# A Social Media Study on the Effects of Psychiatric Medication Use

**Koustuv Saha[†], Benjamin Sugar[†], John Torous[‡], Bruno Abrahao[*], Emre Kıcıman[§], Munmun De Choudhury[†]**

[†]Georgia Tech, [‡]Harvard Medical School, [*]NYU Shanghai, [§]Microsoft Research

[†]{koustuv.saha,bsugar,munmund}@gatech.edu, [‡]jtorous@bidmc.harvard.edu, [*]bd58@nyu.edu, [§]emrek@microsoft.com

## Abstract

Understanding the effects of psychiatric medications during mental health treatment constitutes an active area of inquiry. While clinical trials help evaluate the effects of these medications, many trials suffer from a lack of generalizability to broader populations. We leverage social media data to examine psychopathological effects subject to self-reported usage of psychiatric medication. Using a list of common approved and regulated psychiatric drugs and a Twitter dataset of 300M posts from 30K individuals, we develop machine learning models to first assess effects relating to mood, cognition, depression, anxiety, psychosis, and suicidal ideation. Then, based on a stratified propensity score based causal analysis, we observe that use of specific drugs are associated with characteristic changes in an individual's psychopathology. We situate these observations in the psychiatry literature, with a deeper analysis of pre-treatment cues that predict treatment outcomes. Our work bears potential to inspire novel clinical investigations and to build tools for digital therapeutics.

## Introduction

Psychiatric medications are key to treat many mental health conditions, including mood, psychotic, and anxiety disorders. 1 in 6 Americans take psychiatric medications and they account for 5 of the top 50 drugs sold in the U.S (*drugs.com*). These drugs[1] are designed to correct underlying neuro-pathological disease processes by restoring neural communication by modulating the brains chemical messengers and neurotransmitters (Barchas and Altemus 1999). These changes can be accompanied by debilitating neurological impairments and life-threatening effects as severe as suicidal ideation (Coupland et al. 2011) which reduce psychosocial functioning, and make social capital and vocational development less available to these individuals. Given the pervasiveness of their use, psychiatric medications can either alleviate or exacerbate mental illness burden on both personal and societal levels (Rosenblat et al. 2016).

One reason behind the mixed success of psychiatric medications stems from the fact that the mechanisms by which they modify the brain operation are poorly understood. In practice, their effects vary across individuals, and often do not achieve the intended result. Without any biological markers to match patients with the most appropriate medication, the selection of drug treatments is based primarily on trial-and-error (Cipriani et al. 2018; Trivedi et al. 2006). Unsurprisingly, frustration with treatment and side effects often causes treatment discontinuation (Bull et al. 2002).

Consequently, literature in precision psychiatry has emphasized the need to understand the psychiatric effects of these medications (Cipriani et al. 2009). Presently, most knowledge of drug reactions comes from clinical trials and reports of adverse events; e.g., the FDA's Adverse Event Reporting System (*open.fda.gov/data/faers*) clinical trial database. However, these trials can be biased, being conducted and funded by pharmaceutical companies, and are rarely replicated in large populations (Lexchin et al. 2003). In addition, these clinical trials suffer from limitations such as non-standardized study design, confounding factors, and restrictive eligibility criteria (Lexchin et al. 2003). For example, an analysis found that existing inclusion criteria for most trials would exclude 75% of individuals with major depressive disorders (Blanco et al. 2008). Even well-designed clinical trials can suffer from low statistical power, or limited observability of effects due to short monitoring and study periods, spanning just weeks or months

**Contributions** Our work seeks to address these gaps and complements existing methodologies for understanding the effects of psychiatric medications. We report a large-scale social media study of the effects of 49 FDA approved antidepressants across four major families (SSRIs, SNRIs, TCAs, and TeCAs) (descriptions in (Lopez-Munoz and Alamo 2009)). Our analysis is conducted using two years of Twitter data from two populations: 112M posts from 30K self-reported users of psychiatric medications and 707M posts from 300K users who did not. Adopting a patient-centered approach (Shippee et al. 2012), in this paper, we seek to study the effects of these drugs as reflected and self-reported in the naturalistic social media activities of individuals.

Accomplishing this goal involves meeting several technical challenges, importantly addressing causality, and our work offers robust and validated computational methods for the purpose. We first develop expert-validated machine learning models to assess psychopathological states



[1]This paper uses *medications* and *drugs* interchangeably, referring to U.S. FDA regulated psychiatric drugs only.

known to be affected by psychiatric medications, including mood, cognition, depression, anxiety, psychosis, and suicidal ideation, as given in the literature (Coupland et al. 2011). Using initial social media mentions of drug intake, we then identify individuals likely beginning treatment. Based on a stratified propensity score analysis (Olteanu et al. 2017), we compare post-treatment symptoms in treated individuals to large untreated control population. With an individual treatment effect analysis, we study the relationship between pre-treatment mental health signals and post-treatment response.

**Findings** Our results show that most drugs are linked to a post-treatment increase in negative affect and decrease in positive affect and cognition. We find varying effects both within and between the drug families on psychopathological symptoms (depression, anxiety, psychosis, and suicidal ideation). Clinically speaking, SSRIs are associated with worsening symptoms, whereas TCAs lead to improvements. Studying the individual-specific outcomes, our analyses help associate drug effectiveness with individuals' psycholinguistic attributes on social media.

Clinically, our findings reveal signals of the most common effects of the psychiatric medications over a large population, with the potential for improved characterization of their occurrence. Technologically, we show the potential of novel technologies in digital therapeutics, powered by large-scale social media analyses, to support digital therapeutics (Vieta 2015). These tools can improve the identification of adverse outcomes, as well as the behavioral and lifestyle changes in the heterogeneous outcomes of psychiatric drugs.

**Privacy, Ethics, and Disclosure** Given the sensitive nature of our work, despite working with public social media data, we are committed to securing the privacy of the individuals in our dataset. We use paraphrased examples of content and avoid personally identifiable information. Our findings were corroborated with our co-author who is a board-certified psychiatrist. *However, our work is not intended to replace clinical evaluation by a medical professional, and should not be used to compare or recommend medications.*

## Background and Related Work

**Psychiatric Drug Research and Prescriptions** The mechanisms of action of many psychiatric drugs and the basis for specific therapeutic interventions, are not fully understood. Among other hypotheses, the monoanime hypothesis postulates that these drugs target the neurotransmitters serotonin, norepinephrine and dopamine, associated with feelings of well-being, alertness, and pleasure (Barchas and Altemus 1999). From the monoamine standpoint, medications are classified into families, based on their brain receptor affinities, which distinguish their mechanism of action.

Antidepressant research has grown tremendously, ever since Imipramine, and other Tricyclic Antidepressants (TCAs) were discovered and found to be effective (Gillman 2007). However, TCAs have a broad spectrum of neurotransmitter affinities, which may often lead to undesirable side effects, such as liver toxicity, excessive sleepiness, and sexual dysfunction (Frommer et al. 1987). Several other compounds have since been introduced whose development was guided by the idea that increasing the selectivity of the target of action to individual neurotransmitters would, in theory, limit the incidence of side effects while maintaining the effectiveness of TCAs (Lopez-Munoz and Alamo 2009). These include Tetracyclic Antidepressants (TeCA), Serotonin Norepinephrine Reuptake Inhibitors (SNRI), and Selective Serotonin Reuptake Inhibitors (SSRI).

Given these biochemical underpinnings, historically psychiatric care has adopted a "Disease-Centered Model" (Moncrieff and Cohen 2009), one that justifies prescribing medications on the assumption that they help correct the biological abnormalities related to psychiatric symptoms. However, this model neglects the psychoactive effects of the drugs. Consequently, a "Drug-Centered Model" has been advocated (Moncrieff and Cohen 2009), enabling patients to exercise more control over their pharmacotherapy, and moving treatment in a collaborative direction between clinicians and patients. Our work builds on this notion towards a "Patient-Centered Model" (Shippee et al. 2012), where psychiatrists could leverage complementary techniques (such as stratifying users on their naturalistic digital footprints) to prescribe medications.

**Understanding Effects of Psychiatric Drugs** The efficacy, safety, and approval of psychiatric drugs are typically established through clinical trials. In one such trial, the randomized controlled trial (RCT), participants are randomly assigned to a treatment or a control group, where the former receives a particular drug, and the latter receives a placebo (eg. a sugar pill with no drug content). Then, the effects of the treatment are measured as a difference in the two groups following the drug intake. A major weakness of these trials is that they are often conducted on individuals who may significantly differ from actual patients, and often, they are not externally validated to a larger and a more representative population (Hannan 2008). As an alternative, a study design that has gained interest is observational study (Hannan 2008). The advantage here is that they enable the researchers to conduct subset analyses that can help to precisely identify which patients benefit from each treatment. Similarly, we use large-scale longitudinal data and a causal approach to not only examine the effects of psychiatric drugs, but also to provide a framework that finds insights about their effectiveness across strata of populations.

### Pharmacovigilance, Web, and Social Media

Pharmacovigilance is "the science and activities relating to the detection, assessment, understanding, and prevention of adverse effects or any other drug-related problem" (WHO 2002). Over the years, pharmacovigilance has become centered around data mining of clinical trial databases and patient-reported data. Recently, patient-generated activity online has also been used to understand pharmacological effects in large populations (Harpaz et al. 2017). White et al. (2016) found that web search logs improve detection of adverse effects by 19%, compared to an offline approach.

Social media studies of drug and substance use, including behavioral changes, adverse reactions, and recovery have garnered significant attention in HCI (Chancellor et al. 2019,

Kıcıman et al. 2018, Liu et al. 2017). Recent research has studied the abuse of prescription drugs, by leveraging drug forums (MacLean et al. 2015), Twitter (Sarker et al. 2015), and Reddit (Gaur et al. 2018). Social media has also facilitated the identification of adverse drug reactions at the population level using self-reports (Lardon et al. 2015) as well as the mentions of side effects of adverse drug reaction on Twitter (Nikfarjam et al. 2015).

Social media enables individuals to candidly share their personal and social experiences (Kıcıman et al. 2018, Olteanu et al. 2017, Saha et al. 2019b), thereby providing low-cost, large-scale, non-intrusive data to understand naturalistic patterns of mood, behavior, cognition, social milieu, and even mental and psychological states, both in real-time and longitudinally (Chancellor et al. 2016, Coppersmith et al. 2014, De Choudhury et al. 2013, Dos Reis and Culotta 2015, Saha et al. 2019a, Yoo and De Choudhury 2019). In characterizing drug use, being able to quantify these psychopathological attributes is extremely powerful.

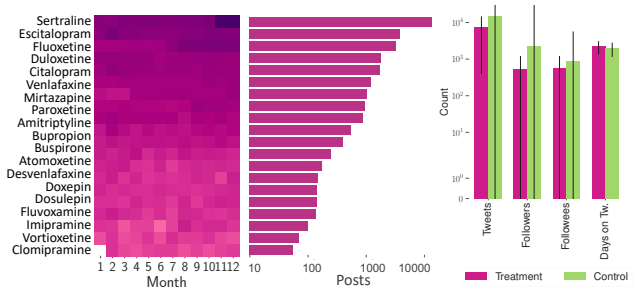
Nevertheless, we observe a gap that digital pharmacovigilance studies, particularly those using social media, have largely targeted the named adverse effects of drugs (e.g., "headache", "palpitations", "nausea"), and have not measured broader forms of symptomatic changes longitudinally. To fill this gap, our work draws on theoretically grounded methodologies, including lexicon-based and machine learning approaches, to measure the symptomatic outcomes of psychiatric drug use longitudinally, including mood, cognition, depression, anxiety, psychosis, and suicidal ideation.

## Data

This work leverages Twitter timeline data of individuals who self-report their use of psychiatric medications. The data collection involve: 1) curating a list of psychiatric medications; 2) using this list to collect Twitter posts that mentioned these medications; 3) identifying and filtering for only those posts where users self-reported about personal medication intake (using a personal medication intake classifier, and 4) collecting the timeline datasets of these individuals who self-reported psychiatric medication intake, and additionally doing that for another set of users who did not self-report psychiatric medication intake. We explain these steps here:

**Psychiatric Medication List** We scope our work to a list of FDA approved antidepressants and antidepressant augmentation drugs. We crawl a hand-curated set of Wikipedia pages of these drugs, to collect brand names, generic names, and drug family information to obtain a list of 297 brand names mapped to 49 generic names, grouped into four major families: SNRI, SSRI, TCA, TeCA. Our clinician co-author established the validity and relevance of this final list.

**Twitter Data of Psychiatric Medication Usage** We query the Twitter API for public English posts mentioning these drugs (brand or generic name) between January 01, 2015 and December 31, 2016 to obtain 601,134 posts by 230,573 unique users. A two year period balances concerns about being long enough to avoid confounds by idiosyncratic events and seasonal changes, but short enough to avoid major changes in social media use and drug prescription policies. This also enables us to collect sufficient pre- and post- medication usage timeline data for our ensuing analyses.




(a) (b)

Figure 1: (a) Monthly distribution and the number of posts in logarithmic scale for the top 20 medications (darker colors correspond to greaeter density); (b) Mean distribution of User Attributes in *Treatment* and *Control* datasets.

| |
|---|
| *I'm taking my first dose of X tonight.* |
| *I was depressed & psychiatrist gave me X, slept for two days!* |
| *First day on X. Dose 1 taken, and I already feel weird from it.* |
| *Just took X for the first time. Let's see how it goes* |
| *I got brain zaps if I took $X_1$ even an hour late. Changed to $X_2$ now!* |
| *My no-med experiment went horribly awry, so I'm starting X today* |

Table 1: Example paraphrased self-reports of psychiatric medication usage. Drug names are masked.

**Personal Medication Intake Classifier** Since mentioning a medication in a tweet does not necessarily indicate its usage, we filter out those posts that were first-person reports of using these medications. For this purpose, we employ a machine learning classifier built in a recent work (Klein et al. 2017). This classifier distinguishes Twitter posts into the binary classes (yes or no) if there is a self-report about personal medication intake. We replicate this model and train it on an expert-annotated dataset of 7,154 Twitter posts (dataset published in Klein et al. 2017). The classifier uses an SVM model with linear kernel and shows a mean k-fold (k=5) cross-validation accuracy and F1-score of 0.82 each.

We use this classifier to label the 601,134 medication-mention posts to find that 93,275 of these posts indicate medication self-intake (example posts in Table 1). Figure 1a shows the monthly and overall distribution of the top 20 drugs in our dataset. We find that SSRIs (eg. Sertraline, Escitalopram, Fluoxetine) rank highest in the distribution. This aligns with external surveys on the most prescribed psychiatric drugs in that time which found that the top 5 antidepressants captured over 70% of the prescription volumes[2].

**Compiling Treatment and Control Datasets** The above 93,275 medication usage posts were posted by 52,567 unique users from whom we then collect Twitter metadata such as the number of tweets, followers, followees, and account creation date. To limit our analyses to typical Twitter users, we remove users (e.g., celebrities or typically inactive users) with more than 5000 followers or followees or posted

[2] *psychcentral.com/blog/top-25-psychiatric-medications-for-2016*, *lab.express-scripts.com/lab/drug-trend-report/media/29f13dee4e7842d6881b7e034fc0916a.ashx*

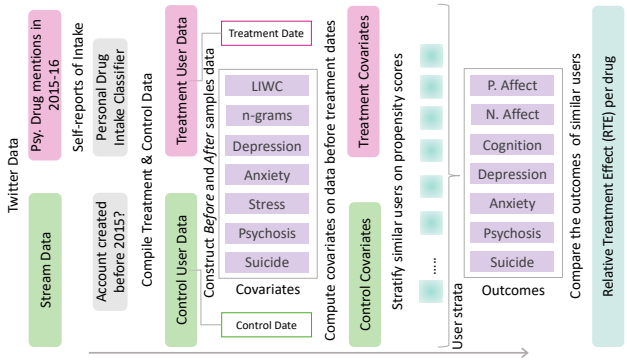



Figure 2: Schematic diagram of propensity score analysis.

outside the range of 200 to 30,000 tweets—a choice motivated from prior work (Pavalanathan and Eisenstein 2016). For the remaining 34,518, we collect the timeline data between January 01, 2014 and February 15, 2018, to obtain a total of 112,025,496 posts. Finally, we limit our dataset to those users who posted both before and after their first self-reported use of medication and did not self-report the use before 2015. The resultant timeline dataset of 23,191 users is referred to hereon as the *Treatment* dataset.

Additionally, we build a *Control* dataset of users who did not self-report using psychiatric medication. We obtain 495,419 usernames via the Twitter streaming API and prune this list (as above) and remove accounts that did not exist pre-2015. We collect the timelines of the remaining 283,374 users, for a total 707,475,862 posts. Figure 1b shows the mean distribution of Twitter attributes in our two datasets.

## Methods

**Study Design and Rationale** Recall that our research objective is to examine the effects of psychiatric medications in terms of the changes in mental health symptoms. Effectively answering this question necessitates the use of causal methods to reduce biases associated with the observed effects following the reported medication usage. The effects of drugs are most often measured through Randomized Controlled Trials (RCTs) in clinical settings (Cipriani et al. 2018; Szegedi et al. 2005). Due to the limitations of this approach, noted in the "Background and Related Work" section, and because of the potential advantages of a "Patient Centered Model" that focuses on using the naturalistic self-reports of individuals regarding their psychiatric medication use, this work adopts an observational study design. We do acknowledge that observational studies are weaker than RCTs in making conclusive causal claims like ones needed to accomplish the goals of this paper, but they provide complementary advantages over RCTs in many aspects (Hannan 2008). Literature in statistics also provides support for these methods and similar frameworks have been leveraged in previous quantitative social media studies (De Choudhury et al. 2016, Kıcıman et al. 2018, Olteanu et al. 2017, Saha et al. 2018).

Specifically, we adopt a causal inference framework based on matching, which simulates an RCT setting by controlling for as many covariates as possible (Imbens and Rubin 2015). This approach is built on the potential outcomes framework, which examines whether an outcome is caused by a treatment $T$, by comparing two potential outcomes: 1) $Y_i(T = 1)$ when exposed to $T$, and 2) $Y_i(T = 0)$ if there was no $T$. However, it is impossible to obtain both of these outcomes for the same individual. To overcome this challenge of missing data, this framework estimates the missing counterfactual outcome for an individual based on the outcomes of other similar (matched) individuals (in terms of their covariate distribution). In particular, we employ stratified propensity score analysis (Olteanu et al. 2017) to match and then to examine the symptomatic outcomes in the *Treatment* and *Control* individuals by measuring the relative treatment effect of the drugs (see Figure 2 for an overview).

### Constructing *Before* and *After* Samples

As our setting concerns measuring the changes *post* reported usage of the medications, we divide our datasets into *Before* and *After* samples around their dates of treatment. For every *Treatment* user, we assign the date of their first medication-intake post as their treatment date. We assign each individual in the *Control* dataset a placebo date, matching the non-parametric distribution of treatment dates of the *Treatment* dataset, to mitigate the effects of any temporal confounds. For this, we ensure that the treatment and placebo dates follow similar distribution by non-parametrically simulating placebo dates from the pool of treatment dates. We measure the similarity in their distribution using Kolmogorov–Smirnov test to obtain an extremely low statistic of 0.06, indicating similarity in the probability distribution of treatment and placebo dates (Figure 3b). We then divided our *Treatment* and *Control* datasets into *Before* and *After* samples based on the treatment and placebo dates.

### Defining and Measuring Symptomatic Outcomes

Drawing on the psychiatry and psychology literature (Pennebaker et al. 2003, Rosenblat et al. 2016), next, we measure mental health symptomatic outcomes, subject to the reported usage of the medications in the above-constructed user samples, based on the changes in mood, cognition, depression, anxiety, stress, psychosis, and suicidal ideation. We use the following approaches:

**Affect and Cognition** To measure the affective and cognitive outcomes, similar to prior work (Ernala et al. 2017, Saha et al. 2018), we quantify psycholinguistic shifts in affect and cognition. In particular, we use the changes in the normalized occurrences of words in these categories per the well-validated Linguistic Inquiry and Word Count (LIWC) lexicon (Tausczik and Pennebaker 2010). These categories include *positive* and *negative affect* for affect, and *cognition mechanics*, *causation*, *certainty*, *inhibition*, *discrepancies*, *negation*, and *tentativeness* for cognition.

**Depression, Anxiety, Stress, Psychosis, Suicidal Ideation** We quantitatively estimate these measures from social media by building several supervised learning based classifiers of mental health attributes. Our approach is inspired by recent work where mental health attributes have been inferred in unlabeled data by transferring a classifier trained on a different labeled dataset (Saha and De Choudhury 2017). To

train such classifiers for use in our work, we identify several Reddit communities that are most closely associated with these measures. That is, the positive examples in our training data comprise posts shared on *r/depression* for depression, *r/anxiety* for anxiety, *r/stress* for stress, *r/psychosis* for psychosis, and *r/SuicideWatch* for suicidal ideation. On the other hand, negative examples are extracted from the collated sample of 20M Reddit posts gathered from 20 subreddits that appear on the landing page of Reddit during the same period of our Twitter data sample, such as *r/AskReddit*,*r/aww*, *r/movies*, and others.

These classifiers are SVM models with linear kernels and use 5000 $n$-grams ($n$=1,2,3) as features. We use a *balanced* number of examples for the two classes in training, and we tune the parameters of the classifiers using $k$-fold ($k$=5) cross-validation (Chandrasekharan et al. 2018). Table 2 summarizes the size of the datasets and the accuracy metrics. Figure 3a shows the ROC curves of these classifiers. These classifiers show a mean cross-validation accuracy ranging between 0.79 and 0.88 and mean test accuracy ranging between 0.81 and 0.91. Table 3 reports the top 10 features in each of the classifiers. Several top $n$-gram features such as *depression*, *stress*, *hope*, *help*, and *feel*, are contextually related to mental health.

*Establishing Model Validity.* Since our next goal is to employ these classifiers, trained on Reddit data, to automatically infer the symptomatic outcomes in the Twitter user samples—a platform with distinct norms and posting style, we present a series of evaluation tests to demonstrate the validity of the transfer approach and the transferred classifiers. 1) First, motivated from prior work (Saha et al. 2017), we conduct a linguistic equivalence test between the Reddit training dataset, and the Twitter unseen dataset based on a word-vector similarity approach. Using word-vectors (pre-trained on Google News dataset of over 100 billion tokens), we find the vector similarity of the top 500 $n$-grams in the Reddit and Twitter corpuses to be 0.95. This shows high content similarity across the two platforms, in turn justifying the transfer approach. 2) 2) Second, we find that the top features of these classifiers align with that of similar mental health classifiers built on Twitter to identify depression (De Choudhury et al. 2013), anxiety (Dutta et al. 2018), stress (Lin et al. 2014), psychosis (Birnbaum et al. 2017), and suicidal ideation (Burnap et al. 2015). This indicates the construct validity of the transferred classifiers. 3) Third, we demonstrate convergence and divergence validity and present a qualitative validation of the outputs of these classifiers. Two researchers manually inspected 170 randomly selected Twitter posts on mental health symptoms, spanning both user samples. Using the methodology outlined in Bagroy et al. (2017) that draws up the DSM-5 clinical framework, they rated each Twitter post on a binary Likert scale (high/low) to assess levels of expressed depression, anxiety, stress, psychosis, or suicidal ideation. We find high (87%) agreement between the manual ratings and the classifiers' respective labels. This aligns with prior work where similar agreements have been reached between classifier outcomes and annotations of mental health experts (a Fleiss' $\kappa$=0.84 was reported in Bagroy et al. (2017)).

## Matching For Causal Inference

**Matching Covariates** When conditioned on high-dimensional covariate data, matching is known to significantly minimize bias compared to naive correlational analyses (Imbens and Rubin 2015). Our approach controls for a variety of covariates so that the compared *Control* and *Treatment* groups show similar pre-treatment online behavior. The 1st set of covariates includes users' *social attributes* (count of tweets, followers, followees, duration on the platform and frequency of posting). The 2nd set corresponds to the distribution of word usage in the Twitter timelines, where for every user, we build a vector model on the top 2,000 unigrams. The 3rd set consists of normalized use of psycholinguistic attributes in the posts, i.e, distribution across 50 categories in the LIWC lexicon (Tausczik and Pennebaker 2010), across *affective*, *cognitive*, *lexical*, *stylistic*, and *social* attributes.

Finally, to minimize the confounding effects of an individual's mental health conditions prior to treatment, in the 4th set we control for the users' mean aggregated usage of posts indicative of *depression*, *anxiety*, *stress*, *psychosis*, and *suicidal ideation*, assessed using the classifiers described above. Note that there is typically a significant time-lag between the onset of mental illness and the first treatment people receive (Hasin et al. 2005; Oliver et al. 2018). Therefore, matching on these pre-treatment symptoms should capture and account for the individual's already existing mental health condition. That is, our matched comparisons should on average be comparing people with a given mental illness who receive treatment to their counterparts who have the same symptoms but did not receive treatment.

**Propensity Score Analysis** We use matching to find pairs (generalizable to groups) of *Treatment* and *Control* users whose covariates are statistically very similar to one another, but where one was *treated*, and the other was not. The propensity score model matches users based on their *likelihood* of receiving the treatment, or the propensity scores. Our stratified matching approach groups individuals with similar propensity scores into strata (Kıcıman et al. 2018). Every stratum, therefore, consist of individuals with similar covariates. This helps us to isolate and estimate the effects of the treatment within each stratum.

To compute the propensity scores, we build a logistic regression model that predicts a user's treatment status based on their covariates. Next, we discard the outliers in the propensity scores (outside the range of 2 standard deviations from the mean), and segregate the remaining distribution into 100 strata of equal width. To further ensure that our causal analysis per stratum remains restricted to a sufficient number of similar users, we remove those strata with very few *Treatment* or *Control* users, as is common practice in causal inference research (De Choudhury et al. 2016). With a threshold of at least 50 users per group in a stratum, this approach gave 63 strata that consisted of 23,163 *Treatment* and 122,941 *Control* individuals (Figure 3c).

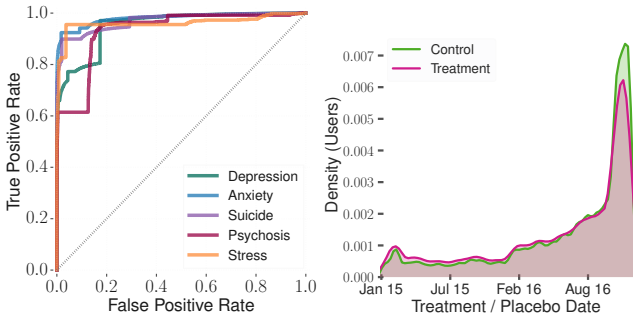
**Quality of Matching** To ensure that we matched statistically comparable *Treatment* and *Control* users, we evaluate

| Precision | | Recall | | Accuracy | |
|---|---|---|---|---|---|
| CV | Test | CV | Test | CV | Test |
| Depression (40,000; 555,955) | | | | | |
| .88 | .86 | .88 | .82 | .88 | .82 |
| Anxiety (40,000; 238,689) | | | | | |
| .82 | .91 | .82 | .90 | .82 | .91 |
| Stress (5,000; 5,969) | | | | | |
| .79 | .92 | .79 | .91 | .79 | .92 |
| Psychosis (5,000; 3,439) | | | | | |
| .87 | .85 | .87 | .81 | .87 | .81 |
| Suicidal Idn. (40,000; 276,769) | | | | | |
| .78 | .91 | .78 | .91 | .78 | .91 |

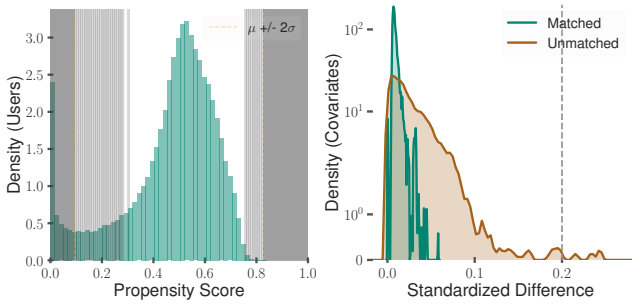
Table 2: Mental health classifiers (training and test data size), cross-validation and test accuracies.

| Depression | | Anxiety | | Stress | | Psychosis | | Suicidal Idn. | |
|---|---|---|---|---|---|---|---|---|---|
| Feature | Score | Feature | Score | Feature | Score | Feature | Score | Feature | Score |
| *concerns* | .6 | *forgetting* | .6 | *stress* | .4 | *psychosis* | .5 | *help* | .4 |
| *it looks like* | .5 | *it looks* | .6 | *help* | .4 | *song* | .4 | *friends* | .4 |
| *here are* | .5 | *does it* | .6 | *try* | .4 | *psychotic* | .4 | *anymore* | .4 |
| *forgetting* | .4 | *looks like* | .6 | *work* | .3 | *hope* | .3 | *never* | .4 |
| *know* | .4 | *concerns* | .6 | *feel* | .3 | *experience* | .3 | *family* | .4 |
| *all really* | .4 | *posting* | .5 | *things* | .2 | *help* | .3 | *suicide* | .4 |
| *depression* | .4 | *anxiety* | .4 | *you can* | .3 | *schizophenia* | .3 | *people* | .4 |
| *have spaces* | .3 | *around* | .4 | *life* | .2 | *symptoms* | .3 | *end* | .4 |
| *suicidal* | .3 | *feel* | 14.5 | *take* | .2 | *medication* | .2 | *think* | .3 |
| *feeling* | .2 | *attack* | .3 | *need to* | .2 | *weed* | .2 | *around* | .3 |

Table 3: Top 10 Features in the mental health outcome classifiers.




(a) (b)

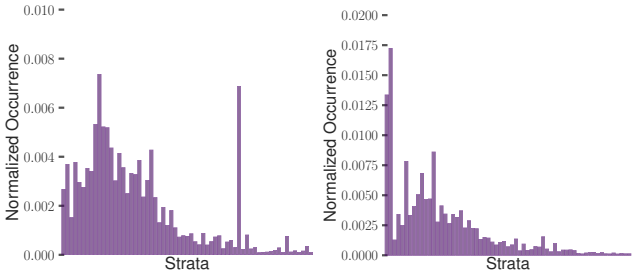



(c) (d)

Figure 3: (a) ROC curves of the classifiers that measure symptomatic outcome, (b) *Treatment* dates distribution, (c) Propensity score distribution (shaded region represents those dropped in our analysis), (d) Quality of matching

the balance of their covariates. We compute the standardized mean difference (SMD) across covariates in the *Treatment* and *Control* groups in each of the 63 valid strata. SMD calculates the difference in the mean covariate values between the two groups as a fraction of the pooled standard deviation of the two groups. Two groups are considered to be balanced if all the covariates reveal SMD lower than 0.2 (Kıcıman et al. 2018), a condition which all our covariates satisfied. We also find a significant drop in the mean SMD from 0.029 (max=0.31) in the unmatched datasets to 0.009 (max=0.05) in the matched datasets (Figure 3d).

**Characterizing the Propensity Strata of Users** To understand how the subpopulations across the several strata




(a) (b)

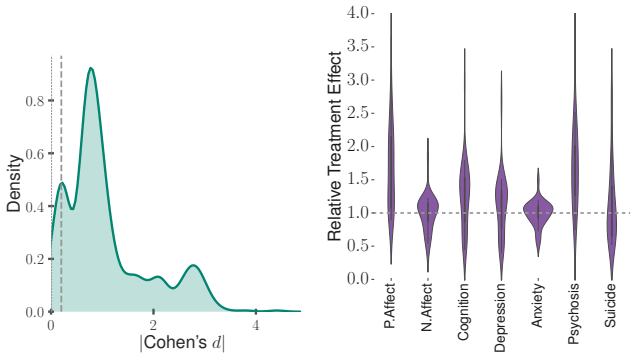
Figure 4: Distribution of words by users across strata by psycholinguistic categories of: a) affect, b) cognition.

vary, we characterize their psycholinguistic attributes. Figure 4 plots the usage of affective and cognitive words across all the strata. The propensity score model distributed these users in such a way that the users with a greater tendency to use affective and cognitive words mostly occur in the lower and middle strata, whereas those with a lower tendency to use these words predominantly occur in the higher strata.

**Measuring Changes in the Outcomes.** To quantify the effects of self-reported psychiatric medication use, we compute the change in the symptomatic outcomes, weighted on the number of *Treatment* users in each stratum. For this, we first determine the Relative Treatment Effect (*RTE*) of the drugs per outcome measure in every stratum, as a ratio of the likelihood of an outcome measure in the *Treatment* group to that in the *Control* group (Kıcıman et al. 2018). Next, using a weighted average across the strata, we obtain the mean *RTE* of the medications per outcome measure. We compute the mean *RTE* for all the drugs and aggregate that for the drug families. An outcome *RTE* greater than 1 suggests that the outcome *increased* in the *Treatment* users, whereas an *RTE* lower than 1 suggests that it decreased in the *Treatment* users, following the reported use of psychiatric medication.

## Exploring Individual-Specific Effects

We finally aim to study how the drugs affect individuals who vary in their pre-existing psychological state. So once we calculated the treatment effect of the drugs, we explore its relationship with the individuals' psycholinguistic attributes (as obtained by LIWC). For this, in every stratum, we first build separate linear regression models for all the outcomes




(a) (b)

Figure 5: (a) Distribution of effect size magnitude in the outcome change between *Treatment* and *Control* users; (b) Distribution of RTE across all the *Treatment* users.

of *Control* users with their covariates as predictors. Using these models we predict the counterfactual outcomes of the *Treatment* users in the strata – that is, the outcome for each treated user if they had not taken the drug. Next, for every user, we obtain the ratio of the predicted and actual value of the outcome. This essentially quantifies how much a *Treatment* user is individually effected by treatment, and is referred to as the Individual Treatment Effect (*ITE*) in individualized and precision medicine literature (Lamont et al. 2018). Finally, we measure the association between pre-treatment psycholinguistic attributes and the *ITE* values per drug, by fitting a linear regression model. This characterizes the directionality and the effect of a drug on an individual based on their pre-existing psycholinguistic attributes.

## Results

### Observations about Symptomatic Outcomes

Our first set of results investigates if self-reported psychiatric drug use had a statistically significant effect on the *Treatment* users. For this, we measure the effect size (Cohen's $d$) in the outcome changes between the *Treatment* and *Control* users, per drug, per outcome, and per valid strata. We find that the magnitude of Cohen's $d$ averages at 0.75 (see Figure 5a). A cohen's $d$ magnitude lower than 0.2 suggests small differences between two distributions. We find that 91% of our values fall outside this range, suggesting the *Treatment* significantly differed from the *Control* group. An independent sample $t$-test further reveals statistical significance in these differences ($t\epsilon$[-9.87,10.96]; $p_{\dot{\iota}}0.001$), confirming that after the self-reported use of medications, the *Treatment* users showed significant changes in outcomes.

We then compute the Relative Treatment Effect (*RTE*) of the psychiatric medications. Figure 5b shows the distribution of *RTE* across the symptomatic outcomes for the matched *Treatment* and *Control* users. We find that the *RTE* across the outcomes averages at 1.28 (stdev=0.61). We dig deeper into the effects per drug. Figure 6 presents the *RTE* of the 20 most popular generic drugs and the 4 drug families. We observe many interesting patterns here, such as most medications lead to similar directionality of effects on all the

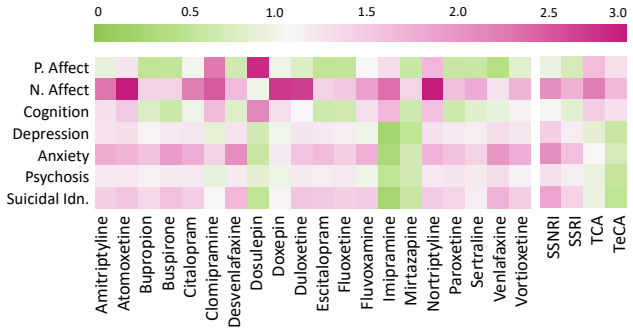



Figure 6: Relative Treatment Effect on the outcomes per 20 most popular drugs (left), and drug families (right).

(a) Sertraline

(b) Escitalopram

(c) Fluoxetine

(d) Duloxetine

Figure 7: *RTE* per propensity stratum for the top four drugs (For colorbar, refer to the one in Figure 6).

outcomes, e.g., all of the outcomes, depression, anxiety, psychosis, and suicidal ideation increase for the *Treatment* users in the *After* period of reported medication use. The similarity in effects across outcomes could be attributed to the comorbidity of the symptomatic outcomes and the clinical presentation of many moods and psychotic disorders (Rosenblat et al. 2016). We also observe that those drugs with similar pharmacological composition, such as Escitalopram and Citalopram, and Desvenlafaxine and Venlafaxine show similar trends in the symptomatic outcomes.

Table 4 summarizes the proportion of *Treatment* users who showed an increased outcome per drug family. For all these outcomes other than *positive affect* and *cognition* (in which case it is the opposite), an increase in the outcome measure also translates to *worsened observable mental health condition* of the individuals, whereas a decrease suggests an *improvement in their mental health condition*, as gleaned from Twitter. To study the strata-wise variation for each of these outcomes, we present Figure 7, which shows the *RTE* per stratum for the four most popular medications.

**Effects on Affect and Cognition** Figure 6 and Table 4 together indicate that the top medications and families are associated with an increase in the likelihood of negative affect. However, that the likelihood of positive affect and cognition also decrease for most of these medications, aligns with lit-

| Family | Users | P.A | N.A | Cog. | Dep. | Anx. | Psy. | S.I |
|---|---|---|---|---|---|---|---|---|
| SNRI | 2535 | 21 | 57 | 33 | 81 | 93 | 76 | 83 |
| SSRI | 16388 | 19 | 59 | 30 | 78 | 98 | 79 | 94 |
| TCA | 2535 | 47 | .52 | 51 | 35 | 62 | 33 | 36 |
| TeCA | 763 | 13 | 55 | 25 | 17 | 24 | 23 | 18 |

Table 4: Outcome measures per drug family, showing the percentage of users in strata showing *RTE* greater than 1.

erature about the inverse relationship observed in the occurrence of these attributes and mental health symptoms (Pennebaker, Mehl, and Niederhoffer 2003). Among the drug families, we find that the TCAs show the greatest improvement in these measures, with about half of their users showing increased positive affect and cognition.

Next, Figure 7 shows that these outcome measures decrease mostly in the lower-valued strata and increase in the higher valued ones (Figure 7). Note that these measures are not mutually exclusive. That is, an individual can see both increasing positive affect and increasing negative affect if they are using more affective words overall. The higher strata included users who typically showed lower affect and cognition than the rest (see Figure 4). Together, our findings suggest that the self-reported use of these medications is associated with ineffective (or worsening) effects on individuals with lower affective expressiveness and cognitive processing. Interestingly, these symptoms are also comorbid with mood disorders (Rosenblat et al. 2016), and the observed ineffectiveness of the drugs is likely influenced by the severity of their mental illness. However, to disentangle that requires further investigation, beyond the scope of our work.

**Effects on Depression, Anxiety, Psychosis, and Suicidal Ideation** For these second set of outcomes, we observe varied changes across medications. We observe that reported use of most of the medications are associated with worsening of these outcomes. These also include the most popular medications such as Sertraline, Escitalopram, and Fluoxetine. All of these are classified as SSRIs—the family which shows the most worsening in these outcomes among the drug families. In fact, our dataset reveals that within SSRIs, over 90% of the users were in strata that showed increased anxiety and suicidal ideation. On the other hand, we find improving symptoms in TCAs such as Dosulepin, Imipramine, and Clomipramine. From the perspective of drug families, the TCAs and the TeCAs show the greatest improving effects, with the majority of their users belonging to strata with decreased effects in the outcome measures.

Although most medications show similar effects at an aggregated level, we find differences in their strata-wise effects distributions (Figure 7). For example, in case of Duloxetine, we find minimal effects in the middle region, the one that showed high cognition (Figure 4). In contrast, Fluoxetine showed improving effects in a few lower valued strata. This observation—that the strata-wise effects can be different, inspired our next set of post-hoc analyses, wherein we examine individual-specific effects and drug-specific changes associated with the reported use of the medications.

## Understanding Individual-Treatment Effects

To understand how pre-treatment psycholinguistic signals correlate with post-treatment response to the drugs, we examine the effects at the individual level. For every *Treatment* user, we obtained their Individual Treatment Effect (*ITE*) values for all outcomes. Next, we fit several linear regression models per psychiatric medication to obtain the relationship between the *ITE*s and the psycholinguistic (LIWC) attributes of the users who reported using the medication. To simplify interpretability, corresponding to every psycholinguistic attribute, we averaged the coefficients of outcomes (preserving their directionality of improvement). For the four most popular drugs, Table 5 reports the coefficients of five psycholinguistic attributes with the greatest magnitudes in improvement or worsening. We summarize a few distinct patterns below, noted by our clinician coauthor to be most salient, based on the clinical literature and experience:

| Attribute | Coefficient | Attribute | Coefficient |
|---|---|---|---|
| Sertraline | | Fluoxetine | |
| *Past Tense* | 0.52 | *Cognitive Mech.* | 0.35 |
| *Tentativeness* | 0.35 | *Present Tense* | 0.34 |
| *1st P. Singular* | -0.18 | *Relative* | 0.31 |
| *Aux. Verbs* | -0.23 | *Percept* | 0.30 |
| *Cognitive Mech.* | -0.25 | *Conjunction* | -0.10 |
| Escitalopram | | Duloxetine | |
| *Article* | 0.22 | *Cognitive Mech.* | 0.46 |
| *1st P. Singular* | 0.10 | *Relative* | 0.44 |
| *Social* | -0.07 | *1st P. Singular* | 0.41 |
| *Bio* | -0.13 | *Social* | -0.20 |
| *2nd Person* | -0.18 | *Work* | -0.26 |

Table 5: Individual Treatment Effects: Relationship between pre-treatment attributes and improvement coefficient (Positive indicates *improvement*, Negative indicates *worsening*).

For Sertraline, the use of first person singular and auxiliary verb shows negative coefficients, indicating that this drug might not be effective in those with greater preoccupation and self-attentional focus—the known characteristics of these two attribute usage, typically prevalent in depressed individuals (De Choudhury et al. 2013). In contrast, Escitalopram and Duloxetine shows better efficacy in those individuals who have greater pre-occupation and lower social integration. Similarly, Fluoxetine and Duloxetine shows better efficacy in those individuals with greater usage of cognitive words—typically those who show lower cognitive impairment, but Sertraline shows the opposite effect in them.

# Discussion

Our work presents two significant contributions: 1) By detecting the effects of drug use and that these changes are sensitive to drug families, we show a proof of concept that social media is useful as an effective sensor to scalably detect behavioral changes in individuals who initiate treatment via (self-reported) use of psychiatric medication; and 2) our empirical findings include the discovery that people's online behaviors change in some unexpected ways following drug intake, and these may differ from the named side-effects of these drugs. We discuss the significance and implications of these contributions in the remainder of this section.

## Contextualizing the Findings in Psychiatry

As highlighted earlier, there are complexities in determining the effects of psychiatric medications in individuals; but at the same time, there are discrepancies in the claims made

by clinical studies. For example, Geddes et al. found no major differences in the efficacy of SSRIs and TCAs, whereas other studies found one kind to perform better than others (Cipriani et al. 2018). Other studies found placebos or non-pharmacological care to have outperformed certain antidepressants (Szegedi et al. 2005). These conflicting findings in the literature prevent us from drawing conclusive claims about the validity of our findings.

From the perspective of clinical literature, our results offer varied interpretations. Figure 6 indicates a small impact of antidepressants on cognitive symptoms—an observation consistent with clinical experience and studies (Rosenblat et al., 2016). It is more difficult to explain the variable impact of the drugs on depressive symptoms. For instance, in our post-hoc analysis, Sertraline showed poor effects for individuals exhibiting attributes of depression, despite clinical evidence suggesting the opposite. On the other hand, Duloxetine was associated with positive symptomatic outcomes, as also found in clinical studies (Cipriani et al. 2018). Nevertheless, that these antidepressants have varying effects on individuals across strata finds support in clinical trials which report varying efficacy of antidepressants on different cohorts (Coupland et al. 2011)

Notwithstanding these varied findings, our work highlights the potential of older antidepressants. While TCAs (Imipramine, Clomipramine) are not often prescribed today because of serious toxicity issues that may be fatal in overdose (Kerr, McGuffie, and Wilkie 2001), our results demonstrate their effectiveness with the most favorable responses reported, compared to the other classes of anti-depressants.

## Clinical Implications

**Patient-Centered Approach to Pharmacological Care** Our findings show that social media can provide valuable complementary insights into the effects of psychiatric drugs. This can complement clinical trials, allowing observations in larger populations and over longer time spans. Further, in psychiatry, medications are still prescribed by trial-and-error, or based on side effect profiles of these medications (Trivedi et al. 2006). Our analysis of individual treatment effects shows that the pre-treatment signals of mental health states appear to be linked to or predictive of individual drug success, raising the possibility of using such signals for **precision psychiatry** (Vieta 2015). While we use social media to demonstrate that this relationship exists, other sources of mental health signals might be used to complement our analyses, that are reliable and more broadly available.

**Drug Repurposing** Our results offer a novel opportunity to advance **drug repurposing**. Presently the pipeline for new pharmacological agents for mental illnesses is sparse (Dubovsky 2018), apart from ongoing research on ketamine and other potential new antidepressants (Dubovsky 2018). Drug repurposing— finding new clinical applications for currently approved medications, offering the potential of low cost and quicker to market treatments (Corsello et al. 2017). So far drug repurposing efforts in psychiatric illnesses like depression have focused on biological targets (Powell et al. 2017); this is the first research to explore how social media may serve to identify novel targets as well. Further, although these approaches have been successful in identifying plausible repositioning candidates, a key challenge is providing direct evidence of candidate efficacy in people, rather than relying on surrogate biomarkers or indirect evidence. Our methods highlight how large quantities of real-time data can offer low cost and high volume assessments of people's own reports and perceptions related to antidepressants' use.

## Technological Implications

**Technologies for Regulatory Bodies** Our results offer an important tool in generating "real-world evidence" for incorporation into technologies that can be used by regulatory bodies like the FDA. The FDA seeks to advance its approach to regulate and rely more on real-world evidence in addition to pre-market clinical studies data. As the FDA currently writes its novel digital health software program certification plan, where medical software such as smartphone applications will receive FDA approval without extensive clinical research—it has stated that a key component is stated to be "monitoring real-world performance", though noting that they are still "considering how to best work to collect and interpret information about the product's safety and effectiveness" (*fda.gov*). This paper offers a novel technological approach that may meet the evolving needs of the FDA, by being able to identify the uses and effects of various medication as self-reported by people on social media.

**Technologies for Drug Safety Surveillance** From a public health perspective, our methods offer the potential to build technologies that surface early warning signs of adverse effects related to psychiatric drug use. The FDA's current Sentinel Initiative which aims to apply big data methods to medical claims data from over 5.5 billion patient encounters in an effort to flag previously unrecognized drug safety issues and tackle issues of under-reporting of drug effects, has still not superseded traditional reporting directly from physicians or pharmaceutical companies (Kuehn 2016). The data gathered in this paper–even if it only represents a subpopulation of those who use social media, offers a new lens onto specific groups of people who may have less or more extreme reactions to medications. Including this information in technologies for drug safety monitoring can thus complement traditional sources, and improve awareness regarding emerging safety issues in a spontaneous fashion —serving as sentinels prompting further exploration in pharmacovigilance research.

**Technologies to Support Digital Therapeutics** Psychiatrists' view and knowledge of a patient's health is often limited to self-reports and information gathered during in-person therapeutic visits (Vieta 2015). This paper provides a new source of collateral information to support digital therapeutics (Fisher and Appelbaum 2017) and enhance evidence-based, personalized pharmacological treatment. Specifically, it reveals the potential to build technologies that can augment information seeking practices of clinicians. E.g., with patient consent, clinicians can learn about the specific effects and symptomatic expressions shared by

patients in the natural course of their lives, and beyond the realms of the therapeutic setting. Further, the awareness of the effects of psychiatric medications in specific patients can lead to improved toxicovigilance related interventions. Further, given the risks posed by prescription drug overdose and abuse (McKenzie and McFarland 2007), increased and finer-grained awareness of the effects of psychiatric medications in specific patients as well as identifying any medication abuse patterns can lead to improved toxicovigilance related interventions.

## Policy and Ethics

Despite the potential highlighted above to build novel technologies for regulatory authorities, guidelines on how social media signals should be handled, and their use in the surveillance of the effects of drugs do not yet exist. Although the FDA has released two guidelines on the use of social media for the risk-benefit analysis of prescription drugs (Sarker et al. 2015), they focus on product promotion and "do not establish legally enforceable rights or responsibilities" (FDA 2014). Therefore, the potential (unintended) negative consequences of this work must be considered.

Note that the clinical and technological implications rest upon the names of the medications not being anonymized. We recognize that this surfaces new ethical complexities. For example, while understanding what medications work for which individuals may facilitate "patient-centered" insurance coverage decisions, it can also be (mis)used to decline coverage of specific drugs resulting in "health inequality". Additionally, patients may blindly adopt these findings creating tension in their therapeutic relationship with their clinicians, causing a decrease in medication adherence. We suggest further research investigating and mitigating such potential unintended consequences of the work.

## Limitations and Conclusion

We recognize that this study suffers from limitations, and some of these suggest promising directions for future work. Our results on the varied effects of psychiatric medications are likely to be influenced by *selection bias* in those who choose to publicly self-report their medication use on social media. This is especially true given the stigma around mental illness (Corrigan 2004), which is a known obstacle to connecting individuals with mental healthcare. We cannot verify if self-reports of medication use corresponded to their actual use (Ernala et al. 2019). Therefore, the users in our data who chose to self-report their medication usage may represent unique populations with lowered inhibitions. Self-report bias further complicates the types of effects that we observed—different individuals respond differently, as shown in our results, however, our observations are limited to only the types of effects that characterize the individuals in our data. For these same reasons of sampling bias, *we caution against drawing population-wide generalizations* of the effects of psychiatric medication usage.

Despite adopting a causal framework that minimizes confounding effects, *we cannot establish true causality*, and our results are plausibly influenced by the severity in the clinical condition of the individuals. While we considered many confounders in our propensity score matching approach, there are other latent factors that could impact the effects considered here; e.g., duration, history, dosage, and compliance of using self-reported medications; additional medications or adjuvant treatments one might be using. Further, future work can adopt methods such as location-based filtering to better account for geo-cultural and linguistic confounds. Additionally, self-reporting bias about medications can lead to treatment leakage, where some control individuals may be taking medications, but not mentioning it on Twitter.

*Our work is not intended as a replacement for clinical trials.* In fact, social media lacks many features that clinical trials possess. First, we do not have the notion of a placebo, used to eliminate the confound that simply the perception of receiving a treatment produces non-specific effects. Second, even though we match users based on several characteristics, we do not pre-qualify individuals as potential beneficiaries of a medication. Last, social media analysis does not allow us to closely monitor the treatment, unlike a clinical trial, which results in high variance in the number of measurements that each individual contributes.

Despite corroboration by a psychiatrist, we are limited by what can be observed from an individual's social media data. Without complementary offline information (e.g., the people's physiologies), we cannot ascertain the clinical nature of the mental health outcomes in our data. Further, the symptomatic outcomes themselves, such as measures of depression or suicidal ideation, need additional clinical validation, e.g., based on DSM-5 criteria (APA and others 2013), or the Research Domain Criteria (RDoC) introduced by the National Institutes of Mental Health (Insel et al. 2010). *Without dampening the clinical potentials, we caution against making direct clinical inferences.* Still, while we acknowledge that the medical community rarely adopts the most innovative approaches for immediate use, this work can inspire replication studies in patient populations.

In conclusion, our work represents a novel dynamic viewpoint onto mental health— limitations notwithstanding, it captures the real-time variation and accounts for dynamic systems theory, network theory, and instability mechanisms (Nelson et al. 2017). Such a new window onto the field clearly contrasts the traditional static viewpoint on the effects of psychiatric medications. It warrants further research in this evolving space and opens up interesting opportunities beyond existing reporting methodologies.


## Acknowledgement

We thank the members of the Social Dynamics and Well-being Lab at Georgia Tech for their valuable feedback. Saha and De Choudhury were partly supported by NIH grant #R01GM112697. Torous was supported by a patient-oriented research career development award (K23) from NIMH #1K23MH116130-01. Abrahao was supported by a National Natural Science Foundation of China (NSFC) grant #61850410536 and developed part of this research while affiliated with Microsoft Research AI, Redmond.

## References


APA, et al. 2013. *Diagnostic and statistical manual of mental disorders, (DSM-5)*. American Psychiatric Pub.

Bagroy, S.; Kumaraguru, P.; and De Choudhury, M. 2017. A social media based index of mental well-being in college campuses. In *Proc. CHI*.

Barchas, J., and Altemus, M. 1999. Monoamine hypotheses of mood disorders. *Basic Neurochemistry*.

Birnbaum, M. L.; Ernala, S. K.; Rizvi, A. F.; De Choudhury, M.; and Kane, J. M. 2017. A collaborative approach to identifying social media markers of schizophrenia by employing machine learning and clinical appraisals. *J Med Internet Res*.

Blanco, C.; Okuda, M.; Wright, C.; Hasin, D. S.; Grant, B. F.; Liu, S.-M.; and Olfson, M. 2008. Mental health of college students and their non-college-attending peers: results from the national epidemiologic study on alcohol and related conditions. *Arch Gen Psy*.

Bull, S. A.; Hu, X. H.; Hunkeler, E. M.; Lee, J. Y.; Ming, E. E.; Markson, L. E.; and Fireman, B. 2002. Discontinuation of use and switching of antidepressants: influence of patient-physician communication. *Jama*.

Burnap, P.; Colombo, W.; and Scourfield, J. 2015. Machine classification and analysis of suicide-related communication on twitter. In *Proc. ACM conference on hypertext & social media*.

Chancellor, S.; Lin, Z.; Goodman, E. L.; Zerwas, S.; and De Choudhury, M. 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*, 1171–1184. ACM.

Chancellor, S.; Nitzburg, G.; Hu, A.; Zampieri, F.; and De Choudhury, M. 2019. Discovering alternative treatments for opioid use recovery using social media. In *Proc. CHI*.

Chandrasekharan, E.; Samory, M.; Jhaver, S.; Charvat, H.; Bruckman, A.; Lampe, C.; Eisenstein, J.; and Gilbert, E. 2018. The internet's hidden rules: An empirical study of reddit norm violations at micro, meso, and macro scales. *PACM HCI* (CSCW).

Cipriani, A.; Furukawa, T. A.; Salanti, G.; Geddes, J. R.; Higgins, J. P.; Churchill, R.; Watanabe, N.; Nakagawa, A.; Omori, I. M.; McGuire, H.; et al. 2009. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. *The lancet* 373(9665):746–758.

Cipriani, A.; Furukawa, T. A.; Salanti, G.; Chaimani, A.; Atkinson, L. Z.; Ogawa, Y.; Leucht, S.; Ruhe, H. G.; Turner, E. H.; Higgins, J. P.; et al. 2018. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *The Lancet* 391(10128):1357–1366.

Coppersmith, G.; Harman, C.; and Dredze, M. 2014. Measuring post traumatic stress disorder in twitter. In *ICWSM*.

Corrigan, P. 2004. How stigma interferes with mental health care. *American Psychologist* 59(7):614.

Corsello, S. M.; Bittker, J. A.; Liu, Z.; Gould, J.; McCarren, P.; Hirschman, J. E.; Johnston, S. E.; Vrcic, A.; Wong, B.; Khan, M.; et al. 2017. The drug repurposing hub: a next-generation drug library and information resource. *Nature medicine* 23(4):405.

Coupland, C.; Dhiman, P.; Morriss, R.; Arthur, A.; Barton, G.; and Hippisley-Cox, J. 2011. Antidepressant use and risk of adverse outcomes in older people: population based cohort study. *Bmj*.

De Choudhury, M.; Gamon, M.; Counts, S.; and Horvitz, E. 2013. Predicting depression via social media. In *ICWSM*.

De Choudhury, M.; Kiciman, E.; Dredze, M.; Coppersmith, G.; and Kumar, M. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *CHI*.

Dos Reis, V. L., and Culotta, A. 2015. Using matched samples to estimate the effects of exercise on mental health from twitter.

Dubovsky, S. L. 2018. What is new about new antidepressants? *Psychotherapy and psychosomatics*.

Dutta, S.; Ma, J.; and De Choudhury, M. 2018. Measuring the impact of anxiety on online social interactions. In *ICWSM*.

Ernala, S. K.; Rizvi, A. F.; Birnbaum, M. L.; Kane, J. M.; and De Choudhury, M. 2017. Linguistic markers indicating therapeutic outcomes of social media disclosures of schizophrenia. *CSCW*.

Ernala, S. K.; Birnbaum, M. L.; Candan, K. A.; Rizvi, A. F.; Sterling, W. A.; Kane, J. M.; and De Choudhury, M. 2019. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *ACM CHI*.

FDA, et al. 2014. Guidance for industry: internet/social media platforms with character space limitations; presenting risk and benefit information for prescription drugs and medical devices.

Fisher, C. E., and Appelbaum, P. S. 2017. Beyond googling: The ethics of using patients' electronic footprints in psychiatric practice. *Harvard review of psychiatry* 25(4):170–179.

Frommer, D. A.; Kulig, K. W.; Marx, J. A.; and Rumack, B. 1987. Tricyclic antidepressant overdose: a review. *Jama*.

Gaur, M.; Kursuncu, U.; Alambo, A.; Sheth, A.; Daniulaityte, R.; Thirunarayan, K.; and Pathak, J. 2018. Let me tell you about your mental health!: Contextualized classification of reddit posts to dsm-5 for web-based intervention. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 753–762. ACM.

Geddes, J.; Freemantle, N.; Mason, J.; Eccles, M.; and Boynton, J. 2000. Selective serotonin reuptake inhibitors (ssris) versus other antidepressants for depression. *Cochrane Database Syst Rev*.

Gillman, P. K. 2007. Tricyclic antidepressant pharmacology and therapeutic drug interactions updated. *Br. J. Pharmacol.*

Hannan, E. L. 2008. Randomized clinical trials and observational studies: guidelines for assessing respective strengths and limitations. *JACC: Cardiovascular Interventions* 1(3):211–217.

Harpaz, R.; DuMouchel, W.; Schuemie, M. J.; Bodenreider, O.; Friedman, C.; Horvitz, E.; et al. 2017. Toward multimodal signal detection of adverse drug reactions. *J. Biomed. Inform.*

Hasin, D. S.; Goodwin, R. D.; Stinson, F. S.; and Grant, B. F. 2005. Epidemiology of major depressive disorder: results from the national epidemiologic survey on alcoholism and related conditions. *Arch. Gen. Psychiatry*.

Imbens, G. W., and Rubin, D. B. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge.

Insel, T.; Cuthbert, B.; Garvey, M.; Heinssen, R.; Pine, D. S.; Quinn, K.; Sanislow, C.; and Wang, P. 2010. Research domain criteria (rdoc): toward a new classification framework for research on mental disorders.

Kerr, G.; McGuffie, A.; and Wilkie, S. 2001. Tricyclic antidepressant overdose: a review. *Emergency Medicine Journal*.

Kiciman, E.; Counts, S.; and Gasser, M. 2018. Using longitudinal social media analysis to understand the effects of early college alcohol use. In *ICWSM*.

Klein, A.; Sarker, A.; Rouhizadeh, M.; O'Connor, K.; and Gonzalez, G. 2017. Detecting personal medication intake in twitter: An annotated corpus and baseline classification system. *BioNLP 2017*.

Kuehn, B. M. 2016. Fdaś foray into big data still maturing. *Jama*.

Lamont, A.; Lyons, M. D.; Jaki, T.; Stuart, E.; Feaster, D. J.; et al. 2018. Identification of predicted individual treatment effects in randomized clinical trials. *Stat. Methods Med. Res*.

Lardon, J.; Abdellaoui, R.; Bellet, F.; Asfari, H.; Souvignet, J.; Texier, N.; Jaulent, M.-C.; Beyens, M.-N.; Burgun, A.; and Bousquet, C. 2015. Adverse drug reaction identification and extraction in social media: A scoping review. *J. Med. Internet Res*.

Lexchin, J.; Bero, L. A.; Djulbegovic, B.; and Clark, O. 2003. Pharmaceutical industry sponsorship and research outcome and quality: systematic review. *Bmj* 326(7400):1167–1170.

Lin, H.; Jia, J.; Guo, Q.; Xue, Y.; Li, Q.; Huang, J.; Cai, L.; and Feng, L. 2014. User-level psychological stress detection from social media using deep neural network. In *Proc. ACM Multimedia*.

Liu, J.; Weitzman, E. R.; and Chunara, R. 2017. Assessing behavioral stages from social media data. In *CSCW*.

Lopez-Munoz, F., and Alamo, C. 2009. Monoaminergic neurotransmission: The history of the discovery of antidepressants from 1950s until today. *Current Pharmaceutical Design*.

MacLean, D.; Gupta, S.; Lembke, A.; Manning, C.; and Heer, J. 2015. Forum77: An analysis of an online health forum dedicated to addiction recovery. In *CSCW*.

McKenzie, M. S., and McFarland, B. H. 2007. Trends in antidepressant overdoses. *Pharmacoepidemiology and drug safety*.

Moncrieff, J., and Cohen, D. 2009. How do psychiatric drugs work? *BMJ* 338:1535.

Nelson, B.; McGorry, P. D.; Wichers, M.; Wigman, J. T.; and Hartmann, J. A. 2017. Moving from static to dynamic models of the onset of mental disorder: a review. *JAMA psychiatry*.

Nikfarjam, A.; Sarker, A.; Oconnor, K.; Ginn, R.; and Gonzalez, G. 2015. Pharmacovigilance from social media: mining adverse drug reaction mentions using sequence labeling with word embedding cluster features. *J. Am. Med. Inform.*

Oliver, D.; Davies, C.; Crossland, G.; Lim, S.; Gifford, G.; McGuire, P.; and Fusar-Poli, P. 2018. Can we reduce the duration of untreated psychosis? a systematic review and meta-analysis of controlled interventional studies. *Schizophrenia bulletin*.

Olteanu, A.; Varol, O.; and Kiciman, E. 2017. Distilling the outcomes of personal experiences: A propensity-scored analysis of social media. In *Proc. CSCW*.

Pavalanathan, U., and Eisenstein, J. 2016. More emojis, less:) the competition for paralinguistic function in microblog writing.

Pennebaker, J. W.; Mehl, M. R.; and Niederhoffer, K. G. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54(1):547–577.

Powell, T. R.; Murphy, T.; Lee, S. H.; Price, J.; Thuret, S.; and Breen, G. 2017. Transcriptomic profiling of human hippocampal progenitor cells treated with antidepressants and its application in drug repositioning. *Journal of Psychopharmacology*.

Rosenblat, J. D.; Kakar, R.; and McIntyre, R. S. 2016. The cognitive effects of antidepressants in major depressive disorder: a systematic review and meta-analysis of randomized clinical trials. *International Journal of Neuropsychopharmacology* 19(2).

Saha, K., and De Choudhury, M. 2017. Modeling stress with social media around incidents of gun violence on college campuses.

Saha, K.; Chan, L.; De Barbaro, K.; Abowd, G. D.; and De Choudhury, M. 2017. Inferring mood instability on social media by leveraging ecological momentary assessments. *Proc. ACM IMWUT*.

Saha, K.; Bayraktaraglu, A. E.; Campbell, A.; Chawla, N. V.; De Choudhury, M.; D'Mello, S. K.; Dey, A. K.; Gao, G.; Gregg, J. M.; Jagannath, K.; Mark, G.; Martinez, G. J.; Mattingly, S. M.; Moskal, E.; Sirigiri, A.; Striegel, A.; and Yoo, D. W. 2019a. Social media as a passive sensor in longitudinal studies of human behavior and wellbeing. In *CHI Ext. Abstracts*. ACM.

Saha, K.; Torous, J.; Ernala, S. K.; Rizuto, C.; Stafford, A.; and De Choudhury, M. 2019b. A computational study of mental health awareness campaigns on social media. *Translational behavioral medicine*.

Saha, K.; Weber, I.; and De Choudhury, M. 2018. A social media based examination of the effects of counseling recommendations after student deaths on college campuses. In *ICWSM*.

Sarker, A.; Ginn, R.; Nikfarjam, A.; OConnor, K.; Smith, K.; Jayaraman, S.; Upadhaya, T.; and Gonzalez, G. 2015. Utilizing social media data for pharmacovigilance: a review. *J. Biomed. Inform.*

Shippee, N. D.; Shah, N. D.; May, C. R.; Mair, F. S.; and Montori, V. M. 2012. Cumulative complexity: a functional, patient-centered model of patient complexity can improve research and practice. *Journal of clinical epidemiology* 65(10):1041–1051.

Szegedi, A.; Kohnen, R.; Dienel, A.; and Kieser, M. 2005. Acute treatment of moderate to severe depression with hypericum extract ws 5570 (st john's wort): randomised controlled double blind non-inferiority trial versus paroxetine. *Bmj* 330(7490):503.

Tausczik, Y. R., and Pennebaker, J. W. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology* 29(1):24–54.

Trivedi, M.; Rush, A.; Wisniewski, S.; et al. 2006. Evaluation of outcomes with citalopram for depression using measurement-based care in star* d: implications for clinical practice. *Am. J. Psychiatry*.

Vieta, E. 2015. Personalised medicine applied to mental health: precision psychiatry. *Revista de psiquiatria y salud mental*.

White, R. W.; Wang, S.; Pant, A.; Harpaz, R.; Shukla, P.; Sun, W.; DuMouchel, W.; and Horvitz, E. 2016. Early identification of adverse drug reactions from search log data. *J. Biomed. Inform.*

WHO. 2002. The importance of pharmacovigilance. *WHO*.

Yoo, D. W., and De Choudhury, M. 2019. Designing dashboard for campus stakeholders to support college student mental health. In *Pervasive Health*.