

# BMJ Open

BMJ Open is committed to open peer review. As part of this commitment we make the peer review history of every article we publish publicly available.

When an article is published we post the peer reviewers' comments and the authors' responses online. We also post the versions of the paper that were used during peer review. These are the versions that the peer review comments apply to.

The versions of the paper that follow are the versions that were submitted during the peer review process. They are not the versions of record or the final published versions. They should not be cited or distributed as the published version of this manuscript.

BMJ Open is an open access journal and the full, final, typeset and author-corrected version of record of the manuscript is available on our site with no access controls, subscription charges or pay-per-view fees (<http://bmjopen.bmj.com>).

If you have any questions on BMJ Open's open peer review process please email [info.bmjopen@bmj.com](mailto:info.bmjopen@bmj.com)

# BMJ Open

## Characterizing and Measuring Expressions of Loneliness in Individuals using Twitter

Journal:	<i>BMJ Open</i>
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, Rachele; 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

SCHOLARONE™  
Manuscripts

## Characterizing and Measuring Expressions of Loneliness in Individuals using Twitter

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

<sup>1</sup>Penn Medicine Center for Digital Health, Philadelphia, PA 19104

<sup>2</sup>Penn Medicine Center for Healthcare Innovation, Philadelphia, PA 19104

<sup>3</sup>Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104

<sup>4</sup>Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

<sup>5</sup>Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

<sup>6</sup>The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

<sup>7</sup>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

## 27 **Abstract**

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

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

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

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

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

50

51

52

1  
2  
3 **53 Strengths and Limitations of this study**  
4

- 5 **54** • Study's novel focus on timelines of social media users to study expressions of loneliness  
6  
7 and correlation with predictors of mental health.  
8  
9  
10 **56** • The study sample consists of social media users and is not representative of the general  
11  
12 **57** population.  
13  
14  
15 **58** • Though we manually annotated a subset of posts mentioning loneliness, some may have  
16  
17 **59** been metaphorical or non sequiturs.  
18

19 **60**

20  
21 **61**

22  
23 **62**

24  
25 **63**

26  
27 **64**

28  
29 **65**

30  
31 **66**

32  
33 **67**

34  
35 **68**

36  
37 **69**

38  
39 **70**

40  
41 **71**

42  
43 **72**

44  
45 **73**

46  
47 **74**

48  
49 **75**  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 76 Introduction

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

89  
90 Online data on social networks is growing exponentially. More than 2.3 billion individuals use  
91 social media regularly (e.g. Facebook 1.71 billion, Twitter 320 million, Instagram 400 million).<sup>20</sup>  
92 Increasingly, individuals are using social media as a platform to post about their thoughts,  
93 feelings, perceptions, and experiences.<sup>21-22</sup> The regular production of data on online platforms  
94 also allows for tracking of health in real-time. These data offer promise as they provide different  
95 insights than data from traditional surveys. Another opportunity is in the ability of digital  
96 platforms to not only provide markers of health but also serve as platforms that can be used for  
97 direct intervention.<sup>23-24</sup> Users on social media often post about how they are coping (or not) with  
98 life stressors and their support networks. Specifically, expressions of loneliness have been

1  
2  
3 99 associated with feeling unloved, depressed, bored, and not having friends.<sup>21-22</sup> Prior research has  
4  
5 100 also demonstrated that users' mental health conditions, such as depression and anxiety, can be  
6  
7 101 predicted from their social media language.<sup>25-26</sup>  
8  
9  
10 102

11  
12 103 Social media seeks to 'connect' people, yet several studies have reported an association between  
13  
14 104 social media use and increased perceived social isolation.<sup>27</sup> As loneliness can impact health  
15  
16 105 outcomes, identifying ways to track prevalence and manifestations of loneliness online would be  
17  
18 106 useful for developing approaches for identifying and offering support for these individuals.  
19  
20  
21 107

22  
23 108 We sought to identify data from a widely used publicly available social network, Twitter, to  
24  
25 109 characterize what and when individuals post about loneliness, association of posts with mental  
26  
27 110 health, and how manifestations of loneliness can be predicted across individuals.  
28  
29  
30 111

## 31 112 **Methods**

32  
33 113 This was a retrospective analysis of publicly available data on users posting about loneliness on  
34  
35 114 Twitter in Pennsylvania. This study was approved by the University of Pennsylvania Institutional  
36  
37 115 Review Board.  
38  
39  
40 116

### 41 117 *Twitter Data*

42 118 Twitter is a popular social media platform which allows users to send and receive short 140  
43  
44 119 character messages, or 'tweets' (at the time of this study; the character limit was later increased  
45  
46 120 to 280). First, the Twitter Streaming API was used to collect a random 1% sample of public  
47  
48 121 tweets from 2012-2016. This initial dataset was then filtered to contain only geolocated tweets or  
49  
50 122 tweets originating from users with nonempty location fields in their profile. The county of origin  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 123 of each tweet user was determined, and the dataset was filtered to obtain only tweets for users in  
4  
5 124 Pennsylvania. To increase the sample size of tweets from the state, all unique user IDs were  
6  
7  
8 125 recorded, and the Twitter search API was used to extract timelines (each user's prior 3200  
9  
10 126 tweets) filtered by timestamps ranging from 2012-2016.

11  
12 127

### 13 14 128 *Study Sample*

15  
16  
17 129 We identified users who posted the word "alone" or "lonely" at least once in their timeline  
18  
19 130 (25,966 users). Of these, 6,202 users posted messages with "alone" or "lonely" at least 5 times.  
20  
21 131 As social media includes colloquial, metaphorical, and light-hearted language (eg. "If I see Justin  
22  
23 132 Bieber, I will have a heart attack")<sup>28</sup> we sought to identify the proportion of tweets in which  
24  
25 133 lonely seemed to refer to the public health meaning rather than other uses of the term (e.g.  
26  
27 134 metaphor, joke). Two co-authors independently coded a random set of 100 tweets to identify  
28  
29 135 them as presumed to be associated with the feeling of loneliness or other. The Kappa was 0.70  
30  
31 136 and we identified that 76% of users' tweets indicate presumably feeling lonely. A few examples  
32  
33 137 are as follows: "i'm feelin real depressed, confused, & lonely", "im always the only up around  
34  
35 138 this time, feeling a lil lonely" and "I'm so Lonely in life :( I just wish I can have love again it  
36  
37 139 feels so go to be in love with someone whom loves you.". This research was done without  
38  
39 140 patient involvement.

40  
41  
42 141

### 43 44 142 *Control group*

45  
46  
47 143 We then identified a control group of users by matching each user in the above dataset to another  
48  
49 144 user by age, gender and period of activity (dates of first and last posting on twitter). We obtained  
50  
51 145 the age and gender estimates by using lexica developed previously.<sup>17</sup> Then, we selected users  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 146 with a minimum of 500 words across all their posts to have sufficient language for linguistic  
4  
5 147 analyses.<sup>29</sup> We excluded non-English, non-US tweets, retweets, and tweets that were used to  
6  
7 148 identify users in the lonely group in all analyses. Hereafter, we use ‘lonely’ group to indicate  
8  
9 149 users who had more than 5 posts with the words ‘lonely’ or ‘alone’, and ‘control’ group to  
10  
11 150 represent the matched set of users who had no such posts.  
12  
13  
14

15 151  
16  
17 152 Deriving language features to characterize individuals expressing loneliness  
18  
19 153 We used three sets of language features: a) open-vocabulary topics,<sup>30</sup> b) dictionary-based  
20  
21 154 psycholinguistic features,<sup>31</sup> c) mental well-being attributes such as anxiety, depression by  
22  
23 155 applying previously developed statistical models,<sup>32</sup> d) number of drug words and time of posts as  
24  
25 156 past research <sup>11</sup> has shown an association between loneliness and substance use. These language  
26  
27 157 features have been shown to be predictive of several health outcomes, such as depression,  
28  
29 158 schizophrenia, attention deficit hyperactivity disorder (ADHD), and general well-being.<sup>33; 26</sup>  
30  
31  
32

33 159  
34  
35 160 Open-vocabulary: As closed-vocabulary approaches like LIWC include only a small subset of  
36  
37 161 the entire language used on social media, we use an open-vocabulary approach to improve the  
38  
39 162 coverage and find topics that people who express being lonely talk about. Two hundred topics  
40  
41 163 (groups of co-occurring words) were generated using tweets across all users in the dataset of  
42  
43 164 lonely and control users using the Mallet implementation of Latent Dirichlet Allocation (LDA).<sup>34</sup>  
44  
45 165 The topic distribution of each user aggregated across all the messages was then calculated.  
46  
47  
48

49 166  
50  
51 167 Dictionary-based: From each post, we extracted the relative frequency of single words and  
52  
53 168 phrases (consisting of two or three consecutive words). Then, all words used by less than 1% of  
54  
55  
56  
57  
58  
59

1  
2  
3 169 users were removed from analysis so as to remove uncommonly used words (outliers).  
4

5 170 Additionally, all messages used to identify our study group were removed prior to further  
6  
7 171 analysis. The distribution of Linguistic Inquiry Word Count (LIWC) dictionary features are also  
8  
9  
10 172 extracted for each post. For each user, we measure the proportion of word tokens that fall into a  
11  
12 173 given LIWC category. Then, we compare it against the word tokens from the control data using  
13  
14 174 an empirical distribution of the proportion of language attributable to each LIWC category.  
15  
16

17 175  
18  
19 176 Mental well-being attributes: We used automatic text-regression methods to assign to each user  
20  
21 177 scores on the Depression, Anxiety and Anger facets for users.<sup>32</sup> This model was trained on a  
22  
23 178 sample of over 28,749 users who had taken the International Personality Item Pool Neuroticism-  
24  
25 179 Extraversion-Openness Personality Inventory Revised (IPIP NEO-PI-R) survey that contains the  
26  
27 180 Depression, Anxiety and Anger Facets of the Neuroticism Factor.<sup>32</sup> The text model was trained  
28  
29 181 using tokens and topics extracted from status updates as features. In the original validation, the  
30  
31 182 model achieved a Pearson correlation of  $r = .32$  predictive performance, which is considered high  
32  
33 183 in psychology, especially when measuring internal states.<sup>35</sup>  
34  
35  
36  
37

38 184  
39  
40 185 Use of Drug-words: We also extracted the frequency (aggregated to every user) of most common  
41  
42 186 drug words as used on social media.<sup>36</sup>  
43  
44

45 187  
46  
47 188 Temporal patterns: We determined the frequency of posts across different hours of the day by  
48  
49 189 users in both the lonely and control groups to understand the diurnal patterns in posting.  
50

51 190  
52  
53 191 Identifying differentially expressed language features in the lonely group  
54  
55  
56  
57  
58  
59

1  
2  
3 192 We isolated the patterns in users' loneliness expressions using the linguistic attributes and user  
4  
5 193 traits by correlating them with the lonely and control groups. We used Benjamini-Hochberg p-  
6  
7 194 correction and use  $p < 0.001$  for indicating meaningful correlations and the effect size was  
8  
9 195 measured using Cohen's d. The statistical analysis, data synthesis, and model creation was  
10  
11 196 conducted in 2018-2019.  
12  
13  
14

15 197

16  
17 198 Predicting the likelihood of posting about loneliness online

18  
19 199 We then looked at the feasibility of predicting whether a user is likely to express that they are  
20  
21 200 lonely or not based on their social media language. Automated analysis of social media is  
22  
23 201 accomplished by building predictive models, which use 'features,' or variables that have been  
24  
25 202 extracted from social media data. For this analysis we used LIWC and topics as features.  
26  
27 203 Features are then treated as independent variables in an algorithm (Random Forests) to predict  
28  
29 204 the dependent variable of an outcome of interest (e.g., users' saying that they are lonely or not).  
30  
31 205 For cross validation, the predictive model was trained, using Random Forests, on the training set  
32  
33 206 and then evaluated on a test set to avoid overfitting. The prediction performances are reported as  
34  
35 207 one of several possible metrics on an out-of-sample 5-fold cross validation setting.  
36  
37  
38  
39

40 208

## 41 42 209 **Results**

43  
44 210 Of the 408,296,620 tweets posted by users geo-located in Pennsylvania, USA, 25,966 users with  
45  
46 211 46,160,774 posts in their timelines, had at least one post with the words 'lonely' or 'alone', and  
47  
48 212 6,202 users (referred to as 'lonely' group hereafter) with 17,995,084 posts in their timelines, had  
49  
50 213 more than five such posts (Table 1). The lonely group had 1.9 times more posts in the study time  
51  
52  
53  
54  
55  
56  
57  
58  
59

214 period as the control (Table 1). The median estimated age of this cohort was 21 years, and 69%  
215 female.

216 **Table 1:** Descriptive statistics for the lonely group about loneliness and the control group  
217

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

218  
219 \*the lonely group is defined as any user posting at least 5 times about loneliness and the control  
220 group is defined as any user who does not have any posts about loneliness

221  
222 Identifying differentially expressed language features in the lonely group  
223 Open vocabulary approach: Analyzing differences in individual words and phrases used across  
224 both groups, we observed (Figure 1a) that users in the lonely group referred to themselves  
225 ('myself' (d=.18), 'I' (d=.16)) in their Twitter posts significantly more than the control group.  
226 They also posted about relationship issues ('want\_somebody' (d=.08), 'no\_one\_to' (d=.1), needs  
227 and feelings ('i\_just\_wanna (d=.12), 'in\_my\_feelings' (d=.1), 'i\_need' (d=.12), 'i\_cant' (d=.1)),  
228 and included more expletives. Users in the control group (Figure 1b) engaged in a lot more  
229 conversations as indicated by '<user>' (d=-.2) (we anonymize '@' mentions in users tweets as  
230 '<user>') compared to the lonely group. The control group also posted more about games

1  
2  
3 231 ('season' (d=-.09), 'coach' (d=-.07), 'team' (d=-.1)) and positivity ('!' (d=-.13), 'awesome' (d=-  
4  
5 232 .09), ':' (d=-.08)). Figure 1 illustrates the words and phrases most prominently associated with  
6  
7 233 the lonely and control groups.  
8  
9

10 234  
11  
12 235 Using topics generated from LDA, we identified the themes which occur more frequently in  
13  
14 236 posts in the lonely group. Posts were about interpersonal relationships (d=.28) (and associated  
15  
16 237 issues (d=.22)), self-reflection (d=.21) (accompanied with wondering about the future (d=.12)),  
17  
18 238 drug/alcohol use (d=.29) (considering them to be the 'only friend'), insomnia (d=.27),  
19  
20 239 uncontrolled emotions (d=.28) (accompanied by confusion (d=.11)), and psychosomatic  
21  
22 240 symptoms (d=.29). Table 2 shows the effect sizes between most prominent topic distributions  
23  
24 241 and the users who have more than 5 posts with the words lonely or alone.  
25  
26 242  
27  
28  
29

30  
31 243 Dictionary-based: Association of LIWC categories with the posts by users in the lonely group are  
32  
33 244 shown in Table 3. Individuals who posted about being alone or lonely used increased self-  
34  
35 245 references (first person pronouns, d=.18), words indicating cognitive processes (including  
36  
37 246 certainty, d=.15, discrepancies, d=.14, differentiation, d=.13 and tentativeness, d=.13), and  
38  
39 247 negative emotions (anger, d=.12 and swearing, d=.11).  
40  
41  
42  
43

44 249 Mental well-being: Users in the lonely group were more likely to have posts associated with  
45  
46 250 anger (d=.95), depression (d=.81) and anxiety (d=.75) when compared to the control group.  
47  
48

49 251  
50  
51 252 Use of Drug Words: We also identified the distribution of words pertaining to drugs in the posts  
52  
53 253 of users in the lonely group, and these were more likely to reference a blunt (d=.16), smoke  
54  
55  
56  
57  
58  
59

254 (d=.13), and heroin (d=.1), and included prescribed medications for treatment, recreational drug  
 255 use, and recreational drugs.

256  
 257 Temporal patterns: Users in the lonely group were found to post more during the night (d=.1).  
 258 We also see themes associated with night-time posting and having difficulty sleeping (d=.27) in  
 259 the open-vocabulary analysis.

260 **Table 2:** Highly correlated topics with expressions of loneliness.

261  
 262  
 263 \* Effect size is measured using Cohen's d. Only significant topics after Benjamini-Hochberg p-  
 264 correction and use  $p < 0.001$  are shown.

265  
 266 Predictive Analysis: A random forest model predicted language associated with lonely  
 267 expressions with an AUC of .86 using a combination of LIWC and LDA topics as linguistic  
 268 features.

269  
 270 **Table 3:** LIWC categories with expressions of loneliness

271

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

Swearing	0.11
----------	------

272

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

274

275

## 276 Discussion

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

281

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

291

292 Association of the lonely group with linguistic estimates of anger, depression, and anxiety  
 293 corroborate prior research.<sup>5-6</sup> Specifically, anxiety, anger, and negative mood were reported as  
 294 higher in lonely young adults.<sup>38</sup> Tweets by users in the lonely group were more self-focused  
 295 compared to the control group. Prior researchers have found that “first person singular pronouns

1  
2  
3 296 are a modest linguistic marker of depression.”<sup>39</sup> This presents the potential for early  
4  
5 297 identification and assessment to intervene on loneliness as well as mental health conditions for  
6  
7  
8 298 this group.  
9  
10 299  
11  
12 300 Trends in temporal variation in posting may reflect difficulties in terms of engaging in online  
13  
14 301 activity and doing so during hours typically devoted to sleep. Prior work has shown that sleep  
15  
16 302 deprivation can contribute to social withdrawal and loneliness.<sup>40</sup> A better understanding of the  
17  
18 303 temporality of posting could inform timing of interventions designed to address loneliness, as  
19  
20 304 well as provide insight for other researchers to test the inter-relationships between loneliness and  
21  
22 305 the motivations for using social media during nighttime.  
23  
24  
25  
26 306  
27  
28 307 Loneliness is known to be one of the primary underlying causes and correlates for chronic  
29  
30 308 mental health conditions.<sup>5-6</sup> As loneliness is becoming increasingly recognized as a public health  
31  
32 309 issue associated with chronic mental and physical health problems, several groups have taken  
33  
34 310 action to address it. For example, the United Kingdom appointed a Minister for Loneliness who  
35  
36 311 is responsible for addressing loneliness within communities.<sup>41</sup> CareMore, a health plan and  
37  
38 312 delivery system providing care for enrollees in Medicare Advantage and Medicaid health plans  
39  
40 313 in seven states across the U.S., launched the “Togetherness Program” in a clinical setting to  
41  
42 314 address loneliness in elderly patients.<sup>42</sup> Through this work, CareMore reported that participation  
43  
44 315 in exercise programs increased by 56.6%, emergency room utilization decreased by 3.3%, and  
45  
46 316 hospital admissions among participants were 20.8% lower per thousand compared to the “intent  
47  
48 317 to treat population.”<sup>43</sup> Additionally, social network interventions targeting loneliness have been  
49  
50 318 found to be effective in reducing social isolation among individuals with severe mental health  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 319 conditions but these interventions are not included in the treatment plans for individuals with a  
4  
5 320 mental illness.<sup>44</sup> Using natural language processing and machine learning to automatically  
6  
7  
8 321 identify a person expressing loneliness on Twitter could inform interventions targeted at early  
9  
10 322 identification and support for affected and at risk individuals.

11  
12 323

13  
14 324 Future work that builds off this study could be to validate whether the characteristics of people  
15  
16 325 who are using the words ‘lonely’ or ‘alone’ on Twitter can be used to track community health  
17  
18  
19 326 risks, particularly, the risk of social isolation. Our methods can potentially be used to identify  
20  
21 327 problematic loneliness for community public health monitoring.

22  
23  
24 328

### 25 26 329 **Limitations and Ethics**

27  
28 330 The study sample consists of social media users and is not representative of the general  
29  
30 331 population. An estimated 40% of US adults using Twitter are between the ages of 18 and 29, so  
31  
32 332 our analysis is skewed towards younger people.<sup>45</sup> Posts mentioning loneliness may have been  
33  
34 333 metaphorical or non sequiturs.

35  
36  
37 334

38  
39  
40 335 The feasibility of social media-based assessments of loneliness expressions (and mental health  
41  
42 336 more broadly) needs further assessment. Privacy of individuals is an ongoing concern, especially  
43  
44 337 with social media users not fully realizing the amount of health insights that can be gleaned by  
45  
46 338 their online posts. Employers and insurance companies, for example, may be motivated to derive  
47  
48  
49 339 these assessments, but could use these insights against those suffering from mental illness. As  
50  
51 340 mental illnesses carry social stigma and may engender discrimination, data protection and  
52  
53 341 ownership frameworks are needed to make sure the data is not used against the users’ interest.<sup>46</sup>

1  
2  
3 342 Further, transparency about which indicators are derived by whom for what purpose should be  
4  
5 343 part of ethical and policy discourse.  
6  
7

8 344  
9  
10 345 There are also open questions around the impact of misclassifications, and how derived mental  
11  
12 346 health indicators can be responsibly integrated into systems of care.<sup>47</sup>  
13  
14

15 347

## 16 17 348 **Conclusions**

18  
19 349 In this study we characterized expressions of loneliness on Twitter at the individual level.  
20  
21 350 Furthermore, we identified specific contexts, themes, and traits in the posts of individuals  
22  
23 351 expressing loneliness on Twitter. As loneliness is a public health challenge, a better  
24  
25 352 understanding of how loneliness is described online can inform tracking of loneliness and  
26  
27 353 interventions targeted at addressing this important public health problem in regards to the  
28  
29 354 behavior of lonely individuals that may be at risk of developing a severe mental health  
30  
31 355 condition.<sup>42</sup>  
32  
33  
34

35 356

## 36 37 357 **Acknowledgements**

38  
39  
40 358 **Funding:** This project is funded, in part, under a grant with the Pennsylvania Department of  
41  
42 359 Health (ORG-FUND-PROG-CREF is 4290-567862-2446-2049). No sponsor of funding source  
43  
44 360 played a role in: “study design and the collection, analysis, and interpretation of data and the  
45  
46 361 writing of the article and the decision to submit it for publication.” All researchers are  
47  
48 362 independent from funders  
49

50 363  
51 364 **Data Sharing Statement:** Because of our IRB requirements, data will be shared upon request  
52  
53 365 from the corresponding author.  
54

55 366  
56 367 **Contributors:** S.C. Guntuku and R. Merchant originated the study. S.C. Guntuku, R. Schneider,  
57  
58 368 A. Pelullo, L.H. Ungar, and R. Merchant developed methods, interpreted analysis, and  
59  
60 369 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.

371

372 **Disclosures:** None

373

374

375 **References:**

376

377 (1) Gerst-Emerson K, Jayawardhana J.

378 Loneliness as a Public Health Issue: The Impact of Loneliness on Health Care

379 Utilization Among Older Adults. *American*380 *Journal of Public Health*.

381 2015;105(5):1013-1019. doi:10.2105/ajph.2014.302427.

382

383 (2) New Cigna Study Reveals Loneliness at

384 Epidemic Levels in America. Cigna, a Global Health Insurance and Health Service  
385 Company.386 <https://www.cigna.com/newsroom/news-releases/2018/new-cigna-study-reveals->387 [loneliness-at-epidemic-levels-in-america.](https://www.cigna.com/newsroom/news-releases/2018/new-cigna-study-reveals-)

388 Accessed February 18, 2019.

389

390 (3) Jong-Gierveld JD. Developing and

391 testing a model of loneliness. *Journal*392 *of Personality and Social Psychology*.

393 1987;53(1):119-128. doi:10.1037//0022-3514.53.1.119.

394

395 (4) Peplau LA, Perlman D. *Loneliness: a Sourcebook of Current Theory,*396 *Research, and Therapy*. New York:

397 Wiley; 1982.

398

399 (5) Stravynski A, Boyer R. Loneliness in Relation to Suicide Ideation and Parasuicide: A

400 Population-Wide Study. *Suicide and Life-Threatening Behavior*. 2001;31(1):32-40.

401 doi:10.1521/suli.31.1.32.21312.

402

403 (6) Blai B. Health Consequences of

404 Loneliness: A Review of the Literature. *Journal of American College Health*.

405 1989;37(4):162-167. doi:10.1080/07448481.1989.9938410.

406

407 (7) Heinrich LM, Gullone E. The clinical

408 significance of loneliness: A literature review. *Clinical Psychology Review*.

409 2006;26(6):695-718.

410 doi:10.1016/j.cpr.2006.04.002.

411

- 1  
2  
3 412 (8) Richard A, Rohrmann S, Vandeleur CL,  
4 413 Schmid M, Barth J, Eichholzer M. Loneliness is adversely associated with  
5 414 physical and mental health and lifestyle factors: Results from a Swiss national  
6 415 survey. *Plos One*. 2017;12(7). doi:10.1371/journal.pone.0181442.  
7  
8 416  
9  
10 417 (9) Rico-Urbe  
11 418 LA, Caballero FF, Olaya B, et al. Loneliness, Social Networks, and Health: A  
12 419 Cross-Sectional Study in Three Countries. *Plos One*. 2016;11(1).  
13 420 doi:10.1371/journal.pone.0145264.  
14  
15 421  
16 422 (10) Pinquart M, Sorensen S. Influences on  
17 423 Loneliness in Older Adults: A Meta-Analysis. *Basic and Applied Social Psychology*.  
18 424 2001;23(4):245-266. doi:10.1207/S15324834BASP2304\_2.  
19  
20 425  
21  
22 426 (11) Rokach A. Determinants of Loneliness  
23 427 of Young Adult Drug Users. *The*  
24 428 *Journal of Psychology*.  
25 429 2002;136(6):613-630. doi:10.1080/00223980209604823.  
26  
27 430  
28 431 (12) Vanhalst J, Luyckx K, Scholte RHJ,  
29 432 Engels RCME, Goossens L. Low Self-Esteem as a Risk Factor for Loneliness in  
30 433 Adolescence: Perceived - but not Actual - Social Acceptance as an Underlying  
31 434 Mechanism. *Journal of Abnormal*  
32 435 *Child Psychology*.  
33 436 2013;41(7):1067-1081. doi:10.1007/s10802-013-9751-y.  
34  
35 437  
36  
37 438 (13) Seeman T. How Do Others Get under Our  
38 439 Skin? *Emotion, Social*  
39 440 *Relationships, and Health*.  
40 441 2001:189-220. doi:10.1093/acprof:oso/9780195145410.003.0006.  
41  
42 442  
43  
44 443 (14) Steptoe A, Shankar A, Demakakos P,  
45 444 Wardle J. Social isolation, loneliness, and all-cause mortality in older men  
46 445 and women. *Proceedings of the*  
47 446 *National Academy of Sciences*.  
48 447 2013;110(15):5797-5801. doi:10.1073/pnas.1219686110.  
49  
50 448  
51  
52 449 (15) Berkman LF, Glass T, Brissette I,  
53 450 Seeman TE. From social integration to health: Durkheim in the new millennium. *Social*

- 1  
2  
3 451 *Science & Medicine*. 2000;51(6):843-857.  
4 452 doi:10.1016/s0277-9536(00)00065-4.  
5  
6 453  
7 454 (16) Hawkley LC, Thisted RA, Masi CM,  
8 455 Cacioppo JT. Loneliness predicts increased blood pressure: 5-year cross-lagged  
9 456 analyses in middle-aged and older adults. *Psychology and Aging*.  
10 457 2010;25(1):132-141. doi:10.1037/a0017805.  
11  
12 458  
13  
14 459 (17) Sap M, Park G, Eichstaedt J, et al. Developing Age and Gender Predictive Lexica over  
15 460 Social Media. *Proceedings of the 2014 Conference on Empirical Methods in Natural*  
16 461 *Language Processing (EMNLP)*. 2014. doi:10.3115/v1/d14-1121.  
17  
18 462  
19 463 (18) Knox S, Uvnäs-Moberg K. Social  
20 464 isolation and cardiovascular disease: an atherosclerotic pathway?  
21 465 *Psychoneuroendocrinology*. 1998;23(8):877-890.  
22  
23 466  
24 467 (19) Cacioppo JT, Hughes ME, Waite LJ,  
25 468 Hawkley LC, Thisted RA. Loneliness as a specific risk factor for depressive  
26 469 symptoms: Cross-sectional and longitudinal analyses. *Psychology and Aging*.  
27 470 2006;21(1):140-151.  
28 471 doi:10.1037/0882-7974.21.1.140.  
29  
30 472  
31  
32 473 (20) U.S. population with a social media  
33 474 profile 2018. Statista.  
34 475 [https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-](https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/)  
35 476 [network-profile/](https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/).  
36 477 Accessed February 18, 2019.  
37  
38 478  
39  
40 479 (21) Kivran-Swaine F, Ting J, Naaman M. "Understanding  
41 480 Loneliness in Social Awareness Streams: Expressions and Responses. *ICWSM 2014*.  
42 481 2014.  
43 482 [https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-](https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef)  
44 483 [and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef](https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef).  
45 484  
46  
47  
48 485 (22) Kamvar S, Kamvar S, Harris JJ. *We Feel Fine: an Almanac of Human Emotion*. New York:  
49 486 Scribner; 2009.  
50  
51 487  
52 488 (23) Coppersmith G, Ngo K, Leary R, Wood A.  
53 489 Exploratory Analysis of Social Media Prior to a Suicide Attempt. *Proceedings of the*  
54 490 *Third Workshop on Computational*  
55  
56  
57  
58  
59  
60

- 1  
2  
3 491 *Lingusitics and Clinical Psychology*.  
4 492 2016. doi:10.18653/v1/w16-0311.  
5  
6 493  
7 494 (24) Qntfy, Inc. OurDataHelps.  
8 495 OurDataHelps. <https://ourdatahelps.org/>. Accessed February 18, 2019.  
9 496  
10  
11 497 (25) Guntuku SC, Buffone A, Jaidka K,  
12 498 Eichstaedt J, Ungar L. Understanding and Measuring Psychological Stress using  
13 499 Social Media. *ICWSM 2019*. 2018.  
14  
15 500  
16 501 (26) Guntuku SC, Yaden DB, Kern ML, Ungar  
17 502 LH, Eichstaedt JC. Detecting depression and mental illness on social media: an  
18 503 integrative review. *Current*  
19 504 *Opinion in Behavioral Sciences*.  
20 505 2017;18:43-49. doi:10.1016/j.cobeha.2017.07.005.  
21  
22 506  
23  
24 507 (27) Primack BA, Shensa A, Sidani JE, et  
25 508 al. Social Media Use and Perceived Social Isolation Among Young Adults in the  
26 509 U.S. *American Journal of*  
27 510 *Preventive Medicine*.  
28 511 2017;53(1):1-8. doi:10.1016/j.amepre.2017.01.010.  
29  
30 512  
31  
32 513 (28) Sinnenberg L, DiSilvestro CL, Mancheno C, et al. Twitter as a Potential Data Source for  
33 514 Cardiovascular Disease Research. *JAMA Cardiology*. December 2016.  
34 515 <https://jamanetwork.com/journals/jamacardiology/fullarticle/2556216>.  
35  
36 516  
37  
38 517 (29) Jaidka K, Guntuku SC, Buffone A, Schwartz HA, Ungar L. Facebook versus Twitter:  
39 518 Differences in Self-Disclosure and Trait Prediction. *ICWSM*. 2018.  
40  
41 519  
42 520 (30) Schwartz HA, Eichstaedt JC, Kern ML, et al. Personality, Gender, and Age in the Language  
43 521 of Social Media: The Open-Vocabulary Approach. *PLoS ONE*. 2013;8(9).  
44 522 doi:10.1371/journal.pone.0073791.  
45  
46 523  
47 524 (31) Pennebaker Jw, Jordan K, Blackburn K. The development and psychometric properties of  
48 525 LIWC2015. *UT Faculty/Researcher Works*. 2014:1-22.  
49  
50 526  
51 527 (32) Schwartz HA, Eichstaedt J, Kern ML, et al. Towards Assessing Changes in Degree of  
52 528 Depression through Facebook. *Proceedings of the Workshop on Computational*  
53 529 *Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014.  
54 530 doi:10.3115/v1/w14-3214.  
55  
56  
57  
58  
59  
60

- 1  
2  
3 531  
4 532 (33) Guntuku SC, Ramsay JR, Merchant RM, Ungar LH. Language of ADHD in Adults on  
5 533 Social Media. *Journal of Attention Disorders*. 2017;108705471773808.  
6 534 doi:10.1177/1087054717738083.  
7  
8 535  
9  
10 536 (34) Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *The Journal of Machine Learning*  
11 537 *Research*. 2003;3.  
12 538  
13  
14 539 (35) Meyer GJ, Finn SE, Eyde LD, et al. Psychological testing and psychological assessment. A  
15 540 review of evidence and issues. *The American psychologist*. 2001;56(2):128-165.  
16 541  
17  
18 542 (36) Jones R. Drug Slang Dictionary - Words Starting With G.  
19 543 <http://www.noslang.com/drugs/dictionary.php>. Accessed February 19, 2019.  
20 544  
21  
22 545 (37) Booth R. Toward an Understanding of  
23 546 Loneliness. *Social Work*. 1983;28(2):116-119. doi:10.1093/sw/28.2.116.  
24 547  
25  
26 548 (38) Cacioppo JT, Hawkley LC, Ernst JM, et  
27 549 al. Loneliness within a nomological net: An evolutionary perspective. *Journal of*  
28 550 *Research in Personality*. 2006;40(6):1054-1085.  
29 551 doi:10.1016/j.jrp.2005.11.007.  
30 552  
31  
32 553 (39) Edwards T, Holtzman NS. A  
33 554 meta-analysis of correlations between depression and first person singular  
34 555 pronoun use. *Journal of Research*  
35 556 *in Personality*. 2017;68:63-68.  
36 557 doi:10.1016/j.jrp.2017.02.005.  
37 558  
38  
39 559 (40) Simon EB, Walker MP. Sleep loss causes  
40 560 social withdrawal and loneliness. *Nature*  
41 561 *Communications*. 2018;9(1).  
42 562 doi:10.1038/s41467-018-05377-0.  
43 563  
44  
45 564 (41) Yeginsu C. U.K. Appoints a Minister for Loneliness. *The New York Times*.  
46 565 <https://www.nytimes.com/2018/01/17/world/europe/uk-britain-loneliness.html>. Published  
47 566 January 17, 2018.  
48 567  
49  
50 568 (42) Rubin R. Loneliness Might Be a Killer,  
51 569 but What's the Best Way to Protect Against It? *Jama*.  
52 570 2017;318(19):1853. doi:10.1001/jama.2017.14591.  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 571  
4 572 (43) CareMore Health Announces New Outcomes  
5  
6 573 Data from First-of-its-Kind Togetherness Program. Business Wire A Berkshire  
7 574 Hathaway Company . <https://www.businesswire.com/news/home/20181218005059/en/>.  
8 575 Published December 18, 2018. Accessed February 18, 2019.  
9 576  
10  
11 577 (44) Perese EF, Wolf M. Combating  
12 578 Loneliness Among Persons With Severe Mental Illness: Social Network  
13 579 Interventions Characteristics, Effectiveness, And Applicability. *Issues in Mental Health*  
14 580 *Nursing*. 2005;26(6):591-609.  
15 581 doi:10.1080/01612840590959425.  
16 582  
17  
18  
19 583 (45) U.S. Twitter reach by age group 2018 |  
20 584 Statistic. Statista.  
21 585 [https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-](https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/)  
22 586 [age-group/](https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/).  
23 587 Accessed February 18, 2019.  
24  
25 588  
26  
27 589 (46) Mckee R. Ethical issues in using  
28 590 social media for health and health care research. *Health Policy*. 2013;110(2-3):298-301.  
29 591 doi:10.1016/j.healthpol.2013.02.006.  
30 592  
31  
32 593 (47) Inkster B, Stillwell D, Kosinski M,  
33 594 Jones P. A decade into Facebook: where is psychiatry in the digital age? *The Lancet*  
34 595 *Psychiatry*. 2016;3(11):1087-1090.  
35 596 doi:10.1016/s2215-0366(16)30041-4.  
36  
37 597  
38 598  
39 599  
40 600  
41  
42  
43 601

## 602 **Figure legends**

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



1  
2  
3 606 Word size indicates the strength of the correlation and word color indicates relative word  
4  
5 607 frequency. ( $p < 0.01$ , Bonferroni p-corrected)  
6  
7

8 608

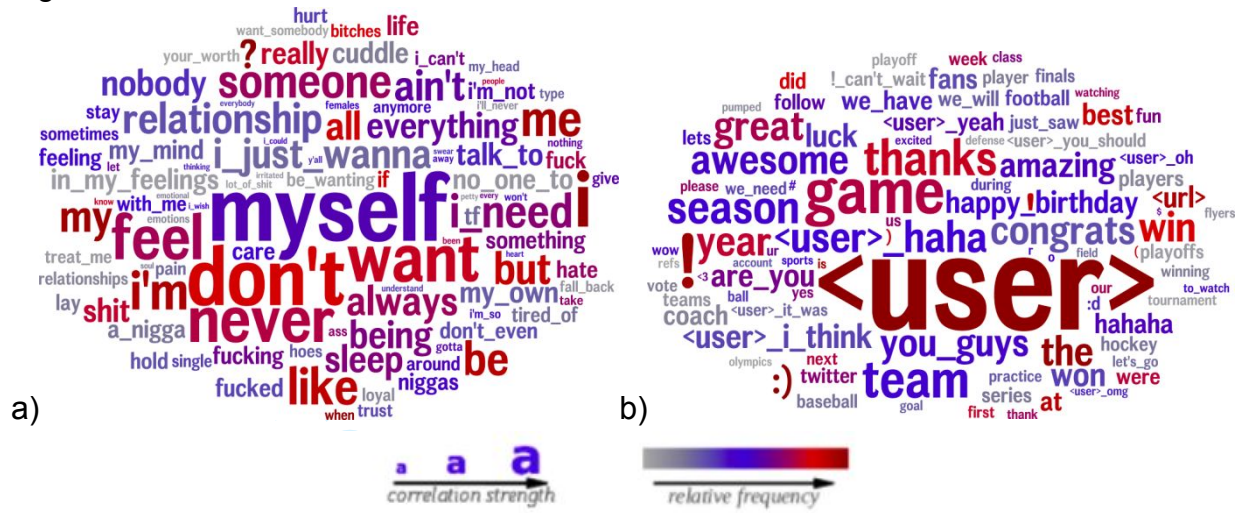
9  
10 609 **Figure 2: Temporal variation showing diurnal patterns of post frequency of both the**  
11  
12 610 **'lonely' and 'control' groups.**  
13

14 611 The dotted line indicates the percentage of posts at different hours of the day by the group of  
15  
16 612 users with at least 5 posts containing the word 'lonely' or 'alone' and the solid line indicates  
17  
18 613 users who do not have any posts about loneliness. The x-axis represents the hour of the day each  
19  
20 614 post occurs and the y-axis indicates the number of posts for each group.  
21  
22  
23

24 615

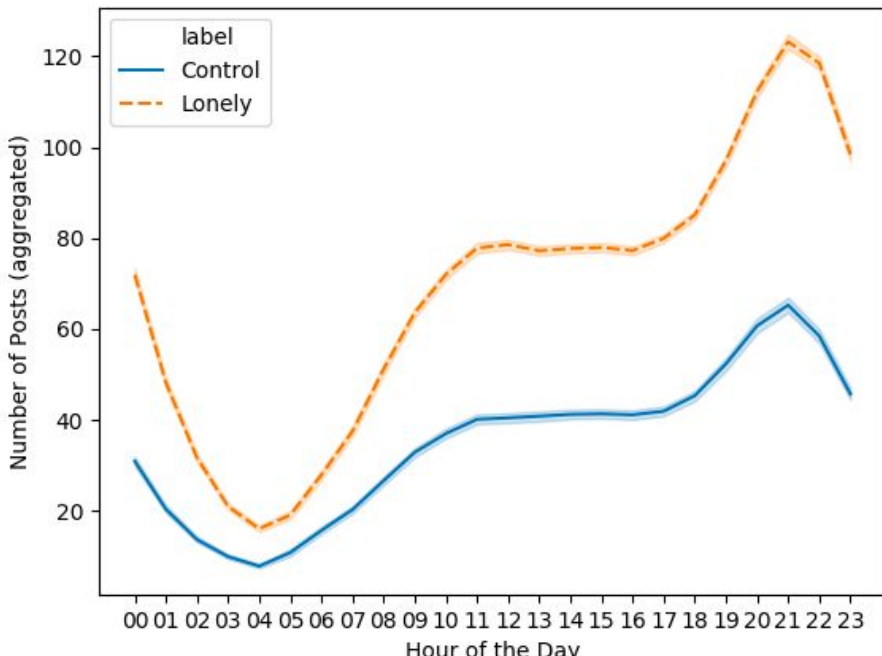
25  
26 616  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

Figure 1:



peer review only

Figure 2



For peer review only

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

	Item No	Recommendation
<b>Title and abstract</b>	1	(a) Indicate the study's design with a commonly used term in the title or the abstract ( <b>pg.2</b> ) (b) Provide in the abstract an informative and balanced summary of what was done and what was found ( <b>pg.2</b> )
<b>Introduction</b>		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported ( <b>pg.4</b> )
Objectives	3	State specific objectives, including any prespecified hypotheses ( <b>pg. 4</b> )
<b>Methods</b>		
Study design	4	Present key elements of study design early in the paper ( <b>pg.5</b> )
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection ( <b>pg. 5</b> )
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up ( <b>pg. 6</b> ) (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 ( <b>pg. 6</b> )
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 ( <b>pg. 6</b> )
Bias	9	Describe any efforts to address potential sources of bias ( <b>pg. 6</b> )
Study size	10	Explain how the study size was arrived at ( <b>pg. 6</b> )
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why ( <b>pg. 7</b> )
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding ( <b>pg. 9</b> ) (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
<b>Results</b>		
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 ( <b>pg. 9</b> ) (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 ( <b>pg. 9</b> ) (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 ( <b>pgs 10,11</b> )
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 ( <b>pgs 10,11</b> )

1		(b) Report category boundaries when continuous variables were categorized
2		(c) If relevant, consider translating estimates of relative risk into absolute risk for a
3		meaningful time period
4		
5	Other analyses	17 Report other analyses done—eg analyses of subgroups and interactions, and
6		sensitivity analyses (pgs 10,11)
7		
8	<b>Discussion</b>	
9	Key results	18 Summarise key results with reference to study objectives (pgs. 12, 13)
10	Limitations	19 Discuss limitations of the study, taking into account sources of potential bias or
11		imprecision. Discuss both direction and magnitude of any potential bias (pgs. 14)
12		
13	Interpretation	20 Give a cautious overall interpretation of results considering objectives, limitations,
14		multiplicity of analyses, results from similar studies, and other relevant evidence
15		(pg. 13, 14)
16	Generalisability	21 Discuss the generalisability (external validity) of the study results (pgs. 13, 14)
17		
18	<b>Other information</b>	
19	Funding	22 Give the source of funding and the role of the funders for the present study and, if
20		applicable, for the original study on which the present article is based (pg. 15)
21		

22  
23 \*Give information separately for exposed and unexposed groups.

24  
25  
26 **Note:** An Explanation and Elaboration article discusses each checklist item and gives methodological background and  
27 published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely  
28 available on the Web sites of PLoS Medicine at <http://www.plosmedicine.org/>, Annals of Internal Medicine at  
29 <http://www.annals.org/>, and Epidemiology at <http://www.epidem.com/>). Information on the STROBE Initiative is  
30 available at <http://www.strobe-statement.org>.  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

# BMJ Open

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

Journal:	<i>BMJ Open</i>
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, Rachele; 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
<b>Primary Subject Heading</b>:	Public health
Secondary Subject Heading:	Communication, Mental health
Keywords:	loneliness, social media, natural language processing, MENTAL HEALTH, STATISTICS & RESEARCH METHODS

SCHOLARONE™  
Manuscripts

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

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

<sup>1</sup>Penn Medicine Center for Digital Health, Philadelphia, PA 19104

<sup>2</sup>Penn Medicine Center for Healthcare Innovation, Philadelphia, PA 19104

<sup>3</sup>Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104

<sup>4</sup>Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

<sup>5</sup>Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

<sup>6</sup>The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

<sup>7</sup>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.).<sup>1</sup> 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%).<sup>2</sup> 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.”<sup>1, 3, 4</sup> Loneliness is also one of the primary underlying causes and correlates for chronic mental health conditions and physician visits in some populations.<sup>2, 5-9</sup>

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.<sup>10-12</sup> Studies have also looked at the co-occurrence of substance use and loneliness as a risk factor in adolescent.<sup>13</sup> 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).<sup>14</sup> 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.<sup>15</sup> 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,

1  
2  
3 preferences, and interests, social media has emerged as a powerful tool to passively measure  
4 behaviors of people<sup>16-17</sup>.  
5  
6  
7  
8  
9

10 Moreover, social media is being increasingly used for communicating about mental health.<sup>18-19</sup>,  
11 opening an avenue to uncover insights that might be different from data using traditional surveys  
12 considering the passive data collection on social media. For example, stressed and depressed  
13 individuals use more first-person singular pronouns suggesting higher self-focus and  
14 communities with heart disease discuss hate more frequently.<sup>18-20</sup> Natural language processing  
15 and machine learning automate the analysis of posts that would have been too hard to evaluate  
16 without that automation, and have revealed their value in using social media posts to predict  
17 mental health. For example, individual's Facebook posts 6 months immediately preceding the  
18 first documented diagnosis of depression yielded a prediction AUC of 0.72.<sup>21</sup> Further,  
19 preliminary work studying expressions of loneliness on social media have found associations  
20 with feeling unloved, depressed, bored, and not having friends.<sup>16-17</sup> Another opportunity is in the  
21 ability of digital platforms to not only provide markers of health but also serve as platforms that  
22 can be used for direct intervention.<sup>22-23</sup>  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

42 While social media use has also been associated with increased perceived social isolation<sup>24</sup>, in  
43 this study, we are interested to understand expressions of loneliness as they manifest on social  
44 media. Specifically, we sought to characterize individuals' posts about loneliness on Twitter.  
45 Studying the language of users who express being lonely or alone, we analyze the correlations  
46 between loneliness and users' mental health attributes, and several psycholinguistic attributes  
47 inspired from prior work at the intersection of mental health and natural language processing on  
48 social media. Privacy of individuals has to be at the forefront of this research to shield  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 unintended use of this data, specifically with the amount of health insights that can be gleaned  
4  
5 from social media.  
6  
7  
8  
9

10 We hypothesize that language usage patterns would both confirm existing understanding of  
11 loneliness and give new insights into the daily lives of those who express being lonely. As  
12 loneliness can impact health outcomes, identifying ways to track prevalence and manifestations  
13 of loneliness online would be useful for developing approaches for identifying and offering  
14 support for these individuals.  
15  
16  
17  
18  
19  
20  
21  
22

## 23 **Methods**

24  
25  
26 This was a retrospective analysis of publicly available data on users posting about loneliness on  
27 Twitter. This study was exempt by the University of Pennsylvania Institutional Review Board.  
28  
29  
30  
31

### 32 *Twitter Data*

33  
34  
35 Twitter is a popular social media platform which allows users to send and receive short 140  
36 character messages, or ‘tweets’ (at the time of this study; the character limit was later increased  
37 to 280). First, from the Twitter Streaming API, we collected tweets from the 1% sample using a  
38 bounding box of location coordinates around Pennsylvania. The county of origin of each tweet  
39 user was determined. To increase the sample size of tweets from the state, all unique user IDs  
40 were recorded, and the Twitter search API was used to extract timelines (each user’s prior 3200  
41 tweets) filtered by timestamps ranging from 2012-2016 geolocated in Pennsylvania.  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51

### 52 *Patient and Public Involvement*

1  
2  
3 Patients and public were not involved in the development of the research question and outcome  
4  
5 measures.  
6  
7  
8  
9

### 10 *Study Sample*

11  
12 We identified users who posted the word “alone” or “lonely” at least once in their timeline  
13  
14 (25,966 users). Of these, 6,202 users posted messages with “alone” or “lonely” at least 5 times.  
15  
16 As social media includes colloquial, metaphorical, and light-hearted language (eg. “If I see Justin  
17  
18 Bieber, I will have a heart attack”) we sought to identify the proportion of tweets in which lonely  
19  
20 seemed to refer to the public health meaning rather than other uses of the term (e.g. metaphor,  
21  
22 joke).<sup>25</sup> Two co-authors independently coded a random set of 100 tweets from individuals who  
23  
24 used the words lonely/alone at least 5 times in their timeline to identify them as presumed to be  
25  
26 associated with the feeling of loneliness or other. The Kappa was 0.70 and we identified that  
27  
28 76% of users’ tweets indicate presumably feeling lonely. A few examples are as follows: “i’m  
29  
30 feelin real depressed, confused, & lonely”, “im always the only up around this time, feeling a lil  
31  
32 lonely” and “I’m so Lonely in life :( I just wish I can have love again it feels so go to be in love  
33  
34 with someone whom loves you.” Distribution of users with different number of lonely/alone  
35  
36 words in their Twitter timeline and the temporal distribution of tweets containing these words is  
37  
38 shown in supplementary file.  
39  
40  
41  
42  
43  
44  
45  
46

### 47 *Control group*

48  
49 We then identified a control group of users by matching each user in the above dataset to another  
50  
51 user by age, gender and period of activity (dates of first and last posting on twitter). We obtained  
52  
53 the age and gender estimates by using lexica developed previously.<sup>26</sup> Then, we selected users  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 with a minimum of 500 words across all their posts to have sufficient language for linguistic  
4 analyses.<sup>27</sup> We excluded non-English, non-US tweets, retweets, and tweets containing ‘alone’  
5 and/or ‘lonely’ that were used to identify users in the lonely group in all analyses to identify  
6 linguistic features that are actually characteristics of lonelier people -- looking at their entire  
7 timeline of tweets. Hereafter, we use ‘lonely’ group to indicate users who had more than 5 posts  
8 with the words ‘lonely’ or ‘alone’, and ‘control’ group to represent the matched set of users who  
9 had no such posts.  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20

### 21 *Deriving language features to characterize individuals expressing loneliness*

22 We used four sets of language features: a) open-vocabulary topics,<sup>28</sup> b) dictionary-based  
23 psycholinguistic features,<sup>29</sup> c) mental well-being attributes such as anxiety, depression by  
24 applying previously developed statistical models,<sup>30</sup> d) number of drug words and time of posts as  
25 past research has shown an association between loneliness and substance use.<sup>11; 13</sup> These  
26 language features have been shown to be predictive of several health outcomes, such as  
27 depression, schizophrenia, attention deficit hyperactivity disorder (ADHD), and general well-  
28 being.<sup>31; 19</sup>  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

42 *Open-vocabulary:* As closed-vocabulary approaches like LIWC include only a small subset of  
43 the entire language used on social media, we use an open-vocabulary approach to improve the  
44 coverage and find topics that people who express being lonely talk about. Topics consist of  
45 clusters of co-occurring words created using Latent Dirichlet Allocation (LDA).<sup>32</sup> The LDA  
46 generative model assumes that tweets contain a combination of topics, and that topics are a  
47 distribution of words. Since the words in a tweet are known, topics, which are latent variables,  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 can be estimated through Gibbs sampling.<sup>33</sup> We use the Mallet implementation of the LDA  
4  
5 algorithm, adjusting one parameter (alpha=5) to favor fewer topics per tweet.<sup>34</sup> All other  
6  
7 parameters were kept at their default. An example of such a model is the following sets of words  
8  
9 ('tuesday', 'monday', 'wednesday', ...) which clusters together days of the week by exploiting  
10  
11 their similar distributional properties across tweets. In our study, two hundred topics were  
12  
13 generated using tweets across all users in the dataset of lonely and control users.  
14  
15  
16  
17  
18

19 *Dictionary-based:* From each post, we extracted the relative frequency of single words and  
20  
21 phrases (consisting of two or three consecutive words). Then, all words used by less than 1% of  
22  
23 users were removed from analysis so as to remove uncommonly used words (outliers).  
24  
25

26 Additionally, all messages used to identify our study group were removed prior to further  
27  
28 analysis. The Linguistic Inquiry Word Count (LIWC) dictionary is a language-specific, many-to-  
29  
30 many mapping of tokens (including words and word stems) and psychologically validate  
31  
32 categories. Each category (a curated list of words) is found to be correlated with and also  
33  
34 predictive of several psychological traits and outcomes. For each user, we measure the  
35  
36 proportion of word tokens that fall into a given LIWC category.  
37  
38  
39  
40  
41

42 *Mental well-being attributes:* We used automatic text-regression methods to assign to each user  
43  
44 scores on the depression, anxiety and anger facets for users.<sup>30</sup> This model was trained on a  
45  
46 sample of over 28,749 users who had taken the International Personality Item Pool Neuroticism-  
47  
48 Extraversion-Openness Personality Inventory Revised (IPIP NEO-PI-R) survey that contains the  
49  
50 depression, anxiety and anger Facets of the Neuroticism Factor.<sup>30</sup> The machine learning model  
51  
52 trained on words and phrases from Facebook posts to predict survey measure of depression,  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 anger and anxiety resulted in a performance of  $r = .32$ , which is considered high in psychology,  
4 especially when measuring internal states.<sup>35</sup> The model was trained using status updates of users  
5 from another study<sup>30</sup>, and has been shown to generalize to Twitter users.<sup>36</sup>  
6  
7  
8  
9

10  
11  
12 *Use of Drug-words:* We also extracted the frequency (aggregated to every user) of most common  
13 drug words as used on social media.<sup>37</sup>  
14  
15  
16  
17

18  
19 *Temporal patterns:* We determined the frequency of posts across different hours of the day by  
20 users in both the lonely and control groups to understand the diurnal patterns in posting.  
21  
22  
23  
24  
25

### 26 *Identifying differentially expressed language features in the lonely group*

27  
28 We isolated the patterns in users' loneliness expressions using the linguistic attributes and user  
29 traits by correlating them with the lonely and control groups. We use logistic regression to  
30 distinguish open-vocabulary words, phrases, LIWC categories and topics associated with lonely  
31 and control groups and measure the effect size using Cohen's D. Details of the method are  
32 described in a previous work<sup>28</sup>. We used Benjamini-Hochberg p-correction and use  $p < 0.001$  for  
33 indicating meaningful correlations and the effect size was measured using Cohen's D. The  
34 statistical analysis, data synthesis, and model creation was conducted in 2018-2019.  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46

### 47 *Predicting the likelihood of posting about loneliness online*

48  
49 We then looked at the feasibility of predicting whether a user is likely to express that they are  
50 lonely or not based on their social media language. Automated analysis of social media is  
51 accomplished by building predictive models, which use 'features', or variables that have been  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 extracted from social media data. For this analysis we used LIWC and topics as features.  
4  
5 Features are then treated as independent variables in an algorithm (Random Forests) to predict  
6  
7 the dependent variable of an outcome of interest (e.g., users' saying that they are lonely or not).  
8  
9  
10 For cross validation, the predictive model was trained, using Random Forests, on the training set  
11  
12 and then evaluated on a test set to avoid overfitting. The prediction performances are reported as  
13  
14 Area Under the Receiver Operating Curves (AUC) and several performance metrics on an out-  
15  
16 of-sample 5-fold cross validation setting.  
17  
18  
19  
20

## 21 **Results**

22  
23 Of the 408,296,620 tweets posted by users geo-located in Pennsylvania, USA, 25,966 users with  
24  
25 46,160,774 posts in their timelines, had at least one post with the words 'lonely' or 'alone', and  
26  
27 6,202 users (referred to as 'lonely' group hereafter) with 17,995,084 posts in their timelines, had  
28  
29 more than five such posts (Table 1). The lonely group had 1.9 times more posts in the study time  
30  
31 period as the control (Table 1). The median estimated age of this cohort was 21 years, and 69%  
32  
33 female.  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**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	0.281
	hurt, feelings, trust, forget	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*
<b>Pronouns</b>	
1st Person Pronouns	0.18
<b>Cognitive Processes</b>	
Certainty	0.15
Discrepancies	0.15
Differentiation	0.14
Tentativeness	0.13
<b>Negative Emotions</b>	

Swearing	0.11
<b>Mental Well-being</b>	
Depression	0.81
Anger	0.95
Anxiety	0.75
<b>Drug words</b>	
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.<sup>38</sup> This expression of negativity related to relationships is likely related to a

1  
2  
3 hypervigilance to social threat, associated with loneliness.<sup>39</sup> Lonely individuals were also  
4 reported to focus on overcoming past events as well as showing feelings of helplessness.<sup>38</sup>  
5  
6  
7  
8  
9

10 Association of the lonely group with linguistic estimates of anger, depression, and anxiety  
11 corroborate prior research, showing that loneliness and social isolation influence psychological  
12 functioning , specifically the ability to self-regulate emotion.<sup>5-6; 40</sup> Specifically, anxiety, anger,  
13 and negative mood were reported as higher in lonely young adults.<sup>41</sup> Tweets by users in the  
14 lonely group were more self-focused compared to the control group. Prior researchers have  
15 found that “first person singular pronouns are a modest linguistic marker of depression.”<sup>42</sup> Also,  
16 previous research has shown that loneliness has been associated with greater self-disclosure in  
17 Facebook posts.<sup>43</sup> This presents the potential for early identification and assessment to intervene  
18 on loneliness as well as mental health conditions for this group.  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32

33 Trends in temporal variation in posting may reflect that sleep deprivation can contribute to social  
34 withdrawal and loneliness.<sup>44</sup> This finding corroborates prior research associating loneliness with  
35 diminished sleep quality.<sup>40</sup> A better understanding of the temporality of posting could inform  
36 timing of interventions designed to address loneliness, as well as provide insight for other  
37 researchers to test the inter-relationships between loneliness and the motivations for using social  
38 media during nighttime.  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48

49 Loneliness is known to be one of the primary underlying causes and correlates for chronic  
50 mental health conditions.<sup>5-6; 45</sup> As loneliness is becoming increasingly recognized as a public  
51 health, several groups have taken action to address it. For example, the United Kingdom  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 appointed a Minister for Loneliness who is responsible for addressing loneliness within  
4  
5 communities.<sup>46</sup> CareMore, a health plan and delivery system providing care for enrollees in  
6  
7 Medicare Advantage and Medicaid health plans in seven states across the U.S., launched the  
8  
9 “Togetherness Program” in a clinical setting to address loneliness in elderly patients.<sup>47</sup> Through  
10  
11 this work, CareMore reported that participation in exercise programs increased by 56.6%,  
12  
13 emergency room utilization decreased by 3.3%, and hospital admissions among participants were  
14  
15 20.8% lower per thousand compared to the “intent to treat population.”<sup>48</sup> Additionally, social  
16  
17 network interventions targeting loneliness have been found to be effective in reducing social  
18  
19 isolation among individuals with severe mental health conditions but these interventions are not  
20  
21 included in the treatment plans for individuals with a mental illness.<sup>49-50</sup>  
22  
23  
24  
25  
26  
27

28 Considering the advantage of large sample sizes and also the association between increased  
29  
30 social media usage and individuals expressions of loneliness, it is promising to use natural  
31  
32 language processing and machine learning to automatically identify a person expressing  
33  
34 loneliness on Twitter to inform interventions targeted at early identification and support for  
35  
36 affected and at risk individuals with the caveat that social media users are not representative of a  
37  
38 random sample of individuals. To address loneliness will require being able to identify it  
39  
40 passively, remotely, and over time. Many people rarely visit a healthcare provider so would miss  
41  
42 the opportunity for screening. Approaches for treatment will also need to harness the tools and  
43  
44 technologies that are accessible and integrated with the things people use every day (e.g. mobile  
45  
46 phones). Future interventions would have to potentially rely on digital phenotyping of loneliness  
47  
48 and using digital platforms (e.g. text messaging) to complement human-to-human interaction  
49  
50 strategies to treat loneliness.  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 In this first study, our aim was to characterize loneliness expressions based on users' entire  
4 timelines. Future studies could perform a time-series analysis of the temporal variations  
5 associated with loneliness expressions. Further, works should also validate whether the  
6 characteristics of people who are using the words 'lonely' or 'alone' on Twitter can be used to  
7 track community health risks, particularly, the risk of social isolation. Other studies should  
8 replicate the findings in this study using more formal ground truth such as surveys and extend  
9 this work to investigate if Twitter can potentially map regional hotspots of loneliness to identify  
10 problematic loneliness for community public health monitoring.  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23

### 24 **Limitations and Ethics**

25  
26 The study sample consists of social media users and is not representative of the general  
27 population. An estimated 40% of US adults using Twitter are between the ages of 18 and 29, so  
28 our analysis is skewed towards younger people.<sup>51</sup> Considering we identified that 76% of users'  
29 tweets indicated presumably feeling lonely in the sample we hand coded, posts mentioning the  
30 words alone or lonely may have been metaphorical or non sequiturs. Also, considering the  
31 inclusion criteria based on number of tweets mentioning alone or lonely, we are potentially  
32 selecting users with more posts than the average twitter user. Additionally, Twitter is far from  
33 perfect to be used as a diagnostic tool. However, an automated machine learning tool could be a  
34 low-cost method to potentially detect elevated loneliness levels in a person who could then be  
35 referred to more formal screening methods. Further, the effects presented in this dataset may not  
36 be specific to loneliness considering the potential comorbidity with mental health conditions  
37 such as depression in this dataset.  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 The feasibility of social media-based assessments of loneliness expressions (and mental health  
4 more broadly) needs further assessment. Privacy of individuals is an ongoing concern, especially  
5 with social media users not fully realizing the amount of health insights that can be gleaned by  
6 their online posts. Employers and insurance companies, for example, may be motivated to derive  
7 these assessments, but could use these insights against those suffering from mental illness. As  
8 mental illnesses carry social stigma and may engender discrimination, data protection and  
9 ownership frameworks are needed to make sure the data is not used against the users' interest.<sup>52</sup>  
10  
11 Further, transparency about which indicators are derived by whom for what purpose should be  
12 part of ethical and policy discourse.  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25

26 There are also open questions around the impact of misclassifications, and how derived mental  
27 health indicators can be responsibly integrated into systems of care.<sup>53</sup>  
28  
29  
30  
31  
32

### 33 **Conclusions**

34  
35 In this study we characterized expressions of loneliness on Twitter at the individual level.  
36 Furthermore, we identified specific contexts, themes, and traits in the posts of individuals  
37 expressing loneliness on Twitter. As loneliness is a public health challenge, a better  
38 understanding of how loneliness is described online can inform tracking of loneliness and  
39 interventions targeted at addressing this important public health problem in regards to the  
40 behavior of lonely individuals that may be at risk of developing a severe mental health  
41 condition.<sup>47</sup>  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 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.

- 1  
2  
3  
4  
5 (5) Stravynski A, Boyer R. Loneliness in Relation to Suicide Ideation and Parasuicide: A  
6 Population-Wide Study. *Suicide and Life-Threatening Behavior*. 2001;31(1):32-40.  
7 doi:10.1521/suli.31.1.32.21312.  
8  
9  
10 (6) Blai B. Health Consequences of  
11 Loneliness: A Review of the Literature. *Journal of American College Health*.  
12 1989;37(4):162-167. doi:10.1080/07448481.1989.9938410.  
13  
14  
15 (7) Heinrich LM, Gullone E. The clinical  
16 significance of loneliness: A literature review. *Clinical Psychology Review*.  
17 2006;26(6):695-718.  
18 doi:10.1016/j.cpr.2006.04.002.  
19  
20  
21 (8) Richard A, Rohrmann S, Vandeleur CL,  
22 Schmid M, Barth J, Eichholzer M. Loneliness is adversely associated with  
23 physical and mental health and lifestyle factors: Results from a Swiss national  
24 survey. *Plos One*. 2017;12(7). doi:10.1371/journal.pone.0181442.  
25  
26  
27  
28 (9) Rico-Uribe  
29 LA, Caballero FF, Olaya B, et al. Loneliness, Social Networks, and Health: A  
30 Cross-Sectional Study in Three Countries. *Plos One*. 2016;11(1).  
31 doi:10.1371/journal.pone.0145264.  
32  
33  
34  
35 (10) Pinquart M, Sorensen S. Influences on  
36 Loneliness in Older Adults: A Meta-Analysis. *Basic and Applied Social Psychology*.  
37 2001;23(4):245-266. doi:10.1207/S15324834BASP2304\_2.  
38  
39  
40 (11) Rokach A. Determinants of Loneliness  
41 of Young Adult Drug Users. *The*  
42 *Journal of Psychology*.  
43 2002;136(6):613-630. doi:10.1080/00223980209604823.  
44  
45  
46  
47 (12) Vanhalst J, Luyckx K, Scholte RHJ,  
48 Engels RCME, Goossens L. Low Self-Esteem as a Risk Factor for Loneliness in  
49 Adolescence: Perceived - but not Actual - Social Acceptance as an Underlying  
50 Mechanism. *Journal of Abnormal*  
51 *Child Psychology*.  
52 2013;41(7):1067-1081. doi:10.1007/s10802-013-9751-y.  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 (13) Page RM, Allen O, Moore L, *et al.* Co-Occurrence of Substance Use and Loneliness as a  
4 Risk Factor for Adolescent Hopelessness. *Journal of School Health* 1993;**63**:104–8.  
5 doi:10.1111/j.1746-1561.1993.tb06090.x  
6  
7  
8 (14) U.S. population with a social media  
9 profile 2018. Statista.  
10 [https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-](https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/)  
11 [network-profile/](https://www.statista.com/statistics/273476/percentage-of-us-population-with-a-social-network-profile/).  
12 Accessed February 18, 2019.  
13  
14  
15  
16 (15) Lenhart A. Mobile Access Shifts Social Media Use and Other Online Activities. Pew  
17 Research Center: Internet, Science & Tech.  
18 2015. [http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-](http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-other-online-activities/)  
19 [other-online-activities/](http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-other-online-activities/) (accessed 9 May2019).  
20  
21  
22  
23 (16) Kivran-Swaine F, Ting J, Naaman M. Understanding  
24 Loneliness in Social Awareness Streams: Expressions and Responses. *ICWSM 2014*.  
25 2014.  
26 [https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-](https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef)  
27 [and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef](https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef).  
28  
29  
30  
31 (17) Kamvar S, Kamvar S, Harris JJ. *We Feel Fine: an Almanac of Human Emotion*.  
32 New York: Scribner; 2009.  
33  
34  
35 (18) Guntuku SC, Buffone A, Jaidka K,  
36 Eichstaedt J, Ungar L. Understanding and Measuring Psychological Stress using  
37 Social Media. *ICWSM 2019*. 2018.  
38  
39  
40 (19) Guntuku SC, Yaden DB, Kern ML, Ungar  
41 LH, Eichstaedt JC. Detecting depression and mental illness on social media: an  
42 integrative review. *Current*  
43 *Opinion in Behavioral Sciences*.  
44 2017;18:43-49. doi:10.1016/j.cobeha.2017.07.005.  
45  
46  
47  
48 (20) Eichstaedt JC, Schwartz HA, Kern ML, *et al.* Psychological Language on Twitter Predicts  
49 County-Level Heart Disease Mortality. *Psychological Science* 2015;**26**:159–69.  
50 doi:10.1177/0956797614557867  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 (21) Eichstaedt JC, Smith RJ, Merchant RM, *et al.* Facebook language predicts depression in  
4 medical records. *Proceedings of the National Academy of Sciences* 2018;**115**:11203–8.  
5 doi:10.1073/pnas.1802331115  
6  
7  
8 (22) Coppersmith G, Ngo K, Leary R, Wood A.  
9 Exploratory Analysis of Social Media Prior to a Suicide Attempt. *Proceedings of the*  
10 *Third Workshop on Computational*  
11 *Linguistics and Clinical Psychology*.  
12 2016. doi:10.18653/v1/w16-0311.  
13  
14  
15 (23) Qntfy, Inc. OurDataHelps.  
16 OurDataHelps. <https://ourdatahelps.org/>. Accessed February 18, 2019.  
17  
18  
19 (24) Primack BA, Shensa A, Sidani JE, *et*  
20 *al.* Social Media Use and Perceived Social Isolation Among Young Adults in the  
21 U.S. *American Journal of*  
22 *Preventive Medicine*.  
23 2017;53(1):1-8. doi:10.1016/j.amepre.2017.01.010.  
24  
25  
26 (25) Sinnenberg L, DiSilvestro CL, Mancheno C, *et al.* Twitter as a Potential Data Source  
27 for Cardiovascular Disease Research. *JAMA Cardiology*. December 2016.  
28 <https://jamanetwork.com/journals/jamacardiology/fullarticle/2556216>.  
29  
30  
31 (26) Sap M, Park G, Eichstaedt J, *et al.* Developing Age and Gender Predictive Lexica  
32 over Social Media. *Proceedings of the 2014 Conference on Empirical Methods in*  
33 *Natural Language Processing (EMNLP)*. 2014. doi:10.3115/v1/d14-1121.  
34  
35  
36 (27) Jaidka K, Guntuku SC, Buffone A, Schwartz HA, Ungar L. Facebook versus  
37 Twitter: Differences in Self-Disclosure and Trait Prediction. *ICWSM*. 2018.  
38  
39  
40 (28) Schwartz HA, Eichstaedt JC, Kern ML, *et al.* Personality, Gender, and Age in the  
41 Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*. 2013;8(9).  
42 doi:10.1371/journal.pone.0073791.  
43  
44  
45 (29) Pennebaker Jw, Jordan K, Blackburn K. The development and psychometric  
46 properties of LIWC2015. *UT Faculty/Researcher Works*. 2014:1-22.  
47  
48  
49 (30) Schwartz HA, Eichstaedt J, Kern ML, *et al.* Towards Assessing Changes in Degree  
50 of Depression through Facebook. *Proceedings of the Workshop on Computational*  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 *Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014.  
4 doi:10.3115/v1/w14-3214.  
5  
6  
7 (31) Guntuku SC, Ramsay JR, Merchant RM, Ungar LH. Language of ADHD in Adults  
8 on Social Media. *Journal of Attention Disorders*. 2017;108705471773808.  
9 doi:10.1177/1087054717738083.  
10  
11  
12 (32) Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *The Journal of Machine*  
13 *Learning Research*. 2003;3.  
14  
15  
16 (33) Gelfand AE, Smith AF. Sampling Based Approaches to Calculating Marginal  
17 Densities. *Journal of the American Statistical Association* 1990;**85**:398–409.  
18 doi:10.21236/ada208388  
19  
20  
21 (34) McCallum A. Mallet: A machine learning for language toolkit. MALLET Machine  
22 Learning for Language Toolkit. 2002.<http://mallet.cs.umass.edu/> (accessed 9 May2019).  
23  
24  
25 (35) Meyer GJ, Finn SE, Eyde LD, et al. Psychological testing and psychological  
26 assessment. A review of evidence and issues. *The American psychologist*.  
27 2001;**56**(2):128-165.  
28  
29  
30  
31 (36) Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, *et al*. What Twitter Profile and Posted  
32 Images Reveal About Depression and Anxiety. *ICWSM* Published Online First:  
33 2019.<https://arxiv.org/pdf/1904.02670.pdf>  
34  
35  
36 (37) Jones R. Drug Slang Dictionary - Words Starting With G.  
37 <http://www.noslang.com/drugs/dictionary.php>. Accessed February 19, 2019.  
38  
39  
40 (38) Booth R. Toward an Understanding of  
41 Loneliness. *Social Work*. 1983;**28**(2):116-119. doi:10.1093/sw/28.2.116.  
42  
43  
44 (39) Qualter P, Vanalst J, Harris R, *et al*. Loneliness across the life span. *Perspectives on*  
45 *psychological science : a journal of the Association for Psychological Science*  
46 2015;**10**:250–  
47 64.[https://journals.sagepub.com/doi/full/10.1177/1745691615568999?url\\_ver=Z39.88-](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)  
48 [2003&rfr\\_id=ori:rid:crossref.org&rfr\\_dat=cr\\_pub=pubmed](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)  
49  
50  
51  
52 (40) Hawkey LC, Cacioppo JT. Loneliness Matters: A Theoretical and Empirical Review of  
53 Consequences and Mechanisms. *Annals of Behavioral Medicine* 2010;**40**:218–27.  
54 doi:10.1007/s12160-010-9210-8  
55  
56  
57  
58  
59  
60

- 1  
2  
3  
4  
5 (41) Cacioppo JT, Hawkley LC, Ernst JM, et  
6 al. Loneliness within a nomological net: An evolutionary perspective. *Journal of*  
7 *Research in Personality*. 2006;40(6):1054-1085.  
8 doi:10.1016/j.jrp.2005.11.007.  
9
- 10  
11 (42) Edwards T, Holtzman NS. A  
12 meta-analysis of correlations between depression and first person singular  
13 pronoun use. *Journal of Research*  
14  
15  
16  
17
- 18 (43) Al-Saggaf Y, Nielsen S. Self-disclosure on Facebook among female users and its  
19 relationship to feelings of loneliness. *Computers in Human Behavior* 2014;**36**:460–8.  
20 doi:10.1016/j.chb.2014.04.014  
21  
22
- 23 (44) Simon EB, Walker MP. Sleep loss causes  
24 social withdrawal and loneliness. *Nature*  
25 *Communications*. 2018;9(1).  
26 doi:10.1038/s41467-018-05377-0.  
27  
28
- 29 (45) Cacioppo JT, Hawkley LC, Thisted RA. Perceived social isolation makes me sad: 5-year  
30 cross-lagged analyses of loneliness and depressive symptomatology in the Chicago  
31 Health, Aging, and Social Relations Study. *Psychology and Aging* 2010;**25**:453–63.  
32 doi:10.1037/a0017216  
33  
34  
35
- 36 (46) Yeginsu C. U.K. Appoints a Minister for Loneliness. *The New York Times*.  
37 <https://www.nytimes.com/2018/01/17/world/europe/uk-britain-loneliness.html>. Published  
38 January 17, 2018.  
39  
40
- 41 (47) Rubin R. Loneliness Might Be a Killer,  
42 but What’s the Best Way to Protect Against It? *Jama*.  
43 2017;318(19):1853. doi:10.1001/jama.2017.14591.  
44  
45  
46
- 47 (48) CareMore Health Announces New Outcomes  
48 Data from First-of-its-Kind Togetherness Program. Business Wire A Berkshire  
49 Hathaway Company . <https://www.businesswire.com/news/home/20181218005059/en/>.  
50 Published December 18, 2018. Accessed February 18, 2019.  
51  
52
- 53 (49) Perese EF, Wolf M. Combating  
54 Loneliness Among Persons With Severe Mental Illness: Social Network  
55  
56  
57  
58  
59  
60



1  
2  
3 Interventions Characteristics, Effectiveness, And Applicability. *Issues in Mental Health*  
4 *Nursing*. 2005;26(6):591-609.  
5 doi:10.1080/01612840590959425.  
6  
7

8 (50) Webber M, Fendt-Newlin M. A review of social participation interventions for  
9 people with mental health problems. *Social Psychiatry and Psychiatric Epidemiology*  
10 2017;**52**:369–80. doi:10.1007/s00127-017-1372-2  
11  
12

13 (51) U.S. Twitter reach by age group 2018 |  
14 Statistic. Statista.  
15 [https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-](https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/)  
16 [age-group/](https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/).  
17  
18 Accessed February 18, 2019.  
19  
20

21 (52) Mckee R. Ethical issues in using  
22 social media for health and health care research. *Health Policy*. 2013;110(2-3):298-301.  
23 doi:10.1016/j.healthpol.2013.02.006.  
24  
25

26 (53) Inkster B, Stillwell D, Kosinski M,  
27 Jones P. A decade into Facebook: where is psychiatry in the digital age? *The Lancet*  
28 *Psychiatry*. 2016;3(11):1087-1090.  
29 doi:10.1016/s2215-0366(16)30041-4.  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 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)

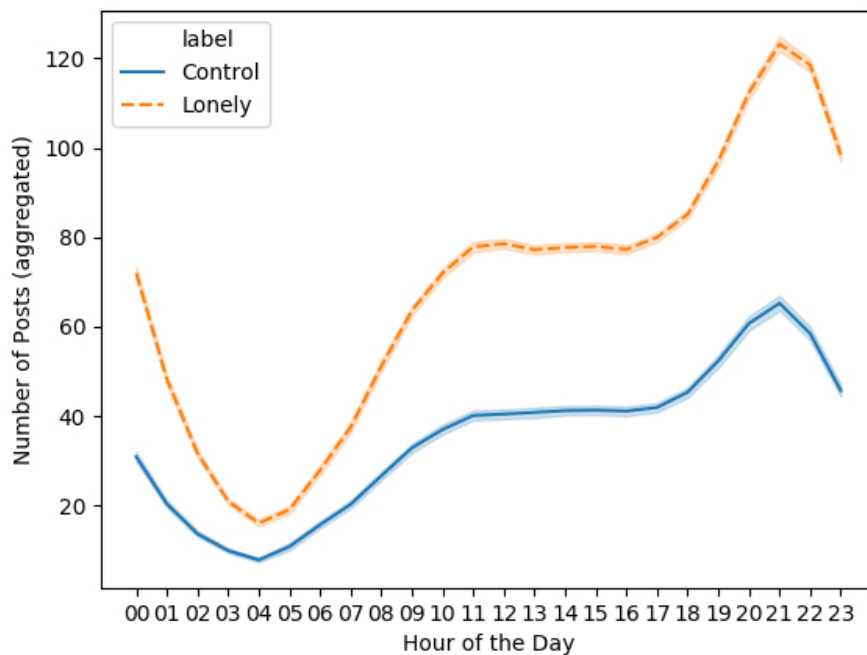
**Figure 2: Temporal variation showing diurnal patterns of post frequency of both the ‘lonely’ and ‘control’ groups.**

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.

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



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
<b>Title and abstract</b>	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)
<b>Introduction</b>		
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)
<b>Methods</b>		
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
<b>Results</b>		
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 ( <b>pgs 10,11</b> )
<b>Discussion</b>		
Key results	18	Summarise key results with reference to study objectives ( <b>pgs. 12, 13</b> )
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 ( <b>pgs. 14</b> )
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence ( <b>pg. 13, 14</b> )
Generalisability	21	Discuss the generalisability (external validity) of the study results ( <b>pgs. 13, 14</b> )
<b>Other information</b>		
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 ( <b>pg. 15</b> )

\*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:	<i>BMJ Open</i>
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, Rachele; 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
<b>Primary Subject Heading</b>:	Public health
Secondary Subject Heading:	Communication, Mental health
Keywords:	loneliness, social media, natural language processing, MENTAL HEALTH, STATISTICS & RESEARCH METHODS

SCHOLARONE™  
Manuscripts

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

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

<sup>1</sup>Penn Medicine Center for Digital Health, Philadelphia, PA 19104

<sup>2</sup>Penn Medicine Center for Healthcare Innovation, Philadelphia, PA 19104

<sup>3</sup>Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104

<sup>4</sup>Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

<sup>5</sup>Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

<sup>6</sup>The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

<sup>7</sup>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%.

1  
2  
3 **Conclusions:** Posts with the words lonely or alone often include psychosocial features and can  
4 potentially have associations with how individuals presumably express and experience  
5 loneliness. This can inform online surveillance for high risk individuals experiencing loneliness  
6 and interventions focused on addressing morbidity in this condition.  
7  
8  
9  
10  
11  
12  
13  
14

### 15 **Strengths and Limitations of this study**

- 16  
17 ● Novel focus on timelines of social media users to study mentions of loneliness and  
18 correlation with predictors of mental health.
- 19  
20 ● The study sample consists of social media users and is not representative of the general  
21 population.  
22  
23
- 24 ● Though we manually annotated a subset of posts mentioning loneliness, some may have  
25 been metaphorical or non sequiturs.  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 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.<sup>1</sup> 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.<sup>1-5</sup> Loneliness is also one of the primary underlying causes and correlates for chronic mental health conditions and physician visits in some populations.<sup>1, 5-9</sup>

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.<sup>10</sup> 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.<sup>11-12</sup> Social media use seeks to connect people but it also has been associated with increased perceived social isolation.<sup>13</sup> 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.<sup>14-15</sup> For example, individuals who are stressed and depressed use more first-person singular pronouns suggesting higher self-focus and communities with heart

1  
2  
3 disease discuss hate more frequently.<sup>11-12; 16</sup> Natural language processing and machine learning  
4  
5 have revealed their value in using social media posts to predict first documented diagnosis of  
6  
7 depression using posts 6 months prior yielding an AUC of 0.72.<sup>17</sup>  
8  
9

10  
11  
12 While the use of social media is increasingly common, less is known about how often individuals  
13  
14 use the platform to explicitly share about feelings of loneliness or being alone.<sup>13</sup> In this study, we  
15  
16 sought to characterize Twitter timelines of individuals' whose posts include the words lonely or  
17  
18 alone. Studying the language of users who use these terms, we analyzed the correlations between  
19  
20 posting about loneliness and users' mental health and psycholinguistic attributes (e.g. anger and  
21  
22 depression). This has the potential to further our understanding of how social media platforms  
23  
24 are used for mentions of loneliness and if there is an opportunity to use these platforms for  
25  
26 surveillance of an important but hard to track and measure condition that impacts public health.  
27  
28 However, privacy of individuals has to be at the forefront of this research to shield unintended  
29  
30 use of this data, specifically with the amount of health insights that can be gleaned from social  
31  
32 media.  
33  
34  
35  
36  
37  
38  
39

40 We hypothesize that language usage patterns would both confirm existing understanding of  
41  
42 loneliness and give new insights into the daily lives of those who express being lonely. As  
43  
44 loneliness can impact health outcomes, identifying ways to track prevalence and manifestations  
45  
46 of loneliness online would be useful for developing approaches for identifying and offering  
47  
48 support for these individuals. This presents the opportunity of digital platforms to not only  
49  
50 provide markers of health but also potentially serve as platforms that can be used for developing  
51  
52 interventions.<sup>18-19</sup>  
53  
54  
55  
56  
57  
58  
59  
60

## 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,

1  
2  
3 joke).<sup>20</sup> Two co-authors independently coded a random set of 100 tweets from individuals who  
4 used the words lonely/alone at least 5 times in their timeline to identify them as presumed to be  
5 associated with the feeling of loneliness or other. The Kappa was 0.70 and we identified that  
6  
7  
8  
9  
10 76% of users' tweets indicate presumably feeling lonely. A few examples are as follows: "i'm  
11  
12 feelin real depressed, confused, & lonely", "im always the only up around this time, feeling a lil  
13  
14 lonely" and "I'm so Lonely in life :-( I just wish I can have love again it feels so go to be in love  
15  
16 with someone whom loves you."

### 21 *Control group*

22  
23  
24 We then identified a control group of users by matching each user in the above dataset to another  
25  
26 user by age, gender and period of activity (dates of first and last posting on twitter). We obtained  
27  
28 the age and gender estimates by using lexica developed previously.<sup>21</sup> Then, we selected users  
29  
30 with a minimum of 500 words across all their posts to have sufficient language for linguistic  
31  
32 analyses.<sup>22</sup> We excluded non-English, non-US tweets, retweets, and tweets containing 'alone'  
33  
34 and/or 'lonely' that were used to identify users who had more than 5 posts with the words  
35  
36 'lonely' or 'alone' in all analyses to identify linguistic features that are actually characteristics of  
37  
38 lonelier people -- looking at their entire timeline of tweets. Hereafter, we indicate users who had  
39  
40 more than 5 posts with the words 'lonely' or 'alone' as 'users with posts including the words  
41  
42 lonely or alone', and 'control' group to represent the matched set of users who had no such posts.  
43  
44  
45  
46  
47  
48

### 49 *Deriving language features to characterize individuals expressing loneliness*

50  
51 We used four sets of language features: a) open-vocabulary topics,<sup>23</sup> b) dictionary-based  
52  
53 psycholinguistic features,<sup>24</sup> c) mental well-being attributes such as anxiety, depression by  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 applying previously developed statistical models,<sup>25</sup> d) number of drug words and time of posts as  
4  
5 past research has shown an association between loneliness and substance use.<sup>26; 12</sup> These  
6  
7 language features have been shown to be predictive of several health outcomes, such as  
8  
9 depression, schizophrenia, attention deficit hyperactivity disorder (ADHD), and general well-  
10  
11 being.<sup>27; 28</sup>  
12  
13  
14  
15  
16

17 *Open-vocabulary:* As closed-vocabulary approaches like LIWC include only a small subset of  
18  
19 the entire language used on social media, we use an open-vocabulary approach to improve the  
20  
21 coverage and find topics that people who mention loneliness. Topics consist of clusters of co-  
22  
23 occurring words created using Latent Dirichlet Allocation (LDA).<sup>29</sup> The LDA generative model  
24  
25 assumes that tweets contain a combination of topics, and that topics are a distribution of words.  
26  
27 Since the words in a tweet are known, topics, which are latent variables, can be estimated  
28  
29 through Gibbs sampling.<sup>30</sup> We use the Mallet implementation of the LDA algorithm, adjusting  
30  
31 one parameter ( $\alpha=5$ ) to favor fewer topics per tweet.<sup>31</sup> All other parameters were kept at their  
32  
33 default. An example of such a model is the following sets of words ('tuesday', 'monday',  
34  
35 'wednesday', ...) which clusters together days of the week by exploiting their similar  
36  
37 distributional properties across tweets. In our study, two hundred topics were generated using  
38  
39 tweets across all users in the dataset of users with posts including the words lonely or alone and  
40  
41 control users.  
42  
43  
44  
45  
46  
47  
48

49 *Dictionary-based:* From each post, we extracted the relative frequency of single words and  
50  
51 phrases (consisting of two or three consecutive words). Then, all words used by less than 1% of  
52  
53 users were removed from analysis so as to remove uncommonly used words (outliers).  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Additionally, all messages used to identify our study group were removed prior to further  
4 analysis. The Linguistic Inquiry Word Count (LIWC) dictionary is a language-specific, many-to-  
5 many mapping of tokens (including words and word stems) and psychologically validate  
6 categories. Each category (a curated list of words) is found to be correlated with and also  
7 predictive of several psychological traits and outcomes. For each user, we measure the  
8 proportion of word tokens that fall into a given LIWC category.  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18

19 *Mental well-being attributes:* We used automatic text-regression methods to assign to each user  
20 scores on the depression, anxiety and anger facets for users.<sup>25</sup> This model was trained on a  
21 sample of over 28,749 users who had taken the International Personality Item Pool Neuroticism-  
22 Extraversion-Openness Personality Inventory Revised (IPIP NEO-PI-R) survey that contains the  
23 depression, anxiety and anger Facets of the Neuroticism Factor.<sup>25</sup> The machine learning model  
24 trained on words and phrases from Facebook posts to predict survey measure of depression,  
25 anger and anxiety resulted in a performance of  $r = .32$ , which is consistent with other reports of  
26 mental health states identified via social media.<sup>32</sup> The model was trained using status updates of  
27 users from another study<sup>25</sup>, and has been shown to generalize to Twitter users.<sup>33</sup>  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

42 *Use of Drug-words:* We also extracted the frequency (aggregated to every user) of most common  
43 drug words as used on social media.<sup>34</sup>  
44  
45  
46  
47  
48

49 *Temporal patterns:* We determined the frequency of posts across different hours of the day by  
50 users in both users with posts including the words lonely or alone and control groups to  
51 understand the diurnal patterns in posting.  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3  
4  
5  
6 *Identifying differentially expressed language features in users with posts including the words*  
7  
8 *lonely or alone*

9  
10 We isolated the patterns in users' loneliness mentions using the linguistic attributes and user  
11 traits by correlating them with users with posts including the words lonely or alone and control  
12 groups. We used logistic regression to distinguish open-vocabulary words, phrases, LIWC  
13 categories and topics associated with lonely and control groups and measure the effect size using  
14 Cohen's D. The models were set up to predict the group of users with posts including the words  
15 lonely or alone against the control group (e.g., group was the dependent variable). Details of the  
16 method are described in a previous work<sup>23</sup>. For identifying themes from topics, researchers  
17 looked at 20 messages each with the highest topic prevalence to identify themes. We used  
18 Benjamini-Hochberg p-correction and use  $p < 0.001$  for indicating meaningful correlations and  
19 the effect size was measured using Cohen's D. The statistical analysis, data synthesis, and model  
20 creation was conducted in 2018-2019.  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36

37  
38 *Predicting the likelihood of posting about loneliness online*

39  
40 We then looked at the feasibility of predicting whether a user is likely to mention loneliness or  
41 not based on their social media language. Automated analysis of social media is accomplished by  
42 building predictive models, which use 'features', or variables that have been extracted from  
43 social media data. For this analysis we used LIWC and topics as features. Features are then  
44 treated as independent variables in an algorithm (Random Forests) to predict the dependent  
45 variable of an outcome of interest (e.g., users' saying that they are lonely or not). For cross  
46 validation, the predictive model was trained, using Random Forests, on the training set and then  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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

1  
2  
3 \*users with posts including the words lonely or alone is defined as any user posting at least 5  
4 times about loneliness and the control group is defined as any user who does not have any posts  
5 about loneliness  
6  
7

8  
9 *Identifying differentially expressed language features in users with posts including the words*  
10 *lonely or alone*  
11

12  
13  
14 *Open vocabulary approach:* Analyzing differences in individual words and phrases used across  
15 both groups, we observed (Figure 1a) that users with posts including the words lonely or alone  
16 referred to themselves ('myself' (d=.18), 'I' (d=.16)) in their Twitter posts significantly more  
17 than the control group. They also posted about relationship issues ('want\_somebody' (d=.08),  
18 'no\_one\_to' (d=.1), needs and feelings ('i\_just\_wanna (d=.12), 'in\_my\_feelings' (d=.1), 'i\_need'  
19 (d=.12), 'i\_cant' (d=.1)), and included more expletives. Users in the control group (Figure 1b)  
20 engaged in a lot more conversations as indicated by '<user>' (d=-.2) (we anonymize '@'  
21 mentions in users tweets as '<user>') compared to users with posts including the words lonely or  
22 alone. The control group also posted more about games ('season' (d=-.09), 'coach' (d=-.07),  
23 'team' (d=-.1)) and positivity ('!' (d=-.13), 'awesome' (d=-.09), ':' (d=-.08)). Figure 1  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

44 Using topics generated from LDA, we identified the themes which occur more frequently in  
45 posts of users with posts including the words lonely or alone. Posts were about interpersonal  
46 relationships (d=.28) (and associated issues (d=.22)), self-reflection (d=.21) (accompanied with  
47 wondering about the future (d=.12)), drug/alcohol use (d=.29) (considering them to be the 'only  
48 friend'), insomnia (d=.27), uncontrolled emotions (d=.28) (accompanied by confusion (d=.11)),  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

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	0.28
	hurt, feelings, trust, forget	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*
<b>Pronouns</b>	
1st Person Pronouns	0.18
<b>Cognitive Processes</b>	
Certainty	0.15
Discrepancies	0.15
Differentiation	0.14
Tentativeness	0.13
<b>Negative Emotions</b>	
Swearing	0.11
<b>Mental Well-being</b>	
Depression	0.81
Anger	0.95
Anxiety	0.75
<b>Drug words</b>	
Blunt	0.16
Smoke	0.13
Heroin	0.1

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

1  
2  
3 *Mental well-being:* Users with posts including the words lonely or alone were more likely to  
4  
5 have posts associated with anger ( $d=.95$ ), depression ( $d=.81$ ) and anxiety ( $d=.75$ ) when  
6  
7 compared to the control group.  
8  
9

10  
11  
12 *Use of Drug Words:* We also identified the distribution of words pertaining to drugs in the posts  
13  
14 of users with posts including the words lonely or alone, and these were more likely to reference a  
15  
16 blunt ( $d=.16$ ), smoke ( $d=.13$ ), and heroin ( $d=.1$ ), and included prescribed medications for  
17  
18 treatment, recreational drug use, and recreational drugs.  
19  
20  
21  
22  
23

24 *Temporal patterns:* Users with posts including the words lonely or alone were found to post  
25  
26 more during the night ( $d=.1$ ), shown in Figure 2. We also see themes associated with night-time  
27  
28 posting and having difficulty sleeping ( $d=.27$ ) in the open-vocabulary analysis (Table 2).  
29  
30  
31  
32

33 *Predictive Analysis:* Table 4 shows that random forest model using Topics as input features  
34  
35 predicted mentions of loneliness in users with an AUC of .854 (F1 score = 0.778) and LIWC  
36  
37 features resulted in AUC of 0.859 (F1 score = 0.777). A combination of LIWC and Topics  
38  
39 resulted in the best AUC of 0.863 (F1 score = 0.782).  
40  
41  
42  
43

44 **Table 4:** Performance of different features at predicting mentions of loneliness, reported on an  
45 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.<sup>35</sup> This expression of negativity related to relationships is likely related to a hypervigilance to social threat, associated with loneliness.<sup>36</sup> Lonely individuals were also reported to focus on overcoming past events as well as showing feelings of helplessness.<sup>35</sup>

1  
2  
3 Researchers who coded the topics were attempting to identify these associations by looking at 20  
4 messages each with the highest topic prevalence to identify themes, and we acknowledge that  
5  
6 this can be subjective.  
7  
8  
9

10  
11 Association of users with posts including the words lonely or alone with linguistic estimates of  
12 anger, depression, and anxiety corroborate prior research, showing that loneliness and social  
13 isolation influence psychological functioning , specifically the ability to self-regulate emotion.<sup>5-6;</sup>  
14  
15  
16  
17  
18  
19 <sup>37</sup> Specifically, anxiety, anger, and negative mood were reported as higher in lonely young  
20 adults.<sup>38</sup> Tweets by users with posts including the words lonely or alone were more self-focused  
21 compared to the control group. Prior researchers have found that “first person singular pronouns  
22 are a modest linguistic marker of depression.”<sup>39</sup> Also, previous research has shown that  
23 loneliness has been associated with greater self-disclosure in Facebook posts.<sup>40</sup> This presents the  
24 potential for early identification and assessment to intervene on loneliness as well as mental  
25 health conditions for this group.  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37

38 Trends in temporal variation in posting may reflect that sleep deprivation can contribute to social  
39 withdrawal and loneliness.<sup>41</sup> This finding corroborates prior research associating loneliness with  
40 diminished sleep quality.<sup>37</sup> A better understanding of the temporality of posting could inform  
41 timing of interventions designed to address loneliness, as well as provide insight for other  
42 researchers to test the inter-relationships between loneliness and the motivations for using social  
43 media during nighttime.  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 Loneliness is known to be one of the primary underlying causes and correlates for chronic  
4 mental health conditions.<sup>5-6; 42</sup> As loneliness is becoming increasingly recognized as a public  
5 health, several groups have taken action to address it. For example, the United Kingdom  
6 appointed a Minister for Loneliness who is responsible for addressing loneliness within  
7 communities.<sup>43</sup> CareMore, a health plan and delivery system providing care for enrollees in  
8 Medicare Advantage and Medicaid health plans in seven states across the U.S., launched the  
9 “Togetherness Program” in a clinical setting to address loneliness in elderly patients.<sup>44</sup> Through  
10 this work, CareMore reported that participation in exercise programs increased by 56.6%,  
11 emergency room utilization decreased by 3.3%, and hospital admissions among participants were  
12 20.8% lower per thousand compared to the “intent to treat population.”<sup>45</sup> Additionally, social  
13 network interventions targeting loneliness have been found to be effective in reducing social  
14 isolation among individuals with severe mental health conditions but these interventions are not  
15 included in the treatment plans for individuals with a mental illness.<sup>46-47</sup>

16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35 Considering the advantage of large sample sizes and also the association between increased  
36 social media usage and individuals mentions of loneliness, it is promising to use natural language  
37 processing and machine learning to automatically identify a person mentions the words alone or  
38 lonely on Twitter to inform interventions targeted at early identification and support for affected  
39 and at risk individuals with the caveat that social media users are not representative of a random  
40 sample of individuals. To address loneliness will require being able to identify it passively,  
41 remotely, and over time. Many people rarely visit a healthcare provider so would miss the  
42 opportunity for screening. Approaches for treatment will also need to harness the tools and  
43 technologies that are accessible and integrated with the things people use every day (e.g. mobile  
44 phones). Future interventions would have to potentially rely on digital phenotyping of loneliness

1  
2  
3 and using digital platforms (e.g. text messaging) to complement human-to-human interaction  
4  
5 strategies to treat loneliness.  
6  
7  
8  
9

10 In this first study, our aim was to characterize loneliness mentions based on users' entire  
11 timelines. Future studies could perform a time-series analysis of the temporal variations  
12 associated with loneliness mentions. Further, works should also validate whether the  
13 characteristics of people who are using the words 'lonely' or 'alone' on Twitter can be used to  
14 track community health risks, particularly, the risk of social isolation. Other studies should  
15 replicate the findings in this study using more formal ground truth such as surveys and extend  
16 this work to investigate if Twitter can potentially map regional hotspots of loneliness to identify  
17 problematic loneliness for community public health monitoring.  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29

### 30 **Limitations and Ethics**

31  
32 The study sample consists of social media users and is not representative of the general  
33 population. An estimated 40% of US adults using Twitter are between the ages of 18 and 29, so  
34 our analysis is skewed towards younger people.<sup>48</sup> An automated machine learning tool could be a  
35 low-cost method to potentially detect posts about loneliness or being alone that may occur with  
36 other concerning signals from digital sensors (e.g. changes in sleep, activity, purchases). These  
37 signals could trigger could then be referred to more formal screening methods or support  
38 resources.<sup>49</sup>  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50

51 Considering we identified that 76% of users' tweets indicated presumably feeling lonely in the  
52 sample we hand coded, posts mentioning the words alone or lonely may have been metaphorical  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 or non sequiturs. Also, considering the inclusion criteria based on number of tweets mentioning  
4 alone or lonely, we are potentially selecting users with more posts than the average twitter user.  
5  
6 Additionally, Twitter is far from perfect to be used as a diagnostic tool. However, an automated  
7  
8 machine learning tool could be a low-cost method to potentially detect elevated loneliness levels  
9  
10 in a person who could then be referred to more formal screening methods. Further, the effects  
11  
12 presented in this dataset may not be specific to loneliness considering the potential comorbidity  
13  
14 with mental health conditions such as depression in this dataset.  
15  
16  
17  
18  
19  
20

21 The feasibility of social media-based assessments of loneliness mentions (and mental health  
22 more broadly) needs further assessment. Privacy of individuals is an ongoing concern, especially  
23  
24 with social media users not fully realizing the amount of health insights that can be gleaned by  
25  
26 their online posts. Employers and insurance companies, for example, may be motivated to derive  
27  
28 these assessments, but could use these insights against those suffering from mental illness. As  
29  
30 mental illnesses carry social stigma and may engender discrimination, data protection and  
31  
32 ownership frameworks are needed to make sure the data is not used against the users' interest.<sup>50</sup>  
33  
34 Further, transparency about which indicators are derived by whom for what purpose should be  
35  
36 part of ethical and policy discourse.  
37  
38  
39  
40  
41  
42  
43

44 There are also open questions around the impact of misclassifications, and how derived mental  
45  
46 health indicators can be responsibly integrated into systems of care.<sup>51</sup>  
47  
48  
49  
50

## 51 **Conclusions**

52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 In this study we characterized mentions of loneliness on Twitter at the individual level.  
4  
5 Furthermore, we identified specific contexts, themes, and traits in the posts of individuals  
6  
7 mentioning loneliness on Twitter. As loneliness is a public health challenge, a better  
8  
9 understanding of how loneliness is described online can inform tracking of loneliness and  
10  
11 interventions targeted at addressing this important public health problem in regards to the  
12  
13 behavior of lonely individuals that may be at risk of developing a severe mental health  
14  
15 condition.<sup>44</sup>  
16  
17  
18  
19  
20  
21

## 22 **Acknowledgements**

23  
24 **Funding:** This project is funded, in part, under a grant with the Pennsylvania Department of  
25 Health (ORG-FUND-PROG-CREF is 4290-567862-2446-2049). No sponsor of funding source  
26 played a role in: “study design and the collection, analysis, and interpretation of data and the  
27 writing of the article and the decision to submit it for publication.” All researchers are  
28 independent from funders  
29  
30

31  
32 **Data Sharing Statement:** Because of our IRB requirements, data will be shared upon request  
33 from the corresponding author.  
34  
35

36 **Contributors:** S.C. Guntuku and R. Merchant originated the study. S.C. Guntuku, R. Schneider,  
37 A. Pelullo, L.H. Ungar, and R. Merchant developed methods, interpreted analysis, and  
38 contributed to the writing of the article. J.F. Young, V. Wong, D. Polsky, and K. Volpp assisted  
39 with the interpretation of the findings and contributed to the writing of the article.  
40  
41

42 **Disclosures:** None  
43  
44  
45  
46  
47  
48  
49

## 50 **References:**

51 (1) Hyland P, Shevlin M, Cloitre M, *et al.* Quality not quantity: loneliness subtypes,  
52 psychological trauma, and mental health in the US adult population. *Social Psychiatry*  
53 *and Psychiatric Epidemiology* Published Online First: 2018. doi:10.1007/s00127-018-  
54 1597-8  
55  
56  
57  
58  
59

- 1  
2  
3  
4  
5 (2) Gerst-Emerson K, Jayawardhana J.  
6 Loneliness as a Public Health Issue: The Impact of Loneliness on Health Care  
7 Utilization Among Older Adults. *American*  
8 *Journal of Public Health*.  
9 2015;105(5):1013-1019. doi:10.2105/ajph.2014.302427.  
10  
11  
12  
13  
14 (3) Jong-Gierveld JD. Developing and  
15 testing a model of loneliness. *Journal*  
16 *of Personality and Social Psychology*.  
17 1987;53(1):119-128. doi:10.1037//0022-3514.53.1.119.  
18  
19  
20 (4) Peplau LA, Perlman D. *Loneliness: a Sourcebook of Current Theory,*  
21 *Research, and Therapy*. New York:  
22 Wiley; 1982.  
23  
24  
25 (5) Stravynski A, Boyer R. Loneliness in Relation to Suicide Ideation and Parasuicide: A  
26 Population-Wide Study. *Suicide and Life-Threatening Behavior*. 2001;31(1):32-40.  
27 doi:10.1521/suli.31.1.32.21312.  
28  
29  
30  
31 (6) Blai B. Health Consequences of  
32 Loneliness: A Review of the Literature. *Journal of American College Health*.  
33 1989;37(4):162-167. doi:10.1080/07448481.1989.9938410.  
34  
35  
36 (7) Heinrich LM, Gullone E. The clinical  
37 significance of loneliness: A literature review. *Clinical Psychology Review*.  
38 2006;26(6):695-718.  
39 doi:10.1016/j.cpr.2006.04.002.  
40  
41  
42  
43 (8) Richard A, Rohrmann S, Vandeleur CL,  
44 Schmid M, Barth J, Eichholzer M. Loneliness is adversely associated with  
45 physical and mental health and lifestyle factors: Results from a Swiss national  
46 survey. *Plos One*. 2017;12(7). doi:10.1371/journal.pone.0181442.  
47  
48  
49 (9) Rico-Urbe  
50 LA, Caballero FF, Olaya B, et al. Loneliness, Social Networks, and Health: A  
51 Cross-Sectional Study in Three Countries. *Plos One*. 2016;11(1).  
52 doi:10.1371/journal.pone.0145264.  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 (10) Draugalis JR, Coons SJ, Plaza CM. Best Practices for Survey Research Reports: A  
4 Synopsis for Authors and Reviewers. *American Journal of Pharmaceutical Education*  
5 2008;**72**:11. doi:10.5688/aj720111  
6  
7  
8 (11) Guntuku SC, Buffone A, Jaidka K,  
9 Eichstaedt J, Ungar L. Understanding and Measuring Psychological Stress using  
10 Social Media. *ICWSM 2019*. 2018.  
11  
12  
13 (12) Guntuku SC, Yaden DB, Kern ML, Ungar  
14 LH, Eichstaedt JC. Detecting depression and mental illness on social media: an  
15 integrative review. *Current*  
16 *Opinion in Behavioral Sciences*.  
17 2017;**18**:43-49. doi:10.1016/j.cobeha.2017.07.005.  
18  
19  
20  
21 (13) Primack BA, Shensa A, Sidani JE, et  
22 al. Social Media Use and Perceived Social Isolation Among Young Adults in the  
23 U.S. *American Journal of*  
24 *Preventive Medicine*.  
25 2017;**53**(1):1-8. doi:10.1016/j.amepre.2017.01.010.  
26  
27  
28  
29 (14) Kivran-Swaine F, Ting J, Naaman M. Understanding  
30 Loneliness in Social Awareness Streams: Expressions and Responses. *ICWSM 2014*.  
31 2014.  
32 [https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-  
33 and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef](https://www.semanticscholar.org/paper/Understanding-Loneliness-in-Social-Awareness-and-Kivran-Swaine-Ting/6b2921eb65968fb68974b7701e0f3101fdf92eef).  
34  
35  
36  
37 (15) Kamvar S, Kamvar S, Harris JJ. *We Feel Fine: an Almanac of Human Emotion*.  
38 New York: Scribner; 2009.  
39  
40  
41 (16) Eichstaedt JC, Schwartz HA, Kern ML, *et al*. Psychological Language on Twitter  
42 Predicts County-Level Heart Disease Mortality. *Psychological Science* 2015;**26**:159–69.  
43 doi:10.1177/0956797614557867  
44  
45  
46 (17) Eichstaedt JC, Smith RJ, Merchant RM, *et al*. Facebook language predicts  
47 depression in medical records. *Proceedings of the National Academy of Sciences*  
48 2018;**115**:11203–8. doi:10.1073/pnas.1802331115  
49  
50  
51 (18) Coppersmith G, Ngo K, Leary R, Wood A.  
52 Exploratory Analysis of Social Media Prior to a Suicide Attempt. *Proceedings of the*  
53 *Third Workshop on Computational*  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 *Linguistics and Clinical Psychology*.

4 2016. doi:10.18653/v1/w16-0311.  
5  
6

7 (19) Qntfy, Inc. OurDataHelps.

8 OurDataHelps. <https://ourdatahelps.org/>. Accessed February 18, 2019.  
9  
10

11 (20) Sinnenberg L, DiSilvestro CL, Mancheno C, et al. Twitter as a Potential Data Source  
12 for Cardiovascular Disease Research. *JAMA Cardiology*. December 2016.

13 <https://jamanetwork.com/journals/jamacardiology/fullarticle/2556216>.  
14  
15

16 (21) Sap M, Park G, Eichstaedt J, et al. Developing Age and Gender Predictive Lexica  
17 over Social Media. *Proceedings of the 2014 Conference on Empirical Methods in*  
18 *Natural Language Processing (EMNLP)*. 2014. doi:10.3115/v1/d14-1121.  
19  
20

21 (22) Jaidka K, Guntuku SC, Buffone A, Schwartz HA, Ungar L. Facebook versus  
22 Twitter: Differences in Self-Disclosure and Trait Prediction. *ICWSM*. 2018.  
23  
24

25 (23) Schwartz HA, Eichstaedt JC, Kern ML, et al. Personality, Gender, and Age in the  
26 Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*. 2013;8(9).  
27 doi:10.1371/journal.pone.0073791.  
28  
29

30 (24) Pennebaker Jw, Jordan K, Blackburn K. The development and psychometric  
31 properties of LIWC2015. *UT Faculty/Researcher Works*. 2014:1-22.  
32  
33

34 (25) Schwartz HA, Eichstaedt J, Kern ML, et al. Towards Assessing Changes in Degree  
35 of Depression through Facebook. *Proceedings of the Workshop on Computational*  
36 *Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 2014.  
37 doi:10.3115/v1/w14-3214.  
38  
39  
40

41 (26) Lenhart A. Mobile Access Shifts Social Media Use and Other Online Activities. Pew  
42 Research Center: Internet, Science & Tech.  
43 2015. [http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-](http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-other-online-activities/)  
44 [other-online-activities/](http://www.pewinternet.org/2015/04/09/mobile-access-shifts-social-media-use-and-other-online-activities/) (accessed 9 May2019).  
45  
46  
47

48 (27) Guntuku SC, Ramsay JR, Merchant RM, Ungar LH. Language of ADHD in Adults  
49 on Social Media. *Journal of Attention Disorders*. 2017:108705471773808.  
50 doi:10.1177/1087054717738083.  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

- 1  
2  
3 (28) Pinquart M, Sorensen S. Influences on  
4 Loneliness in Older Adults: A Meta-Analysis. *Basic and Applied Social Psychology*.  
5 2001;23(4):245-266. doi:10.1207/S15324834BASP2304\_2.  
6  
7  
8 (29) Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *The Journal of Machine*  
9 *Learning Research*. 2003;3.  
10  
11  
12 (30) Gelfand AE, Smith AF. Sampling Based Approaches to Calculating Marginal  
13 Densities. *Journal of the American Statistical Association* 1990;85:398–409.  
14 doi:10.21236/ada208388  
15  
16  
17 (31) McCallum A. Mallet: A machine learning for language toolkit. MALLET Machine  
18 Learning for Language Toolkit. 2002.<http://mallet.cs.umass.edu/> (accessed 9 May2019).  
19  
20  
21 (32) Meyer GJ, Finn SE, Eyde LD, et al. Psychological testing and psychological  
22 assessment. A review of evidence and issues. *The American psychologist*.  
23 2001;56(2):128-165.  
24  
25  
26  
27 (33) Guntuku SC, Preotiuc-Pietro D, Eichstaedt JC, et al. What Twitter Profile and Posted  
28 Images Reveal About Depression and Anxiety. *ICWSM* Published Online First:  
29 2019.<https://arxiv.org/pdf/1904.02670.pdf>  
30  
31  
32 (34) Jones R. Drug Slang Dictionary - Words Starting With G.  
33 <http://www.noslang.com/drugs/dictionary.php>. Accessed February 19, 2019.  
34  
35  
36 (35) Booth R. Toward an Understanding of  
37 Loneliness. *Social Work*. 1983;28(2):116-119. doi:10.1093/sw/28.2.116.  
38  
39  
40 (36) Qualter P, Vanalst J, Harris R, et al. Loneliness across the life span. *Perspectives on*  
41 *psychological science : a journal of the Association for Psychological Science*  
42 2015;10:250–  
43 64.[https://journals.sagepub.com/doi/full/10.1177/1745691615568999?url\\_ver=Z39.88-](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)  
44 [2003&rfr\\_id=ori:rid:crossref.org&rfr\\_dat=cr\\_pub=pubmed](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)  
45  
46  
47  
48 (37) Hawkey LC, Cacioppo JT. Loneliness Matters: A Theoretical and Empirical Review of  
49 Consequences and Mechanisms. *Annals of Behavioral Medicine* 2010;40:218–27.  
50 doi:10.1007/s12160-010-9210-8  
51  
52  
53 (38) Cacioppo JT, Hawkey LC, Ernst JM, et  
54 al. Loneliness within a nomological net: An evolutionary perspective. *Journal of*  
55  
56  
57  
58  
59



1  
2  
3 *Research in Personality*. 2006;40(6):1054-1085.  
4 doi:10.1016/j.jrp.2005.11.007.  
5

6  
7 (39) Edwards T, Holtzman NS. A  
8 meta-analysis of correlations between depression and first person singular  
9 pronoun use. *Journal of Research*  
10

11  
12  
13  
14 (40) Al-Saggaf Y, Nielsen S. Self-disclosure on Facebook among female users and its  
15 relationship to feelings of loneliness. *Computers in Human Behavior* 2014;**36**:460–8.  
16 doi:10.1016/j.chb.2014.04.014  
17

18  
19 (41) Simon EB, Walker MP. Sleep loss causes  
20 social withdrawal and loneliness. *Nature*  
21 *Communications*. 2018;9(1).  
22 doi:10.1038/s41467-018-05377-0.  
23

24  
25  
26 (42) Cacioppo JT, Hawkley LC, Thisted RA. Perceived social isolation makes me sad: 5-year  
27 cross-lagged analyses of loneliness and depressive symptomatology in the Chicago  
28 Health, Aging, and Social Relations Study. *Psychology and Aging* 2010;**25**:453–63.  
29 doi:10.1037/a0017216  
30

31  
32  
33 (43) Yeginsu C. U.K. Appoints a Minister for Loneliness. *The New York Times*.  
34 <https://www.nytimes.com/2018/01/17/world/europe/uk-britain-loneliness.html>. Published  
35 January 17, 2018.  
36  
37

38  
39 (44) Rubin R. Loneliness Might Be a Killer,  
40 but What's the Best Way to Protect Against It? *Jama*.  
41 2017;318(19):1853. doi:10.1001/jama.2017.14591.  
42

43  
44 (45) Jain S. CareMore Health Tackles the Unmet Challenges of the Aging Population.  
45 *American Society on Aging* 2018;**42**:14–  
46 8.[https://www.ingentaconnect.com/contentone/asag/gen/2018/00000042/00000001/art00](https://www.ingentaconnect.com/contentone/asag/gen/2018/00000042/00000001/art00003)  
47 003  
48  
49

50  
51 (46) Perese EF, Wolf M. Combating  
52 Loneliness Among Persons With Severe Mental Illness: Social Network  
53 Interventions Characteristics, Effectiveness, And Applicability. *Issues in Mental Health*  
54  
55  
56  
57  
58  
59

1  
2  
3 *Nursing*. 2005;26(6):591-609.  
4 doi:10.1080/01612840590959425.  
5  
6

7 (47) Webber M, Fendt-Newlin M. A review of social participation interventions for  
8 people with mental health problems. *Social Psychiatry and Psychiatric Epidemiology*  
9 2017;**52**:369–80. doi:10.1007/s00127-017-1372-2  
10  
11

12 (48) U.S. Twitter reach by age group 2018 |  
13 Statistic. Statista.  
14 [https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-](https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/)  
15 [age-group/](https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/).  
16  
17 Accessed February 18, 2019.  
18  
19

20 (49) Barnett I, Torous J. Ethics, Transparency, and Public Health at the Intersection of  
21 Innovation and Facebooks Suicide Prevention Efforts. *Annals of Internal Medicine*  
22 2019;**170**:565. doi:10.7326/m19-0366  
23  
24

25 (50) Mckee R. Ethical issues in using  
26 social media for health and health care research. *Health Policy*. 2013;110(2-3):298-301.  
27 doi:10.1016/j.healthpol.2013.02.006.  
28  
29

30 (51) Inkster B, Stillwell D, Kosinski M,  
31 Jones P. A decade into Facebook: where is psychiatry in the digital age? *The Lancet*  
32 *Psychiatry*. 2016;3(11):1087-1090.  
33 doi:10.1016/s2215-0366(16)30041-4.  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44

## 45 **Figure legends**

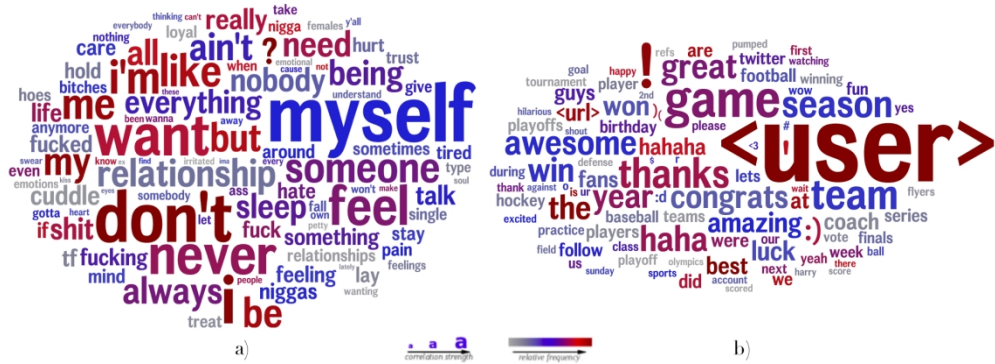
46  
47 **Figure 1: Words/Phrases more likely to be posted by Twitter users with a) self-reported**  
48 **loneliness (Individuals with at least 5 posts with the words ‘lonely’ or ‘alone’ group**  
49 **compared to the b) control group.**  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Word size indicates the strength of the correlation and word color indicates relative word  
4  
5 frequency. ( $p < 0.01$ , Bonferroni p-corrected)  
6  
7  
8  
9

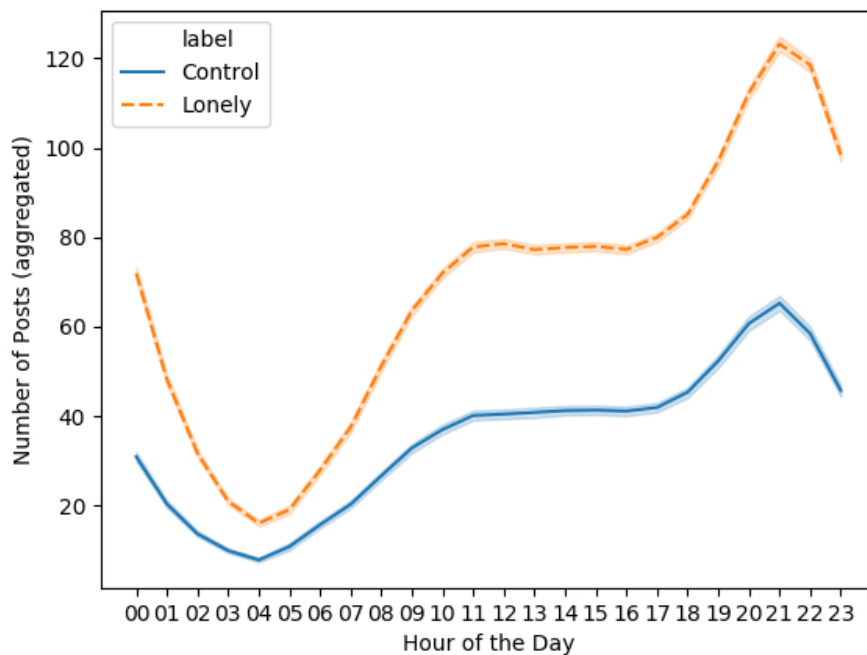
10 **Figure 2: Temporal variation showing diurnal patterns of post frequency of both the** users  
11  
12 with posts including the words lonely or alone **and control group.**

13  
14 The dotted line indicates the percentage of posts at different hours of the day by the group of  
15 users with at least 5 posts containing the word ‘lonely’ or ‘alone’ and the solid line indicates  
16 users who do not have any posts about loneliness. The x-axis represents the hour of the day each  
17 post occurs and the y-axis indicates the number of posts for each group.  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



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
<b>Title and abstract</b>	1	(a) Indicate the study's design with a commonly used term in the title or the abstract ( <b>pg.2</b> ) (b) Provide in the abstract an informative and balanced summary of what was done and what was found ( <b>pg.2</b> )
<b>Introduction</b>		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported ( <b>pg.4</b> )
Objectives	3	State specific objectives, including any prespecified hypotheses ( <b>pg. 4</b> )
<b>Methods</b>		
Study design	4	Present key elements of study design early in the paper ( <b>pg.5</b> )
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection ( <b>pg. 5</b> )
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up ( <b>pg. 6</b> ) (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 ( <b>pg. 6</b> )
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 ( <b>pg. 6</b> )
Bias	9	Describe any efforts to address potential sources of bias ( <b>pg. 6</b> )
Study size	10	Explain how the study size was arrived at ( <b>pg. 6</b> )
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why ( <b>pg. 7</b> )
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding ( <b>pg. 9</b> ) (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
<b>Results</b>		
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 ( <b>pg. 9</b> ) (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 ( <b>pg. 9</b> ) (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 ( <b>pgs 10,11</b> )
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 ( <b>pgs 10,11</b> )

		(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 ( <b>pgs 10,11</b> )
<b>Discussion</b>		
Key results	18	Summarise key results with reference to study objectives ( <b>pgs. 12, 13</b> )
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 ( <b>pgs. 14</b> )
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence ( <b>pg. 13, 14</b> )
Generalisability	21	Discuss the generalisability (external validity) of the study results ( <b>pgs. 13, 14</b> )
<b>Other information</b>		
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 ( <b>pg. 15</b> )

\*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:	<i>BMJ Open</i>
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, Rachele; 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
<b>Primary Subject Heading</b>:	Public health
Secondary Subject Heading:	Communication, Mental health
Keywords:	loneliness, social media, natural language processing, MENTAL HEALTH, STATISTICS & RESEARCH METHODS

SCHOLARONE™  
Manuscripts



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

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

<sup>1</sup>Penn Medicine Center for Digital Health, Philadelphia, PA 19104

<sup>2</sup>Penn Medicine Center for Healthcare Innovation, Philadelphia, PA 19104

<sup>3</sup>Positive Psychology Center, University of Pennsylvania, Philadelphia, PA 19104

<sup>4</sup>Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104

<sup>5</sup>Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104

<sup>6</sup>The Wharton School, University of Pennsylvania, Philadelphia, PA 19104

<sup>7</sup>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.

1  
2  
3 **Conclusions:** Posts with the words lonely or alone often include psychosocial features and can  
4 potentially have associations with how individuals express and experience loneliness. This can  
5 inform low-resource online assessment for high risk individuals experiencing loneliness and  
6 interventions focused on addressing morbidities in this condition.  
7  
8  
9  
10  
11  
12  
13  
14

### 15 **Strengths and Limitations of this study**

- 16  
17 ● Novel focus on timelines of social media users to study mentions of loneliness and  
18 correlation with predictors of mental health.
- 19  
20 ● The study sample consists of social media users and is not representative of the general  
21 population.  
22  
23
- 24 ● Though we manually annotated a subset of posts mentioning loneliness, some may have  
25 been metaphorical or non sequiturs.  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

## 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.<sup>1</sup> 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.<sup>1-6</sup> It has also been linked with an increased risk of heart disease, stroke, dementia, depression, and anxiety.<sup>1,2,7-9</sup>

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.<sup>10</sup> 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.<sup>11</sup> 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,<sup>12,13</sup> social media has emerged as a potentially relevant tool to passively measure health states and behaviors of people.<sup>14,15</sup> 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.<sup>13,16</sup> 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.<sup>17</sup>

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

1  
2  
3 sought to characterize Twitter timelines of individuals' whose posts include the words lonely or  
4  
5 alone. Based on the language of such Twitter users, we analyzed the correlations between  
6  
7 posting about loneliness and users' mental health and psycholinguistic attributes (e.g. anger and  
8  
9 depression).

10  
11  
12  
13  
14 We hypothesize that language usage patterns would both confirm existing understanding of  
15  
16 loneliness and give new insights into the daily lives of those who express being lonely. As  
17  
18 loneliness can impact health outcomes, identifying ways to track prevalence and manifestations  
19  
20 of loneliness online would be useful for developing approaches for identifying and offering  
21  
22 support for these individuals. While prioritizing the privacy of individuals, specifically with the  
23  
24 amount of health insights that can be gleaned from social media, this research presents the  
25  
26 opportunity of digital platforms to not only provide markers of health but also potentially serve  
27  
28 as platforms that can be used for developing interventions.<sup>18,19</sup>  
29  
30  
31  
32  
33  
34

## 35 **Methods**

36  
37 This was a retrospective analysis of publicly available data on users posting about loneliness on  
38  
39 Twitter. This study was exempt by the University of Pennsylvania Institutional Review Board.  
40  
41  
42  
43

### 44 *Twitter Data*

45  
46 Twitter is a popular social media platform which allows users to send and receive short 140-  
47  
48 character messages, or 'tweets' (at the time of this study; the character limit was later increased  
49  
50 to 280). First, from the Twitter Streaming API, we collected tweets from the 1% sample using a  
51  
52 bounding box of location coordinates around Pennsylvania. To increase the sample size of tweets  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 from the state, all unique user IDs were recorded, and the Twitter API was used to extract  
4  
5 timelines (each user's prior 3200 tweets) filtered by timestamps ranging from 2012-2016.  
6  
7  
8  
9

### 10 *Patient and Public Involvement*

11  
12 Patients and public were not involved in the development of the research question and outcome  
13  
14 measures.  
15  
16  
17  
18

### 19 *Study Sample*

20  
21 We identified users who posted the word "alone" or "lonely" at least once in their timeline  
22  
23 (25,966 users). As social media includes colloquial, metaphorical, and light-hearted language  
24  
25 (eg. "If I see Justin Bieber, I will have a heart attack") we sought to identify the proportion of  
26  
27 tweets in which lonely seemed to refer to the public health meaning rather than other uses of the  
28  
29 term (e.g. metaphor, joke).<sup>20</sup> Two co-authors independently coded a random set of 100 tweets  
30  
31 from individuals who used the words lonely/alone at least 5 times in their timeline to identify  
32  
33 them as presumed to be associated with the feeling of loneliness or other (Cohen's  $\kappa = 0.70$ , and  
34  
35 76% of users' tweets indicate presumably feeling lonely). A few examples are as follows: "*i'm*  
36  
37 *feelin real depressed, confused, & lonely*", "*im always the only up around this time, feeling a lil*  
38  
39 *lonely*" and "*I'm so Lonely in life :( I just wish I can have love again it feels so go to be in love*  
40  
41 *with someone whom loves you.*" 6,202 users posted messages with "alone" or "lonely" at least 5  
42  
43  
44  
45  
46  
47 times.  
48  
49  
50

### 51 *Control group*

1  
2  
3 We then identified a control group of users by matching each user in the above dataset to another  
4 user by age, gender and period of activity (dates of first and last posting on twitter). We obtained  
5 the age and gender estimates by using lexica developed previously.<sup>21</sup> Then, we selected users  
6 with a minimum of 500 words across all their posts to have sufficient language for linguistic  
7 analyses.<sup>11</sup> We excluded non-English tweets, re-tweets, and tweets containing ‘alone’ and/or  
8 ‘lonely’ that were used to identify users in all analyses. Hereafter, we indicate users who had  
9 more than 5 posts with the words ‘lonely’ or ‘alone’ as ‘users with posts including the words  
10 lonely or alone’, and ‘control’ group to represent the matched set of users who had no such posts.  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23

#### 24 *Deriving language features to characterize individuals expressing loneliness*

25  
26 We used four sets of language features: a) dictionary-based psycholinguistic features,<sup>22</sup> b) open-  
27 vocabulary topics,<sup>23</sup> c) mental well-being attributes such as anxiety, depression by applying  
28 previously developed statistical models,<sup>24,25</sup> d) number of drug words and time of posts as past  
29 research has shown an association between loneliness and substance use.<sup>26,27</sup> These language  
30 features have been shown to be predictive of several health outcomes, such as depression,  
31 schizophrenia, attention deficit hyperactivity disorder (ADHD), and general well-being.<sup>17,26,28</sup>  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

42 *Dictionary-based:* From each post, we extracted the relative frequency of single words and  
43 phrases (consisting of two or three consecutive words). Then, all words used by less than 1% of  
44 users were removed from analysis so as to remove uncommonly used words (outliers).  
45  
46  
47

48  
49 Additionally, all tweets used to identify our study group were removed prior to further analysis.  
50  
51 The Linguistic Inquiry Word Count (LIWC) dictionary is a language-specific, many-to-many  
52 mapping of tokens (including words and word stems) and psychologically validate categories.  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Each category (a curated list of words) is found to be correlated with and also predictive of  
4 several psychological traits and outcomes. For each user, we measure the proportion of word  
5 tokens that fall into a given LIWC category.  
6  
7  
8  
9

10  
11  
12 *Open-vocabulary:* As closed-vocabulary approaches like LIWC include only a subset of the  
13 entire language used on social media, we use an open-vocabulary approach to improve the  
14 coverage and find topics in users' timelines mentioning loneliness. Topics consist of clusters of  
15 co-occurring words created using Latent Dirichlet Allocation (LDA).<sup>29</sup> The LDA generative  
16 model assumes that tweets contain a combination of topics, and that topics are a distribution of  
17 words. Since the words in a tweet are known, topics, which are latent variables, can be estimated  
18 through Gibbs sampling.<sup>30</sup> We use the Mallet implementation of the LDA algorithm, adjusting  
19 one parameter ( $\alpha=5$ ) to favor fewer topics per tweet.<sup>31</sup> All other parameters were kept at their  
20 default. An example of such a model is the following sets of words ('tuesday', 'monday',  
21 'wednesday', ...) which clusters together days of the week by exploiting their similar  
22 distributional properties across tweets. In our study, two hundred topics were generated using  
23 tweets across all users in the dataset including the words lonely or alone and control users.  
24  
25  
26  
27  
28  
29  
30  
31  
32  
33  
34  
35  
36  
37  
38  
39  
40  
41

42 *Mental well-being attributes:* We used automatic text-regression methods to assign to each user  
43 scores on the depression, anxiety and anger facets for users.<sup>24,25</sup> This model was trained on a  
44 sample of over 28,749 users who had taken the International Personality Item Pool Neuroticism-  
45 Extraversion-Openness Personality Inventory Revised (IPIP NEO-PI-R) survey that contains the  
46 depression, anxiety and anger Facets of the Neuroticism Factor.<sup>32,33</sup> The machine learning model  
47 trained on words and phrases from Facebook posts to predict survey measure of depression,  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 anger and anxiety resulted in a performance of  $r = .32$ , which is consistent with other reports of  
4 mental health states identified via social media.<sup>13</sup> The model was trained using status updates of  
5 users from another study<sup>24</sup>, and has been shown to generalize to Twitter users.<sup>25</sup>  
6  
7  
8  
9  
10

11  
12 *Use of Drug-words:* We also extracted the frequency of most common drug words as used on  
13 social media for every user in our analysis.<sup>27</sup>  
14  
15  
16  
17  
18

19 *Temporal patterns:* We determined the frequency of posts across different hours of the day by  
20 users in both users with posts including the words lonely or alone and control groups to  
21 understand the diurnal patterns in posting.  
22  
23  
24  
25  
26  
27

28 *Identifying differentially expressed language features in users with posts including the words*  
29 *lonely or alone*  
30  
31  
32

33 We isolated the patterns in users' loneliness mentions using the linguistic attributes and mental  
34 health attributes by correlating them with users with posts including the words lonely or alone  
35 and control groups. We used logistic regression to distinguish the different features associated  
36 with lonely and control groups and measure the effect size using Cohen's D. The models were  
37 set up to predict the group of users with posts including the words lonely or alone against the  
38 control group (e.g., group was the dependent variable). Details of the method are described in a  
39 previous work<sup>23</sup>. For identifying themes from topics, researchers looked at 20 messages each  
40 with the highest topic prevalence. We used Benjamini-Hochberg p-correction and  $p < 0.001$  for  
41 indicating meaningful correlations and the effect size was measured using Cohen's D. We also  
42 tested that the results hold if frequency of posting is used as an additional variable on which to  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 match the users with lonely expressions and the control subject. The statistical analysis, data  
4  
5 synthesis, and model creation was conducted in 2018-2019.  
6  
7  
8  
9

### 10 *Predicting the likelihood of posting about loneliness online*

11  
12 We then looked at the feasibility of predicting whether a user is likely to mention expressions of  
13  
14 loneliness or not based on their social media language. Automated analysis of social media is  
15  
16 accomplished by building predictive models, which use linguistic features that have been  
17  
18 extracted from social media data. For this analysis we used LIWC and topics as features.  
19  
20 Features are then treated as independent variables in an algorithm (Random Forests) to predict  
21  
22 the dependent variable of an outcome of interest (e.g., users' expressing that they are lonely or  
23  
24 not). For cross validation, the predictive model was trained, using Random Forests, on the  
25  
26 training set and then evaluated on a test set to avoid overfitting. The prediction performances are  
27  
28 reported as Area Under the Receiver Operating Curves (AUC) on an out-of-sample 5-fold cross  
29  
30 validation setting.  
31  
32  
33  
34  
35  
36  
37

### 38 **Results**

39  
40 Of the 408,296,620 tweets posted by users geo-located in Pennsylvania, USA, 25,966 users with  
41  
42 46,160,774 posts in their timelines, had at least one post with the words 'lonely' or 'alone', and  
43  
44 6,202 users with 17,995,084 posts in their timelines, had more than five such posts (Table 1).  
45  
46 Users with posts including the words lonely or alone had 1.9 times more posts in the study time  
47  
48 period as the control (Table 1). The median estimated age of this cohort was 21 years, and 69%  
49  
50 female.  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

**Table 1:** Descriptive statistics for users in the dataset

<b>Descriptive Statistics of the Dataset</b>		
	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	0.28
	hurt, feelings, trust, forget	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*
<b>Pronouns</b>	
1st Person Pronouns	0.18
<b>Cognitive Processes</b>	
Certainty	0.15

Discrepancies	0.15
Differentiation	0.14
Tentativeness	0.13
<b>Negative Emotions</b>	
Swearing	0.11
<b>Mental Well-being</b>	
Depression	0.81
Anger	0.95
Anxiety	0.75
<b>Drug words</b>	
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

1  
2  
3 analyzing the large volumes of social media data with the aim of better understanding the  
4  
5 varying manifestations of loneliness. This paper has three main findings. First, we identified  
6  
7 themes and contexts associated with users posting about loneliness on Twitter. Second, we  
8  
9 observed that users posting about loneliness used language associated with linguistic models for  
10  
11 anger, depression, and anxiety. Third, posts about loneliness were more likely to occur in the  
12  
13 evening or night.  
14  
15  
16  
17  
18

19 Themes associated with people mentioning loneliness on Twitter are consistent with prior  
20  
21 literature about substance use, emotional dysregulation, and troubles with relationships. For  
22  
23 example, in one study, a high positive correlation was found between alcoholism and groups of  
24  
25 lonely people, and lonely people were also found to express negative feelings towards  
26  
27 relationships.<sup>34</sup> This expression of negativity related to relationships is likely related to a  
28  
29 hypervigilance to social threat, associated with loneliness.<sup>35</sup> Lonely individuals were also  
30  
31 reported to focus on overcoming past events as well as showing feelings of helplessness.<sup>35</sup>  
32  
33  
34  
35  
36  
37

38 Association of users with posts including the words lonely or alone with linguistic estimates of  
39  
40 anger, depression, and anxiety corroborate prior research, showing that loneliness and social  
41  
42 isolation influence psychological functioning , specifically the ability to self-regulate  
43  
44 emotion.<sup>2,3,36</sup> Specifically, anxiety, anger, and negative mood were reported as higher in lonely  
45  
46 young adults.<sup>37</sup> Tweets by users with posts including the words lonely or alone were more self-  
47  
48 focused compared to the control group. Prior researchers have found that “first person singular  
49  
50 pronouns are a modest linguistic marker of depression.”<sup>38</sup> Also, previous research has shown that  
51  
52 loneliness has been associated with greater self-disclosure in Facebook posts.<sup>39</sup> This presents the  
53  
54  
55  
56  
57  
58  
59  
60



1  
2  
3 potential for early identification and assessment to intervene on loneliness as well as mental  
4  
5 health conditions for this group.  
6  
7  
8  
9

10 Trends in temporal variation in posting may reflect that sleep deprivation can contribute to social  
11 withdrawal and loneliness.<sup>40</sup> This finding corroborates prior research associating loneliness with  
12 diminished sleep quality.<sup>36</sup> A better understanding of the temporality of posting could inform  
13  
14 timing of interventions designed to address loneliness, as well as provide insight for other  
15  
16 researchers to test the inter-relationships between loneliness and the motivations for using social  
17  
18 media during nighttime.  
19  
20  
21  
22  
23  
24  
25

26 Loneliness is known to be one of the primary underlying causes and correlates for chronic  
27 mental health conditions.<sup>41</sup> As loneliness is becoming increasingly recognized as a public health,  
28 several groups have taken action to address it. For example, the United Kingdom appointed a  
29  
30 Minister for Loneliness who is responsible for addressing loneliness within communities.<sup>42</sup>  
31  
32 CareMore, a health plan and delivery system providing care for enrollees in Medicare Advantage  
33 and Medicaid health plans in seven states across the U.S., launched the “Togetherness Program”  
34  
35 in a clinical setting to address loneliness in elderly patients.<sup>43</sup> Through this work, CareMore  
36  
37 reported that participation in exercise programs increased by 56.6%, emergency room utilization  
38  
39 decreased by 3.3%, and hospital admissions among participants were 20.8% lower per thousand  
40  
41 compared to the “intent to treat population.”<sup>44</sup> Additionally, social network interventions  
42  
43 targeting loneliness have been found to be effective in reducing social isolation among  
44  
45 individuals with severe mental health conditions but these interventions are not included in the  
46  
47 treatment plans for individuals with a mental illness.<sup>45,46</sup>  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 Considering the advantage of large sample sizes and also the association between increased  
4 social media usage and individuals mentions of loneliness, it is promising to use natural language  
5 processing and machine learning to automatically identify a person mentions the words alone or  
6 lonely on Twitter to inform interventions targeted at early identification and support for affected  
7 and at risk individuals with the caveat that social media users are not representative of a random  
8 sample of individuals. To address loneliness will require being able to identify it passively,  
9 remotely, and over time. Many people rarely visit a healthcare provider so would miss the  
10 opportunity for screening. Approaches for treatment will also need to harness the tools and  
11 technologies that are accessible and integrated with the things people use every day (e.g. mobile  
12 phones). Future interventions would have to potentially rely on digital phenotyping of loneliness  
13 and using digital platforms (e.g. text messaging) to complement human-to-human interaction  
14 strategies to treat loneliness.  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30  
31  
32

33 In this first study, our aim was to characterize loneliness mentions based on users' entire  
34 timelines. Future studies could perform a time-series analysis of the temporal variations  
35 associated with loneliness mentions. Further, works should also validate whether the  
36 characteristics of people who are using the words 'lonely' or 'alone' on Twitter can be used to  
37 track community health risks, particularly, the risk of social isolation. Other studies should  
38 replicate the findings in this study using more formal ground truth such as surveys and extend  
39 this work to investigate if Twitter can potentially map regional hotspots of loneliness to identify  
40 problematic loneliness for community public health monitoring.  
41  
42  
43  
44  
45  
46  
47  
48  
49  
50  
51  
52  
53

## 54 **Limitations and Ethics**

55  
56  
57  
58  
59  
60

1  
2  
3 The study sample consists of social media users and is not representative of the general  
4  
5 population. An estimated 40% of US adults using Twitter are between the ages of 18 and 29, so  
6  
7 our analysis is skewed towards younger people.<sup>47</sup> An automated machine learning tool could be a  
8  
9 low-cost method to potentially detect posts about loneliness or being alone that may occur with  
10  
11 other concerning signals from digital sensors (e.g. changes in sleep, activity, purchases). Though  
12  
13 Twitter is far from perfect to be used as a diagnostic tool, these signals could trigger could then  
14  
15 be referred to more formal screening methods or support resources.<sup>48</sup>  
16  
17  
18  
19  
20

21  
22 Considering we identified that 76% of users' tweets indicated presumably feeling lonely in the  
23  
24 sample we hand coded, posts mentioning the words alone or lonely may have been metaphorical  
25  
26 or non sequiturs. Researchers who coded the topics were attempting to identify these associations  
27  
28 by looking at twenty messages each with the highest topic prevalence to identify themes, and we  
29  
30 acknowledge that this can be subjective. Also, considering the inclusion criteria based on number  
31  
32 of tweets mentioning alone or lonely, we are potentially selecting users with more posts than the  
33  
34 average twitter user. Further, the effects presented in this dataset may not be specific to  
35  
36 loneliness considering the potential comorbidity with mental health conditions such as  
37  
38 depression.  
39  
40  
41  
42  
43

44  
45 Social media use seeks to connect people but it also has been associated with increased perceived  
46  
47 social isolation.<sup>49</sup> It is unclear if social media use causes perceived social isolation or if perceived  
48  
49 social isolation causes social media use. The feasibility of social media-based assessments of  
50  
51 loneliness mentions (and mental health more broadly) needs further assessment. Privacy of  
52  
53 individuals is an ongoing concern, especially with social media users not fully realizing the  
54  
55  
56  
57  
58  
59  
60

1  
2  
3 amount of health insights that can be gleaned by their online posts. Employers and insurance  
4 companies, for example, may be motivated to derive these assessments, but could use these  
5 insights against those suffering from mental illness. As mental illnesses carry social stigma and  
6 may engender discrimination, data protection and ownership frameworks are needed to make  
7 sure the data is not used against the users' interest.<sup>50</sup> Further, transparency about which indicators  
8 are derived by whom for what purpose should be part of ethical and policy discourse. There are  
9 also open questions around the impact of misclassifications, and how derived mental health  
10 indicators can be responsibly integrated into systems of care.<sup>51</sup>  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23

## 24 **Conclusions**

25  
26 In this study we characterized mentions of loneliness on Twitter at the individual level.  
27  
28 Furthermore, we identified specific contexts, themes, and traits in the posts of individuals  
29 mentioning loneliness on Twitter. As loneliness is a public health challenge, a better  
30 understanding of how loneliness is described online can inform tracking of loneliness and  
31 interventions targeted at addressing this important public health problem in regards to the  
32 behavior of lonely individuals that may be at risk of developing a severe mental health condition.  
33  
34  
35  
36  
37  
38  
39  
40  
41

## 42 **Acknowledgements**

43  
44 **Funding:** This project is funded, in part, under a grant with the Pennsylvania Department of  
45 Health (ORG-FUND-PROG-CREF is 4290-567862-2446-2049). No sponsor of funding source  
46 played a role in: "study design and the collection, analysis, and interpretation of data and the  
47 writing of the article and the decision to submit it for publication." All researchers are  
48 independent from funders  
49  
50  
51

52 **Data Sharing Statement:** Data will be shared upon request from the corresponding author. The  
53 code used for analysis is made public at <http://dlatk.wwbp.org>  
54  
55  
56  
57  
58  
59  
60

**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-Urbe 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/6b2921eb65968fb68974b7701e0f3101fd92eef>.
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 Linguistics 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.

- 1  
2  
3 32. Costa Jr PT, McCrae RR. The Revised NEO Personality Inventory (NEO-PI-R). 2008.  
4 33. Bienvenu OJ, Samuels JF, Costa PT, Reti IM, Eaton WW, Nestadt G. Anxiety and  
5 depressive disorders and the five-factor model of personality: A higher-and lower-order  
6 personality trait investigation in a community sample. *Depress Anxiety*. 2004;20(2):92-97.  
7 34. Booth R. Toward an Understanding of Loneliness. *Soc Work*. 1983;28(2):116-119.  
8 doi:10.1093/sw/28.2.116  
9 35. Qualter P, Vanalst J, Harris R, others. Loneliness across the life span. *Perspect Psychol*  
10 *Sci*. 2015;10:250-264.  
11 [https://journals.sagepub.com/doi/full/10.1177/1745691615568999?url\\_ver=Z39.88-](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)  
12 [2003&rfr\\_id=ori:rid:crossref.org&rfr\\_dat=cr\\_pub=pubmed](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).  
13 36. Hawkey LC, JT. C. Loneliness Matters: A Theoretical and Empirical Review of  
14 Consequences and Mechanisms. *Ann Behav Med*. 2010;40:218-227. doi:10.1007/s12160-  
15 010-9210-8  
16 37. Cacioppo JT, Hawkey LC, Ernst JM, others. Loneliness within a nomological net: An  
17 evolutionary perspective. *J Res Pers*. 2006;40(6):1054-1085.  
18 doi:10.1016/j.jrp.2005.11.007  
19 38. Edwards T, Holtzman NS. A. *Meta-Analysis of Correlations between Depression and*  
20 *First Person Singular Pronoun Use*. Journal of Research  
21 39. Al-Saggaf Y, Nielsen S. Self-disclosure on Facebook among female users and its  
22 relationship to feelings of loneliness. *Comput Human Behav*. 2014;36:460-468.  
23 doi:10.1016/j.chb.2014.04.014  
24 40. Simon EB, MP. W. Sleep loss causes social withdrawal and loneliness. *Nat Commun*.  
25 2018;9:1. doi:10.1038/s41467-018-05377-0  
26 41. Cacioppo JT, Hawkey LC, RA. T. Perceived social isolation makes me sad: 5-year cross-  
27 lagged analyses of loneliness and depressive symptomatology in the Chicago Health,  
28 Aging, and Social Relations Study. *Psychol Aging*. 2010;25:453-463.  
29 doi:10.1037/a0017216  
30 42. Yeginsu CUK. Appoints a Minister for Loneliness. *New York Times*. 2018;17.  
31 <https://www.nytimes.com/2018/01/17/world/europe/uk-britain-loneliness.html>.  
32 43. Rubin R. Loneliness Might Be a Killer, but What's the Best Way to Protect Against It.  
33 *Jama*. 2017;318:19. doi:10.1001/jama.2017.14591  
34 44. Jain S. CareMore Health Tackles the Unmet Challenges of the Aging Population. *Am Soc*  
35 *Aging*. 2018;42:14-18.  
36 <https://www.ingentaconnect.com/contentone/asag/gen/2018/00000042/00000001/art00003>  
37 .  
38 45. Perese EF, Wolf M. Combating Loneliness Among Persons With Severe Mental Illness:  
39 Social Network Interventions Characteristics, Effectiveness, And Applicability. *Issues*  
40 *Ment Health Nurs*. 2005;26(6):591-609. doi:10.1080/01612840590959425  
41 46. Webber M, Fendt-Newlin M. A review of social participation interventions for people  
42 with mental health problems. *Soc Psychiatry Psychiatr Epidemiol*. 2017;52:369-380.  
43 doi:10.1007/s00127-017-1372-2  
44 47. *U. S. Twitter Reach by Age Group 2018 | Statista*. Statista  
45 [https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-](https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/)  
46 [age-group/](https://www.statista.com/statistics/265647/share-of-us-internet-users-who-use-twitter-by-age-group/).  
47 48. Barnett I, Ethics TJ. Transparency, and Public Health at the Intersection of Innovation and  
48 Facebooks Suicide Prevention Efforts. *Ann Intern Med*. 2019;170:565. doi:10.7326/m19-

- 1  
2  
3 0366  
4 49. Primack BA, Shensa A, Sidani JE, others. Social Media Use and Perceived Social  
5 Isolation Among Young Adults in the U. *S Am J Prev Med*. 2017;53(1):1-8.  
6 doi:10.1016/j.amepre.2017.01.010  
7  
8 50. Mckee R. Ethical issues in using social media for health and health care research. *Health*  
9 *Policy (New York)*. 2013;110(2-3):298-301. doi:10.1016/j.healthpol.2013.02.006  
10  
11 51. Inkster B, Stillwell D, Kosinski M, P. J. A decade into Facebook where is psychiatry in  
12 the digital age? *The Lancet Psychiatry*. 2016;3(11):1087-1090. doi:10.1016/s2215-  
13 0366(16)30041-4  
14  
15  
16  
17  
18  
19  
20  
21

## 22 **Figure legends**

23  
24 **Figure 1: Words/Phrases more likely to be posted by Twitter users with a) posts including**  
25 **the words lonely or alone compared to the b) control group.**  
26  
27

28  
29 Word size indicates the strength of the correlation and word color indicates relative word  
30 frequency. ( $p < 0.001$ , Bonferroni p-corrected)  
31  
32  
33

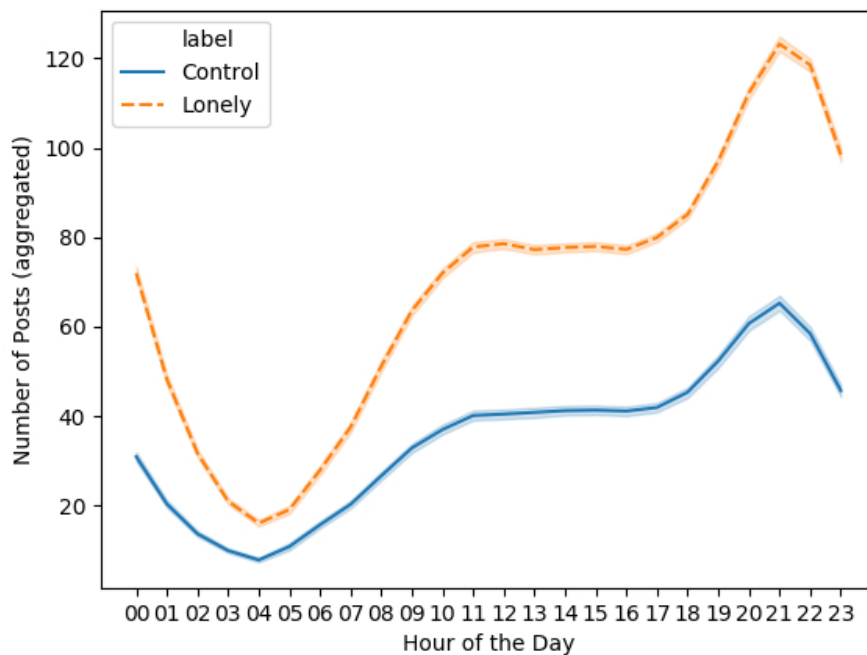
34  
35  
36 **Figure 2: Temporal variation showing diurnal patterns of post frequency of both the users**  
37 **with posts including the words lonely or alone and control group.**  
38  
39

40  
41 The dotted line indicates the percentage of posts at different hours of the day by the group of  
42 users with at least 5 posts containing the word 'lonely' or 'alone' and the solid line indicates  
43 users who do not have any posts about loneliness. The x-axis represents the hour of the day each  
44 post occurs and the y-axis indicates the number of posts for each group.  
45  
46  
47  
48  
49  
50  
51  
52  
53  
54  
55  
56  
57  
58  
59  
60





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
<b>Title and abstract</b>	1	(a) Indicate the study's design with a commonly used term in the title or the abstract ( <b>pg.2</b> ) (b) Provide in the abstract an informative and balanced summary of what was done and what was found ( <b>pg.2</b> )
<b>Introduction</b>		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported ( <b>pg.4</b> )
Objectives	3	State specific objectives, including any prespecified hypotheses ( <b>pg. 4</b> )
<b>Methods</b>		
Study design	4	Present key elements of study design early in the paper ( <b>pg.5</b> )
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection ( <b>pg. 5</b> )
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up ( <b>pg. 6</b> ) (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 ( <b>pg. 6</b> )
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 ( <b>pg. 6</b> )
Bias	9	Describe any efforts to address potential sources of bias ( <b>pg. 6</b> )
Study size	10	Explain how the study size was arrived at ( <b>pg. 6</b> )
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why ( <b>pg. 7</b> )
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding ( <b>pg. 9</b> ) (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
<b>Results</b>		
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 ( <b>pg. 9</b> ) (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 ( <b>pg. 9</b> ) (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 ( <b>pgs 10,11</b> )
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 ( <b>pgs 10,11</b> )

		(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 ( <b>pgs 10,11</b> )
<b>Discussion</b>		
Key results	18	Summarise key results with reference to study objectives ( <b>pgs. 12, 13</b> )
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 ( <b>pgs. 14</b> )
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence ( <b>pg. 13, 14</b> )
Generalisability	21	Discuss the generalisability (external validity) of the study results ( <b>pgs. 13, 14</b> )
<b>Other information</b>		
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 ( <b>pg. 15</b> )

\*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>.