Robust language-based mental health assessments in time and space through social media

Siddharth Mangalik[a,1], **Johannes C. Eichstaedt**[b,c,1], **Salvatore Giorgi**[e], **Jihu Mun**[a], **Farhan Ahmed**[a], **Gilvir Gill**[a], **Adithya V. Ganesan**[a], **Shashanka Subrahmanya**[c], **Nikita Soni**[a], **Sean A. P. Clouston**[d], **and H. Andrew Schwartz**[a,1]

[a]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA; [b]Department of Psychology, Stanford University, Stanford, CA, USA; [c]Institute for Human-Centered A.I., Stanford University, CA, USA; [d]Department of Family, Population, and Preventive Medicine, Renaissance School of Medicine, Stony Brook University, Stony Brook, NY, USA; [e]Department of Computer and Information Science, University of Pennsylvania

This manuscript was compiled on May 1, 2023

Compared to physical health, population mental health assessment in the U.S. is very coarse-grained. Currently, in the largest population surveys by the Centers for Disease Control and Gallup, mental health is only broadly captured through surveys as "mentally unhealthy days" or "sadness", and estimates can only be aggregated infrequently to state or metropolitan estimates. Through the large-scale analysis of social media, robust estimation of population mental health is feasible at finer resolutions. In the present work, we validated a pipeline that used 1.2 billion Tweets from 2 million geo-located users to estimate mental health changes for the two leading mental health conditions, depression and anxiety. First, we found that language-based mental health assessments (LBMHAs) had substantially higher levels of reliability across space and time than surveys, down to the level of county weeks. Further, where surveys were available, we found moderate to large associations between the LBMHAs and survey scores from Gallup for multiple levels of granularity, from the national level down to weekly county measurements (fixed effects $\beta = .25$ to 1.58; $p < .001$). Additionally, LBMHAs demonstrated temporal validity, showing clear absolute increases after a list of major events (+23% increase over average weekly change for depression). Further, LBMHAs showed greater cross-sectional correlations with external health and socioeconomic county variables than Gallup surveys. This study suggests that the careful aggregation of social media data yields spatiotemporal estimates of population mental health that exceed surveys in resolution and may exceed them in reliability and validity.

Depression | Anxiety | Social Media Analysis | Population Health

Mental health is a large public health concern, causing large economic impact and loss of quality of life. Recent estimates suggest that depression affects 19.4 million Americans (7.8% of the population, 2020 est.) each year (1), while generalized anxiety disorder affects approximately 6% of the US population (19.8 million people, 2010 est.) (2). Globally, mental health conditions are the fifth-most common cause of reduced quality of life (3). Critically, poor mental health is thought to play a central role driving recent increases in prevalence and severity of "deaths of despair" (4, 5) in part due to the influence of poorer mental health on suicide attempts and suicide mortality obesity (6), and opioid-related overdoses (7, 8).

Public health researchers and policymakers seek to understand and actively respond to emerging and changing conditions (9, 10). Yet, current standards for monitoring mental health outcomes rely on subjective surveys responses that have limited temporal or regional resolution. For example, yearly changes in depression are measured only by annual Gallup polling (11) and a handful of national surveys (12) while anxiety is not regularly assessed in any of these surveys (13). Nevertheless, improving geospatial resolution can provide researchers with tools to more reliably assess the distribution (14) and determinants of disease (15). Similarly, a wealth of small studies using ecological momentary assessment suggest that observations made on shorter timescales routinely identifies symptoms and correlates that are otherwise inaccessible to researchers (16, 17).

Applying validated measures of depression and anxiety, assessed objectively at regular time-intervals at the county-level could transform research in population mental health, allowing researchers for the first time to locate clusters and reasons for changes to poorer mental health (18). Since originally proposed, language-based assessments have developed to become a flexible source of observed emotions and behaviors from individuals (19), often with greater accuracy and predictive power than existing survey-based measures (20). Further, recent work has found significant increases in convergent validity via post-stratification techniques (21) to address known selection biases (22, 23).

Here, we integrated a series of recent advances into a single pipeline capable of generating *language-based mental health assessments* (*LBMHAs*: Figure 1), to produce appraisals of anxiety and depression over regions and time. We first eval-

Significance Statement

Measuring mental health of communities across time is essential for population health research and practice. However, predominant methods to examine how mental health varies across time and by location are relatively limited due to reliance on expensive self-report surveys. Here, we propose a measurement pipeline that brings together a decade of progress in social media-based well-being assessment, evaluating the technique's reliability and validity over 1 billion recent observations of social media language for assessing weekly depression and anxiety at the community level. Our results show that social media-based assessment can have comparable, and in some aspects superior, validity and reliability to contemporaneous polling-based efforts, providing reliable resolution down to the county-week level.

Author contributions: S.M., H.A.S., J.C.E., and S.A.P.C. designed research; S.M. J.C.E. and H.A.S. performed research; S.G. J.M. F.A., G.G., S.S., A.V.G., and N.S. contributed new reagents/analytic tools; and S.M., J.C.E., S.G., J.M., F.A., G.G., A.V.G., S.S., N.S., S.A.P.C., and H.A.S. analyzed data and wrote the paper.

The author declares no competing interests.

[1] To whom correspondence should be addressed: E-mail: smangalik@cs.stonybrook.edu, johannes.stanford@gmail.com, and has@cs.stonybrook.edu

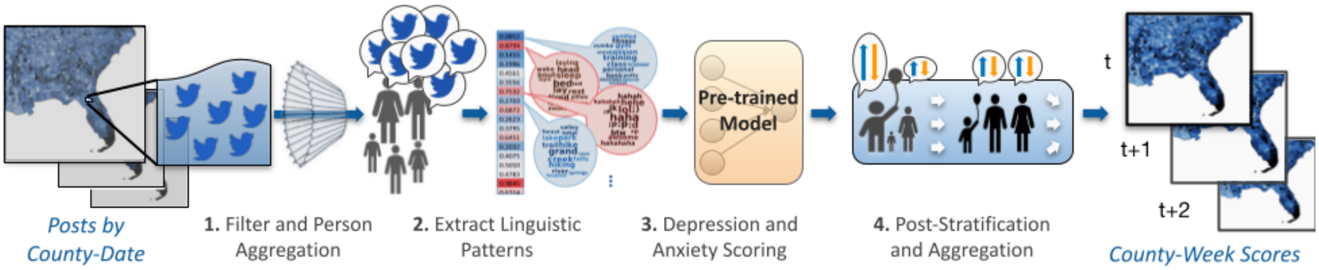


Fig. 1. A brief visual overview of how language is captured and tagged per county and week from social media platforms, and also explains how these data are then used to generate weighted depression and anxiety scores. County mapped messages are filtered to represent self-written language, the language extracted from these messages is used to generate user scores, then those scores are reweighted to better represent county demographics and are then aggregated to communities in time.

uated the reliability of LBMHAs, contrasted with standard survey approaches, while varying the time and space units (from annual, national down to daily, townships) as well as minimum thresholds for time-space specific observations. We then evaluated the convergent and external validity of the measurements as compared to the most extensively collected mental health related surveys available for the same time-period, both cross-sectionally and longitudinally. To facilitate open scientific inquiry we are releasing the LBMHA measurements as well as an open-source toolkit for running the pipeline and deriving mental health estimates.

Results

The nation-week depression and anxiety scores from our language based mental health assessments in 2020 adjusting for 2019 can be found in Figure 3A. The results as shown cover all weeks in 2020, and depict the included counties alongside the national average result in bold. Assessments have been generated for all counties that demonstrated sufficient posting history to be considered reliable per the thresholds determined in the reliability portion of this work, for this visualization a county must have at least 200 unique users in a given week to be included.

CTLB Data Descriptives

	Count
Word Instances	15,731,763,265
Posts	1,229,668,531
Unique Words	40,033,259
Users	2,045,124
Counties	1490
	Mean (S.D.)
Posts per User/Year	161.8 (246.2)
Posts per User/Week	6.9 (11.5)
Users per County	1391.4 (4,859.4)

Table 1. Coverage included in the filtered County Tweet Lexical Bank dataset from 2019 to 2020. Filtering consisted of excluding non-English posts, reposts, posts containing a hyperlink, and duplicated posts from users. Standard deviations are included next to mean measurements.

Reliability of Spatio-Temporal Resolutions. Figure 2 shows the relationship between different resolutions of time and space on the split-half reliability of our measurements. Underlying all measurements we use depression scores within the given spatio-temporal cohorts. The threshold (Cohen's d $= 0.1$) was crossed for all township-level measurements, all but one county-level measurement, and all of the MSA-level measurements. Looking across time for counties we determine that the week level is the smallest time resolutions with our smallest accepted space resolution to have a reliability (1 - Cohen's d) that is ≥ 0.9. Using this county-week finding we observed that once there were at least 50 users (user threshold [UT]=50), reliability exceeded 0.8. In this context, the UT can be understood as the minimum number of unique users needed by a county to be included in our analysis. At a UT of 200 it is possible to obtain a reliability measurement of 0.9 indicating no effect. This analysis lead us to create standard county-week threshold guidelines at UT of 50 and 200. The use of a 50 UT (720 distinct counties) reflects the highest number of counties that are directly usable versus the more restrictive 200 UT (370 distinct counties).

Convergent Validity. Figure 4 depicts the outcomes of our multi-level fixed effects model between Gallup self-reported sadness and worry against our language based assessments of depression and anxiety. At all levels evaluated for fixed effects we find our t-test p value to be significant to 0.01. At the nation-week level, we find that the survey and language are correlated (Pearson's r $= 0.39$) with depression and sadness, and anxiety and worry (r $= 0.68$). Fixed-effects coefficients between survey and language findings indicate higher agreement in analyses using larger spatial and temporal units, with the highest coefficients coming from a national-week analysis. At finer resolutions we nevertheless still identify statistically significant positive values leading us to conclude that county-week level measurements may reflect greater local sensitivity that might not as consistently correspond to the greater national trends.

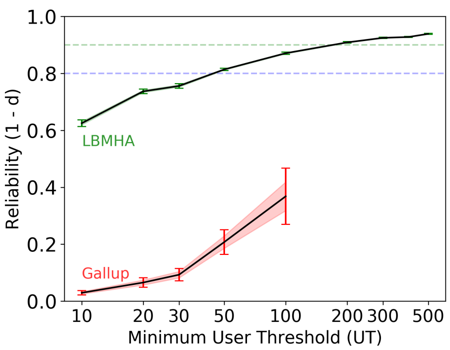
External Criteria. In Figure 4D we graphically represent the validity of our measures against other established county measures. The source of external county level data is the County Health Rankings (25) which track PESH (Political, Economic, Societal, and Health) outcomes on a county-year scale. We observe strong agreement between the correlations of our LBMHA scores and the Gallup self-reported results with these PESH variables.

In Figure 3B we examine the difference between event weeks and non-event weeks. We find an increase on average of the mean absolute difference of both depression (23%) and anxiety (16%) during weeks in which major US events occur. Likewise we see a "resetting" effect wherein non-event weeks on average decrease the general level of both anxiety (6%) and depression (8%), however nationally across 2020 the absolute

Reliability per Spatiotemporal Unit

	MSA (1)	County (23)	Township (155)
Year (2)	0.993	0.933	0.802
Quarter (3)	0.996	0.948	0.816
Month (8)	0.987	0.938	0.753
Week (36)	0.986	**0.921**	0.765
Day (252)	0.977	0.888	0.684

(A) Reliability by spatial and temporal units for LBMHAs.



(B) Reliability vs. Minimum User Threshold for All County-Weeks

Counts as Function of Minimum User Thresholds

	$n > 200$	$n > 50$	Full
County-Weeks	36,260	72,928	150,670
Distinct Counties	370	720	1,490
Distinct States	51	51	51
Means (S.D.) for County-Weeks			
	$n > 200$	**$n > 50$**	**Full**
Users/County-Week	1,585 (3,042)	815 (2,297)	399 (1,650)
Depression Score	2.41 (0.076)	2.42 (0.098)	2.42 (0.34)
Anxiety Score	2.74 (0.073)	2.74 (0.097)	2.76 (0.35)

(C) County-Week Data Descriptives

Fig. 2. Spatiotemporal reliability of language based mental health assessments of depression across different granularities of space and time in the New York metropolitan area. The heatmap in Table 2A shows the $1 -$ Cohen's d reliability of select New York metropolitan depression data, at each space and time unit ≥ 20 unique users were required. From this heatmap we target the smallest time unit from the smallest space unit greater than 0.9, which is county-week. The plot in Figure 2B shows how the reliability of a county-week measurement of depression increases with the minimum number of unique users required to consider that county-week. In the case of Gallup data, after a UT of 100 none of the county measurements can meet the minimum criteria to be reported. Horizontal lines are drawn at 0.8 and 0.9 reliability, which were used to select a 50 and a 200 county user threshold. Standard error of the reliability is shown with red shading, and the 95% confidence interval is shown with error bars. The county-year Intraclass Correlations, test-length corrected (ICC2; (24)) at a UT of 50 are $ICC2 = 0.33$ for Gallup Sadness and $ICC2 = 0.97$ for LBMHA depression, while at a UT of 200 are $ICC2 = 0.87$ for Gallup and $ICC2 = 0.99$ for LBMHA. Table 2C shows data descriptives for the county-week dataset after applying a user threshold of 50 and 200 as per the reliability findings and applying all other thresholds.

unadjusted level of both measures is increasing. These results over a comparison of event and non-event weeks for several counties suggest that changes in community mental health can be attributed to specific events.

American Communities Comparison. Figure 5 shows how anxiety differs across American community types. We select the five communities for which we the greatest representation in our final dataset of county-week LBMHAs. We observe that the Exurbs, defined as communities that "lie on the fringe of major metro areas in the spaces between suburban and rural America", score as the most anxious and most depressed of observed U.S. communities. Although overall difference between community types are modest, we anticipate that examinations of factorized measures of anxiety and depression may show larger discrepancies.

Discussion

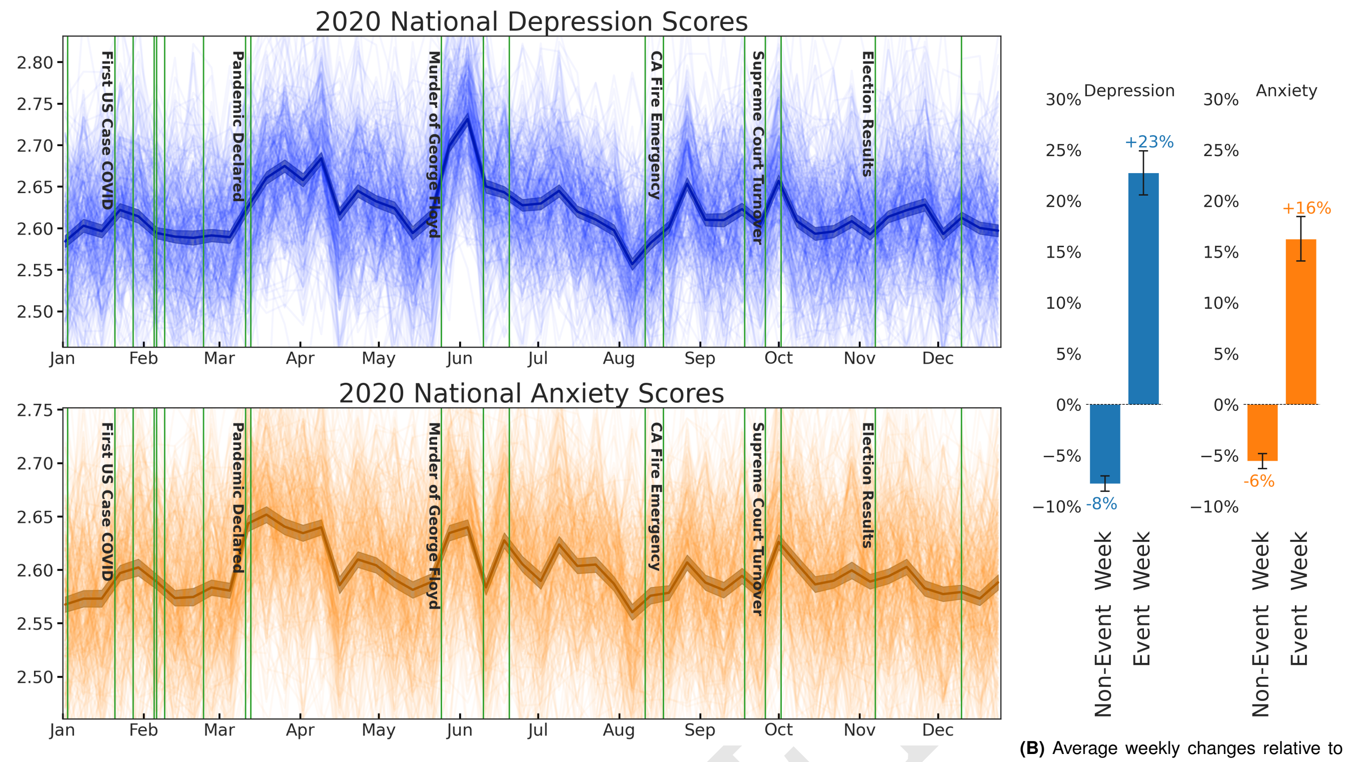
Anxiety and depression are costly, underdiagnosed, undertreated, and, while common overall, their prevalence varies across time and location. Depression alone has been attributed as the second highest mechanism for loss of disability-adjusted life years, more than cancer and diabetes (26). The present study used 15.7 billion words from 2.05 million people living across the U.S. to evaluate a modern approach for measuring public mental health, from behavioral patterns (language use). We found this approach achieved much greater regional and temporal resolution (e.g., within U.S. counties each week) while also achieving high convergent validity for the limited amount of high resolution survey-based assessments available.

We put together many recent developments in the best practices for social media-based well-being assessments. First, we utilized the notion of a digital cohort, whereby documents are aggregated through people, mirroring modern surveys (27). Then, we utilized new computational methods to mitigate epidemiological selection biases using *robust poststratification*(21). Additionally, we adapted anxiety and depression models, in the form of weighted lexica, to the specific domain of 2019 and 2020 Twitter, these adjustments have shown large gains when adapting models designed for new target domains (28). We also contributed a novel analysis on the statistical reliability of LBMHAs in order to establish minimal sampling thresholds. Finally building on epidemiological work we controlled for seasonality effects by adjusting using previous data to find the changes attributable to events occurring in 2020(29).

Our LBMHA pipeline reported similar temporal patterns, both nationally and at the county-level, to existing U.S. weekly data from Gallup, while also demonstrating the ability to report reliable results for a far larger number of counties and weeks. Further, LBMHAs captured changes in depression and generalized anxiety that corresponded to major events in 2020, including those of the COVID-19 pandemic declaration.

Symptom presence and severity cannot be readily measured for mental illness because unlike physical illnesses, they have no highly sensitive biomarkers. Furthermore self-reports are suspected to be hindered by stigmatization associated with mental illness. To improve the assessment process, this work joins recent research focused on identifying behavior-based or objective measures including functional (30) or structural neuroimaging (31), as well as those capturing cellular changes (32). Instead of relying on putative biomarkers to identify behavioral disorders, this study instead determines levels of depression and anxiety by observing individuals' natural unedited communications.

The shared geographic and temporal resolution presented in this study could enable the ability to understand the role of social, economic, or natural events and mental health at unprecedented resolutions. This study shows that improved resolution of mental health outcomes reflect the presence of major national events. For example, following the murder of George Floyd, language-estimated depression prevalence showed a clear increase, mirroring similar trends observed in



(A) Main Measurements

(B) Average weekly changes relative to mean change in event/non-event weeks

Fig. 3. Shown in Figure 3A are depression (blue) and anxiety (orange) measured at the nation-week level for all of 2020, controlling for 2019 measurements. All scores shown are based on aggregated user scores that are scaled from 0 to 5, 5 representing the highest level of depression/anxiety. Labeled green vertical markers are placed on the start of major events. In dark blue/orange, we have plotted nation-week averages alongside x 95% confidence intervals, and in thinner lines we show similar trends for individual counties. This figure requires counties to contain at least unique 200 (UT=200) users in a given week to be included, this gives distinct 370 counties spanning 2020. Figure 3B contains an analysis of the impact of weeks containing major US events against weeks without similar events. Shown are the z-scored percent differences from the prior week in LBMHAs between weeks that do contain major US events and those weeks that do not. Confidence interval bars are generated from Monte Carlo bootstrapping on 10,000 samples from the pool of either event weeks or non-event weeks and re-calculating mean z-scored percent differences between the drawn samples.

Gallup survey data (33).

COVID-19 first arrived in the U.S. during the data collection period (2019 to 2020). Consistent with prior research, we found that COVID-19 caused a rapid increase in depressive symptoms and generalized anxiety across the U.S. that did not dissipate before 2021. The distribution of poorer mental health was widespread and included large increases in regions with relatively low pre-pandemic levels of depression and anxiety. For example, the average level of anxiety increased from the lowest to the highest levels in Kansas in the months after the pandemic. These mental health shocks also began late in 2019, when COVID-19 was first being identified globally, and spiked in early March 2020 when much of the Northeastern U.S. was shuttered and people in open states chose to self-isolate. While these effects show the value of the approach for understanding how public mental health changes in a pandemic, these data also show that anxiety and depressive symptoms had not yet returned to pre-pandemic norms by the end of the observational window.

As with any social media platform, many users will self-present – deliberately behaving in ways that influence how others perceive them. Here that might look like a user sharing posts that emphasize the positive qualities of themselves that they would like their audience to see. It is important to understand that the language-based assessments we use treat language use as a behavior and do not rely on a priori assumptions of what language should signal a psychological trait. Rather, the LBMHAs we used are data-driven. Past work has shown that social media language behavior, whether motivated by self-presentation or not, is predictive of psychological traits and states (34).

We observed mental health using posts from geo-located Twitter users, as this allowed us to examine rapid changes in mental health at scale. LBMHAs have been reliably used outside of social media. For example, studies of psychological stress have noted that LBMHAs can aid in identifying individuals with at risk of poorer postpartum mental health when relying on mothers' diaries (35) and for identifying poorer long-term prognosis in post-traumatic stress disorder when relying on oral histories (20).

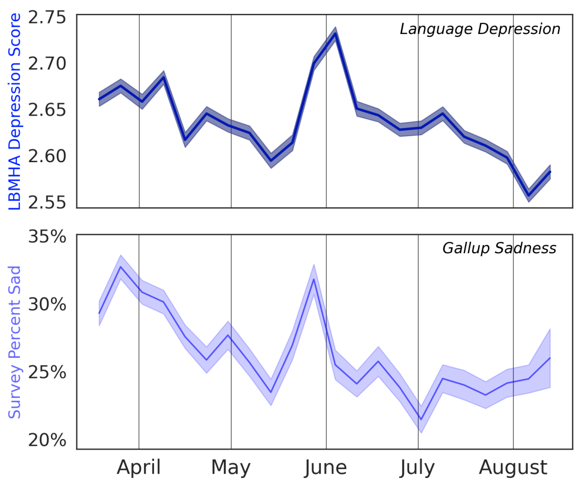
Limitations. Results from this study should be interpreted in light of a number of limitations. First, many U.S. counties with small populations or a small numbers of social media users had to be combined into super-counties to provide reliable estimates. Accounting for a only a small percentage of the total US population, these are regions that are often under-represented in research studies. This approach allowed for their inclusion but nevertheless resulted in units covering large geographic areas.

Additionally, social media platforms aren't rigid organizations and can change ownership, policies, and user populations. Twitter recently changed ownership resulting in new content

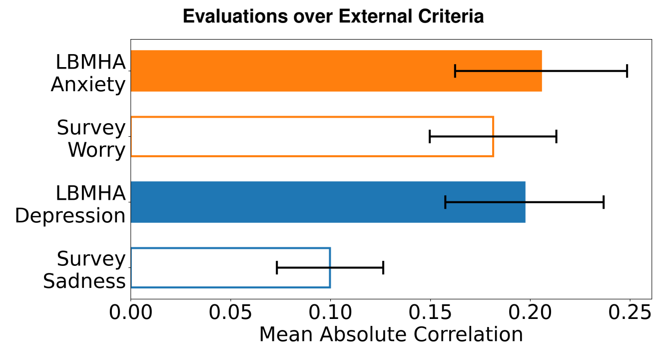
Evaluations for Convergent Validity

Space (N)	Time (N)	Depression β	Anxiety β
National (1)	Weeks (22)	$0.583^{\dagger}$	$1.582^{\ddagger}$
Regions (4)	Weeks (22)	$0.613^{\ddagger}$	$1.533^{\ddagger}$
Counties (132)	Quarters (3)	$0.346^{\ddagger}$	$1.178^{\ddagger}$
Counties (132)	Weeks (22)	$0.255^{\ddagger}$	$0.371^{\ddagger}$

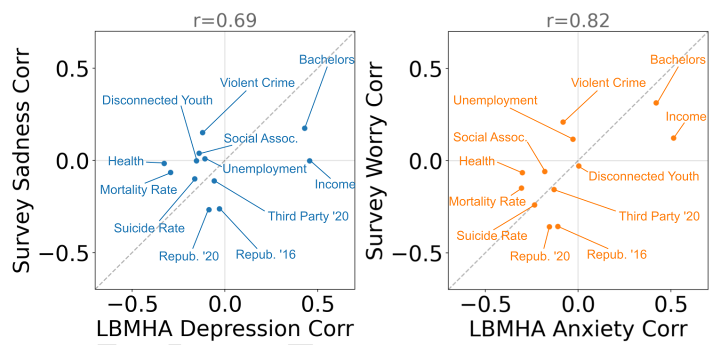
(A) Language and Survey fixed-effects Associations



(B) Language Depression and Survey Sadness, Nationally



(C) Average of Association Strength Across PESH Criteria



(D) Associations with specific PESH criteria

Fig. 4. Left-hand Column: Convergent validity between language-based mental health assessments and survey-based measures at different resolutions, as well as longitudinally. Table 4A shows fixed-effects coefficients between language based mental health assessments and measurements collected by the Gallup COVID-19 Panel Questionnaire. Depression β compares our language-based depression scores to Gallup's surveyed sadness scores via hierarchical linear modeling coefficients. Anxiety β compares our language-based anxiety scores to surveyed Gallup's worry scores. Figure 4B shows the national plots of depression as measured by LBMHAs and sadness as measured by Gallup. Both Questionnaire and LBMHA measures are held to reliability constraints as described in our section on reliability. Between the two national-week plots shown there is a $\beta = 0.583$. Results significant at: $^{\ddagger}p < .001$, $^{\dagger}p < .01$
Right-hand Column: Cross-sectional associations between language based mental health assessments (LBMHAs) of Anxiety/Depression and survey based assessments of Worry/Sadness against external criteria from Political, Economic, Social, and Health (PESH) variables across $N = 256$ counties. Figure 4C compares the average absolute effect Pearson correlations of LBMHA and Survey measures against external PESH variables. Figure 4D shows scatterplots of correlations between external criteria and our scoring method on one axis and the surveyed results on the other axis. All counties included meet our reliability requirements. Perfect agreement is shown as a diagonal dashed line. Association is measured using Pearson correlation. For the limited sample of PESH variables examined we observe a Pearson correlation of Pearson correlations of 0.82 for Anxiety-Worry and 0.69 for Depression-Sadness, both of these findings are significant to $p < 0.01$.

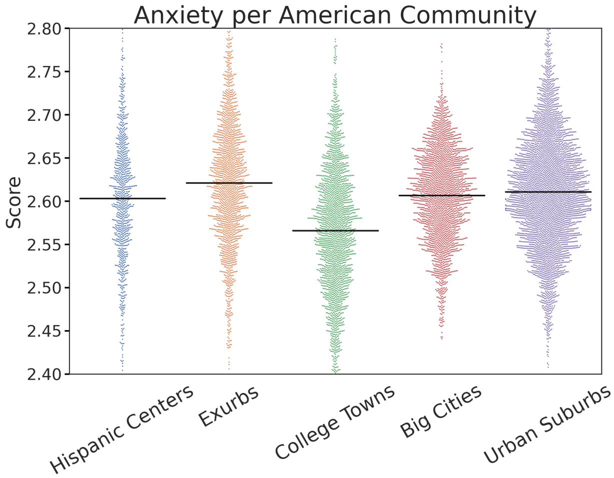
moderation strategies and data sharing practices. While other sources of public language exists, such as Mastodon or Reddit, the evaluations of this paper are focused on prior years of Twitter and any application after the recent ownership change or to other platforms require further validation.

This work centered around 2019 and 2020 data. Using 2019 as a control addresses some effects of having a short time-frame, such as seasonal effects. However, language evolves over time. Social media has a so-called "semantic drift" whereby words slowly begin to take on differing meanings (36–38). Thus, analyses of LBMHAs to future years should include convergent validations, reliability testing, and potentially apply further model adaptations.

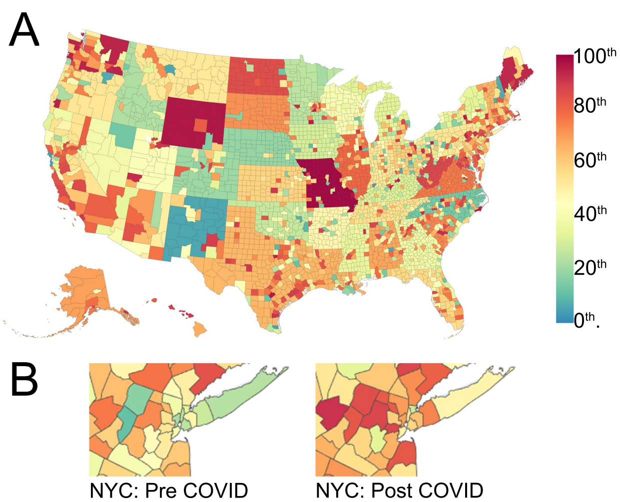
This work utilized lexicon-based models (i.e. weighted dictionaries). Recent work has shown that transformer-based language models (i.e. those used by programs like ChatGPT) can result in performance gains in assessing mental health from language (39, 40). Lexical models had two main advantages when we began this project: First, they have a longer history of use and the models we used have been through a wider range of validations at the person-level (41, 42). Second, they are much faster to run, requiring much fewer computing resources than large language models. As large language models (LLMs) become further validated at the person-level and more efficient to run across billions of texts, we anticipate that LBMHAs will begin to utilize them. We would expect LLM approaches to implicitly handle semantic drift and other word-context issues. The completion of this work supports future pipelines that can be recreated with transformer-based models.

Implications for Population Health. The strength of this epidemiological study is that it applied scalable methods meant to improve generalizability on a sample that included over 1 billion observations on 2 over million individuals (0.6% of the U.S. population) across more than 1,400 U.S. counties. These results are to our knowledge the first to validate temporal results only previously available from U.S. polling sites interested in tracking mental health.

To date, most efforts to profile the mental health of people in the U.S. and globally rely on subjective responses to survey prompts. These surveys may be biased by the tendency for people to under-report less desirable or stigmatized traits, such as the presence of mental illness. Up to date access to objective measures of changing mental health could improve in



(A) American Community anxiety scores



(B) Map of anxiety before and after pandemic declaration

Fig. 5. Scores within communities in 2020 and county mapped anxiety before and after COVID-19 is declared a pandemic. In 5A the 5 communities most represented in our data, out of 15 possible communities as defined by the American Communities Project, are shown ordered by the number of measurements captured. A black horizontal mean line is overlaid on swarm plots of the county-week measurements for each community type. In 5B percentile county-level measurements of anxiety are shown, where red shows where anxiety is highest and blue where anxiety is lowest. Pre-declaration is defined as two months before the declaration (3/13/2020) and post-declaration is defined as two months after the declaration. Section (A) depicts national anxiety per county in the post-declaration time window, while Section (B) shows a zoomed-in view of the NYC Metropolitan Area in each time window. Super-county binning is performed to report results for counties that are not individually reliable.

the ability to allocate scarce mental health treatment resources in a time of great need, and will facilitate new analyses that can help us to better understand the risk factors and consequences of depression and anxiety in population health.

This work lays a foundation to expand on the AI-based population assessment process to both refine the tools and improve the generalizability of assessments as we move this work into public mental health monitoring programs. Furthermore, quasi-experimental designs using rich temporal data have shown potential in revealing deeper facets of longitudinal effects suffered by those struggling with depression (43)

Beyond population health, applications of language based mental health assessments from social media in more the localized health in educational, professional, and medical organizations may be possible (44). For example, integrating a system using the pipeline described here into an opt-in program for communications platforms for high burnout professions, such as hospitals, WHO employees, or legal offices. This study suggests that the careful analysis and aggregation of social media data can yield spatiotemporal estimates of population mental health that exceed surveys in resolution and potentially in reliability and validity.

Materials and Methods

2019–2020 County Tweet Lexical Bank. As our main source of social media data we introduce an updated version of the original *County Tweet Lexical Bank* (27) which we refer to as *CTLB-19-20*. This new version contains a cohort of county mapped Twitter accounts and their posts spanning from 2019 to 2020. These county-user pairs were derived from posts with either explicit longitude/latitude pairs or the first instance of a self-reported user location in the account public profile. Previous work mapping location strings to counties was found to be 93% accurate compared to human assessments (45). The unprocessed CTLB-19-20 contained 2.7 billion total posts from a cohort of 2.6 million users over 2019 and 2020, after filtering this would result in 1.2 billion posts from 2 million users (see Table 1). For each post in this dataset we retain the date it was posted, a unique user identifier, the original text body, and the US county that the poster is from.

Filter and People Aggregation. Following Giorgi and colleagues (27), preprocessing steps filtered out posts to increase the accuracy of social media based population assessments (34). Posts are only included if they are marked likely to be English according to the langid package (46), and then they are further filtered to remove reposts, posts containing URLs (i.e. posts likely of non-original content), and finally any duplicate messages from individual users. The final processed dataset contains nearly 1 billion posts across of 2 million unique accounts for all 104 weeks in 2019 and 2020. At this point 1,490 counties (whose total population equals ∼92.5% of the US population) are captured. Further statistics about the filtered CTLB are described in more detail in Table 1.

To maintain a minimum level of reliability for our depression and anxiety measurements users must post at least 3 times in a given week to be included in that week, and from our reliability testing we determined that counties must contain at least 200 unique users per week to be considered for any given week. The 3 user posting threshold (3-UPT) was determined to balance diversity of users while minimizing noise from infrequent users. The 3-UPT approach resulted in a 37% decreased in unique user-week pairs retained, as opposed to a 23.4% decrease for 2-UPT and a 53% loss for 5-UPT . The 200 user post threshold (UPT) was determined by a reliability analysis whose results are shown in Figure 2B. Counties that fail to report a score for 10 weeks consecutively are dropped from the dataset to remove the influence they pose to findings for a single week.

After applying our 3-UPT, UT, and max gap filtering many posts belonging to mostly rural counties are necessarily excluded from our analysis. Since the target of this work is to better meet mental health reporting needs we implement a super-county binning strategy to reincorporate those "unreliable" county findings. All county-week findings that fail to meet the UT filter are weighted-mean aggregated by state into a super county-week result. Weights for the mean aggregation are assigned based on the reporting population of users of the included counties. Super counties must then pass the same UT set for regular counties to be included. In the case of UT=200 this results in a gain of 4,714 super county-week results over the original 30,899 county-week results. Figure 5 visually demonstrates how super-county binning reincorporates findings from unreliable counties.

The final post-processing step in our county-week pipeline is to run linear interpolation on a per county basis between missing weeks. For reference, at UT=200 this translates to an increase from 35,613 to 36,260 county-weeks. When running our analyses in this work we opt to adjust 2020 county-week findings by removing periodicity effects by subtracting means for 2019. This adjustment highlights 2020-specific movement from week to week.

Extract Linguistic Patterns. To extract language based assessments of well-being from posts, we used existing lexical models of depression and anxiety (41, 42) that we adapted to 2019-2020 Twitter vocabularies using target-side domain adaptation (28) which removes lexical signals that have different usage patterns (see target domain adaptation). The process for applying the model consists of extracting words from posts using the social media-aware tokenizer from *dlatk* (47). Following (48), the relative frequency of the words per user and unit of time are then Anscombe transformed to stabilize the variance of power law distribution. The approach then applies a linear model that is pretrained to produce anxiety and depression prediction scores from the word frequencies (42, 49). This produces a degree of depression (DEP_SCORE) and degree of anxiety (ANX_SCORE) for each user-time unit pair in the processed dataset, for this work that pair is user-week.

Depression and Anxiety Scoring. The calculation of a language based mental health scoring, for example the depression score for a user-week, is defined as:

$$LBMHA_{DEP}(x) = L(x) \times \text{demographics}(x)$$

$$L(x) = \sum_{w \,\in\, lex} [(A_{ns}(freq_w(x))) \times lex_{wt}(w)] + lex_i(DEP)$$

where $LBMHA_{measure}(x)$ is the Language Based Mental Health Assessment of an entity in time. x, is the sum of the summation of the lexicon weights $lex_{wt}()$ of all words w in the lexicon lex times that word's Anscombe transformed frequency,$A_{ns}(freq_w())$, and the overall lexicon intercept $lex_i()$ for that particular assessment. This outcome is multiplied by demographics(), which maps to a per user-week post-stratified weight correcting for the socio-economics of the community before aggregation.

It is noted that Twitter is a biased sample of the American populace, we find that their users are younger, more educated, and more male than the average American (50). In order to correct for these discrepancies from the true socioeconomic diversity of US counties we apply a post-stratified weighting scheme to emphasize the language of voices that are under-represented in social media.

Robust post-stratification (21) is a pipeline for generating post-stratification weights from sparse and noisy data (i.e., demographic estimates from machine learning models applied to social media text). These weights allow us to aggregate biased samples to accurately represent target populations being studied by adaptively removing selection biases. Calculating these weights starts with estimator redistribution where socio-demographic estimates are shifted per user such that the sample distribution matches the national-level target socio-demographic distributions. An adaptive binning process is then applied to these resulting sparse bin distributions to create merged bins that meet minimum observation thresholds. Finally, informed smoothing is applied by padding weights with a sample of users from a known distribution of demographics. In this work user-time-place mental health scores from social media are being redistributed through a weight that is assigned per county user-week LBMHA measurement.

The final aggregated community-time scores for depression and anxiety are then clipped to be between 0 and 5 for ease of interpretation. From these final scores, weighted aggregates can be generated at higher space and time resolutions.

Target Domain Adaptation. The mental health lexicon used in this work was originally trained for use on Facebook posts in the late 2000s so the following target-side domain adaptation steps were taken to adapt the lexicon to Twitter language in 2019-2020. In comparing the language use of Facebook versus Twitter we first trimmed the original lexicon's vocabulary which contained 7,680 unique words, to a set of 5,765 words for the target set where the *word usage* and *mean word frequency* between the two domains fell within certain ranges of each other.

To adapt lexical patterns to the target domain, we remove words which display different usage patterns in the target domain. Specifically, words that appeared with significantly different distributions in terms of sparsity or mean frequency. We then retrained the lexical model of mental health (41, 42) based on this filtered set of words to generate our domain-adapted well-being lexica (28).

More precisely, usage and frequency filters were used to address the phenomena of words and phrases that are used with different frequencies between two domains of text being more likely to have significant differences in their semantics between those two domains (28, 51). As the correlation between the frequency of a phrase and outcomes for a given lexicon may differ for semantically different usages of a phrase, filtering words with different usages and frequencies limits our set of tokens to those that are more likely to carry similar semantics (and thus, similar correlations). We modify (28)'s frequency filter for the source to target adjustment to instead normalize by standard deviation across the source Facebook users, and introduce a usage filter (what percent of users in each domain used a specific token even once).

Specifically, for each of our two domains (the target Twitter domain and the source Facebook domain), we computed each user's frequency for each word, and stored the results in frequency matrices C^S of dimension $n \times m$ and C^T of dimension $k \times m$, where n is the number of users in our source domain, k is the number of users in our target domain, and m is the cardinality of the set of words that appear either in the Twitter or Facebook domain. For each word, we then computed the average relative frequency across all users (word frequencies f^S for Facebook and f^T for Twitter), and the percent of users who used the word at least once (word usage percentages u^S and u^T).

First, only words with word usage percentages within a multiplicative factor 10 across domains were kept ($-1 < \log_{10}(u^T/u^S) < 1$), leaving 6,214 words. Then, for each word we take a Cohen's d filter of f^S versus f^T in the range $[-0.2, 0.2]$ on the word frequency using the larger source domain's standard deviation. A mathematical definition of this process is given in the supplement materials.

Finally we dropped common US names found in the United States' Social Security list of Popular Baby Names by Decade (e.g. Emma, Noah, Olivia, Liam)(52). The resulting Twitter adapted lexicon vocabulary after these three filters is 5,469 words long.

Using the Differential Language Analysis ToolKit's (DLATK) (47) regression-to-lexicon feature a new lexicon was trained using ridge regression, we note that the option to not standardize is selected since it better suits the lexicon creation task.

The final retrained lexicon contained 5,765 words and an intercept each with a weight for depression (DEP_SCORE) and anxiety (ANX_SCORE).

Statistical Analysis

Reliability vs. Resolution. At this point, we can begin to aggregate to a larger spatial or temporal resolution as necessary for analysis. To determine an appropriate resolution, we examine the finest resolution we can achieve while retaining reliable depression and anxiety score measurements.

To evaluate the reliability of a given spatio-temporal resolution, for each space-time pair in the resolution, we gather the set of users who posted at least 3 messages in this time period. If there are at least 20 such users, we randomly split the set into two approximately equally sized subsets and compute the split-half reliability ($R = 1 -$ Cohen's d) using their depression scores. Finally, the reliability is averaged across all space-time pairs.

Figure 2 shows the reliability scores of different spatio-temporal resolutions from running the procedure with counties in the New York City metropolitan area.

It is possible to generate reliable measures ($R > 0.9$) at the county-week level. We also analyze the effect of the threshold for the number of users per county-week pair on reliability.

Figure 2 shows the reliability scores from running the aforementioned procedure with the entire CTLB data and with different thresholds for the number of users.

When relying on regional data, we report data that exceed a final group frequency threshold placed at 50 or 200 to match repeated split-half reliability (RSR) where RSR > 0.7, 0.8, and 0.9 for these thresholds respectively. RSR is calculated as the mean Cohen's d of N repeated split-half samples into equal length a and b halves from the data belonging to a given region in time.

$$RSR = \frac{1}{N}\sum_{i=1}^{N} 1 - \frac{\mu_a - \mu_b}{\sigma_{a\,\cup\,b}}$$

Convergent Validity. Figure 4 we look to the Gallup COVID Panel (53) to compare the validity of our measure and determine if these assessments are tracking the same underlying construct. Note that we do not treat the Gallup poll as a gold standard to exactly align with since the poll is a survey based measure of self-reported sadness and worry, while our language based assessments are scores of depression and anxiety. The purpose of this particular study is to show common alignment between a traditional survey method and an observational social media method. The Gallup data is based on individual responses to a survey which are then tagged with a week and a county of the respondent. This dataset covers 2617 counties with an average of $\sim$4,601 measurements per week across all counties. To this end we use fixed effect multi-level modeling to remove the effects of endogeneity bias stemming from inherent between-county differences. While LBMHA scores are already held to a baseline 1-Cohen's d reliability of 0.9, Gallup results are held to a standard of 0.7. If this adjustment is not made there are no counties collected by Gallup for which county-week results are reliable for the full 22 weeks the survey covered.

External Criteria. To compare our assessments cross-sectionally against other external measurements we look to the County Health Rankings (CHR) (25). From CHR 2020 we look to political, economic, social, and health based outcomes at the county level. For political variables we evaluate the proportion of county voters who voted Republican in 2016 and 2020 and Third party in 2020. For economic variables, the logged median household income, the unemployment rate, and the proportion of people over age 24 holding bachelors degrees. For social variables, the per capita number of social associations, the violent crime rate, and the percent of youth unaffiliated with school or a similar organization. For health variables, the surveyed percent of people reporting fair or poor health, the age-adjusted suicide rate, and the age-adjusted mortality rate. LBMHAs were limited to the same cross-sectional period as was covered by the Gallup survey, reported correlations controlled for geographic effects at the state level. Figure 4D extends the cross-sectional test of validity to conduct a longitudinal study of major events on measurements across counties. For this work we examine the weekly changes in county measurements of anxiety and depression during weeks where major US events occurred and weeks where they did not occur. Combining 14 events identified by The Uproar (54) with 18 events from Business Insider (55) we arrived at 14 weeks of 2020 as "major US event weeks" (13 events were in common between the news sources and a single week could contain more than 1 event). We then filtered these to those that happened within the United States (including those applying global, such as pandemic onset) arriving at 14 total event weeks to compare with 38 non-event weeks. An event week is defined as an ISO week which contains the date any of the labelled major events occurred on. A 1 day buffer is added to the date of the event before mapping to a week so that scoring changes caused by the event can be captured. For each sample of event and non-event weeks, we collect the percent change in national-week depression and anxiety scores from the previous week. Using these two samples we compute Cohen's d between the event week and non-event week findings. To establish a confidence interval we use Monte Carlo bootstrapping over 10,000 iterations of event and non-event weeks.

Data Sharing and Availability. To support open science, we provide an open-source toolkit to run the LBMHA pipeline as well as data describing the results per county week. Please see github.com/wwbp/robust_spatiotemp for a repository of code and github.com/wwbp/lbmha_2019-2020 for a repository of data associated with this article. Additional code used for generating robust post-stratified weights can be found at github.com/wwbp/robust-poststratification.

ACKNOWLEDGMENTS. Support for this work was provided by: The National Institutes of Health and National Science Foundation Program, Smart and Connected Health, Grant NIH/NIMH R01 MH125702 (PIs Eichstaedt, Schwartz.), Centers for Disease Control NIOSH Grant U01 OH012476, as well as DARPA Young Faculty Award W911NF-20-1-0306. The conclusions and opinions expressed are attributable only to the authors and should not be construed as those of DARPA, the U.S. Department of Defense, or any other sponsor.

1. S Abuse, MHS Administration, Key substance use and mental health indicators in the united states: results from the 2019 national survey on drug use and health. *HHS Publ. No* **52**, 17–5044 (2020).
2. AJ Baxter, T Vos, KM Scott, AJ Ferrari, HA Whiteford, The global burden of anxiety disorders in 2010. *Psychol. Medicine* **44**, 2363–2374 (2014).
3. HA Whiteford, et al., Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *The lancet* **382**, 1575–1586 (2013).
4. EA Knapp, U Bilal, LT Dean, M Lazo, DD Celentano, Economic insecurity and deaths of despair in us counties. *Am. journal epidemiology* **188**, 2131–2139 (2019).
5. A Case, A Deaton, *Deaths of Despair and the Future of Capitalism.* (Princeton University Press), (2021).
6. Y Milaneschi, WK Simmons, EF van Rossum, BW Penninx, Depression and obesity: evidence of shared biological mechanisms. *Mol. psychiatry* **24**, 18–33 (2019).
7. MA Davis, LA Lin, H Liu, BD Sites, Prescription opioid use among adults with mental health disorders in the united states. *The J. Am. Board Fam. Medicine* **30**, 407–417 (2017).
8. M Matero, S Giorgi, B Curtis, LH Ungar, HA Schwartz, Opioid death projections with AI-based forecasts using social media language. *npj Digit. Medicine* **6**, 35 (2023).
9. P Nsubuga, et al., Public health surveillance: a tool for targeting and monitoring interventions. *Dis. Control. Priorities Dev. Countries. 2nd edition* (2006).
10. G Rose, Sick individuals and sick populations. *Int. journal epidemiology* **30**, 427–432 (2001).
11. Gallup, Health rating remains below pre-pandemic level [internet] (2021).
12. J Hsia, et al., Comparisons of estimates from the behavioral risk factor surveillance system and other national health surveys, 2011- 2016. *Am. journal preventive medicine* **58**, e181–e190 (2020).
13. NIoMH NIMH, *Prevalence of Generalized Anxiety Disorder Among Adults.* (National Institutes of Health, Bethesda, MD), (2021).
14. JT Chen, N Krieger, Revealing the unequal burden of covid-19 by income, race/ethnicity, and household crowding: Us county versus zip code analyses. *J. Public Heal. Manag. Pract.* **27**, S43–S56 (2021).
15. N Krieger, et al., Geocoding and monitoring of us socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter? the public health disparities geocoding project. *Am. journal epidemiology* **156**, 471–482 (2002).
16. AL Kratz, SL Murphy, TJ Braley, Ecological momentary assessment of pain, fatigue, depressive, and cognitive symptoms reveals significant daily variability in multiple sclerosis. *Arch. physical medicine rehabilitation* **98**, 2142–2150 (2017).
17. MA Russell, JM Gajos, Annual research review: Ecological momentary assessment studies in child psychology and psychiatry. *J. Child Psychol. Psychiatry* **61**, 376–394 (2020).
18. MJ Paul, M Dredze, Social monitoring for public health. *Synth. Lect. on Inf. Concepts, Retrieval, Serv.* **9**, 1–183 (2017).
19. K Jaidka, et al., Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proc. Natl. Acad. Sci.* **117**, 10165–10171 (2020).
20. Y Son, et al., World trade center responders in their own words: predicting ptsd symptom trajectories with ai-based language analyses of interviews. *Psychol. medicine* **2021 Jun 22**, 1–9 (2021).
21. S Giorgi, et al., Correcting sociodemographic selection biases for population prediction from social media in *Proceedings of the International AAAI Conference on Web and Social Media.* Vol. 16, pp. 228–240 (2022).
22. AP Christie, et al., Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences. *Nat. communications* **11**, 1–11 (2020).
23. J Mellon, C Prosser, Twitter and facebook are not representative of the general population: Political attitudes and demographics of british social media users. *Res. & Polit.* **4**, 2053168017720008 (2017).
24. PD Bliese, Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. *Multilevel theory, research, methods organizations* (2000).
25. U of Wisconsin Population Health Institute, County health rankings and roadmaps 2022. (2020).
26. C Holden, Global survey examines impact of depression. *Science* **288**, 39–40 (2000).
27. S Giorgi, et al., The remarkable benefit of user-level aggregation for lexical-based population-level predictions in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.* (Association for Computational Linguistics), pp. 1167–1172 (2018).
28. D Rieman, K Jaidka, HA Schwartz, L Ungar, Domain adaptation from user-level facebook models to county-level twitter predictions in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers).* pp. 764–773 (2017).
29. SH Woolf, DA Chapman, RT Sabo, DM Weinberger, L Hill, Excess deaths from covid-19 and other causes, march-april 2020. *Jama* **324**, 510–513 (2020).
30. JR Sato, et al., Machine learning algorithm accurately detects fmri signature of vulnerability to major depression. *Psychiatry Res. Neuroimaging* **233**, 289–291 (2015).
31. M Kritikos, et al., Cortical complexity in world trade center responders with chronic posttraumatic stress disorder. *Transl. Psychiatry* **11**, 1–10 (2021).
32. PF Kuan, et al., Metabolomics analysis of post-traumatic stress disorder symptoms in world trade center responders. *Transl. psychiatry* **12**, 1–7 (2022).
33. JC Eichstaedt, et al., The emotional and mental health impact of the murder of george floyd on the us population. *Proc. Natl. Acad. Sci.* **118**, e2109139118 (2021).
34. K Jaidka, et al., Estimating geographic subjective well-being from twitter: A comparison of dictionary and data-driven language methods. *Proc. Natl. Acad. Sci.* **117**, 10165–10171 (2020).
35. A Bartal, KM Jagodnik, SJ Chan, MS Babu, S Dekel, Identifying women with postdelivery posttraumatic stress disorder using natural language processing of personal childbirth narratives. *Am. J. Obstet. & Gynecol. MFM* **5**, 100834 (2023).
36. V Kulkarni, B Perozzi, S Skiena, Freshman or fresher? quantifying the geographic variation of language in online social media in *Proceedings of the International AAAI Conference on Web and Social Media.* Vol. 10, pp. 615–618 (2016).
37. WL Hamilton, J Leskovec, D Jurafsky, Cultural shift or linguistic drift? comparing two computational measures of semantic change in *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing.* (NIH Public Access), Vol. 2016, p. 2116 (2016).
38. K Jaidka, N Chhaya, L Ungar, Diachronic degradation of language models: Insights from social media in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).* pp. 195–200 (2018).
39. M Matero, et al., Suicide risk assessment with multi-level dual-context language and BERT in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology.* pp. 39–44 (2019).
40. R Martínez-Castaño, A Htait, L Azzopardi, Y Moshfeghi, Bert-based transformers for early detection of mental health illnesses in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings 12.* (Springer), pp. 189–200 (2021).
41. HA Schwartz, 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.* pp. 118–125 (2014).
42. Y Son, et al., World trade center responders in their own words: predicting ptsd symptom trajectories with ai-based language analyses of interviews. *Psychol. Medicine* p. 1–9 (2021).
43. K Saha, J Torous, E Kiciman, M De Choudhury, , et al., Understanding side effects of antidepressants: large-scale longitudinal study on social media data. *JMIR mental health* **8**, e26589 (2021).
44. K Saha, A Yousuf, RL Boyd, JW Pennebaker, M De Choudhury, Social media discussions predict mental health consultations on college campuses. *Sci. reports* **12**, 123 (2022).
45. H Schwartz, et al., Characterizing geographic variation in well-being using tweets in *Proceedings of the International AAAI Conference on Web and Social Media.* Vol. 7;1, pp. 583–591 (2013).
46. M Lui, T Baldwin, langid. py: An off-the-shelf language identification tool in *Proceedings of the ACL 2012 system demonstrations.* pp. 25–30 (2012).
47. HA Schwartz, et al., Dlatk: Differential language analysis toolkit in *Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations.* pp. 55–60 (2017).
48. HA Schwartz, et al., Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* **8**, e73791 (2013).
49. HA Schwartz, et al., Predicting individual well-being through the language of social media in *Biocomputing 2016: Proceedings of the Pacific Symposium.* (World Scientific), pp. 516–527 (2016).
50. G Blank, C Lutz, Representativeness of social media in great britain: investigating facebook, linkedin, twitter, pinterest, google+, and instagram. *Am. Behav. Sci.* **61**, 741–756 (2017).
51. P Resnik, Using information content to evaluate semantic similarity in a taxonomy in *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95. (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA), p. 448–453 (1995).
52. S Security, Popular baby names by decade (year?).
53. Gallup, Covid-19 panel microdata (2021).
54. C Majerac, The 14 most important events of 2020. *The Uproar: https://nashuproar.org/39777/features/the-14-most-important-events-of-2020* (2020).
55. Y Dzhanova, The events that shook and shaped america in 2020. *Bus. Insid. https://www.businessinsider.com/the-stories-of-2020-that-shaped-and-shook-americans-2020-12* (2020).