Mental Health Surveillance over Social Media with Digital Cohorts

Silvio Amir, Mark Dredze and **John W. Ayers** †
Center for Language & Speech Processing, Johns Hopkins University, Baltimore, MD
†Division of Infectious Diseases & Global Public Health, University of California, La Jolla, CA
samir@jhu.edu, mdredze@cs.jhu.edu, ayers.john.w@gmail.com

Abstract

The ability to track mental health conditions via social media opened the doors for large-scale, automated, mental health surveillance. However, inferring accurate population-level trends requires representative samples of the underlying population, which can be challenging given the biases inherent in social media data. While previous work has adjusted samples based on demographic estimates, the populations were selected based on specific outcomes, e.g. specific mental health conditions. We depart from these methods, by conducting analyses over demographically representative *digital cohorts* of social media users. To validated this approach, we constructed a cohort of US based Twitter users to measure the prevalence of depression and PTSD, and investigate how these illnesses manifest across demographic subpopulations. The analysis demonstrates that cohort-based studies can help control for sampling biases, contextualize outcomes, and provide deeper insights into the data.

1 Introduction

The ability of social media analysis to support computational epidemiology and improve public health practices is well established (Culotta, 2010; Paul and Dredze, 2011; Salathe et al., 2012; Paul and Dredze, 2017). The field has seen particular success around the diagnosis, quantification and tracking of mental illnesses (Hao et al., 2013; Schwartz et al., 2014; Coppersmith et al., 2014a, 2015a,c; Amir et al., 2017). These methods have utilized social media (Coppersmith et al., 2014b; Kumar et al., 2015; De Choudhury et al., 2016), as well as other online data sources (Ayers et al., 2017, 2013, 2012; Arora et al., 2016), to obtain population level estimates and trends around mental health topics.

Accurately estimating population-level trends requires obtaining representative samples of the general population. However, social media has many well know biases, e.g. young adults tend be over-represented (demographic bias). Yet, most social media analyses tend ignore these issues, either by assuming that all the data is equally relevant, or by selecting data for specific outcome. For example, studying depression from users who talk about depression instead of first selecting a population and then measuring outcomes. Outcome based data selection can also introduce biases, such as over-representing individuals vocal about the topic of interest (*self-selection* bias). Consequently, trends or insights gleaned from these analyses might not be generalizable to the broader population.

Fortunately, these problems are well understood in traditional health studies, and well-established techniques from polling and survey-based research are routinely used to correct for these biases. For example, medical studies frequently utilize a cohort based approach in which a group is pre-selected to study disease causes or to identify connections between risk factors and health outcomes (Prentice, 1986). We can replicate these universally accepted approaches by conducting analyses over *digital cohorts* of social media users, characterized with respect to key demographic attributes. In this work, we propose to use such a social media based cohort for the purposes of mental health surveillance. We developed a digital cohort by sampling a large number of Twitter users at random (not based on outcomes), and then using demographic inference techniques to infer key demographics for the users namely, the age, gender, location and race/ethnicity. Then, we used the cohort to measure relative rates of both depression and PTSD, using supervised classifiers for each mental health condition. The inferred de-

Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, pages 114–120
Minneapolis, Minnesota, June 6, 2019.

mographic information allowed us to observe clear differences in how these illnesses manifest in the population. Moreover, the analysis demonstrates how social media based cohort studies can help to control for sampling biases and contextualize the outcomes.

2 Methodology

We now briefly describe our approach for cohort-based studies over social media. A more detailed description of the proposed methodology will appear in a forthcoming publication. Most works on social media analysis estimate trends by aggregating document-level signals inferred from arbitrary (and biased) data samples selected to match a predefined outcome. While some recent work has begun incorporating demographic information to contextualize analyses (Mandel et al., 2012; Mitchell et al., 2013; Huang et al., 2017, 2019) and to improve representativeness of the data (Coppersmith et al., 2015b; Dos Reis and Culotta, 2015), these studies still select on specific outcomes.

We depart from these works by constructing a demographically representative digital cohort of social media users *prior* to the analyses, and then conducting cohort-based studies over this pre-selected population. While a significant undertaking in most medical studies, the vast quantities of available social media data make assembling social media cohorts feasible. Such cohorts can be used to support longitudinal and cross-sectional studies, allowing experts to contextualize the outcomes, produce externally valid trends from inherently biased samples and extrapolate those trends to a broader population. Similar strategies have been utilized in online surveys, which can have comparable validity to other survey modalities simply by controlling for basic demographic features such as the location, age, ethnicity and gender (Duffy et al., 2005).

2.1 Building Digital Cohorts

Our cohort construction process entails two key steps: first, randomly selecting a large sample of Twitter users; and second, annotating those users with key demographic attributes. While such attributes are not provided by the API, automated methods can be used to infer such traits from data (Cesare et al., 2017). Following this approach, we develop a demographic inference pipeline to automatically infer **age**, **gender**, **race/ethnicity** and **location** for each cohort candidate.

Age Identifying age based on the content of a user can be challenging, and exact age often cannot be determined based on language use alone. Therefore, we use discrete categories that provide a more accurate estimate of age: *Teenager* (below 19), *20s*, *30s*, *40s*, *50s* (50 years or older).

Gender The gender was inferred using *Demographer*, a supervised model that predicts the (binary) gender of Twitter users with features based on the *name* field on the user profile (Knowles et al., 2016).

Race/Ethnicity The standard formulation of race and ethnicity is not well understood by the general public, so categorizing social media users along these two axes may not be reasonable. Therefore, we use a single measure of multicultural expression that includes five categories: *White* (W), *Asian* (A), *Black* (B), *Hispanic* (H), and *Other*.

Location The location was inferred using *Carmen*, an open-source library for geolocating tweets that uses a series of rules to lookup location strings in a location knowledge-base (Dredze et al., 2013). We use the inferred location to select users that live in the United States.

The age and race/ethnicity attributes were inferred with custom supervised classifiers based on Amir et al. (2017)'s user-level model. The classifiers were trained and evaluated on a dataset of 5K annotated users, attaining performances of 0.28 and 0.41 Average F_1, respectively. See the supplemental notes for additional details on these experiments[1].

2.2 Mental Health Classifiers

We build on prior work on supervised models for mental health inference over social media data. We focus on two mental health conditions — depression and PTSD — and develop classifiers with the *self-reported* datasets created for CLPysch 2015 (Mitchell et al., 2015; Coppersmith et al., 2015b). These labeled datasets derive from users that have publicly disclosed on Twitter a diagnosis of depression (327 users) or PTSD (246 users), with an equal number of randomly selected demographically-matched (with respect to age and gender) users as *controls*. For each user, the asso-

[1] https://samiroid.github.io/assets/demos.pdf

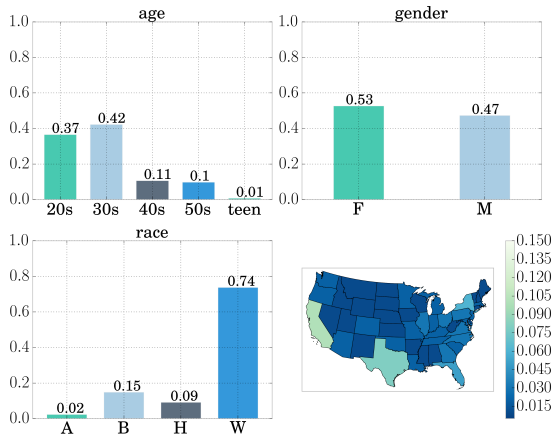


Figure 1: Demographics of the digital cohort.

ciated metadata and posting history was also collected — up to the 3000 most recent *tweets*, per limitations of the Twitter API.

The participants of the task proposed a host of methods ranging from rule-based systems to various supervised models (Pedersen, 2015; Preotiuc-Pietro et al., 2015; Coppersmith et al., 2015b). More recently, the neural user-level classifier proposed by Amir et al. (2017) showed not only good performance on this task, but also the ability to capture implicit similarities between users affected by the same diseases, thus opening the door to more interpretable analyses[2]. Hence, we adopt their model for this analysis.

3 Analysis

We constructed a cohort for our analysis by randomly selecting a sample of Twitter users and processing it with the aforementioned demographic inference pipeline. After discarding accounts from users located outside the United States, we obtained a cohort of 48K Twitter users with the demographic composition shown in Figure 1. Some demographic groups are over-represented (e.g. young adults) while others are grossly under-represented (e.g. teenagers) which illustrates the need for methodologies that can take these disparities into account.

We then processed the cohort through the mental-health classifiers to estimate the prevalence of depression and PTSD, and examine how these illnesses manifest across the population. The analysis revealed that 30.2% of the cohort members are likely to suffer from depression, 30.8% from PTSD, and 20% from both. We observe a significant overlap between people affected by depression **and** PTSD, which is not surprising given that the comorbidity of these disorders is well-known, with approximately half of people with PTSD also having a diagnosis of major depressive disorder (Flory and Yehuda, 2015).

How do these conditions affect different parts of the population? To answer this question, we looked at the affected users and measured how the demographics of individual sub-populations differ from those of the cohort as a whole. Figures 2 and 3 show the estimates for depression, PTSD and both, controlled for the cohort demographics. We observe large generational differences — PTSD seems to be more prevalent among older people whereas depression affects predominantly younger people. We also observe that in all cases Women are more susceptible than Men, and Blacks and Hispanics are more likely to be affected than Whites. This may represent a bias in the underlying data used to construct the classifiers, or a difference in how social media is used by different demographic groups. For example, models that were trained with a majority of data from White users maybe oversensitive to specific dialects used by other communities.

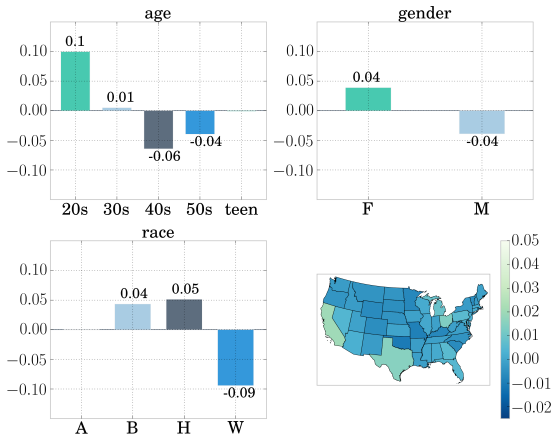
3.1 Discussion

Comparing our estimates with the current statistics provided by the NIH — a prevalence of 6.7% for depression[3] and 3.6% for PTSD[4] —, we can see that ours are much higher. It should be noted however, that the NIH reports refers to Major Depression episodes whereas our classifiers maybe also be sensitive to mild depressions which may never be diagnosed as such. Moreover, these estimates are not directly comparable since the NIH statistics are outdated (the estimates are from 2003 and 2015 for PTSD and depression, respectively) and our cohort was not adjusted to match the demographics of the US population. Nevertheless, it is worth noting that the relative prevalence rates, per demographic group, we obtained correlate with the NIH reports. For example, we observe similar distributions in terms of age and gen-

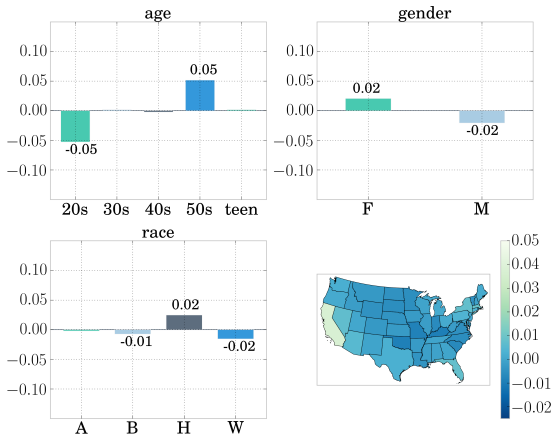
[2] a similar finding to Benton et al. (2017)

[3] https://www.nimh.nih.gov/health/statistics/major-depression.shtml

[4] https://www.nimh.nih.gov/health/statistics/post-traumatic-stress-disorder-ptsd.shtml



(a) Depression



(b) PTSD

Figure 2: The prevalence of two mental health conditions in the cohort.

der. However, we found that Blacks and Hispanics are more likely to be affected by mental illnesses, whereas the NIH reports a higher prevalence among Whites.

One possible reason for these disparities is that racial minorities are more likely to come from communities with lower education rates and socioeconomic status (SES), and to be in a position where they lack proper health coverage and mental-health care. Reports from the NIH and other US governmental agencies show that 46.3% of Whites suffering from a mental-illness were subjected to some form treatment, but this was case for only 29.8% of Blacks and 27.3% of Hispanics[5]. There may also be a bias in reporting within different racial and ethnic groups, as prevalence estimates can be biased by access to mental health care and social stigma. Recent studies show that factors such as discrimination and perceived inequality have a stronger influence on mental-health than it was previously supposed, even when controlling for the SES (Budhwani et al., 2015). Others have found that acute and chronic discrimination causes racial disparities in health to be even more pronounced at the upper ends of the socioeconomic spectrum. One of the reasons being that for Whites, improvements in SES result in improved health and significantly less exposure to discrimination, whereas for Blacks and Hispanics upwards mobility significantly increases the likelihood of discrimination and unfair treatment,

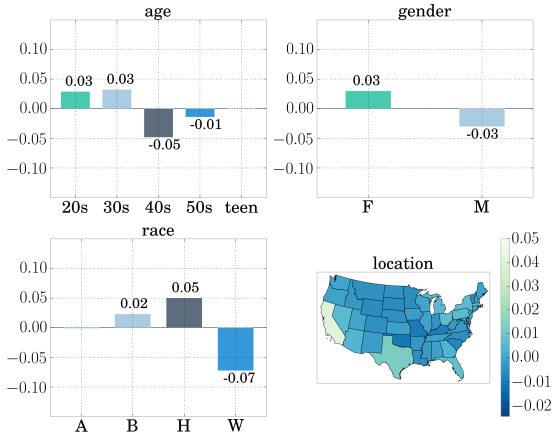


Figure 3: Depression and PTSD

as they move into predominantly White neighborhoods and work environments (Colen et al., 2017).

While an in-depth analysis of this issue is beyond the scope of this work, these results suggest that it deserves further investigation. A follow-up study to investigate the role of discrimination in mental-health could be conducted by adding a model to identify users who reported instances of discrimination and compare the prevalence of mental-illness with a control group.

4 Conclusions

We have presented the first cohort based study of mental health trends on Twitter. Instead of conducting the analysis over arbitrary data samples selected to match a given outcome, we first developed a digital cohort of social media users char-

[5] https://www.integration.samhsa.gov/MHServicesUseAmongAdults.pdf

acterized with respect to key demographic traits. We used this cohort to measure relative rates of depression and PTSD, and examine how these illnesses affect different demographic strata. The ability to disaggregate the estimates per demographic group allowed us to observe clear differences in how these illnesses manifest across different parts of the population — something that would not be possible with typical social media analysis methodologies. This brings social media analysis methodologies closer to universally accepted practices in surveillance based research.

Information about how different subpopulations perceive or are affected by certain health issues, could also improve public health policies and inform intervention campaigns targeted for different demographics. Moreover, the fact that some of our estimates correlate with statistics obtained through traditional methodologies suggests that this might be a promising approach to complement current epidemiology practices. Indeed, this opens the door to more responsive and deliberate public health interventions, and allow experts to track the progress or the effects of targeted interventions, in near real-time.

4.1 Privacy and Ethical Considerations

The majority of social media analysis approaches try to extract signals from individual posts and thus do not need to record any personal information. However, as we start moving towards user-level analyses, we are collecting and storing complete records of social media users communications. Even though this information is publicly available, people might not be consciously aware of the implications of sharing all their data and certainly have not given explicit consent for their data to be analyzed in aggregate. This is even more pertinent for analyses involving sensitive information (e.g. health related issues). As it has been demonstrated by the recent incidents involving companies inadvertently sharing or failing to protect users personal data, there is a serious danger of abuse and exploitation for systems that collect and store large amounts of personal data.

Even though this is in large part an ethical question, there are technical solutions that can be used to partially address this issue. One is to use anonymization techniques to obfuscate any details that allow third parties (even analysts) to identify the individuals that are involved in the study. Another is to store only abstract representations — which can still be updated and consumed by predictive models — , and discard the actual content. In regards to consent, there are initiatives to support voluntary data donation for research purposes, e.g. the *Our Data Helps* program[6].

References

Silvio Amir, Glen Coppersmith, Paula Carvalho, Mario J. Silva, and Bryon C. Wallace. 2017. Quantifying mental health from social media with neural user embeddings. In *Proceedings of the 2nd Machine Learning for Healthcare Conference*, volume 68 of *Proceedings of Machine Learning Research*, pages 306–321, Boston, Massachusetts. PMLR.

Vischal S Arora, David Stuckler, and Martin Mckee. 2016. Tracking search engine queries for suicide in the united kingdom, 2004–2013. *Public health*, 137:147–153.

John W Ayers, Benjamin M Althouse, Jon-Patrick Allem, Matthew A Childers, Waleed Zafar, Carl Latkin, Kurt M Ribisl, and John S Brownstein. 2012. Novel surveillance of psychological distress during the great recession. *Journal of affective disorders*, 142(1-3):323–330.

John W Ayers, Benjamin M Althouse, Jon-Patrick Allem, J Niels Rosenquist, and Daniel E Ford. 2013. Seasonality in seeking mental health information on google. *American journal of preventive medicine*, 44(5):520–525.

John W Ayers, Benjamin M Althouse, Eric C Leas, Mark Dredze, and Jon-Patrick Allem. 2017. Internet searches for suicide following the release of 13 reasons why. *JAMA internal medicine*, 177(10):1527–1529.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multitask learning for mental health conditions with limited social media data. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 152–162.

Henna Budhwani, Kristine Ria Hearld, and Daniel Chavez-Yenter. 2015. Depression in racial and ethnic minorities: the impact of nativity and discrimination. *Journal of racial and ethnic health disparities*, 2(1):34–42.

Nina Cesare, Christan Grant, and Elaine O Nsoesie. 2017. Detection of user demographics on social media: A review of methods and recommendations for best practices. *arXiv preprint arXiv:1702.01807*, 0.

[6] https://ourdatahelps.org/

Cynthia G Colen, David M Ramey, Elizabeth C Cooksey, and David R Williams. 2017. Racial disparities in health among nonpoor african americans and hispanics: the role of acute and chronic discrimination. *Social Science & Medicine*, 0.

Glen Coppersmith, Mark Dredze, and Craig Harman. 2014a. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015a. From ADHD to SAD: Analyzing the language of mental health on twitter through self-reported diagnoses. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 1–10. Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. 2015b. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39. Association for Computational Linguistics.

Glen A Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015c. From adhd to sad: analyzing the language of mental health on twitter through self-reported diagnoses. In *NAACL Workshop on Computational Linguistics and Clinical Psychology*, pages 1–10.

Glen A Coppersmith, Craig Harman, and Mark Dredze. 2014b. Measuring post traumatic stress disorder in twitter. In *International Conference on Weblogs and Social Media (ICWSM)*, pages 579–582.

Aron Culotta. 2010. Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 115–122, New York, NY, USA. ACM.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2098–2110, New York, NY, USA. ACM.

Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health from twitter. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 182–188. AAAI Press.

Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A Twitter geolocation system with applications to public health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

Bobby Duffy, Kate Smith, George Terhanian, and John Bremer. 2005. Comparing data from online and face-to-face surveys. *International Journal of Market Research*, 47(6):615.

Janine D Flory and Rachel Yehuda. 2015. Comorbidity between post-traumatic stress disorder and major depressive disorder: alternative explanations and treatment considerations. *Dialogues in clinical neuroscience*, 17(2):141.

Bibo Hao, Lin Li, Ang Li, and Tingshao Zhu. 2013. Predicting mental health status on social media. In *Cross-Cultural Design. Cultural Differences in Everyday Life.CCD 2013*, Lecture Notes in Computer Science, pages 101–110. Springer, Berlin, Heidelberg.

Xiaolei Huang, Michael Smith, Michael Paul, Dmytro Ryzhkov, Sandra Quinn, David Broniatowski, and Mark Dredze. 2017. Examining patterns of influenza vaccination in social media. In *AAAI Workshops*.

Xiaolei Huang, Michael C Smith, Amelia M Jamison, David A Broniatowski, Mark Dredze, Sandra Crouse Quinn, Justin Cai, and Michael J Paul. 2019. Can online self-reports assist in real-time identification of influenza vaccination uptake? a cross-sectional study of influenza vaccine-related tweets in the usa, 2013–2017. *BMJ open*, 9(1):e024018.

Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely simple name demographics. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 108–113.

Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM conference on Hypertext and hypermedia*. ACM.

Benjamin Mandel, Aron Culotta, John Boulahanis, Danielle Stark, Bonnie Lewis, and Jeremy Rodrigue. 2012. A demographic analysis of online sentiment during hurricane irene. In *Proceedings of the 2nd Workshop on Language in Social Media*, pages 27–36. Association for Computational Linguistics.

Lewis Mitchell, Kameron Decker Harris, Morgan R Frank, Peter Sheridan Dodds, and Christopher M Danforth. 2013. The geography of happiness: connecting Twitter sentiment and expression, demographics, and objective characteristics of place. *PLOS ONE*, 8(5).

Margaret Mitchell, Glen Coppersmith, and Kristy Hollingshead, editors. 2015. *Proceedings of the*

Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. North American Association for Computational Linguistics, Denver, Colorado, USA.

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, ICWSM11. Association for the Advancement of Artificial Intelligence.

Michael J. Paul and Mark Dredze. 2017. *Social Monitoring for Public Health*. Morgan & Claypool Publishers.

Ted Pedersen. 2015. Screening Twitter users for depression and PTSD with lexical decision lists. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 46–53.

Ross L Prentice. 1986. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73(1):1–11.

Daniel Preotiuc-Pietro, Maarten Sap, H Andrew Schwartz, and LH Ungar. 2015. Mental illness detection at the world well-being project for the clpsych 2015 shared task. *NAACL HLT 2015*, page 40.

Marcel Salathe, Linus Bengtsson, Todd J Bodnar, Devon D Brewer, John S Brownstein, Caroline Buckee, Ellsworth M Campbell, Ciro Cattuto, Shashank Khandelwal, Patricia L Mabry, et al. 2012. Digital epidemiology. *PLoS computational biology*, 8(7):e1002616.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. 2014. 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*, pages 118–125. Association for Computational Linguistics.