Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals

Sindhu Kiranmai Ernala
Georgia Tech
sernala3@gatech.edu

Michael L. Birnbaum
Zucker Hillside Hospital, Psychiatry Research
Mbirnbaum@northwell.edu

Kristin A. Candan
Zucker Hillside Hospital, Psychiatry Research
kcandan@northwell.edu

Asra F. Rizvi
Zucker Hillside Hospital, Psychiatry Research
ARizvi3@northwell.edu

William A. Sterling
Zucker Hillside Hospital, Psychiatry Research
wsterling2@northwell.edu

John M. Kane
Zucker Hillside Hospital, Psychiatry Research
JKane2@northwell.edu

Munmun De Choudhury
Georgia Tech
munmund@gatech.edu

ABSTRACT

A growing body of research is combining social media data with machine learning to predict mental health states of individuals. An implication of this research lies in informing evidence-based diagnosis and treatment. However, obtaining clinically valid diagnostic information from sensitive patient populations is challenging. Consequently, researchers have operationalized characteristic online behaviors as "proxy diagnostic signals" for building these models. This paper posits a challenge in using these diagnostic signals, purported to support clinical decision-making. Focusing on three commonly used proxy diagnostic signals derived from social media, we find that predictive models built on these data, although offer strong internal validity, suffer from poor external validity when tested on mental health patients. A deeper dive reveals issues of population and sampling bias, as well as of uncertainty in construct validity inherent in these proxies. We discuss the methodological and clinical implications of these gaps and provide remedial guidelines for future research.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; **Supervised learning by classification**; • **Human-centered computing** → **Social media**.

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300364

KEYWORDS

mental health; social media; machine learning; validity theory; construct validity; population bias; sampling bias

ACM Reference Format:
Sindhu Kiranmai Ernala, Michael L. Birnbaum, Kristin A. Candan, Asra F. Rizvi, William A. Sterling, John M. Kane, and Munmun De Choudhury. 2019. Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3290605.3300364

1 INTRODUCTION

With rising volumes of data and pervasive use, social media has been widely adopted as a lens to provide insights into behaviors [52], mood [42], psychological traits and states [5, 53], and social interactions of individuals [56]. For mental health, a growing body of work, including that in the human computer interaction (HCI) field, is leveraging naturalistic, unobtrusive data from social media to predict mental health states of individuals [21, 25, 28, 29, 31, 34]. Parallel to HCI, in an emergent field called "digital psychiatry" [100], clinicians are exploring the efficacy of diagnostic predictions from online data for early diagnosis, evidence-based treatment, and deploying timely patient-provider interventions [40, 48].

In this line of research, on the methodological front, supervised machine learning techniques have gained prominence, providing promising predictive outcomes of mental health states [66]. The success of these techniques, however, hinges on access to ample and high-quality gold standard labels for model training. In mental health, gold standard labels often comprise *diagnostic signals of people's clinical mental health states*, for instance, whether *an individual might be suffering from a specific mental illness*, or *at the cusp of experiencing an adverse episode* like a relapse or suicidal thoughts.

Unlike conventional machine learning tasks in fields like computer vision and natural language processing, extensive, high quality gold standard data for predicting clinical diagnoses of mental illnesses from social media is not readily available. Literature has advocated the use of clinically validated diagnostic information collected from *patient populations* for building such predictive models [11, 66]. However, undertaking such efforts presents many practical and logistical challenges. These range from the difficulties in recruiting a sensitive and high risk population, to the myriad privacy and ethical concerns that accompany engaging directly with vulnerable individuals. Because of the effort- and time-consuming nature of such data acquisition approaches and the need for deep-seated cross-disciplinary partnerships, particularly with clinicians, researchers have noted such data acquisition efforts to not scale easily and quickly to large and diverse populations [21].

Consequently, researchers have operationalized a variety of *online behaviors* as diagnostic signals to build machine learning approaches that predict mental illness diagnoses. These "proxies" are easily accessible and inexpensively gathered from social media, without the need to directly engage with the individuals themselves.We define binary indicators of the presence or absence of these social media behaviors that might correspond to their clinical mental health state as ***"proxy diagnostic signals"***. One notable example from literature consists of public self-reports of mental illnesses made by individuals in their social media feeds [21, 65].

This paper posits a significant challenge in using these *proxy diagnostic signals* revolving around their lack of clinical grounding, theoretical contextualization, and psychometric validity—concerns noted by psychiatrists and computational researchers alike [27, 48]. In other words, drawing on boyd and Crawford's critique [13], despite gains in scale, gaps exist in our understanding of how these signals are defined, where their theoretical underpinnings are, whether they objectively and accurately measure what they claim to measure (that is, the clinical mental illness diagnosis), and whether the patterns of behaviors they exemplify are truly representative of the behaviors of patients. More generally, our position is situated in criticisms in the broader social media research area, where the use of online data, removed from the individual and the specific offline context, and collected without their direct involvement, poses natural challenges to evaluation [55, 110].

Our Contributions. Toward addressing the above methodological gaps, this paper presents a first empirical study to assess the quality of different social media-derived proxy diagnostic signals in predicting clinical diagnoses of mental illness, for treatment and patient-provider interventions. Focusing on the specific mental illness of schizophrenia and drawing upon an involved partnership between HCI and clinical researchers, we consider three proxy signals widely used in prior literature [11, 63, 65]: 1) behaviors signaling affiliation to mental health resources, 2) self-reported diagnoses, and 3) clinically appraised self-reports of diagnoses. Adopting data triangulation [35] and modern validity theory [62] as methodological foundations, we examine their predictive validity (internal and external). To do so, we design a prediction task to distinguish those with schizophrenia from control populations, and leverage a carefully-curated social media dataset of clinically diagnosed schizophrenia patients seeking treatment at a large health-care organization.

Affiliation
Zhou et al. [111], Jamil et al. [50], McManus et al. [63], Nguyen et al. [68], Shen and Rudzicz [94], Gkotsis et al. [41], Saha and De Choudhury [85], Chancellor et al. [19]
Self-reports
O'Dea et al. [69], Prieto et al. [76], Burnap et al. [15], Coppersmith and colleagues [21, 22, 22, 23, 25], Benton et al. [8], Lin and colleagues [59, 60], Mitchell et al. [65], Loveys et al. [61], Simms et al. [97], Wang and colleagues [106, 107], Resnik et al. [80], Shen et al. [93], Vedula and Parthasarathy [105], Huang et al. [46, 47], Yates et al. [108]
External/Expert validation
Birnbaum et al. [11], Chancellor et al. [17], Park et al. [72], Reece and colleagues [78, 79], Schwartz et al. [90], Saha et al. [84], De Choudhury and colleagues [30, 31], Braithwaite et al. [14], Resnik et al. [81], Tsugawa et al. [102]

Table 1: Prominent proxy diagnostic signals in prior work.

We find that, although the three diagnostic signals demonstrate strong internal validity (reproducing what was established by the original works), they perform poorly on the clinically diagnosed patient data, thus suffering from poor external validity. Among the three signals, we find that the model that uses affiliation behavior as gold standard leads to the poorest performance. Our results also reveal that incorporating clinical judgment via appraisal of social media self-reports of mental illnesses leads to the best performance, among the proxy signals, when tested on clinical patient data. However, we find that all classifiers trained with the proxy diagnostic signals perform significantly poorly when compared to a model built using the data of the schizophrenia patient population. A deep dive in the performance of these classifiers via an error analysis reveals several methodological gaps in the way these diagnostic signals are conceived and employed in the predictive frameworks. These gaps range from uncertainties in the construct validity of the proxy signals, and poor theoretical grounding, to a variety of population and data sampling biases.

Our findings provide remedial guidelines for researchers engaging with prediction of mental health states from social media data, and for clinicians and practitioners interested in incorporating such machine learning based diagnostic assessments in clinical decision-making.

2 BACKGROUND

An Overview of State-of-the-Art

Serving as a mechanism to understand people's psyches and lives in a timely fashion, and spanning large, diverse populations, social media has emerged as an important tool in mental health. These uses, although not all encompassing, have included meeting a variety of social, technical, public health, and clinical goals. Research investigations have ranged from inferring risk to various mental illnesses [24, 29, 31, 34]—the largest body of work in this area; comparison of online and offline mental health behaviors [84]; understanding self-disclosure practices and goals [2, 39]; deciphering social support provisions to promote positive mental health outcomes [3, 33, 92]; discovering community norms and behaviors [18]; and exploring how these platforms can support intervention delivery [48]. Most relevant to the current paper is the line of research focusing on predicting mental health states and diagnoses from social media, which encompasses studies targeting different conditions [18, 22, 65], platforms [18, 32, 94, 107, 111], and disciplines [11, 14, 21, 108].

Burgeoning interest in this topic stems from the fact that social media data is readily available and archived, and can be unobtrusively gathered with low effort and cost [52]. These unique attributes help overcome many challenges in state-of-the-art clinical assessment of mental health that involves subjective recollection of historical facts—a method prone to retrospective recall bias [56]. However, appropriating social media data to inform clinical efforts around early diagnosis, tailoring treatment, or delivering interventions, suffers significant limitations. In a clinical setting, diagnostic information is available to the clinician via self-reported psycho-social signs and symptoms, theoretically and psychometrically validated clinical scales, interviews, questionnaires, and other diagnostic tools [1]. Social media data by itself, however, does not include such clinically validated signals to accurately identify and validate individuals' mental health states. Also, collecting clinically valid diagnostic signals from social media would require engagement with an at-risk patient population, a cohort that is stigmatized, sensitive, and vulnerable. This presents logistical challenges to identification of diagnostic signals, as well as privacy and data protection issues. Such a data collection approach can be difficult to scale, and is effort- and time-consuming, requiring carefully crafted clinical and risk management protocols, and involvement of clinical experts.

To circumvent these challenges, researchers have employed several online behaviors as gold standard information, or what we call *proxy diagnostic signals* to identify individuals' mental illness diagnoses. Through a systematic literature review [58] based on a keyword search of papers on predicting mental health states from social media, we identifiedthree types of proxy diagnostic signals from the literature, which we elaborate below. Table 1 gives a taxonomy of prior works in this area, from the perspective of the proxy diagnostic signals they use.

Proxy Diagnostic Signals in the Literature

Affiliation Behaviors: A first category of research represents behaviors signaling engagement or association (via hashtags, account following, community membership) with content related to mental health resources on social media, as proxy diagnostic signals of an illness [50, 63, 111]. A prominent example is McManus et al. [63] who used following a Twitter account (@schizotribe) dedicated to conversations around lived experiences of schizophrenia as a signal for gold standard information that an individual might be suffering from schizophrenia. A complementary set of papers have operationalized membership in online mental health support communities such as Reddit and Livejournal as proxies for diagnostic information [41, 68, 93, 94].

Self-reports: Next, the most popular form of proxy diagnostic signals, this category operationalizes first-hand, public self-disclosures of diagnosis of a mental illness as indicators of a clinical mental illness [8, 15, 21–23, 25, 32, 46, 47, 59–61, 65, 69, 76, 80, 93, 97, 105–107]. A notable example, Mitchell et. al. [65] used regular expression search queries on Twitter ("I have been diagnosed with schizophrenia") to extract self-reports of schizophrenia diagnoses and then employed them for predicting their presence/absence.

External validation: Finally, this category represents human-in-the-loop, collaborative approaches that either seek self-reported information from the individual, or incorporate diagnostic scales and/or expert appraisal for identification of the proxy diagnostic signals [11, 14, 17, 32, 71, 78, 79, 81, 84, 90, 103]. Most relevantly, Birnbaum et al. [10] incorporated clinical appraisals on self-reports of schizophrenia on Twitter to build machine learning models of diagnoses.

Using Proxy Diagnostic Signals: Critical Challenges

Appropriating these proxy diagnostic signals has overcome many challenges and barriers to gathering clinically valid diagnostic data on social media, particularly around scale and size [21], and these approaches continue to gain traction in the community. However they suffer from significant limitations, which we frame below, drawing upon the critical data literature [13, 51, 55].

Consider the case when affiliation to mental health resources is considered a proxy of a diagnosis. Alongside including genuine patients, it likely also includes other stakeholders like mental health practitioners and experts, non-profits raising awareness campaigns, caregivers etc. As another example, although the act of self-disclosing a mental illness can be an indicator of a person's mental condition, there are gaps in

	Affiliation Data		Self-report Data		Appraised Self-report Data		Gold Standard Patient Data	
	Target Class	Control Class	Target Class	Control Class	Target Class	Control Class	Target Class	Control Class
Total #users	861	539	412	345	153	107	88	55
Total #posts	1,417,688	2,145,319	1,724,237	1,083,790	663,428	233,253	9,821,938	4,958,793
Avg #posts	1646.56	3980.18	4185.04	3141.42	4336.13	2179.93	111,612.93	90159.87
Median #posts	320	1113.0	1682	830	1376	737	28554.5	21178.0

Table 2: Descriptive statistics for the proxy diagnostic signal datasets and their corresponding matched controls.

understanding what an individual chooses to self-report, why, and when they decide to do so, or if they are being truthful.

In other words, there is lack of evidence that these proxy signals are accurately measuring what they intend to measure, also known as construct validity [70] (whether the signals accurately identify and represent individuals at-risk). A lack of contextualization in psychiatric practice [57] or theory [7] additionally reduces confidence in their construct validity—an issue recognized in prior critiques of big data approaches [13, 55]. Although proxy signals with expert validation attempt to tackle some of these theoretical and clinical gaps, because the approach is removed from direct interaction with the individual, their veracity can be questioned, and their "*claims to objectivity and accuracy can be misleading*" [13].

Further, individuals with unique attributes, attitudes, and characteristics, possibly distinct from patient populations, are likely to engage in the specific types of behaviors enumerated by the proxy signals. Apart from the inclusion of "noisy" data, the unique ways in which the proxy diagnostic signals are defined and construed can lead to a variety of biases in the predictions, despite the impressive sample sizes they promise. This resonates with what boyd and Crawford noted, that "*bigger data are not always better data* [13]" and what Olteanu et al. discuss at length surrounding methodological pitfalls of big data [70]. In this work, we systematically examine these methodological gaps in the validity of these proxy diagnostic signals, and explore how validity issues impact their potential application in clinical decision-making.

3 DATA

We use public and non-public data (gathered using appropriate protocols) from two prominent social media sites, Twitter and Facebook, for the purposes of this paper. We begin by introducing four datasets used in this paper, followed by a description of how they were collected.

Gathering Proxy Diagnostic Signal Data

The first three datasets correspond to the three proxy diagnostic signals we adopt based on the topical focus and the existing literature, and which were introduced above. We consider them as proxies (or "proxy datasets") of schizophrenia diagnoses in individuals.

Affiliation Data. Our first dataset is motivated from prior literature that used behaviors signaling affiliation (e.g. following, hashtag usage) to mental health resources, related to schizophrenia, as diagnostic information. Adopting the approach of McManus et al. [63] ($N = 96$), we used a Twitter account named @sardaa (Schizophrenia and Related Disorders Alliance of America), a support organization for people with schizophrenia and their caregivers, as our starting point to build this affiliation dataset. As operationalized by McManus et al. [63] and following verification of the account's trustworthiness with our clinical coauthors, we considered all followers of the account @sardaa as individuals with a schizophrenia diagnosis. Using the official Twitter API, we obtained the list of all followers of @sardaa ($N = 1847$) and consistent with McManus et al. [63] collected their timeline data for the year 2014. We also collected profile information of these individuals including number of posts, chosen language on Twitter (filtering for English), number of followers and number of followees, leading to a final sample of 861 Twitter users. Descriptive statistics of this data are reported in Table 2 and Figure 1(a).

Self-report Data. For the second dataset, we adopt the proxy diagnostic signal of mental illness self-reports utilized in many prior works (e.g., (most prominently [65]), introduced in the previous section. Per Mitchell et al.'s approach [65] ($N = 174$), we used a list of key phrases developed in Ernala et al. [10] to identify self-reports of schizophrenia on Twitter from 2014. Following manual filtering to remove noisy examples, without loss of generality, we collected the historic timeline data of all authors of these self-reports ($N = 412$). We also collected the same metadata information as above, such as, total number of posts, chosen language on Twitter (filtering for English), total number of followers and total number of followees/friends. Descriptive statistics are reported in Table 2 and Figure 1(b).

Clinically Appraised Self-report Data. Our third proxy dataset is inspired from the third body of work that used external expert appraisals on social media data to obtain diagnostic signals of mental illnesses. Following Birnbaum et al.'s approach [11] ($N = 146$), we began with a sample of 635 individuals who had self-reported their diagnosis of schizophrenia on Twitter in 2014. A sample of each individual's timeline, consisting of the self-report post, 10 preceding and 10 succeeding

posts was then passed to two clinical coauthors[1] for appraisal. The experts annotated each self-report sample on a three point scale: *genuine*, *noisy* and *maybe* categories and achieved high inter-rater agreement (Cohen's $\kappa = 0.81$ between genuine and noisy). This third type of diagnostic signal provided Twitter data of 153 individuals whose self-reports were clinically appraised to be genuine. As before, we collected all metadata associated with their Twitter profiles, descriptive statistics of which are given in Table 2 and Figure 1(c).

Matched Control Data

The predictive task of identifying individuals with schizophrenia necessitates comparisons to matched control Twitter users who do not provide an equivalent proxy diagnostic signal. Accordingly, we used the Twitter streaming API to obtain a random sample of public posts and extracted their authors. Then, we gathered their timeline data for 2014 and profile information ($N = 640$). We filtered out any individuals who had mentions of schizophrenia in their posts.

Then, we adopted a statistical matching approach [83] to ensure that the control users and the individuals in each of our proxy datasets are comparable by trait attributes. Since social media behaviors are a reliable indicator of people's personality, psychological states, and even demographic attributes [91], we included the following covariates for the purpose of matching: total number of statuses, chosen language on Twitter, total number of followers and total number of followees. Through an iterative k-nearest-neighbor matching (k=1-15) based on the well validated Mahalanobis distance metric [82, 87], we compared the covariates of each individual's Twitter content in each proxy dataset (affiliation, self-report, appraised self-report) with that of each of the control users obtained above, and identified a set of most similar control users based on a heuristically chosen distance threshold. For the affiliation dataset, we obtained a matched control sample of 539 users. We obtained 345 and 107 matched controls for the self-report and the clinically appraised self-report datasets respectively. The descriptive statistics of these matched controls are given in Table 2.

Schizophrenia Patient Data and Healthy Controls

As the fourth dataset, we include social media data of patients clinically diagnosed with schizophrenia and that of clinically verified healthy controls, based on a clinical examination or DSM-5 [4] criteria. This data was collected as a part of a research study involving the paper's authors, aimed at identifying technology-based health information to provide early identification, intervention and treatment to young adults

[1]The experts are clinical psychiatrists working at a large psychiatry hospital in [blinded city], with extensive expertise providing treatment and counseling to individuals with early stage schizophrenia.

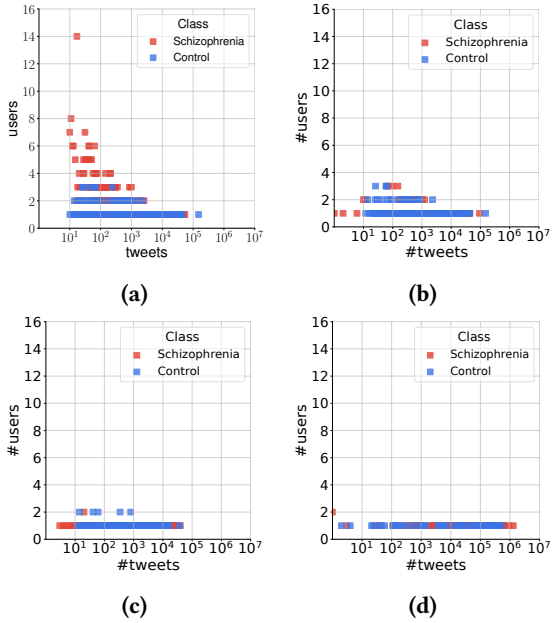


Figure 1: Distribution of #users and #posts for the target and control classes per dataset: (a) Affiliation Data, (b) Self-report Data, (c) Appraised Self-report Data, (d) Patient Data.

with schizophrenia. The research protocol was approved by the Institutional Review Board (IRB) of the coordinating institutions as well as local IRBs at participating sites. Individuals between 15 and 35 years old were recruited from various inpatient and outpatient psychiatric departments at the coordinating and its partner institutions. Participants were eligible if they had a primary psychotic disorder like schizophrenia, based on clinical assessment scales (e.g., the Psychiatric Diagnostic Screening Questionnaire or PDSQ [112]) as well as a formal clinical examination facilitated by Structured Clinical Interview for DSM-5, or SCID [98]—we note that all of these diagnostic tools are backed by sound theoretical underpinnings. Healthy controls who had already been screened for psychiatric disorders and consented to prior studies were also recruited. All participants were asked to request, extract, and share entire archives of their Facebook and Twitter data.

The consented participants included 88 patients who had been diagnosed with schizophrenia. Of these 88, 73 participants consented to provide their Facebook data, whereas 15 provided their Twitter data.Additionally, 55 healthy controls were recruited through the study, out of which 32 provided their Facebook data and 23 participants provided their Twitter data. We use all linguistic content from participants' Facebook and Twitter archives i.e. status updates and comments made on Facebook, and posts shared on Twitter.

As Twitter was the primary data source for all of the proxy diagnostic signals, we conjectured that the small sample of patients and healthy controls with Twitter data ($N = 38$) could pose a challenge to building robust machine learning models.

To combat this issue and to leverage the larger sample sizes with Facebook data, we conducted linguistic equivalence tests between the two data sources, a known approach in the transfer learning literature [45]. As language on Twitter cannot be directly compared to that on Facebook due to the affordances of the platforms [6], establishing linguistic equivalence was a crucial step before both data sources could be combined in building the patient and healthy control datasets.

To test for linguistic equivalence, we use semantic similarity calculated using distributed word vector representations of all linguistic content contained in the Facebook and Twitter archives [75]. Cosine similarity of word vectors is often used to quantify the linguistic similarity between two datasets, and a high value indicated that the content in the two datasets was linguistically equivalent [75]. We found a high cosine similarity between the vector representations of the Facebook and Twitter data across both the schizophrenia patient (=0.98) and healthy control population (=0.84), showing the two data sources to be linguistically equivalent. Thus, our final dataset comprised either Twitter or Facebook archives of 88 schizophrenia patients and 55 healthy controls. The descriptive statistics for the combined patient and healthy control dataset are reported in Table 2. Of the total number of posts, 99% came from the Facebook archives of the patients and the healthy controls, and the remaining 1% were sourced from their Twitter archives.

4 METHODOLOGY

Rationale and Overview

We adopt quantitative data triangulation as our methodological framework. Triangulation is an evaluation approach that uses multiple or heterogeneous methods, or data sources compiled via varied mechanisms, to develop a comprehensive understanding of a phenomenon, or to elucidate its complementary aspects [73]. Specifically, this approach is used to confirm the results of a research, and provide external validation to existing findings [35]. In essence, triangulation is an attempt to map out, or explain more fully, the richness and complexity of human behavior by studying it from more than one standpoint. Using this approach, we assess the efficacy of the three proxy diagnostic signals in identifying diagnoses of individuals with schizophrenia, both within their corresponding proxy datasets, as well in the data of schizophrenia patients. This way, we seek to establish their internal and external validity respectively. Figure 2 gives an overview of our approach.

Classification Framework

We set up a binary classification task to distinguish between individuals with schizophrenia identified by each proxy dataset and its corresponding matched controls. We built four models: three based on the proxy datasets denoted as the *Affiliation*, *Self-report* and *Appraised Self-report Models* and one on the clinically validated patient data known as the *Patient Model*.

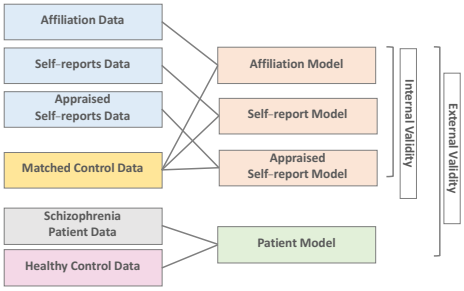


Figure 2: Schematic diagram of our proposed methodology.

Preparing Training and Validation Data: We use the proxy datasets and their corresponding matched control data in their entirety for training and validating the above proxy classifiers. For the *Affiliation Model*, the positive examples (Class 1) comprised the Twitter data of the 861 users while the negative examples (Class 0) consisted of the 539 matched control users. The positive examples for the *Self-report* and *Appraised Self-report Models* spanned the data of 412 and 153 users respectively, while the corresponding negative examples included the Twitter data of 345 and 107 matched controls. For the *Patient Model*, we selected a random sample of 80% of the patient dataset for model training and validation, resulting in 68 patients with schizophrenia in the positive class, and 46 healthy control participants forming the negative class.

Preparing Unseen Test Data: We incorporated the held-out 20% patient data as an unseen test dataset, that could be consistently used across all models (*Affiliation, Self-report, Appraised Self-report* and *Patient*) for triangulation. This comprised 20 patients with schizophrenia and 9 healthy controls.

Features: Linguistic features from text data have been widely adopted and are known to be largely successful in predicting mental health states using social media data [20, 31]. A rich body of literature in psycholinguistics has identified the association of linguistic usage to emotion and behavior, including mental health states of individuals [74]. We adopt two forms of linguistic content as features for classification. First, we build a term-frequency, inverse document-frequency based language model using the most frequent 500 n-grams (n=1-3) from the preprocessed data upon removal of stop words and URLs. Second, we use three categories of psycholinguistic measures: (1) Affective attributes, (2) Cognitive attributes and (3) Linguistic style attributes—from the well-validated psycholinguistic lexicon Linguistic Inquiry and Word Count (LIWC) [20]. Combining the two feature sets together, our overall feature space included 550 numeric features.

We built classifiers for each proxy dataset to predict individuals with schizophrenia on social media from matched control

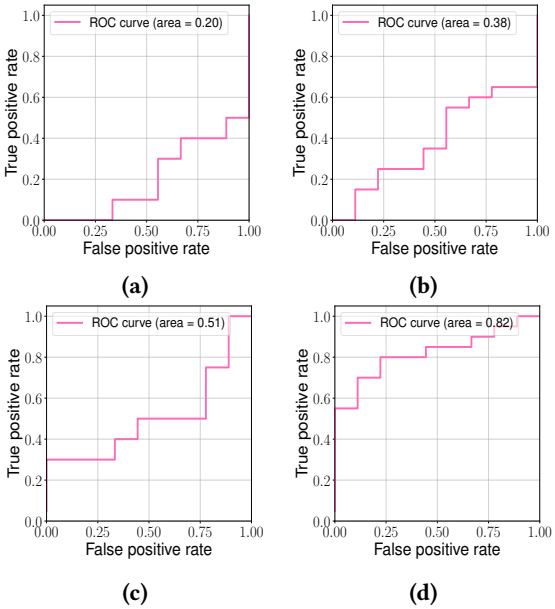


Figure 3: ROC (Receiver Operating Characteristic) curves per classifier (a) *Affiliation Model*, **(b)** *Self-report Model*, **(c)** *Appraised Self-report Model*, **(d)** *Patient Model*.

users. To remove correlated features and to improve the predictive power of the model, we employed feature selection methods [43], eliminating noisy features and identifying the most salient variables in predicting the outcome. Specifically, we use the filter method where features are selected on the basis of their scores in statistical tests for their correlation with the outcome variable. Adopting the ANOVA F-test we reduced the feature space from 550 features to k-best features per classifier.

We experimented with non-linear and ensemble classification algorithms such as Support Vector Machines, Random Forest, and Logistic Regression [36]. For each classifier, we test its performance in two steps: First, for parameter tuning and assessing internal validity, we used stratified k-fold cross validation. We varied model parameters for all classification approaches during the validation step to find the best performing model. Second, choosing this best performing model from the validation step, we evaluated its performance on the unseen test data for external validity. Across the four classifiers, for relative comparison, we report model performance using a variety of metrics: Receiver Operating Characteristic Area Under Curve (ROC AUC), accuracy and F1 scores.

5 RESULTS

Internal Validity

We present in Table 3 the cross validation performance of the four classifiers in distinguishing individuals with schizophrenia from matched controls. Overall, the *Affiliation Model* outperforms the other classifiers with the highest accuracy (Best: 0.94, Mean: 0.88, std: 0.02) and F1 (Best: 0.95, Mean: 0.91, std: 0.02) and a 27% improvement in accuracy over a ZeroR baseline (Accuracy: 0.61). Upon feature selection to top 450 features, a penalized logistic regression classifier led to high model stability. The reported accuracy of this model is close to McManus et al. [63], demonstrating that the trained model can infer distinct patterns between the two classes.

Although both *Self-report* and *Appraised Self-report* models improve over their ZeroR baseline (accuracy: 0.54, 0.44 respectively), the *Appraised Self-report Model* performs better (Best: 0.88, Mean: 0.80, std: 0.03) than the *Self-report Model* (Mean: 0.72, Best:0.79, std: 0.02) across all metrics. The ROC AUC for the *Self-report* and *Appraised Self-report Model* as reported in Table 3 are 0.80 and 0.85 respectively. A penalized, logistic regression classifier again performed best on both of these datasets, based on the K-best features (=350 respectively) that we select for downstream testing.

Comparing the performance of the proxy classifiers on their respective validation sets, we find that the *Appraised Self-report Model* has higher precision than the *Self-report Model*. This was also observed by Birnbaum et al. [10]; the clinician annotation task eliminated inauthentic noisy samples leading to a high precision sample of genuine self-reports. This reveals that contextual cues picked by the experts, such as mentions of medication, mood stability, symptomatic expression provide a strong validation of the self-reports. Since both datasets were sampled from self-reports of schizophrenia in 2014, we conjecture that incorporating clinical appraisals improved performance by eliminating false positives, or ambiguous self-reports from the positive class.

Finally, the *Patient Model* trained on patient data performs modestly, although better, compared to the proxy classifiers, with average accuracy of 0.72 (best: 0.75) and average F1 score of 0.77 (best:0.79) across 5-fold cross validation[2].

External Validity

Next, to examine their external validity on unseen patient test data, we present the performance of the proxy classifiers. Figure 3 (a-c) presents the ROC plots, per proxy classifier, showing the trade-off between true positive rate (sensitivity) against the false positive rate (1-specificity).

Among the three proxy classifiers, the *Affiliation Model* shows poor external validity with the lowest accuracy (0.21), the lowest F1 (0.14), and the lowest AUC (0.2) on the 20% sample of unseen patient data (refer Table 3). The next best performing model is the *Self-report Model* outperforming the *Affiliation Model* with a 27% improvement in the overall accuracy (0.48), 47% improvement in F1 score (0.61) and 18%

[2]Given the relatively small sample sizes, to check for overfitting, we examined model stability through the standard deviation of evaluation metrics across folds. A low standard deviation of 0.02 indicated that despite low sample sizes, the model had stable performance.

	Class 1	Class 0	Cross validation					Testing				
			P	R	f1	Acc	AUC	P	R	f1	Acc	AUC
Affiliation Model	861	539	0.89	0.94	**0.91**	**0.89**	**0.95**	0.28	0.1	0.15	0.21	0.20
Self-report Model	412	345	0.72	0.81	0.76	0.72	0.80	0.63	0.6	0.61	0.48	0.38
Appraised Self-report Model	153	107	0.81	0.88	0.84	0.80	0.85	0.65	0.75	0.70	0.55	0.51
Patient Model	68	46	0.76	0.80	0.77	0.72	0.76	0.93	0.7	**0.8**	**0.76**	**0.82**

Table 3: Average model performance on the validation and unseen test datasets.

improvement (0.38) in the ROC AUC. Although this indicates that self-reports might be a better diagnostic signal than affiliation, the performance of this classifier is still weak compared to its performance during the validation step (test of internal validity). Lastly, among the three proxy classifiers, we see the strongest external validity or best performance for the *Appraised Self-report Model*. This classifier shows a 9% and 55% improvement in F1 (0.70), and 7% and 34% improvement in accuracy (0.55) over the *Self-report* and *Affiliation Model* respectively. Although the *Appraised Self-report Model* demonstrates the strongest external validity so far, there is substantive decrease in its performance compared to the validation phase.

Summarily, testing the proxy classifiers on unseen patient data revealed poor external validity and that relative performance between the validation and testing steps was not preserved when tested in a clinical setting.

Comparison of Classifiers on Unseen Patient Data

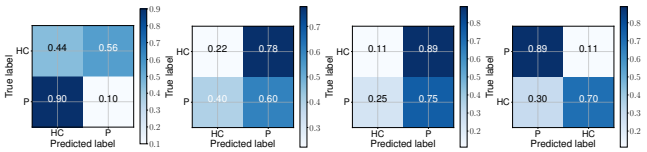
Triangulating the three proxy datasets corresponding to their diagnostic signals, we compare their predictive performance with the *Patient Model*, again trained on the 20% sample of unseen patient test data. Through this, we establish an empirical estimate of the error incorporated by using the proxy classifiers, when applied on patient populations.

First, we report the performance of the *Patient Model*. From Table 3, we see that this model outperforms the proxy classifiers, in distinguishing healthy controls from schizophrenia patients, giving lower false positives and false negatives. We also find that this is a highly precise model (precision: 0.93), correctly predicting schizophrenia patients as the positive class. The performance, however, is affected by low recall, and we find lower precision for the negative class due to the false negatives (=6) wherein schizophrenia patients are wrongly predicted as healthy controls.

We use the performance of the *Patient Model* as gold standard and examine the error incorporated by each of the proxy classifiers. We use F1 and ROC AUC to situate these differences. We note the highest difference in performance exists between the *Patient Model* and the *Affiliation Model*. The *Patient Model* outperforms the *Affiliation Model* by 65% in F1 and 62% in AUC. Comparing the *Patient Model* with the *Self-report Model*, we observe a 19% and 44% gain in F1 and ROC AUC respectively. This indicates that the online behavior of self-reporting a mental illness diagnoses might be a better diagnostic signal than the affiliation behavior. Finally, the *Appraised Self-report Model* shows least difference in performance when compared to the *Patient Model* with 10% and 31% difference in F1 and AUC respectively. This indicates that when using self-reports as a diagnostic signal, clinical appraisal leads to better predictions. In short, the triangulation step reveals variability in predictive performances of the proxy diagnostic signals when tested on unseen patient data, demonstrating trade-offs when proxy signals are used for predicting clinical mental health states, versus when information is gathered directly from patients.

Deep Dive into Performance of Proxy Classifiers

To evaluate beyond performance metrics and to reason about the poor external validity of the proxy classifiers, we present a deeper analysis of the proxy classifiers' performance.



(a) **(b)** **(c)** **(d)**

Figure 4: Confusion matrix per classifier (a) *Affiliation Model*, **(b)** *Self-report Model*, **(c)** *Appraised Self-report Model*, **(d)** *Patient Model*. **Here HC: Healthy controls (Class 0); P: patients with schizophrenia (Class 1).**

Error Analysis. We begin by unpacking mismatches in predictions made by the proxy classifiers on unseen patient data, in terms of example false positives and false negatives.

Unpacking false positive classifications: Consider an example X who is a healthy control, per a clinically validated diagnostic assessment. But, the *Affiliation Model* wrongly predicted them as having schizophrenia. Examining their social media timeline, we find (paraphrased) posts including excerpts such as, "mental screenshot of notes", "are you bad for my mental health" and "use my phone in day mode because I am mentally ill". We note that terms like 'mental' ($\beta = 2.17$), 'health' ($\beta = 1.44$), 'illness' ($\beta = 1.45$) in these excerpts are highly predictive of the positive class in the *Affiliation Model*, leading to

a misclassification of X as a schizophrenia patient. Moreover, because the *Affiliation Model* simply measures engagement, association with, or interest in mental health content and resources, it missed capturing the *context* in which these topics were discussed by X, leading to a misclassification of X as a schizophrenia patient. Now consider a healthy control participant Y's timeline. It includes prolific usage of terms such as 'creepy' (β = 0.241), 'hell' (β = 0.096), 'jesus' (β = 0.091), and 'help' (β = 0.401). These tokens are learned as highly predictive of the positive class by the *Appraised Self-report Model*, thereby leading to a misclassification of Y. Although these tokens reveal symptomatic expression, spirituality and support-seeking behaviors, notable in schizophrenia disclosures made on social media [39], the current example demonstrates varied usage of these tokens by healthy controls, in reference to pop-culture or in casual conversations. We frame these observations as the following methodological gaps: that the outcomes yielded by the proxy classifiers are not valid indicators of a clinical diagnosis of schizophrenia (poor construct validity); and that the behaviors of individuals captured by the proxy signals might not be representative of the behaviors of schizophrenia patients (sampling bias).

Unpacking false negative classifications: Consider a different example A, a clinically diagnosed patient with schizophrenia. Their social media timeline data shows extensive usage of swear terms such as 'fuck' ($\beta = -0.94$), 'ass' ($\beta = -0.63$), 'bitch' ($\beta = -0.67$) that according to the *Affiliation Model* were highly predictive of the negative class, resulting in a false negative classification. Consider example B, a schizophrenia patient whose timeline largely consisted of travel and hobbies related posts with no evidence of schizophrenia experiences. The *Appraised Self-report Model* predicted B as a healthy control, due to lack of explicit disclosures of the illness, like symptomatic expressions and personal struggles (feature importance for LIWC categories: anger (0;0.03), body (0;0.06), swear (0;0.05) anxiety (0;0.03)). These differences reveal that the proxy signals are not measuring what they intend to measure (poor construct validity). Further, that the social media language of patients might not be very different from control users (population bias).

Issues of Dataset Shift & Bias. The population and sampling biases revealed by our error analysis goes on to show that the statistical data distributions might be drastically different between the proxy datasets and the actual patient dataset—a phenomenon referred to as "dataset shift" [99]. As a next step in our deep dive, we present the following analysis to systematically examine this dataset shift and assess its effects. Specifically, to quantify dataset shift, we adopt a measure of semantic distance computation between the linguistic content of proxy and patient datasets [45]. To represent the proxy data distributions, we first identify the most frequent 500 *n*-grams from the positive class, per proxy classifier, and compute the word vector representation [64] for each of these *n*-grams. Similarly, we represent the positive class for the test distribution i.e. the data of all schizophrenia patients in the word vector space. We finally compute the cosine similarity between the proxy and patient data in the vector space. Our results bolster the findings of the error analysis, wherein we observe the farthest distance between the proxy and patient data in case of the affiliation dataset (similarity: 0.907, distance:0.092). The self-report dataset is at a closer semantic distance to the patient data distribution than the affiliation data, with a distance of 0.019 and similarity of 0.980. Finally, confirming the observations thus far, the appraised self-report dataset appears at the closest distance to the patient data with a distance of 0.017 and similarity of 0.982.

Affiliation	β	*Appraised*	β	*Patient*	β
i'm	-0.825	*NegAffect*	0.063	*cog mech*	-0.003
stigma	0.665	negation	0.074	*present*	-0.002
mhchat	0.696	*present*	0.40	*body*	-0.002
body	0.729	help	0.401	*verbs*	-0.002
bipolar	0.774	thought	0.41	*social*	-0.002
work	0.919	i'm	0.44	*aux verbs*	-0.002
self	0.961	die	0.45	help	0.0002
social	1.109	alone	0.45	feeling	0.001
care	1.111	hard	0.457	i'm	0.002
depression	1.116	cry	0.50	gonna	0.002
suicide	1.133	*body*	0.52	angel	0.002
thanks	1.445	feeling	0.523	burning	0.002
illness	1.447	*verbs*	0.58	pray	0.003
help	1.632	sorry	0.662	lifetime	0.005
mental health	1.866	gonna	0.63	attack	0.006

Table 4: Comparing the top features across the Affiliation, Appraised self-report and Patient Model. β weights (significant at the $p = 0.05$ level) denote feature importance. LIWC categories are presented in italics.

Issues of Construct Validity. A second issue revealed by our error analysis was that the behavioral patterns learned by the proxy classifiers were absent in the schizophrenia patient population, raising concerns around construct validity. Therefore, next, we examine the features learned by the proxy classifiers in comparison to the features learned by the *Patient Model*. Table 4 shows the top features, and their feature weights for the worst and best proxy classifiers, and the *Patient Model*.

Overlap of features: Comparing the top features of the *Affiliation Model* with the *Patient Model*, we see little overlap between the two feature spaces, prominently, in terms of use of first person pronouns and LIWC category terms about 'social' and 'body'. We find that these features are predictive of one class in the *Affiliation Model*, whereas predictive of the opposite class in the *Patient Model*. Further comparing

the top features of the *Appraised Self-report Model* with the *Patient Model*, we see a higher overlap than in the case of the *Affiliation Model*. Some of these features such as 'feeling', 'help' and use of first person pronouns are predictive of the positive class in both models, which explains the higher external validity of the *Appraised Self-report Model*. However, we find that there are also a number of features predictive of the positive class in the *Appraised Self-report Model*, that are associated with the negative class in the *Patient Model*, e.g., LIWC categories of present tense, body, verbs and tokens such as 'crazy'. Although the *Appraised Self-report Model* is accurately learning certain patterns specific to the patient population, it misconstrues explicit mental illness disclosure behaviors (symptomatic expressions, combating stigma, and support seeking) as signals of a schizophrenia diagnosis.

Mismatch of features: Finally, we observe that the most predictive features (of the positive class) in the *Affiliation Model* are explicit signals of mental health care and support ('mental health', 'illness', 'depression', 'stigma', 'mhchat'), that have few occurrences in the patient data. Similarly, in the case of the *Appraised Self-report Model*, content related to schizophrenia experiences ('die', 'alone', 'sorry', 'creepy', LIWC categories of negative affect and negation) are either missing or not predictive of the positive class in the *Patient Model*. Therefore, we argue that what these proxy classifiers actually learn is the language use of individuals actively opening up about schizophrenia experiences, seeking informational and emotional support on Twitter. In comparison, our patient population does not exhibit such disclosure or support seeking behaviors on social media.

6 DISCUSSION

In this paper, we presented the first insights into some methodological gaps that exist in using social media derived diagnostic signals for predicting clinical mental health states. We found a lack of external validity when the prediction models developed using the proxy signals were tested on actual patient data. Our triangulation approach further surfaced issues of construct validity, limited theoretical underpinning, and population and sampling biases that permeate in the prediction task, through these diagnostic signals. We discuss the methodological and clinical implications of these findings.

Methodological Implications

Uncertainty in Construct Validity. A first notable limitation of the proxy diagnostic signals we observed is the *uncertainty in their construct validity*. Drawing on the definition of this construct, we explore two methodological implications:

1) *Do these diagnostic signals measure what they claim to measure?* Our results show that the diagnostic signals are not measuring what they claim i.e. the clinical diagnosis of an individual's mental health (schizophrenia) state. This is revealed by the considerable mismatch we observed while comparing the top predictive features of the proxy classifiers and those of the *Patient Model*. Unpacking the context of these features in the actual social media posts, we found that they capture support seeking behaviors, interest in others' lived experiences of the illness, self-reported accounts of stigma and inhibition—patterns absent from the features of the *Patient Model* from the clinical schizophrenia population.

2) *Is what is being measured by a diagnostic signal itself valid?* To the latter point about construct validity, we found a lack of clinical grounding in the diagnostic information (individual's clinical mental health state) that these signals intend to measure. Instead, what these signals presume as diagnostic information are essentially behavioral patterns associated with the appropriation of social media by a wide variety of stakeholders, not necessarily patients, in relation to the illness. These forms of appropriation include individuals posting resources for mental health awareness, individuals seeking therapeutics benefits, or individuals breaking free inhibitions and mental health stigma by disclosing their illness. Although these appropriation patterns can be a valuable resource to understand the experiences of schizophrenia [86], they do not provide clinically grounded information about an individual's diagnosis of a mental illness—thereby making them less suitable for the prediction tasks in this paper.

Although the appraised self-report diagnostic signal attempts to overcome lack of clinical grounding, it suffers from other limitations that affect its construct validity. First of all, the clinical experts who appraised the self-reports of schizophrenia did not have access to the person's clinical history, or symptoms and experiences. Collateral information—information beyond a patient's explicit self-reports of symptoms [40]—are also critical and mainstream in any clinical mental illness diagnosis, and our clinical diagnostic signal derived from the patients factors this information through use of tools like PSDQ [112] and SCID [98]. However, for the appraised diagnostic signal, the clinicians can only gather collateral information from the content of the social media posts. This may not be sufficient for a valid diagnosis, given the limited context of what people consciously or subconsciously choose to share on social media, and vast amounts of collateral information might exist offline, which the appraising clinicians did not have access to.

Theoretical Contextualization. Related to the above two issues lies another limitation, which is *a lack of theoretical underpinning* in the ways the diagnostic signals were identified. All of the scales and questionnaires used for clinical

diagnosis, including the ones used in this paper's patient population, draw upon theoretical frameworks, such as neurobiology, dimensional personality assessment, behavioral science, psychodynamic, and cognitive theories [77]. They undergo rigorous psychometric testing and are continually adjusted as the frameworks around mental illnesses evolve, or as the DSM [4], or more recently the National Institute of Mental Health introduced Research Domain Criteria (RDoC) framework [49] offer newer guidelines for mental health diagnostic and treatment. The proxy diagnostic signals are, however, not inspired by this theory. Instead they focus on online behaviors, which may or may not align with theoretical models, frameworks, or guidelines of mental illnesses.

The other methodological gap we identify in the use of the proxy diagnostic signals for predicting clinical diagnoses relates to *dataset shift* [99]. In the literature, datasets shifts in supervised learning are attributed to population or data sampling biases inherent in the data [70]. We therefore discuss the foundations of this phenomenon in two ways:

Population Biases. We observed that the datasets built using the proxy diagnostic signals include social media data of a unique set of individuals, who may not be representative of schizophrenia patients who are actually diagnosed with the illness and under treatment. Consequently, this population bias may manifest in several different ways: 1) The social media activities of an individual who follows online mental health resources, may be different from someone who publicly discloses their illness and experiences—and these, in turn, might be different from a clinically diagnosed patient's social media usage and behaviors [12]; 2) The diagnostic signals capture subpopulations who may not be truthfully reporting their illnesses or may be reporting about their self-derived assessments of a mental illness experience in an exaggerated fashion, that did not involve the feedback of a clinician; and 3) The diagnostic signals consist of subpopulations who may not be mental illness patients currently under treatment, and the social media activities of those who are under formal care and those who are not, might be considerably different.

Data Sampling Biases. The observed dataset shift between the proxy datasets and the schizophrenia patient data may also be stemming from a type of sampling bias related to boundary regulation preferences of the individuals and the use of public versus private accounts. Individuals identified using these online diagnostic signals have largely public social media accounts; however, the clinically diagnosed patients we considered largely had private accounts. This difference in boundary regulation and privacy choices between the schizophrenia patients and the individuals captured via the diagnostic signals might lead to sampling biases in the proxy datasets.

Identifying and quantifying the biases between the populations targeted by the diagnostic signals, alongside examining their theoretical and construct validities is, therefore, crucial before the signals are deployed to make clinical predictions.

Clinical (Patient-Provider) Implications

Alongside the methodological implications of making predictions of mental illness diagnoses with the proxy diagnostic signals, it is equally important to consider their impact on the key stakeholders such as clinicians and patients.

To the clinician community, whose primary source of diagnostic information comprises clinically validated questionnaires, scales, interviews, and symptoms reported by the patient [1], these new forms of proxy diagnostic signals derived from social media, despite the right intentions, add complexities to the conventional psychiatric assessment method. We highlight some of these complexities in the questions below. For instance, in the absence of supplementary and accessible details of their inner workings and biases, how can clinicians trust these new forms of diagnostic signals and their validity, and thereafter act upon them? How do these new signals complement or even contradict clinicians' mental models of reasoning, or how clinicians pursue diagnosis and treatment of their patients?

Importantly, decision-making by the clinicians (for diagnosis, treatment, or patient-provider interventions) involves both high stakes and high costs. Therefore, incorrect predictions made as a result of data with poor external, construct validity, or those suffering from population and sampling biases can be dangerous and have serious consequences for the patients' well-being, and social and professional life. While personalized patient care is touted as a strong motivation for adopting social media for clinical diagnosis and treatment [66], validity and bias issues may additionally adversely impact patients trust and attitudes towards mental healthcare. When outcomes of these proxy classifiers are incorporated into clinical decisions without the patients' awareness, poor validity can even negatively impact patients' perceived agency in treatment, or the therapeutic relationship they share with their clinicians. These issues may further conflict with patients' preferences, needs, and values in treatment [100]. Thus bridging these methodological gaps with interactions with and involvement of the patient and clinician stakeholders is key to translating the potential of social media to support clinical diagnosis and treatment.

Remedial Guidelines: A Proposal

In the light of the above discussion, we suggest some guidelines for researchers to bolster efforts in examining and establishing the efficacy of social media based signals for prediction of mental health states in clinical populations.

□ *Improving Methodological Rigor and Adopting Alternative Research Designs.* A first set of guidelines center around

reducing or eliminating the issues noted above. We conjecture that combining multiple proxy diagnostic signals, especially those that are complementary to each other, could provide more rigor because of their potential to target more diverse social media populations. However, this warrants empirical investigation. Alternatively, given the stigma around experiences of mental illness [26], some of the proxy diagnostic signals can be leveraged in a respondent-driven sampling framework [44]. This can be a viable mechanism to reach and recruit individuals for clinical studies that seek to collect gold standard patient data. Implementing an online-offline framework [48], that combines social media data with pre-existing offline longitudinal information of comparable sub-populations, can also reduce the dataset shift challenges. Further, issues of dataset shift can be overcome by adopting recent approaches from the machine learning field, such as including importance weighting of training instances based on similarity to test set [95], and employing online learning of prediction models to identify and recover from incorrect predictions [16, 54]. Crowdsourcing based data analysis and replication efforts [38, 96] can also be used to make transparent the impact of proxy dataset biases on predictive models.

□ *Building and Utilizing Shared Infrastructures for Data Collection, and Data Donation Efforts.* One next guideline centers around building, contributing to, and leveraging shared infrastructures and data repositories for conducting this research. Our findings showed the value of using patient data in building predictive models of mental illness diagnosis. However, we recognize that researchers without access to patient populations within large healthcare systems, or without involved collaborations in the clinical field may be at an unfortunate disadvantage. Further, patient data collection can be complex, including technological and ethical dimensions, due to the need to engage with a vulnerable population and gather sensitive (largely non-public) information, that might include HIPAA [67] protected data. Open source, HIPAA compliant infrastructures with customizable data collection functionalities can be helpful to overcome some of these technical challenges. Participatory research efforts such as the Connected and Open Research Ethics (CORE) initiative [101] can be used to develop dynamic and relevant ethical practices to guide and navigate the social and ethical complexities of patient data collection. Initiatives focusing on voluntary data donation approaches, such as the notable OurDataHelps [24] program for suicide prevention research, can be utilized to gather high quality data about people's clinical mental health states, alongside their social media data.

□ *Harnessing Partnerships Between Computational and Clinical Researchers, and Patients.* Finally, this research area can benefit extensively from cross-disciplinary partnerships. Collecting patient data for building the predictive models involves human costs, and suffers from resource and logistical constraints. In working with sensitive population such as patients with mental illnesses, it is important to have appropriate clinical risk management protocols in place [37], especially when the source of data concerns social media activities of patients monitored in a near real-time fashion [109]. Computational researchers by themselves may not be best equipped to define or implement such protocols. Moreover, clinical expertise is needed to identify and navigate the right way and the right time to approach patients for informed consent regarding data sharing, and assess how it would impact their perceptions of clinical care. Partnership of computational researchers and clinicians throughout the research pipeline—e.g., right from establishing validity of measured online behaviors, providing appraisal of the data via qualitative coding tasks, to interpreting and situating large scale data analysis, can also improve rigor and eliminate issues of construct validity and improve theoretical grounding of the approach. Moreover, directly incorporating patients' feedback in the construction and acquisition of the clinical diagnostic signals will not only help represent their voices in the functioning of the predictive models and engage them as partners in treatment, but also support advancing the vision of participatory mental healthcare [89].

7 LIMITATIONS, FUTURE WORK & CONCLUSION

We acknowledge some limitations in this empirical study. We note caveats in our patient data, both in terms of its limited size, as well as in terms of limited diversity—the data was collected among patients suffering from a specific mental illness and seeking treatment within a single healthcare system, albeit one of the largest in the United States. We note here that, schizophrenia is known to manifest uniformly across demographic groups (gender, ethnicity, race) and geography [9] so we conjecture such biases to be minimal. Further, our examination of validity concerns and population biases in the proxy diagnostic signals is limited to English speaking, largely Western populations. The demographics therefore, are skewed and we caution against generalization. We also acknowledge that the observations we derived are limited to largely one social media platform, Twitter. Future work can extend and attempt to replicate these findings, both by focusing on other patient populations as well as additional illnesses and social media sites. Finally, we have considered only three proxy diagnostic signals in this paper, although they are amongst the most widely used in the community. Future work can also present additional investigations on alternative proxy signals and the potential of employing the use of multiple proxy signals in a concerted fashion.

Notwithstanding these limitations, there is a key takeaway in this paper: that *if the broader research agenda is to use social*

media data to inform clinical decision-making, such as early diagnosis, treatment or patient-provider interventions, using patient data to build machine learning models is imperative. That said, the goal of this paper is not to be dismissive of the immense potential that lies in this source of data. To quote Inkster and colleagues [48]: "*While acknowledging that issues are far from settled about the role that social media should play in mental health, we argue that it should no longer be a debate about whether researchers and healthcare providers engage with social networking sites, but rather how best to utilize this technology to promote positive change.*" Our remedial guidelines attempt to chart some of these possible ways of using social media data for positive change in mental health—by pairing offline and online patient data as well as by meaningfully involving researchers, clinicians, and patients, it is possible to extend social media based approaches for early diagnosis, treatment, and for developing novel patient-provider interventions.

In closing, we recognize that the use of proxy diagnostic signals is attractive because of scalability, as well as the low effort and minimal researcher, clinician, and patient burden they pose. However, for the limitations, both methodological and clinical, discovered in this paper, we suggest exercising caution and rigor going forward, and be cognizant of their implications for key stakeholders like clinicians and patients. Research in the HCI field has been instrumental in surfacing the many challenges of transplanting algorithms built with biased data, lacking theoretical and domain-specific validity, in real-world contexts [88, 104]. We believe the remedial guidelines we proposed will augment these efforts by starting conversations in the broader research community interested in leveraging social media data and machine learning predictive techniques to revolutionize mental health-care.

8 ACKNOWLEDGMENTS

We thank the members of the Social Dynamics and Well-Being Lab for their valuable feedback in preparing this manuscript. Ernala and De Choudhury were partly supported by a National Institutes of Health grant #R01GM112697.

REFERENCES

[1] Carlos Rejón Altable. 2012. Logic structure of clinical judgment and its relation to medical and psychiatric semiology. *Psychopathology* 45, 6 (2012), 344–351.

[2] Nazanin Andalibi, Oliver L Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding social media disclosures of sexual abuse through the lenses of support seeking and anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3906–3918.

[3] Nazanin Andalibi, Pinar Öztürk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of# Depression.. In *CSCW*. 1485–1500.

[4] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders, (DSM-5)*. American Psychiatric Pub.

[5] Mitja D Back, Juliane M Stopfer, Simine Vazire, Sam Gaddis, Stefan C Schmukle, Boris Egloff, and Samuel D Gosling. 2010. Facebook profiles reflect actual personality, not self-idealization. *Psychological science* (2010).

[6] Timothy Baldwin, Paul Cook, Marco Lui, Andrew MacKinlay, and Li Wang. 2013. How noisy social media text, how diffrnt social media sources?. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*. 356–364.

[7] Gregory Bateson, Don D Jackson, Jay Haley, and John Weakland. 1956. Toward a theory of schizophrenia. *Behavioral science* 1, 4 (1956), 251–264.

[8] Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538* (2017).

[9] Dinesh Bhugra. 2005. The global prevalence of schizophrenia. *PLoS medicine* 2, 5 (2005), e151.

[10] Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra Rizvi, Munmun De Choudhury, and John Kane. 2017. A Clinician-Machine Collaborative Approach to Identifying Social Media Markers of Schizophrenia. *Journal of medical Internet research* (2017). To appear.

[11] Michael L Birnbaum, Sindhu Kiranmai Ernala, Asra F Rizvi, Munmun De Choudhury, and John M Kane. 2017. A Collaborative Approach to Identifying Social Media Markers of Schizophrenia by Employing Machine Learning and Clinical Appraisals. *Journal of Medical Internet Research* 19, 8 (2017).

[12] Michael L Birnbaum, Asra F Rizvi, Christoph U Correll, John M Kane, and Jamie Confino. 2017. Role of social media and the I nternet in pathways to care for adolescents and young adults with psychotic disorders and non-psychotic mood disorders. *Early intervention in psychiatry* 11, 4 (2017), 290–295.

[13] danah boyd and Kate Crawford. 2012. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society* 15, 5 (2012), 662–679.

[14] Scott R Braithwaite, Christophe Giraud-Carrier, Josh West, Michael D Barnes, and Carl Lee Hanson. 2016. Validating Machine Learning Algorithms for Twitter Data Against Established Measures of Suicidality. *JMIR mental health* 3, 2 (2016), e21.

[15] Pete Burnap, Walter Colombo, and Jonathan Scourfield. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 75–84.

[16] Nicolo Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, learning, and games*. Cambridge university press.

[17] Stevie Chancellor, Yannis Kalantidis, Jessica A Pater, Munmun De Choudhury, and David A Shamma. 2017. Multimodal Classification of Moderated Online Pro-Eating Disorder Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 3213–3226.

[18] Stevie Chancellor, Zhiyuan Lin, Erica L Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and Predicting Mental Illness Severity in Online Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM, 1171–1184.

[19] Stevie Chancellor, Tanushree Mitra, and Munmun De Choudhury. 2016. Recovery amid pro-anorexia: Analysis of recovery in social media. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2111–2123.

[20] Cindy Chung and James W Pennebaker. 2007. The psychological functions of function words. *Social communication* (2007), 343–359.

[21] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in twitter. In *ACL Workshop on Computational Linguistics and Clinical Psychology*.

[22] Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. 2015. 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*. 1–10.

[23] Glen Coppersmith, Craig Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *International Conference on Weblogs and Social Media (ICWSM)*.

[24] Glen Coppersmith, Ryan Leary, Patrick Crutchley, and Alex Fine. 2018. Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights* 10 (2018), 1178222618792860.

[25] Glen Coppersmith, Ryan Leary, Eric Whyne, and Tony Wood. 2015. Quantifying suicidal ideation via language usage on social media. In *Joint Statistics Meetings Proceedings, Statistical Computing Section, JSM*.

[26] Patrick Corrigan. 2004. How stigma interferes with mental health care. *American psychologist* 59, 7 (2004), 614.

[27] Ashlynn R Daughton, Michael J Paul, and Rumi Chunara. 2018. What Do People Tweet When They're Sick? A Preliminary Comparison of Symptom Reports and Twitter Timelines. (2018).

[28] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*. ACM, 47–56.

[29] Munmun De Choudhury, Scott Counts, Eric Horvitz, and Aaron Hoff. 2014. Characterizing and Predicting Postpartum Depression from Facebook Data. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM.

[30] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. ACM, 626–638.

[31] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM* 13 (2013), 1–10.

[32] Munmun De Choudhury, Michael Gamon, Aaron Hoff, and Asta Roseway. 2013. "Moon Phrases": A Social Media Facilitated Tool for Emotional Reflection and Wellness. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '13)*. 41–44. http://dx.doi.org/10.4108/icst.pervasivehealth.2013.252106

[33] Munmun De Choudhury and Emre Kiciman. 2017. The Language of Social Support in Social Media and Its Effect on Suicidal Ideation Risk.. In *ICWSM*. 32–41.

[34] 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*. ACM, 2098–2110.

[35] Norman K Denzin. 2012. Triangulation 2.0. *Journal of mixed methods research* 6, 2 (2012), 80–88.

[36] Richard O Duda, Peter E Hart, and David G Stork. 2012. *Pattern classification*. John Wiley & Sons.

[37] Ezekiel J Emanuel, David Wendler, and Christine Grady. 2000. What makes clinical research ethical? *Jama* 283, 20 (2000), 2701–2711.

[38] T Emmens and A Phippen. 2010. Evaluating Online Safety Programs. *Harvard Berkman Center for Internet and Society.[23 July 2011]* (2010).

[39] Sindhu Kiranmai. Ernala, Asra F. Rizvi, Michael L. Birnbaum, John M. Kane, and Munmun De Choudhury. 2017. Linguistic Markers Indicating Therapeutic Outcomes of Social Media Disclosures of Schizophrenia. *Proceedings of the ACM Human Computer Interaction (PACM)* (2017).

[40] Carl E Fisher and Paul S Appelbaum. 2017. Beyond Googling: The Ethics of Using Patients' Electronic Footprints in Psychiatric Practice. *Harvard Review of Psychiatry* (2017).

[41] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports* 7 (2017), 45141.

[42] Scott A Golder and Michael W Macy. 2011. Diurnal and seasonal mood vary with work, sleep, and daylength across diverse cultures. *Science* 333, 6051 (2011), 1878–1881.

[43] Isabelle Guyon and André Elisseeff. 2003. An introduction to variable and feature selection. *Journal of machine learning research* 3, Mar (2003), 1157–1182.

[44] Douglas D Heckathorn. 1997. Respondent-driven sampling: a new approach to the study of hidden populations. *Social problems* 44, 2 (1997), 174–199.

[45] Eric H Huang, Richard Socher, Christopher D Manning, and Andrew Y Ng. 2012. Improving word representations via global context and multiple word prototypes. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 873–882.

[46] Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. 2015. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*. 553–562.

[47] Xiaolei Huang, Lei Zhang, David Chiu, Tianli Liu, Xin Li, and Tingshao Zhu. 2014. Detecting suicidal ideation in Chinese microblogs with psychological lexicons. In *Ubiquitous Intelligence and Computing, 2014 IEEE 11th Intl Conf on and IEEE 11th Intl Conf on and Autonomic and Trusted Computing, and IEEE 14th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UTC-ATC-ScalCom)*. IEEE, 844–849.

[48] Becky Inkster, David Stillwell, Michal Kosinski, and Peter Jones. 2016. A decade into Facebook: where is psychiatry in the digital age? *The Lancet Psychiatry* 3, 11 (2016), 1087–1090.

[49] Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. 2010. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders.

[50] Zunaira Jamil. 2017. *Monitoring Tweets for Depression to Detect At-risk Users*. Ph.D. Dissertation. Université d'Ottawa/University of Ottawa.

[51] Rob Kitchin. 2014. Big Data, new epistemologies and paradigm shifts. *Big Data & Society* 1, 1 (2014), 2053951714528481.

[52] Michal Kosinski, Sandra C Matz, Samuel D Gosling, Vesselin Popov, and David Stillwell. 2015. Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist* 70, 6 (2015), 543.

[53] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences* 110, 15 (2013), 5802–5805.

[54] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying Unknown Unknowns in the Open World: Representations and Policies for Guided Exploration.. In *AAAI*, Vol. 1. 2.

[55] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google Flu: traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[56] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: the coming age of computational social science. *Science (New York, NY)* 323, 5915 (2009), 721.

[57] Anthony F Lehman, Jeffrey A Lieberman, Lisa B Dixon, Thomas H McGlashan, Alexander L Miller, Diana O Perkins, Julie Kreyenbuhl, John S McIntyre, Sara C Charles, Kenneth Altshuler, et al. 2004. Practice guideline for the treatment of partients with schizophrenia. *American Journal of psychiatry* 161, 2 SUPPL. (2004).

[58] Alessandro Liberati, Douglas G Altman, Jennifer Tetzlaff, Cynthia Mulrow, Peter C Gøtzsche, John PA Ioannidis, Mike Clarke, Pl J Devereaux, Jos Kleijnen, and David Moher. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS medicine* 6, 7 (2009), e1000100.

[59] Huijie Lin, Jia Jia, Quan Guo, Yuanyuan Xue, Qi Li, Jie Huang, Lianhong Cai, and Ling Feng. 2014. User-level psychological stress detection from social media using deep neural network. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 507–516.

[60] Huijie Lin, Jia Jia, Liqiang Nie, Guangyao Shen, and Tat-Seng Chua. [n. d.]. What Does Social Media Say about Your Stress?.

[61] Kate Loveys, Patrick Crutchley, Emily Wyatt, and Glen Coppersmith. 2017. Small but mighty: Affective micropatterns for quantifying mental health from social media language. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*. 85–95.

[62] Keith A Markus and Denny Borsboom. 2013. *Frontiers of test validity theory: Measurement, causation, and meaning*. Routledge.

[63] Kimberly McManus, Emily K Mallory, Rachel L Goldfeder, Winston A Haynes, and Jonathan D Tatum. 2015. Mining Twitter data to improve detection of schizophrenia. *AMIA Summits on Translational Science Proceedings* 2015 (2015), 122.

[64] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[65] Margaret Mitchell, Kristy Hollingshead, and Glen Coppersmith. 2015. Quantifying the language of schizophrenia in social media. *NAACL HLT 2015* (2015), 11.

[66] David C Mohr, Mi Zhang, and Stephen Schueller. 2017. Personal sensing: Understanding mental health using ubiquitous sensors and machine learning. *Annual Review of Clinical Psychology* 13, 1 (2017).

[67] Roberta B Ness, Joint Policy Committee, et al. 2007. Influence of the HIPAA privacy rule on health research. *Jama* 298, 18 (2007), 2164–2170.

[68] Thin Nguyen, Dinh Phung, Bo Dao, Svetha Venkatesh, and Michael Berk. 2014. Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing* 5, 3 (2014), 217–226.

[69] Bridianne O'Dea, Stephen Wan, Philip J Batterham, Alison L Calear, Cecile Paris, and Helen Christensen. 2015. Detecting suicidality on Twitter. *Internet Interventions* 2, 2 (2015), 183–188.

[70] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kiciman. 2016. Social data: Biases, methodological pitfalls, and ethical boundaries. (2016).

[71] Minsu Park, David W McDonald, and Meeyoung Cha. 2013. Perception Differences between the Depressed and Non-depressed Users in Twitter. In *Proceedings of ICWSM*.

[72] Sungkyu Park, Sang Won Lee, Jinah Kwak, Meeyoung Cha, and Bumseok Jeong. 2013. Activities on Facebook reveal the depressive state of users. *Journal of medical Internet research* 15, 10 (2013).

[73] Michael Quinn Patton. 1999. Enhancing the quality and credibility of qualitative analysis. *Health services research* 34, 5 Pt 2 (1999), 1189.

[74] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.

[75] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[76] Víctor M Prieto, Sergio Matos, Manuel Alvarez, Fidel Cacheda, and José Luís Oliveira. 2014. Twitter: a good place to detect health conditions. *PloS one* 9, 1 (2014), e86191.

[77] Fredrick C Redlich and Daniel X Freedman. 1966. The theory and practice of psychiatry. (1966).

[78] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, 1 (2017), 15.

[79] Andrew G Reece, Andrew J Reagan, Katharina LM Lix, Peter Sheridan Dodds, Christopher M Danforth, and Ellen J Langer. 2017. Forecasting the onset and course of mental illness with Twitter data. *Scientific reports* 7, 1 (2017), 13006.

[80] Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 99–107.

[81] Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*. 1348–1353.

[82] Paul R Rosenbaum and Donald B Rubin. 1985. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician* 39, 1 (1985), 33–38.

[83] Donald B Rubin. 1986. Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* 4, 1 (1986), 87–94.

[84] Koustuv Saha, Larry Chan, Kaya De Barbaro, Gregory D Abowd, and Munmun De Choudhury. 2017. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 95.

[85] Koustuv Saha and Munmun De Choudhury. 2017. Modeling stress with social media around incidents of gun violence on college campuses. *Proc. ACM Hum.-Comput. Interact.(CSCW)* 92 (2017), 1–92.

[86] Koustuv Saha, Ingmar Weber, Michael L Birnbaum, and Munmun De Choudhury. 2017. Characterizing Awareness of Schizophrenia Among Facebook Users by Leveraging Facebook Advertisement Estimates. *Journal of medical Internet research* 19, 5 (2017).

[87] Koustuv Saha, Ingmar Weber, and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses. (2018).

[88] Ari Schlesinger, Kenton P O'Hara, and Alex S Taylor. 2018. Let's Talk About Race: Identity, Chatbots, and AI. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 315.

[89] Douglas Schuler and Aki Namioka. 1993. *Participatory design: Principles and practices*. CRC Press.

[90] 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*. 118–125.

[91] H Andrew Schwartz, Maarten Sap, Margaret L Kern, Johannes C Eichstaedt, Adam Kapelner, Megha Agrawal, Eduardo Blanco, Lukasz Dziurzynski, Gregory Park, David Stillwell, et al. 2016. Predicting individual well-being through the language of social media. In *Pac Symp Biocomput*, Vol. 21. 516–527.

[92] Eva Sharma and Munmun De Choudhury. 2018. Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 641.

[93] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. [n. d.]. Depression detection via harvesting social media: A multimodal dictionary learning solution.

[94] Judy Hanwen Shen and Frank Rudzicz. 2017. Detecting anxiety through Reddit. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*. 58–65.

[95] Hidetoshi Shimodaira. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference* 90, 2 (2000), 227–244.

[96] Raphael Silberzahn, Eric Luis Uhlmann, Dan Martin, Pasquale Anselmi, Frederik Aust, Eli C Awtrey, Štěpán Bahník, Feng Bai, Colin Bannard, Evelina Bonnier, et al. 2017. Many analysts, one dataset: Making transparent how variations in analytical choices affect results. (2017).

[97] T Simms, C Ramstedt, M Rich, M Richards, T Martinez, and C Giraud-Carrier. 2017. Detecting cognitive distortions through machine learning text analytics. In *Healthcare Informatics (ICHI), 2017 IEEE International Conference on*. IEEE, 508–512.

[98] Robert L Spitzer, Janet BW Williams, Miriam Gibbon, and Michael B First. 1992. The structured clinical interview for DSM-III-R (SCID): I: history, rationale, and description. *Archives of general psychiatry* 49, 8 (1992), 624–629.

[99] Masashi Sugiyama, Neil D Lawrence, Anton Schwaighofer, et al. 2017. *Dataset shift in machine learning*. The MIT Press.

[100] John Torous, Matcheri Keshavan, and Thomas Gutheil. 2014. Promise and perils of digital psychiatry. *Asian journal of psychiatry* 10 (2014), 120–122.

[101] John Torous and Camille Nebeker. 2017. Navigating ethics in the digital age: introducing Connected and Open Research Ethics (CORE), a tool for researchers and institutional review boards. *Journal of medical Internet research* 19, 2 (2017).

[102] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 3187–3196.

[103] Sho Tsugawa, Yukiko Mogi, Yusuke Kikuchi, Fumio Kishino, Kazuyuki Fujita, Yuichi Itoh, and Hiroyuki Ohsaki. 2013. On estimating depressive tendencies of twitter users utilizing their tweet data. In *Virtual Reality (VR), 2013 IEEE*. IEEE, 1–4.

[104] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 440.

[105] Nikhita Vedula and Srinivasan Parthasarathy. 2017. Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*. ACM, 127–136.

[106] Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM, 91–100.

[107] Xinyu Wang, Chunhong Zhang, Yang Ji, Li Sun, Leijia Wu, and Zhana Bao. 2013. A depression detection model based on sentiment analysis in micro-blog social network. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 201–213.

[108] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848* (2017).

[109] Sean D Young and Renee Garett. 2018. Ethical Issues in Addressing Social Media Posts About Suicidal Intentions During an Online Study Among Youth: Case Study. *JMIR mental health* 5, 2 (2018).

[110] Reza Zafarani and Huan Liu. 2015. Evaluation without ground truth in social media research. *Commun. ACM* 58, 6 (2015), 54–60.

[111] Yiheng Zhou, Jingyao Zhan, and Jiebo Luo. 2017. Predicting Multiple Risky Behaviors via Multimedia Content. In *Social Informatics*, Giovanni Luca Ciampaglia, Afra Mashhadi, and Taha Yasseri (Eds.). Springer International Publishing, Cham, 65–73.

[112] Mark Zimmerman and Jill I Mattia. 2001. The Psychiatric Diagnostic Screening Questionnaire: development, reliability and validity. *Comprehensive psychiatry* (2001).