2015;105(5):1013-1019. doi:10.2105/ajph.2014.302427.

(3) Jong-Gierveld JD. Developing and testing a model of loneliness. *Journal of Personality and Social Psychology*. 1987;53(1):119-128. doi:10.1037//0022-3514.53.1.119.

(4) Peplau LA, Perlman D. *Loneliness: a Sourcebook of Current Theory, Research, and Therapy*. New York: Wiley; 1982.

(5) Stravynski A, Boyer R. Loneliness in Relation to Suicide Ideation and Parasuicide: A Population-Wide Study. *Suicide and Life-Threatening Behavior*. 2001;31(1):32-40. doi:10.1521/suli.31.1.32.21312.

(6) Blai B. Health Consequences of Loneliness: A Review of the Literature. *Journal of American College Health*. 1989;37(4):162-167. doi:10.1080/07448481.1989.9938410.

(7) Heinrich LM, Gullone E. The clinical significance of loneliness: A literature review. *Clinical Psychology Review*. 2006;26(6):695-718. doi:10.1016/j.cpr.2006.04.002.

(8) Richard A, Rohrmann S, Vandeleur CL, Schmid M, Barth J, Eichholzer M. Loneliness is adversely associated with physical and mental health and lifestyle factors: Results from a Swiss national survey. *Plos One*. 2017;12(7). doi:10.1371/journal.pone.0181442.

(9) Rico-Uribe LA, Caballero FF, Olaya B, et al. Loneliness, Social Networks, and Health: A Cross-Sectional Study in Three Countries. *Plos One*. 2016;11(1). doi:10.1371/journal.pone.0145264.

(10) Draugalis JR, Coons SJ, Plaza CM. Best Practices for Survey Research Reports: A Synopsis for Authors and Reviewers. *American Journal of Pharmaceutical Education* 2008;**72**:11. doi:10.5688/aj720111

(11) Guntuku SC, Buffone A, Jaidka K, Eichstaedt J, Ungar L. Understanding and Measuring Psychological Stress using Social Media. *ICWSM 2019*. 2018.

(12) Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*. 2017;18:43-49. doi:10.1016/j.cobeha.2017.07.005.

(13) Primack BA, Shensa A, Sidani JE, et al. Social Media Use and Perceived Social Isolation Among Young Adults in the U.S. *American Journal of Preventive Medicine*. 2017;53(1):1-8. doi:10.1016/j.amepre.2017.01.010.

(14) Kivran-Swaine F, Ting J, Naaman M. Understanding Loneliness in Social Awareness Streams: Expressions and Responses. *ICWSM 2014*. 2014. https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef.

(15) Kamvar S, Kamvar S, Harris JJ. *We Feel Fine: an Almanac of Human Emotion*. New York: Scribner; 2009.

(16) Eichstaedt JC, Schwartz HA, Kern ML, *et al.* Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychological Science* 2015;**26**:159–69. doi:10.1177/0956797614557867

(17) Eichstaedt JC, Smith RJ, Merchant RM, *et al.* Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences* 2018;**115**:11203–8. doi:10.1073/pnas.1802331115

(18) Coppersmith G, Ngo K, Leary R, Wood A. Exploratory Analysis of Social Media Prior to a Suicide Attempt. *Proceedings of the Third Workshop on Computational*

Lingusitics and Clinical Psychology.
2016. doi:10.18653/v1/w16-0311.

(19) Qntfy, Inc. OurDataHelps.
OurDataHelps. https://ourdatahelps.org/. Accessed February 18, 2019.

(20) Sinnenberg L, DiSilvestro CL, Mancheno C, et al. Twitter as a Potential Data Source for Cardiovascular Disease Research. *JAMA Cardiology*. December 2016. https://jamanetwork.com/journals/jamacardiology/fullarticle/2556216.

(21) Sap M, Park G, Eichstaedt J, et al. Developing Age and Gender Predictive Lexica over Social Media. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014. doi:10.3115/v1/d14-1121.

(22) Jaidka K, Guntuku SC, Buffone A, Schwartz HA, Ungar L. Facebook versus Twitter: Differences in Self-Disclosure and Trait Prediction. *ICWSM*. 2018.

(23) 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). doi:10.1371/journal.pone.0073791.

(24) Pennebaker Jw, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works*. 2014:1-22.

(25) Schwartz HA, Eichstaedt J, Kern ML, et al. Towards Assessing Changes in Degree of Depression through Facebook. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014. doi:10.3115/v1/w14-3214.

(26) Lenhart A. Mobile Access Shifts Social Media Use and Other Online Activities. Pew Research Center: Internet, Science & Tech.
2015.http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-other-online-activities/ (accessed 9 May2019).

(27) Guntuku SC, Ramsay JR, Merchant RM, Ungar LH. Language of ADHD in Adults on Social Media. *Journal of Attention Disorders*. 2017:108705471773808. doi:10.1177/1087054717738083.

(28) Pinquart M, Sorensen S. Influences on Loneliness in Older Adults: A Meta-Analysis. *Basic and Applied Social Psychology*. 2001;23(4):245-266. doi:10.1207/S15324834BASP2304_2.

(29) Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *The Journal of Machine Learning Research*. 2003;3.

(30) Gelfand AE, Smith AF. Sampling Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association* 1990;**85**:398–409. doi:10.21236/ada208388

(31) McCallum A. Mallet: A machine learning for language toolkit. MALLET Machine Learning for Language Toolkit. 2002.http://mallet.cs.umass.edu/ (accessed 9 May2019).

(32) Meyer GJ, Finn SE, Eyde LD, et al. Psychological testing and psychological assessment. A review of evidence and issues. *The American psychologist*. 2001;56(2):128-165.

(33) Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, *et al.* What Twitter Profile and Posted Images Reveal About Depression and Anxiety. *ICWSM* Published Online First: 2019.https://arxiv.org/pdf/1904.02670.pdf

(34) Jones R. Drug Slang Dictionary - Words Starting With G. http://www.noslang.com/drugs/dictionary.php. Accessed February 19, 2019.

(35) Booth R. Toward an Understanding of Loneliness. *Social Work*. 1983;28(2):116-119. doi:10.1093/sw/28.2.116.

(36) Qualter P, Vanalst J, Harris R, *et al.* Loneliness across the life span. *Perspectives on psychological science : a journal of the Association for Psychological Science* 2015;**10**:250–64.https://journals.sagepub.com/doi/full/10.1177/1745691615568999?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub=pubmed

(37) Hawkley LC, Cacioppo JT. Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms. *Annals of Behavioral Medicine* 2010;**40**:218–27. doi:10.1007/s12160-010-9210-8

(38) Cacioppo JT, Hawkley LC, Ernst JM, et al. Loneliness within a nomological net: An evolutionary perspective. *Journal of*

Research in Personality. 2006;40(6):1054-1085. doi:10.1016/j.jrp.2005.11.007.

(39) Edwards T, Holtzman NS. A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research*

(40) Al-Saggaf Y, Nielsen S. Self-disclosure on Facebook among female users and its relationship to feelings of loneliness. *Computers in Human Behavior* 2014;**36**:460–8. doi:10.1016/j.chb.2014.04.014

(41) Simon EB, Walker MP. Sleep loss causes social withdrawal and loneliness. *Nature Communications*. 2018;9(1). doi:10.1038/s41467-018-05377-0.

(42) Cacioppo JT, Hawkley LC, Thisted RA. Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago Health, Aging, and Social Relations Study. *Psychology and Aging* 2010;**25**:453–63. doi:10.1037/a0017216

(43) Yeginsu C. U.K. Appoints a Minister for Loneliness. *The New York Times*. https://www.nytimes.com/2018/01/17/world/europe/uk-britain-loneliness.html. Published January 17, 2018.

(44) Rubin R. Loneliness Might Be a Killer, but What's the Best Way to Protect Against It? *Jama*. 2017;318(19):1853. doi:10.1001/jama.2017.14591.

(45) Jain S. CareMore Health Tackles the Unmet Challenges of the Aging Population. *American Society on Aging* 2018;**42**:14–8.https://www.ingentaconnect.com/contentone/asag/gen/2018/00000042/00000001/art00003

(46) Perese EF, Wolf M. Combating Loneliness Among Persons With Severe Mental Illness: Social Network Interventions Characteristics, Effectiveness, And Applicability. *Issues in Mental Health*

Nursing. 2005;26(6):591-609.
doi:10.1080/01612840590959425.

(47) Webber M, Fendt-Newlin M. A review of social participation interventions for people with mental health problems. *Social Psychiatry and Psychiatric Epidemiology* 2017;**52**:369–80. doi:10.1007/s00127-017-1372-2

(48) U.S. Twitter reach by age group 2018 |
Statistic. Statista.
https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/.
Accessed February 18, 2019.

(49) Barnett I, Torous J. Ethics, Transparency, and Public Health at the Intersection of Innovation and Facebooks Suicide Prevention Efforts. *Annals of Internal Medicine* 2019;**170**:565. doi:10.7326/m19-0366

(50) Mckee R. Ethical issues in using
social media for health and health care research. *Health Policy*. 2013;110(2-3):298-301.
doi:10.1016/j.healthpol.2013.02.006.

(51) Inkster B, Stillwell D, Kosinski M,
Jones P. A decade into Facebook: where is psychiatry in the digital age? *The Lancet Psychiatry*. 2016;3(11):1087-1090.
doi:10.1016/s2215-0366(16)30041-4.

Figure legends

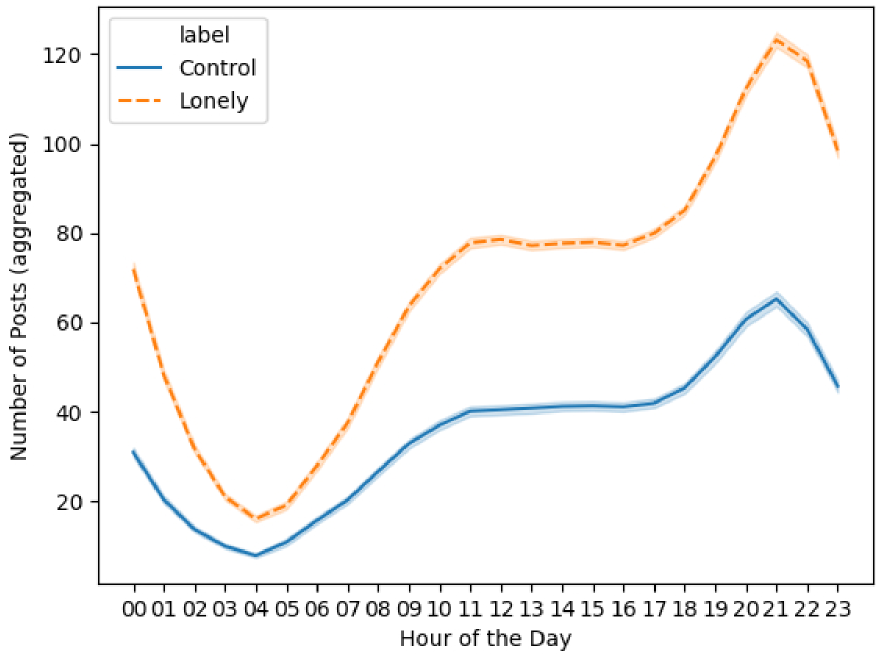
Figure 1: Words/Phrases more likely to be posted by Twitter users with a) self-reported loneliness (Individuals with at least 5 posts with the words ‘lonely’ or ‘alone’ group compared to the b) control group.

Word size indicates the strength of the correlation and word color indicates relative word frequency. ($p<0.01$, Bonferroni p-corrected)

Figure 2: Temporal variation showing diurnal patterns of post frequency of both the users with posts including the words lonely or alone **and control group.**

The dotted line indicates the percentage of posts at different hours of the day by the group of users with at least 5 posts containing the word ‘lonely’ or ‘alone’ and the solid line indicates users who do not have any posts about loneliness. The x-axis represents the hour of the day each post occurs and the y-axis indicates the number of posts for each group.

Words/Phrases more likely to be posted by Twitter users with a) self-reported loneliness (Individuals with at least 5 posts with the words 'lonely' or 'alone' group compared to the b) control group.



Temporal variation showing diurnal patterns of post frequency of both the 'lonely' and 'control' groups.

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **(pg.2)**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **(pg.2)**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **(pg.4)**
Objectives	3	State specific objectives, including any prespecified hypotheses **(pg. 4)**
Methods		
Study design	4	Present key elements of study design early in the paper **(pg.5)**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **(pg. 5)**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up **(pg. 6)**
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **(pg. 6)**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **(pg. 6)**
Bias	9	Describe any efforts to address potential sources of bias **(pg. 6)**
Study size	10	Explain how the study size was arrived at **(pg. 6)**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **(pg. 7)**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **(pg. 9)**
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) If applicable, explain how loss to follow-up was addressed
		(*e*) Describe any sensitivity analyses
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **(pg. 9)**
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **(pg. 9)**
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarise follow-up time (eg, average and total amount)
Outcome data	15*	Report numbers of outcome events or summary measures over time **(pgs 10,11)**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **(pgs 10,11)**

		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **(pgs 10,11)**
Discussion		
Key results	18	Summarise key results with reference to study objectives **(pgs. 12, 13)**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **(pgs. 14)**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **(pg. 13, 14)**
Generalisability	21	Discuss the generalisability (external validity) of the study results **(pgs. 13, 14)**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **(pg. 15)**

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.

BMJ Open

Studying Expressions of Loneliness in Individuals using Twitter: An Observational Study

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2019-030355.R3
Article Type:	Original research
Date Submitted by the Author:	24-Sep-2019

Complete List of Authors:	Guntuku, Sharath Chandra; University of Pennsylvania, Department of Emergency Medicine Schneider, Rachelle; University of Pennsylvania, Department of Emergency Medicine Pelullo, Arthur; University of Pennsylvania, Department of Emergency Medicine Young, Jami; University of Pennsylvania, Department of Emergency Medicine Wong, Vivien; University of Pennsylvania, Department of Emergency Medicine Ungar, Lyle ; University of Pennsylvania, Department of Emergency Medicine Polsky, Daniel ; Leonard Davis Institute of Health Economics, Volpp, Kevin ; University of Pennsylvania Perelman School of Medicine, Medical Ethics and Policy Merchant, Raina; University of Pennsylvania, Department of Emergency Medicine

Primary Subject Heading:	Public health
Secondary Subject Heading:	Communication, Mental health
Keywords:	loneliness, social media, natural language processing, MENTAL HEALTH, STATISTICS & RESEARCH METHODS

Studying Expressions of Loneliness in Individuals using Twitter: An Observational Study

Sharath Chandra Guntuku, PhD[1,4,5], Rachelle C. Schneider, BS [1,5], Arthur Pelullo, MS[1,4,5], Jami F. Young, PhD[5,7], Vivien Wong, BS[1,5], Lyle H. Ungar, PhD[3,4], Daniel Polsky, PhD[5,6], Kevin Volpp, MD, PhD[5,6], Raina M. Merchant, MD, MSHP[1,2,5]

[1]Penn Medicine Center for Digital Health, Philadelphia, PA 19104

[2]Penn Medicine Center for Healthcare Innovation, Philadelphia, PA 19104

[3]Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104

[4]Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

[5]Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

[6]The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

[7]Children's Hospital of Philadelphia, Philadelphia, PA 19146

Corresponding author and request for reprints

Sharath Chandra Guntuku

3400 Civic Center Blvd

Philadelphia PA 19104

Email: sharathg@sas.upenn.edu

Word count: 3723, 31 pages, 4 tables, 2 figures

Keywords: loneliness mentions; social media; twitter; natural language processing; mental health

Abstract

Objectives: Loneliness is a major public health problem and an estimated 17% of adults aged 18-70 in the United States reported being lonely. We sought to characterize the (online) lives of people who mention the words 'lonely' or 'alone' in their Twitter timeline and correlate their posts with predictors of mental health.

Setting and design: From approximately 400 million tweets collected from Twitter in Pennsylvania, USA, between 2012-2016, we identified users whose Twitter posts contained the words 'lonely' or 'alone' and compared them to a control group matched by age, gender, and period of posting. Using natural-language processing, we characterized the topics and diurnal patterns of users' posts, their association with linguistic markers of mental health, and if language can predict manifestations of loneliness. The statistical analysis, data synthesis, and model creation was conducted in 2018-2019.

Primary outcome measures: We evaluated counts of language features in the users with posts including the words lonely or alone compared to the control group. These language features were measured by (a) Linguistic Inquiry Word Count (LIWC) lexicon, (b) open-vocabulary topics, and (c) linguistic markers of anger, depression, and anxiety. Using machine learning, we also evaluated if expressions of loneliness can be predicted compared to the control group, measured by Area Under Curve (AUC).

Results: Twitter timelines of users (N=6202) with posts including the words lonely or alone were found to include themes about difficult interpersonal relationships, psychosomatic symptoms, substance use, wanting change, unhealthy eating, and having troubles with sleep. Their posts were also associated with linguistic markers of anger, depression, and anxiety. A random forest model predicted expressions of loneliness online with an AUC of 0.77.

Conclusions: Posts with the words lonely or alone often include psychosocial features and can potentially have associations with how individuals express and experience loneliness. This can inform low-resource online assessment for high risk individuals experiencing loneliness and interventions focused on addressing morbidities in this condition.

Strengths and Limitations of this study

- Novel focus on timelines of social media users to study mentions of loneliness and correlation with predictors of mental health.
- The study sample consists of social media users and is not representative of the general population.
- Though we manually annotated a subset of posts mentioning loneliness, some may have been metaphorical or non sequiturs.

Introduction

Loneliness is a major public health problem and an estimated 17% of adults aged 18-70 in the United States are reported being lonely.[1] Loneliness is defined as the discrepancy between a person's desired and actual social relationships. Loneliness is also one of the primary underlying causes and correlates for chronic mental health conditions and physician visits in some populations.[1–6] It has also been linked with an increased risk of heart disease, stroke, dementia, depression, and anxiety.[1,2,7–9]

Reducing morbidity from loneliness requires identifying who experiences it. Traditionally this has occurred through surveys but unfortunately this is not common and not scalable to screen large populations.[10] Rather than relying on the traditional screening approach, social media platforms, like Facebook, Twitter, and Instagram are being investigated to shed light on individual's health and well-being.[11] With people increasingly using social media platforms to inform others about their mental states, solicit social support, as well as to keep records of their daily activities, preferences, and interests,[12,13] social media has emerged as a potentially relevant tool to passively measure health states and behaviors of people.[14,15] For example, individuals who are stressed and depressed use more first-person singular pronouns suggesting higher self-focus and communities with heart disease discuss hate more frequently.[13,16] Social media posts have also been used to predict first documented diagnosis of depression using posts 6 months prior yielding an AUC of 0.72.[17]

While the use of social media is increasingly common, less is known about how often individuals use the platform to explicitly share about feelings of loneliness or being alone. In this study, we sought to characterize Twitter timelines of individuals' whose posts include the words lonely or alone. Based on the language of such Twitter users, we analyzed the correlations between posting about loneliness and users' mental health and psycholinguistic attributes (e.g. anger and depression).

We hypothesize that language usage patterns would both confirm existing understanding of loneliness and give new insights into the daily lives of those who express being lonely. As loneliness can impact health outcomes, identifying ways to track prevalence and manifestations of loneliness online would be useful for developing approaches for identifying and offering support for these individuals. While prioritizing the privacy of individuals, specifically with the amount of health insights that can be gleaned from social media, this research presents the opportunity of digital platforms to not only provide markers of health but also potentially serve as platforms that can be used for developing interventions.[18,19]

Methods

This was a retrospective analysis of publicly available data on users posting about loneliness on Twitter. This study was exempt by the University of Pennsylvania Institutional Review Board.

Twitter Data

Twitter is a popular social media platform which allows users to send and receive short 140-character messages, or 'tweets' (at the time of this study; the character limit was later increased to 280). First, from the Twitter Streaming API, we collected tweets from the 1% sample using a bounding box of location coordinates around Pennsylvania. To increase the sample size of tweets from the state, all unique user IDs were recorded, and the Twitter API was used to extract timelines (each user's prior 3200 tweets) filtered by timestamps ranging from 2012-2016.

Patient and Public Involvement

Patients and public were not involved in the development of the research question and outcome measures.

Study Sample

We identified users who posted the word "alone" or "lonely" at least once in their timeline (25,966 users). As social media includes colloquial, metaphorical, and light-hearted language (eg. "If I see Justin Bieber, I will have a heart attack") we sought to identify the proportion of tweets in which lonely seemed to refer to the public health meaning rather than other uses of the term (e.g. metaphor, joke).[20] Two co-authors independently coded a random set of 100 tweets from individuals who used the words lonely/alone at least 5 times in their timeline to identify them as presumed to be associated with the feeling of loneliness or other (Cohen's κ = 0.70, and 76% of users' tweets indicate presumably feeling lonely). A few examples are as follows: "*i'm feelin real depressed, confused, & lonely", "im always the only up around this time, feeling a lil lonely*" and "*I'm so Lonely in life :-(I just wish I can have love again it feels so go to be in love with someone whom loves you*." 6,202 users posted messages with "alone" or "lonely" at least 5 times.

Control group

We then identified a control group of users by matching each user in the above dataset to another user by age, gender and period of activity (dates of first and last posting on twitter). We obtained the age and gender estimates by using lexica developed previously.[21] Then, we selected users with a minimum of 500 words across all their posts to have sufficient language for linguistic analyses.[11] We excluded non-English tweets, re-tweets, and tweets containing ‘alone’ and/or ‘lonely’ that were used to identify users in all analyses. Hereafter, we indicate users who had more than 5 posts with the words ‘lonely’ or ‘alone’ as ‘users with posts including the words lonely or alone’, and ‘control’ group to represent the matched set of users who had no such posts.

Deriving language features to characterize individuals expressing loneliness

We used four sets of language features: a) dictionary-based psycholinguistic features,[22] b) open-vocabulary topics,[23] c) mental well-being attributes such as anxiety, depression by applying previously developed statistical models,[24,25] d) number of drug words and time of posts as past research has shown an association between loneliness and substance use.[26,27] These language features have been shown to be predictive of several health outcomes, such as depression, schizophrenia, attention deficit hyperactivity disorder (ADHD), and general well-being.[17,26,28]

Dictionary-based: From each post, we extracted the relative frequency of single words and phrases (consisting of two or three consecutive words). Then, all words used by less than 1% of users were removed from analysis so as to remove uncommonly used words (outliers). Additionally, all tweets used to identify our study group were removed prior to further analysis. The Linguistic Inquiry Word Count (LIWC) dictionary is a language-specific, many-to-many mapping of tokens (including words and word stems) and psychologically validate categories.

Each category (a curated list of words) is found to be correlated with and also predictive of several psychological traits and outcomes. For each user, we measure the proportion of word tokens that fall into a given LIWC category.

Open-vocabulary: As closed-vocabulary approaches like LIWC include only a subset of the entire language used on social media, we use an open-vocabulary approach to improve the coverage and find topics in users' timelines mentioning loneliness. Topics consist of clusters of co-occurring words created using Latent Dirichlet Allocation (LDA).[29] The LDA generative model assumes that tweets contain a combination of topics, and that topics are a distribution of words. Since the words in a tweet are known, topics, which are latent variables, can be estimated through Gibbs sampling.[30] We use the Mallet implementation of the LDA algorithm, adjusting one parameter (alpha=5) to favor fewer topics per tweet.[31] All other parameters were kept at their default. An example of such a model is the following sets of words ('tuesday', 'monday', 'wednesday', ...) which clusters together days of the week by exploiting their similar distributional properties across tweets. In our study, two hundred topics were generated using tweets across all users in the dataset including the words lonely or alone and control users.

Mental well-being attributes: We used automatic text-regression methods to assign to each user scores on the depression, anxiety and anger facets for users.[24,25] This model was trained on a sample of over 28,749 users who had taken the International Personality Item Pool Neuroticism-Extraversion-Openness Personality Inventory Revised (IPIP NEO-PI-R) survey that contains the depression, anxiety and anger Facets of the Neuroticism Factor.[32,33] The machine learning model trained on words and phrases from Facebook posts to predict survey measure of depression, anger and anxiety resulted in a performance of r = .32, which is consistent with other reports of mental health states identified via social media.[13] The model was trained using status updates of users from another study[24], and has been shown to generalize to Twitter users.[25]

Use of Drug-words: We also extracted the frequency of most common drug words as used on social media for every user in our analysis.[27]

Temporal patterns: We determined the frequency of posts across different hours of the day by users in both users with posts including the words lonely or alone and control groups to understand the diurnal patterns in posting.

Identifying differentially expressed language features in users with posts including the words lonely or alone

We isolated the patterns in users' loneliness mentions using the linguistic attributes and mental health attributes by correlating them with users with posts including the words lonely or alone and control groups. We used logistic regression to distinguish the different features associated with lonely and control groups and measure the effect size using Cohen's D. The models were set up to predict the group of users with posts including the words lonely or alone against the control group (e.g., group was the dependent variable). Details of the method are described in a previous work[23]. For identifying themes from topics, researchers looked at 20 messages each with the highest topic prevalence. We used Benjamini-Hochberg p-correction and $p<0.001$ for indicating meaningful correlations and the effect size was measured using Cohen's D. We also tested that the results hold if frequency of posting is used as an additional variable on which to match the users with lonely expressions and the control subject. The statistical analysis, data synthesis, and model creation was conducted in 2018-2019.

Predicting the likelihood of posting about loneliness online

We then looked at the feasibility of predicting whether a user is likely to mention expressions of loneliness or not based on their social media language. Automated analysis of social media is accomplished by building predictive models, which use linguistic features that have been extracted from social media data. For this analysis we used LIWC and topics as features. Features are then treated as independent variables in an algorithm (Random Forests) to predict the dependent variable of an outcome of interest (e.g., users' expressing that they are lonely or not). For cross validation, the predictive model was trained, using Random Forests, on the training set and then evaluated on a test set to avoid overfitting. The prediction performances are reported as Area Under the Receiver Operating Curves (AUC) on an out-of-sample 5-fold cross validation setting.

Results

Of the 408,296,620 tweets posted by users geo-located in Pennsylvania, USA, 25,966 users with 46,160,774 posts in their timelines, had at least one post with the words 'lonely' or 'alone', and 6,202 users with 17,995,084 posts in their timelines, had more than five such posts (Table 1). Users with posts including the words lonely or alone had 1.9 times more posts in the study time period as the control (Table 1). The median estimated age of this cohort was 21 years, and 69% female.

Table 1: Descriptive statistics for users in the dataset

Descriptive Statistics of the Dataset		
	Users with posts including the words lonely or alone (n= 6,202)	Control group (n= 6,202)
Median Age	21 ± 3 yrs	21 ± 3 yrs
# Messages in timelines	17,995,084	9,219,677
# Females	4,400	4,400
# Males	1,802	1,802

Identifying differentially expressed language features in users with posts including the words lonely or alone

Open vocabulary approach: Analyzing differences in individual words and phrases used across both groups, we observed (Figure 1a) that users with posts including the words lonely or alone referred to themselves ('myself' (d=.18), 'I' (d=.16)) in their Twitter posts significantly more than the control group. They also posted about relationship issues ('want_somebody' (d=.08), 'no_one_to' (d=.1), needs and feelings ('i_just_wanna (d=.12), 'in_my_feelings' (d=.1), 'i_need' (d=.12), 'i_cant' (d=.1)), and included more expletives. Users in the control group (Figure 1b) engaged in a lot more conversations as indicated by '<user>' (d=-.2) (anonymized '@' mentions in users tweets as '<user>') compared to users with posts including the words lonely or alone.

The control group also posted more about games ('season' (d=-.09) ,'coach' (d=-.07), 'team' (d=-.1)) and positivity ('!' (d=-.13), 'awesome' (d=-.09), ':)' (d=-.08)). Figure 1 illustrates the words and phrases most prominently associated with the group of users with posts including the words lonely or alone and the control group.

Using topics generated from LDA, we identified the themes which occur more frequently in posts of users with posts including the words lonely or alone. Table 2 shows the effect sizes between most prominent topic distributions and the users with mentions of loneliness. Posts were about interpersonal relationships (d=.28) (and associated issues (d=.22)), self-reflection (d=.21) (accompanied with wondering about the future (d=.12)), drug/alcohol use (d=.29) (considering them to be the 'only friend'), insomnia (d=.27), uncontrolled emotions (d=.28) (accompanied by confusion (d=.11)), and psychosomatic symptoms (d=.29).

Table 2: Highly correlated topics with mentions of loneliness.

Topic Theme	Highly Correlated Words in Topic	Effect size (Cohen's D)
Interpersonal Relationships	relationships, matter, perfect hurt, feelings, trust, forget	0.28 0.22
Self-Reflection	times, changed, lost, i've	0.21
Drug/Alcohol Use	smoke, weed, blunt, drugs, drunk	0.29
Psychosomatic	bad, stomach, hurt, head, sick	0.29

Symptoms		
Insomnia	sleep, awake, tired, bed	0.27
Emotional Dysregulation	people, f***ing, hate, stupid	0.28
Food/Hunger	food, breakfast, eat, pizza, hungry	0.26

* Effect size is measured using Cohen's D. Only significant topics after Benjamini-Hochberg p-correction and use p<0.001 are shown.

Dictionary-based: Association of LIWC categories of users with posts including the words lonely or alone are shown in Table 3. Individuals who had posts including the word lonely or alone used increased self-references (first person pronouns, d=.18), words indicating cognitive processes (including certainty, d=.15, discrepancies, d=.14, differentiation, d=.13 and tentativeness, d=.13), and negative emotions (swearing, d=.11).

Table 3: Association of LIWC categories, mental health attributes, and drug words with mentions of loneliness

Category	**Cohen's D***
Pronouns	
1st Person Pronouns	0.18
Cognitive Processes	
Certainty	0.15

Discrepancies	0.15
Differentiation	0.14
Tentativeness	0.13
Negative Emotions	
Swearing	0.11
Mental Well-being	
Depression	0.81
Anger	0.95
Anxiety	0.75
Drug words	
Blunt	0.16
Smoke	0.13
Heroin	0.1

*Only significant categories after Benjamini-Hochberg p-correction and p<0.001 are shown.

Mental well-being: Users with posts including the words lonely or alone were more likely to have posts associated with anger (d=.95), depression (d=.81) and anxiety (d=.75) when compared to the control group.

Use of Drug Words: We also identified the distribution of words pertaining to drugs in the posts of users with posts including the words lonely or alone, and these were more likely to reference a blunt (d=.16), smoke (d=.13), and heroin (d=.1), and included prescribed medications for treatment, recreational drug use, and recreational drugs.

Temporal patterns: Users with posts including the words lonely or alone were found to post more during the night (d=.1), shown in Figure 2. We also see themes associated with night-time posting and having difficulty sleeping (d=.27) in the open-vocabulary analysis (Table 2).

Predictive Analysis: Table 4 shows that random forest model using Topics as input features predicted mentions of loneliness in users with an AUC of .854 (F1 score = 0.778) and LIWC features resulted in AUC of 0.859 (F1 score = 0.777). A combination of LIWC and Topics resulted in the best AUC of 0.863 (F1 score = 0.782).

Table 4: Performance of different features at predicting mentions of loneliness, reported on an out-of-sample 5-fold cross validation setting.

Feature	AUC	F1 Score	Accuracy	Precision	Recall
Topics	0.854	0.778	0.778	0.780	0.778
LIWC	0.859	0.777	0.777	0.778	0.777
LIWC + Topics	0.863	0.782	0.783	0.785	0.783

Discussion

From a widely used publicly available social network, Twitter, we characterized what and when individuals post about loneliness, association of posts with mental health, and if manifestations of loneliness can be predicted in individuals. Our fundamental hypothesis was that the language of users with posts including the words lonely or alone would be significantly different from matched controls, that this language would reveal differences in characteristics such as mental health attributes between both groups, and that the language usage patterns would both confirm existing understanding of loneliness and give new insights into the daily lives of those who post the words alone or lonely. Towards this goal, we took an inductive approach of computationally analyzing the large volumes of social media data with the aim of better understanding the varying manifestations of loneliness. This paper has three main findings. First, we identified themes and contexts associated with users posting about loneliness on Twitter. Second, we observed that users posting about loneliness used language associated with linguistic models for anger, depression, and anxiety. Third, posts about loneliness were more likely to occur in the evening or night.

Themes associated with people mentioning loneliness on Twitter are consistent with prior literature about substance use, emotional dysregulation, and troubles with relationships. For example, in one study, a high positive correlation was found between alcoholism and groups of lonely people, and lonely people were also found to express negative feelings towards relationships.[34] This expression of negativity related to relationships is likely related to a hypervigilance to social threat, associated with loneliness.[35] Lonely individuals were also reported to focus on overcoming past events as well as showing feelings of helplessness.[35]

Association of users with posts including the words lonely or alone with linguistic estimates of anger, depression, and anxiety corroborate prior research, showing that loneliness and social isolation influence psychological functioning , specifically the ability to self-regulate emotion.[2,3,36] Specifically, anxiety, anger, and negative mood were reported as higher in lonely young adults.[37] Tweets by users with posts including the words lonely or alone were more self-focused compared to the control group. Prior researchers have found that "first person singular pronouns are a modest linguistic marker of depression." [38]Also, previous research has shown that loneliness has been associated with greater self-disclosure in Facebook posts.[39] This presents the potential for early identification and assessment to intervene on loneliness as well as mental health conditions for this group.

Trends in temporal variation in posting may reflect that sleep deprivation can contribute to social withdrawal and loneliness.[40] This finding corroborates prior research associating loneliness with diminished sleep quality.[36] A better understanding of the temporality of posting could inform timing of interventions designed to address loneliness, as well as provide insight for other researchers to test the inter-relationships between loneliness and the motivations for using social media during nighttime.

Loneliness is known to be one of the primary underlying causes and correlates for chronic mental health conditions.[41] As loneliness is becoming increasingly recognized as a public health, several groups have taken action to address it. For example, the United Kingdom appointed a Minister for Loneliness who is responsible for addressing loneliness within communities.[42] CareMore, a health plan and delivery system providing care for enrollees in Medicare Advantage and Medicaid health plans in seven states across the U.S., launched the “Togetherness Program” in a clinical setting to address loneliness in elderly patients.[43] Through this work, CareMore reported that participation in exercise programs increased by 56.6%, emergency room utilization decreased by 3.3%, and hospital admissions among participants were 20.8% lower per thousand compared to the “intent to treat population.” [44] Additionally, social network interventions targeting loneliness have been found to be effective in reducing social isolation among individuals with severe mental health conditions but these interventions are not included in the treatment plans for individuals with a mental illness.[45,46]

Considering the advantage of large sample sizes and also the association between increased social media usage and individuals mentions of loneliness, it is promising to use natural language processing and machine learning to automatically identify a person mentions the words alone or lonely on Twitter to inform interventions targeted at early identification and support for affected and at risk individuals with the caveat that social media users are not representative of a random sample of individuals. To address loneliness will require being able to identify it passively, remotely, and over time. Many people rarely visit a healthcare provider so would miss the opportunity for screening. Approaches for treatment will also need to harness the tools and technologies that are accessible and integrated with the things people use every day (e.g. mobile phones). Future interventions would have to potentially rely on digital phenotyping of loneliness and using digital platforms (e.g. text messaging) to complement human-to-human interaction strategies to treat loneliness.

In this first study, our aim was to characterize loneliness mentions based on users' entire timelines. Future studies could perform a time-series analysis of the temporal variations associated with loneliness mentions. Further, works should also validate whether the characteristics of people who are using the words 'lonely' or 'alone' on Twitter can be used to track community health risks, particularly, the risk of social isolation. Other studies should replicate the findings in this study using more formal ground truth such as surveys and extend this work to investigate if Twitter can potentially map regional hotspots of loneliness to identify problematic loneliness for community public health monitoring.

Limitations and Ethics

The study sample consists of social media users and is not representative of the general population. An estimated 40% of US adults using Twitter are between the ages of 18 and 29, so our analysis is skewed towards younger people.[47] An automated machine learning tool could be a low-cost method to potentially detect posts about loneliness or being alone that may occur with other concerning signals from digital sensors (e.g. changes in sleep, activity, purchases). Though Twitter is far from perfect to be used as a diagnostic tool, these signals could trigger could then be referred to more formal screening methods or support resources.[48]

Considering we identified that 76% of users' tweets indicated presumably feeling lonely in the sample we hand coded, posts mentioning the words alone or lonely may have been metaphorical or non sequiturs. Researchers who coded the topics were attempting to identify these associations by looking at twenty messages each with the highest topic prevalence to identify themes, and we acknowledge that this can be subjective. Also, considering the inclusion criteria based on number of tweets mentioning alone or lonely, we are potentially selecting users with more posts than the average twitter user. Further, the effects presented in this dataset may not be specific to loneliness considering the potential comorbidity with mental health conditions such as depression.

Social media use seeks to connect people but it also has been associated with increased perceived social isolation.[49] It is unclear if social media use causes perceived social isolation or if perceived social isolation causes social media use. The feasibility of social media-based assessments of loneliness mentions (and mental health more broadly) needs further assessment. Privacy of individuals is an ongoing concern, especially with social media users not fully realizing the amount of health insights that can be gleaned by their online posts. Employers and insurance companies, for example, may be motivated to derive these assessments, but could use these insights against those suffering from mental illness. As mental illnesses carry social stigma and may engender discrimination, data protection and ownership frameworks are needed to make sure the data is not used against the users' interest.[50] Further, transparency about which indicators are derived by whom for what purpose should be part of ethical and policy discourse. There are also open questions around the impact of misclassifications, and how derived mental health indicators can be responsibly integrated into systems of care. [51]

Conclusions

In this study we characterized mentions of loneliness on Twitter at the individual level. Furthermore, we identified specific contexts, themes, and traits in the posts of individuals mentioning loneliness on Twitter. As loneliness is a public health challenge, a better understanding of how loneliness is described online can inform tracking of loneliness and interventions targeted at addressing this important public health problem in regards to the behavior of lonely individuals that may be at risk of developing a severe mental health condition.

Acknowledgements

Funding: This project is funded, in part, under a grant with the Pennsylvania Department of Health (ORG-FUND-PROG-CREF is 4290-567862-2446-2049). No sponsor of funding source played a role in: "study design and the collection, analysis, and interpretation of data and the writing of the article and the decision to submit it for publication." All researchers are independent from funders

Data Sharing Statement: Data will be shared upon request from the corresponding author. The code used for analysis is made public at http://dlatk.wwbp.org

Contributors: S.C. Guntuku and R. Merchant originated the study. S.C. Guntuku, R. Schneider, A. Pelullo, L.H. Ungar, and R. Merchant developed methods, interpreted analysis, and contributed to the writing of the article. J.F. Young, V. Wong, D. Polsky, and K. Volpp assisted with the interpretation of the findings and contributed to the writing of the article.

Disclosures: None

References

1. Hyland P, Shevlin M, Cloitre M, others. *Quality Not Quantity: Loneliness Subtypes, Psychological Trauma, and Mental Health in the US Adult Population*. Social Psychiatry and Psychiatric Epidemiology Published Online First; 2018. doi:10.1007/s00127-018-1597-8
2. Stravynski A, Boyer R. Loneliness in Relation to Suicide Ideation and Parasuicide: A Population-Wide Study. *Suicide Life-Threatening Behav*. 2001;31(1):32-40. doi:10.1521/suli.31.1.32.21312
3. Blai B. Health Consequences of Loneliness: A Review of the Literature. *J Am Coll Heal*. 1989;37(4):162-167. doi:10.1080/07448481.1989.9938410
4. Heinrich LM, Gullone E. The clinical significance of loneliness: A literature review. *Clin Psychol Rev*. 2006;26(6):695-718. doi:10.1016/j.cpr.2006.04.002
5. Richard A, Rohrmann S, Vandeleur CL, Schmid M, Barth J, Eichholzer M. Loneliness is adversely associated with physical and mental health and lifestyle factors: Results from a Swiss national survey. *PLoS One*. 2017;12:7. doi:10.1371/journal.pone.0181442
6. Rico-Uribe LA, Caballero FF, Olaya B, others. Loneliness, Social Networks, and Health: A Cross-Sectional Study in Three Countries. *PLoS One*. 2016;11:1. doi:10.1371/journal.pone.0145264
7. Gerst-Emerson K, Jayawardhana J. Loneliness as a Public Health Issue: The Impact of Loneliness on Health Care Utilization Among Older Adults. *Am J Public Health*. 2015;105(5):1013-1019. doi:10.2105/ajph.2014.302427
8. JD. J-G. Developing and testing a model of loneliness. *J Pers Soc Psychol*. 1987;53(1):119-128. doi:10.1037//0022-3514.53.1.119
9. Peplau LA, D. P. *Loneliness: A Sourcebook of Current Theory, Research, and Therapy*. New York: Wiley; 1982.
10. Draugalis JR, Coons SJ, CM. P. Best Practices for Survey Research Reports: A Synopsis for Authors and Reviewers. *Am J Pharm Educ*. 2008;72:11. doi:10.5688/aj720111
11. Jaidka K, Guntuku SC, Ungar LH. Facebook versus Twitter: Differences in Self-Disclosure and Trait Prediction. *Twelfth Int AAAI Conf Web Soc Media*. June 2018.
12. Guntuku SC, Buffone A, Jaidka K, Eichstaedt JC, Ungar LH. Understanding and measuring psychological stress using social media. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol 13. ; 2019:214-225.
13. Guntuku SC, Yaden DB, Kern ML, Ungar LH, Eichstaedt JC. Detecting depression and mental illness on social media: an integrative review. *Curr Opin Behav Sci*. 2017;18:43-49.
14. Kivran-Swaine F, Ting J, Naaman M. Understanding Loneliness in Social Awareness

Streams: Expressions and Responses. *ICWSM*. 2014;2014. https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef.
15. Kamvar S, Kamvar S, JJ. H. *We Feel Fine an Almanac of Human Emotion*. New York: Scribner; 2009.
16. Eichstaedt JC, Schwartz HA, Kern ML, others. Psychological Language on Twitter Predicts County-Level Heart Disease Mortality. *Psychol Sci*. 2015;26:159-169. doi:10.1177/0956797614557867
17. Eichstaedt JC, Smith RJ, Merchant RM, others. Facebook language predicts depression in medical records. *Proc Natl Acad Sci*. 2018;115:11203-11208. doi:10.1073/pnas.1802331115
18. Qntfy I. *OurDataHelps*. OurDataHelps https://ourdatahelps.org/.
19. Coppersmith G, Ngo K, Leary R, Wood A. Exploratory Analysis of Social Media Prior to a Suicide Attempt. In: *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*. ; 2016. doi:10.18653/v1/w16-0311
20. Sinnenberg L, DiSilvestro CL, Mancheno C, others. *Twitter as a Potential Data Source for Cardiovascular Disease Research*. JAMA Cardiology; 2016. https://jamanetwork.com/journals/jamacardiology/fullarticle/2556216.
21. Sap M, Park G, Eichstaedt J, others. Developing Age and Gender Predictive Lexica over Social Media. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP*. ; 2014. doi:10.3115/v1/d14-1121
22. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. *The Development and Psychometric Properties of LIWC2015*.; 2015.
23. 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.
24. Schwartz HA, Eichstaedt J, Kern ML, et al. Towards assessing changes in degree of depression through facebook. In: *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. ; 2014:118-125.
25. Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, Ungar LH. What twitter profile and posted images reveal about depression and anxiety. In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol 13. ; 2019:236-246.
26. Guntuku SC, Ramsay JR, Merchant RM, LH. U. Language of ADHD in Adults on Social Media. *J Atten Disord*. 2017;108705471773808. doi:10.1177/1087054717738083
27. Jones R. *No Title*. Drug Slang Dictionary - Words Starting With G http://www.noslang.com/drugs/dictionary.php.
28. De Choudhury M, Counts S, Horvitz E. Predicting postpartum changes in emotion and behavior via social media. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '13*. New York, New York, USA: ACM Press; 2013:3267. doi:10.1145/2470654.2466447
29. Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. *J Mach Learn Res*. 2003;3(Jan):993-1022.
30. Gelfand AE, AF. S. Sampling Based Approaches to Calculating Marginal Densities. *J Am Stat Assoc*. 1990;85:398-409. doi:10.21236/ada208388
31. Graham S, Weingart S, Milligan I. *Getting Started with Topic Modeling and MALLET*.; 2012.

32. Costa Jr PT, McCrae RR. The Revised NEO Personality Inventory (NEO-PI-R). 2008.
33. Bienvenu OJ, Samuels JF, Costa PT, Reti IM, Eaton WW, Nestadt G. Anxiety and depressive disorders and the five-factor model of personality: A higher-and lower-order personality trait investigation in a community sample. *Depress Anxiety*. 2004;20(2):92-97.
34. Booth R. Toward an Understanding of Loneliness. *Soc Work*. 1983;28(2):116-119. doi:10.1093/sw/28.2.116
35. Qualter P, Vanalst J, Harris R, others. Loneliness across the life span. *Perspect Psychol Sci*. 2015;10:250-264. https://journals.sagepub.com/doi/full/10.1177/1745691615568999?url_ver=Z39.88-2003&rfr_id=ori:rid:crossref.org&rfr_dat=cr_pub=pubmed.
36. Hawkley LC, JT. C. Loneliness Matters: A Theoretical and Empirical Review of Consequences and Mechanisms. *Ann Behav Med*. 2010;40:218-227. doi:10.1007/s12160-010-9210-8
37. Cacioppo JT, Hawkley LC, Ernst JM, others. Loneliness within a nomological net: An evolutionary perspective. *J Res Pers*. 2006;40(6):1054-1085. doi:10.1016/j.jrp.2005.11.007
38. Edwards T, Holtzman NS. A. *Meta-Analysis of Correlations between Depression and First Person Singular Pronoun Use*. Journal of Research
39. Al-Saggaf Y, Nielsen S. Self-disclosure on Facebook among female users and its relationship to feelings of loneliness. *Comput Human Behav*. 2014;36:460-468. doi:10.1016/j.chb.2014.04.014
40. Simon EB, MP. W. Sleep loss causes social withdrawal and loneliness. *Nat Commun*. 2018;9:1. doi:10.1038/s41467-018-05377-0
41. Cacioppo JT, Hawkley LC, RA. T. Perceived social isolation makes me sad: 5-year cross-lagged analyses of loneliness and depressive symptomatology in the Chicago Health, Aging, and Social Relations Study. *Psychol Aging*. 2010;25:453-463. doi:10.1037/a0017216
42. Yeginsu CUK. Appoints a Minister for Loneliness. *New York Times*. 2018;17. https://www.nytimes.com/2018/01/17/world/europe/uk-britain-loneliness.html.
43. Rubin R. Loneliness Might Be a Killer, but What's the Best Way to Protect Against It. *Jama*. 2017;318:19. doi:10.1001/jama.2017.14591
44. Jain S. CareMore Health Tackles the Unmet Challenges of the Aging Population. *Am Soc Aging*. 2018;42:14-18. https://www.ingentaconnect.com/contentone/asag/gen/2018/00000042/00000001/art00003.
45. Perese EF, Wolf M. Combating Loneliness Among Persons With Severe Mental Illness: Social Network Interventions Characteristics, Effectiveness, And Applicability. *Issues Ment Health Nurs*. 2005;26(6):591-609. doi:10.1080/01612840590959425
46. Webber M, Fendt-Newlin M. A review of social participation interventions for people with mental health problems. *Soc Psychiatry Psychiatr Epidemiol*. 2017;52:369-380. doi:10.1007/s00127-017-1372-2
47. *U. S. Twitter Reach by Age Group 2018 | Statistic*. Statista https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/.
48. Barnett I, Ethics TJ. Transparency, and Public Health at the Intersection of Innovation and Facebooks Suicide Prevention Efforts. *Ann Intern Med*. 2019;170:565. doi:10.7326/m19-

0366
49. Primack BA, Shensa A, Sidani JE, others. Social Media Use and Perceived Social Isolation Among Young Adults in the U. *S Am J Prev Med*. 2017;53(1):1-8. doi:10.1016/j.amepre.2017.01.010
50. Mckee R. Ethical issues in using social media for health and health care research. *Health Policy (New York)*. 2013;110(2-3):298-301. doi:10.1016/j.healthpol.2013.02.006
51. Inkster B, Stillwell D, Kosinski M, P. J. A decade into Facebook where is psychiatry in the digital age? *The Lancet Psychiatry*. 2016;3(11):1087-1090. doi:10.1016/s2215-0366(16)30041-4

Figure legends

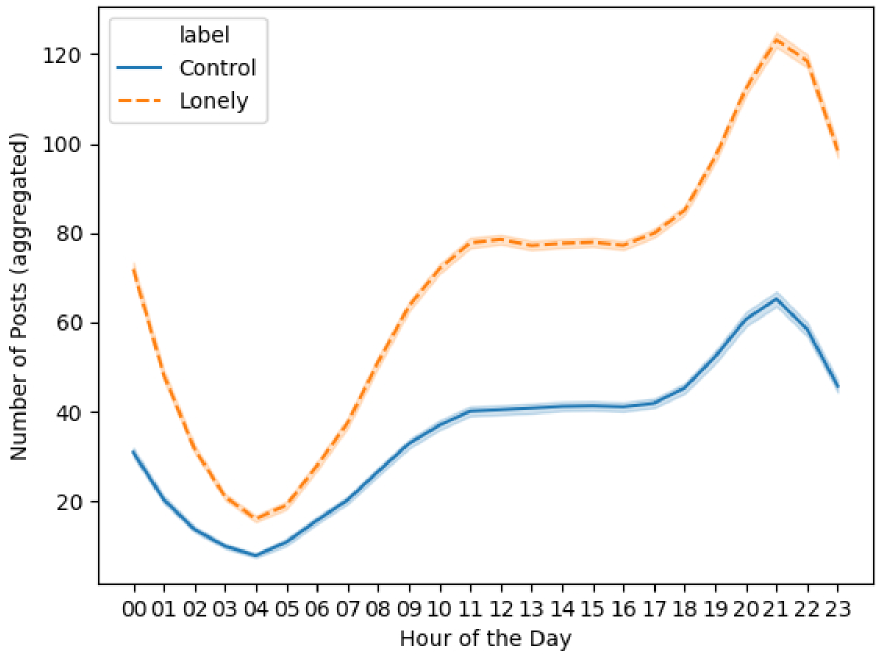
Figure 1: Words/Phrases more likely to be posted by Twitter users with a) posts including the words lonely or alone compared to the b) control group.

Word size indicates the strength of the correlation and word color indicates relative word frequency. ($p<0.001$, Bonferroni p-corrected)

Figure 2: Temporal variation showing diurnal patterns of post frequency of both the users with posts including the words lonely or alone and control group.

The dotted line indicates the percentage of posts at different hours of the day by the group of users with at least 5 posts containing the word 'lonely' or 'alone' and the solid line indicates users who do not have any posts about loneliness. The x-axis represents the hour of the day each post occurs and the y-axis indicates the number of posts for each group.

Words/Phrases more likely to be posted by Twitter users with a) self-reported loneliness (Individuals with at least 5 posts with the words ‘lonely’ or ‘alone’ group compared to the b) control group.



Temporal variation showing diurnal patterns of post frequency of both the 'lonely' and 'control' groups.

STROBE Statement—Checklist of items that should be included in reports of ***cohort studies***

	Item No	Recommendation
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract **(pg.2)**
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found **(pg.2)**
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported **(pg.4)**
Objectives	3	State specific objectives, including any prespecified hypotheses **(pg. 4)**
Methods		
Study design	4	Present key elements of study design early in the paper **(pg.5)**
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection **(pg. 5)**
Participants	6	(*a*) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up **(pg. 6)**
		(*b*) For matched studies, give matching criteria and number of exposed and unexposed
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable **(pg. 6)**
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group **(pg. 6)**
Bias	9	Describe any efforts to address potential sources of bias **(pg. 6)**
Study size	10	Explain how the study size was arrived at **(pg. 6)**
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why **(pg. 7)**
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding **(pg. 9)**
		(*b*) Describe any methods used to examine subgroups and interactions
		(*c*) Explain how missing data were addressed
		(*d*) If applicable, explain how loss to follow-up was addressed
		(*e*) Describe any sensitivity analyses
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed **(pg. 9)**
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders **(pg. 9)**
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarise follow-up time (eg, average and total amount)
Outcome data	15*	Report numbers of outcome events or summary measures over time **(pgs 10,11)**
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included **(pgs 10,11)**

		(*b*) Report category boundaries when continuous variables were categorized
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses **(pgs 10,11)**
Discussion		
Key results	18	Summarise key results with reference to study objectives **(pgs. 12, 13)**
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias **(pgs. 14)**
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence **(pg. 13, 14)**
Generalisability	21	Discuss the generalisability (external validity) of the study results **(pgs. 13, 14)**
Other information		
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based **(pg. 15)**

*Give information separately for exposed and unexposed groups.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at http://www.strobe-statement.org.