Behind the Screen: A Narrative Review of Advancements in Mobile and Wearable Passive Sensing for Mental Health Assessment

Anastasia C. Bryan[1,†], Michael V. Heinz[1,2,†], Abigail J. Salzhauer[1], George D. Price[1,3], M.L. Tlachac[5,6], and Nicholas C. Jacobson[1,2,3,4]

[1]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

[2]Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Hanover, NH, United States

[3]Quantitative Biomedical Sciences Program, Dartmouth College, Hanover, NH, United States

[4]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

[5]Department of Information Systems and Analytics, Bryant University, RI, United States

[6]Center for Health and Behavioral Sciences, Bryant University, RI, United States

† These Authors contributed equally to this work.

Anastasia Bryan https://orcid.org/0009-0005-6603-0266
Michael V. Heinz https://orcid.org/0000-0003-0866-0508
Abigail Salzhauer https://orcid.org/0009-0009-3210-6807
George Price https://orcid.org/0000-0002-9164-4973
M.L. Tlachac 0000-0002-6634-678X
Nicholas C. Jacobson https://orcid.org/0000-0002-8832-4741

Keywords: Passive Sensing, Digital Mental Health, Mental Health Assessment, Digital Phenotyping, Digital Biomarkers, Ubiquitous Sensing

Abstract

Mental health disorders—including depression, anxiety, trauma-related, and psychotic conditions—are pervasive and impairing, representing considerable challenges for both individual well-being and public health. Often the first challenge to treatment can be financial, geographic, and stigmatic barriers which limit the accessibility of traditional assessment measures. Further, compounded by frequent misdiagnosis or delayed detection, there is a need for effective, accessible, and scalable approaches to identification and management. Leveraging advances in computing and the ubiquitous nature of personal mobile and wearable technology, this narrative review examines the utilization of passive sensor data as a screening and diagnostic tool for mental disorders. As an alternative to traditional screening measures, passive sensing offers a tool to overcome barriers which prevent many from seeking services. We critically assess the literature up to September 2023, exploring the use of passive data—such as heart rate variability, movement patterns, and geolocation—to predict mental health outcomes across a spectrum of disorders. Through a translational perspective, our review explores the state of passive sensing science, with special emphasis on the capacity for the science to be implemented in real world clinical and general populations. To this aim, we consider study designs, including participant demographics, data collection methods, sensor modalities, outcome measures, and analytic modeling approaches. Our findings highlight overarching limitations in the field including 1) the use of smaller, specialized sample populations, and 2) the predominant use of Android operating systems, 3) a reliance on self-reported measures as proxies for mental health outcomes which ultimately limit the potential to provide robust mental health assessment in larger population samples. We suggest that further research incorporate larger and more diverse samples, inclusion of further smartphone operating systems, and additional clinical-related assessments to strengthen predictive models and maintain a clinician in the loop. Despite these limitations, passive sensing technologies like GPS, heart rate monitoring, and actigraphy offer promise for enhancing early detection and improving the diagnostic process for mental disorders. We conclude that careful consideration of translational factors in the design of future research will aid in enhancing the potential and scope of future passive sensing studies, ultimately enhancing mental health outcomes on a broad scale.

Introduction

Mental Disorders, including mood, anxiety, trauma-related, and psychotic disorders are highly burdensome to both individuals and society, posing a considerable public health burden worldwide [1]. Mental health disorders pose significant difficulty to effective screening and diagnosis, due to large differences in individual symptom presentation [2]–[4] as well as their heavy reliance on retrospective self report [5], often in the form of surveys. Although categorical diagnosis was a major step forward in the field of mental health, introducing a “common language” for researchers and clinicians to use [6], categorical diagnoses lack the capacity to account for the tremendous heterogeneity in symptom profiles of these disorders. Further, assessment and diagnosis of mental disorders has traditionally relied heavily on retrospective self report, requiring considerable subjective recall from the patient, known to be impacted by multiple biases [5]. Mental health diagnosis and often screening requires speaking with a health professional, which is costly and resource intensive, with the demand for services far outweighing the availability. As well, this process often results in a relatively stagnant or stable diagnosis, an assumption established by the DSM-5, that limits the ability for researchers to study how symptoms and diagnoses change over time–inherently limiting the nuance possible in our understanding of mental health. Given these considerations, it is perhaps not surprising that despite 1 in 5 U.S. adults experiencing mental illness, only 47.2% of U.S. adults received treatment in 2021 [7].

Considering the limitations of traditional mental health assessment, alternatives have been proposed to (1) improve upon the current categorical framework and (2) augment existing methods for screening and assessment. Initiatives such as National Institute of Mental Health (NIMH)’s Research Domain Criteria (RDoC) [8] and the The Hierarchical Taxonomy of Psychopathology (HiTOP) [9] have emphasized the need for a *dimensional approach* to mental disorders, integrating objective, transdiagnostic physiologic and behavioral data to better understand human experience. Complementing these initiatives, *digital phenotyping* [10] has grown in translational research popularity as a means by which to understand an individual's behavior, cognitive function, mood, and overall mental health moment-by-moment using data from personal digital devices. Closely related to digital phenotyping is the concept of *passive sensing* – the utilization of data from sensors, not requiring active user participation or engagement, such as accelerometers or GPS, often built into personal mobile devices. Indeed, digital phenotyping is often accomplished through passively collected streams of data.

Given the ubiquity of mobile devices and growing popularity of consumer wearables [11], [12], such as smartwatches and fitness bands, the ongoing development of digital phenotypes for mental health through passive sensing has greater potential than ever before to improve mental health screening at scale. The increase in consumer grade mobile devices has been accompanied by considerable advances in hardware (including development and deployment of new sensors, made possible by advances in material science) and software [13], which has enabled the real time sensing and inference of many naturalistic behaviors and physiologic changes. These advances combined with the concurrent exponential growth in computing power [14] mean that important mental health outcomes can be modeled using highly dimensional data from personal devices, and such models can be effectively utilized for inference. Many studies to date have explored the use of such data within the mental health domain [15], [16], identifying promising results in mood [17], anxiety [18], [19], trauma-related [20], and psychotic disorders [21], as well for other important mental health outcomes such as suicide [22].

The volume of such studies has recently grown very rapidly [23], with promising results discovered using many passively collected data types. For example, call and text metadata [24], location via GPS [25], passive movement [26], and heart rate [22] have shown promise in mental health assessment. When considering the potential for the findings from such studies to be translated and implemented in general population and clinical samples, we find that there is great variability among studies. Such variability includes

differences across study sample size, generalizability, outcome, model performance, and overall clinical usefulness of the modeled phenomena.

In the current narrative review, we propose to holistically consider studies utilizing passively collected information from personal or wearable devices to predict mental health outcomes. In preemptive defense of the current writing becoming “yet another digital phenotyping review,” we aim to examine such studies in light of their direct translational value, that is, their capacity to directly guide the development of strategies and methodologies for screening and diagnosis in clinical and general population samples. In so doing, we examine the study size, the generalizability of the sample, the modeling approach and performance, the clinical outcomes considered, and the safety and accessibility for users, including privacy considerations. Further, unlike many disorder-specific reviews to date, we intentionally set a broad scope for our review, given our observation that many studies examine multiple pathologies with multiple sensors (hereafter referred to as multimodal) and our supposition that the synthesis of studies across pathologies may reveal important transdiagnostic phenotypes.

Background

Mental disorders touch the lives of countless people, across the boundaries of lifespan and demographics impacting psychological, physiological, and social-well being. As of 2020, 1 in 5 adults had experienced mental illness in the U.S. with only 46.2% of those having received treatment. For those who pursued treatment after disorder onset, the average delay reached 11 years [7][27]. These factors have tremendous cost for society with studies showing that mental illnesses increase the risk of cardiovascular and metabolic diseases [28], [29], increase the risk of substance use disorders [7], and lead to a trillion dollars lost in global productivity every year [30], in addition to the personal and community toll. Despite this need, significant barriers prevent access to mental health care, including financial cost, geographic distance, stigma, and a shortage of care providers. The Health Resources and Services Administration reports that over 164 million individuals in the U.S. alone live in areas identified as mental health care shortage areas, a number that is 64 million more than primary care shortage areas [31]. These statistics underline the global burden of mental illness, as well, which is underestimated by more than a third given the overlap between psychiatric and neurological disorders, the consideration of suicide/self-harm as a separate category, and inadequate consideration of severe mental illness contributing to mortality from associated causes [32]. Further, mental disorders are highly comorbid, both with other non-psychiatric illnesses and with other mental health disorders, such that comorbidity is often the rule and not the exception [33]. Suicide, considered in most cases to be the most serious adverse outcome, is an important consequence of untreated mental illness [34]. Ultimately, the cost of not treating mental disorders is too great, emphasizing the need for changes that will make mental health care more accessible.

Current mental health care practice takes a categorical approach to assessment and diagnosis [6], in which a patient is evaluated by a set of criteria as outlined in standard reference sources [35]. The Diagnostic Statistical Manual-5 (DSM-5) and International Classification of Diseases-11 (ICD-11) are widely used diagnostic references in the United States and internationally, with the key difference being the DSM-5’s focus on mental disorders and the ICD-11’s focus more broadly on medical disorders [6], [36].

Traditionally mental health disorders have been screened for and diagnosed by some combination of subjective patient self report, objective observation by a trained clinician, and sometimes collateral information provided by relations of the patient. However, assessments conducted by these methods are often time consuming, resource intensive and thus have limited potential for scalability. Over the last several decades, self-administered screening questionnaires have been increasingly utilized in large populations as a more efficient screening tool. Certain questionnaires, like those to screen for depression, have been integrated into

many primary care practices to increase detection rates [37]. There are a number of validated screening instruments for important mental health conditions including the 9-item Patient Health depression Questionnaire (PHQ-9) and the 7-item Generalized Anxiety Disorder (GAD-7) questionnaire. Many of these self-report screening measures require patients to recall symptoms over an extended period of time. The PHQ-9, for instance, requires reflecting on symptom frequency over the past two weeks [38]. Although validated on large clinical samples, instruments which require retrospective self report are subject to known reporting biases, such as recall bias [39], limiting their effectiveness. Further, retrospective assessments can be obtrusive and lack temporal resolution for changes in symptoms over time, imposing strong assumptions about how symptoms occur within an individual's daily life, namely that symptoms are fixed over the recall window. This limits the granularity with which an individual or clinician can understand one's symptoms over time, oversights which can minimize opportunities to identify meaningful symptom trends, contextual associations, and utilize targeted interventions.

Potentially mitigating some of the problems with retrospective screeners, ecological momentary assessment (EMA), extensively utilized in mental health passive sensing research [22], [40], repeatedly samples an individual's behaviors and experiences [41]. It offers insights into an individual's real-time experiences over time and is less susceptible to recall bias, providing an ecologically valid measure. An illustration of EMA is an individual self-reporting symptoms of depression three times daily through a mobile device for 90 days [42]. In our review, we do not include studies where EMA serves as a predictor; however, we do examine numerous studies that employ EMA as an outcome. EMA requires more active engagement on the part of an individual to assess their mood and behaviors, which comparatively does not have the same benefit of passive sensing inherently requiring very little active engagement. These studies often use passive sensing to predict an EMA outcome, utilizing a sliding window over time. EMA as an assessment tool, however, is similarly limited by its high obtrusiveness and time burden.

Ultimately, the categorical approach to understanding mental illness renders a dependency on self-reported screening instruments to facilitate broad mental health assessment. While EMAs have also been used heavily in passive sensing research to circumvent some of the bias related issues associated with traditional screeners, they alone lack the unobtrusive and in-real time aspects which favor passive sensing technologies instead. While a number of these screening instruments and EMAs have and continue to be highly useful for mental health assessment, the prevalence of mental health disorders is too high and alternative methods to transform the field are required to begin to address this need for care.

Background on Machine Learning

Machine learning holds great potential value for clinical psychology and psychiatry. Dwyer et al. [43] identified three primary tasks for which machine learning could be helpful: diagnosis, prognosis, and treatment. While all of these tasks are crucial to improving mental health care, the delivery of effective treatment requires the accurate assessment of affected persons, who are often identified initially by screening. There is thus a wide breadth of digital mental health research focused on diverse assessment objectives including binary screening, severity assessment, onset detection, differential diagnosis, and trajectory assessment. Within the digital mental health space, data used to train models for disorder identification, severity assessment, and diagnosis most often utilize labels derived from traditional retrospective screening measures, such as the PHQ-9 [38]. Single retrospective labels, however, are insufficient for nuanced time-related objectives, such as mapping trajectory over time. Thus, EMA is also utilized as an outcome in passive sensing research, e.g., [44], to explore symptom-change over time.

Predictive modeling is essential for achieving these mental disorder assessment objectives. Fundamentally, predictive modeling involves training machine learning models on labeled datasets, followed by evaluating their performance on previously unseen data [5]. A common method for continued revisions of

predictive models is cross-validation. This approach partitions the dataset into multiple folds, each fold taking a turn as the evaluation set, allowing for adjustment of model settings, or hyperparameters. While cross-validation is sometimes the primary evaluation metric reported, especially in cases of limited sample sizes, a more robust approach involves using a completely separate and held-out test set. The held-out test set, which remains unseen during both model training and hyperparameter tuning, offers a stronger assessment of model performance and generalizability. This approach is particularly common in benchmarking deep learning models, often in conjunction with cross-validation. The choice of evaluation strategy in digital health studies typically hinges on sample size, which often tends to be relatively small.

A classification model is most commonly used for binary screening, onset detection, and differential diagnosis. Such models output a true positive, true negative, false positive, or false negative prediction for every instance in the test set. To assess performance over the entirety of the test set, these counts can be aggregated into evaluation metrics such as sensitivity (the true positive rate, and also commonly known as recall) and specificity (the true negative rate), which are common metrics for assessing diagnostic models. Balanced accuracy [45], which is the average of these metrics, is appropriate for both balanced and unbalanced datasets. However, sometimes an outcome is rare which results in a severely unbalanced dataset; the popular F1 score is often used in such cases, though has been criticized as being potentially misleading due to the equal weighting of precision (also known as positive predictive value) and recall (also known as sensitivity) by default [46]. AUC (area under the receiver operating characteristic [ROC] curve) is a broadly applicable classification metric. The ROC curve is constructed by plotting sensitivity against the false positive rate (calculated by 1-specificity) across varying cutpoints for the model's probability score output, providing a more nuanced model evaluation across varying cutpoints. While accuracy is also a common classification metric, it is the ratio of correct predictions which is inappropriate for unbalanced datasets since a high accuracy can be obtained through only predicting the majority class. All the classification metrics presented so far are bounded between 0 and 1, with higher scores indicating greater predictive success. For research that uses multiple test sets, the average and standard deviation of these evaluation metrics are often reported.

In contrast, for severity assessment models, a regression model is used to output numeric predictions. These numeric predictions are then compared to the numeric labels in the test set. Like for classification models, there are multiple metrics to assess the effectiveness of regression models. The most common of these evaluation metrics are root mean squared error (RMSE) and mean absolute error (MAE). The latter, which is the average distance between the predicted and real scores, is the most intuitive of the regression evaluation metrics. As there are no discrete classes, the distribution and range of the numeric labels must inform the selection and interpretation of the regression metrics. In order to compare such metrics across studies, normalized versions, computed by dividing the metric by the range of the outcome measure, can be utilized.

Predictive machine learning models can largely be classified as traditional machine learning models or deep learning models. Traditional machine learning models require explicitly engineered features, which offers explainability but unfortunately limits the models to human engineered features based on existing domain knowledge. Gaussian naive bayes, support vector machine (SVM), k-nearest neighbor (kNN), logistic regression, and decision tree are common classification models [47]. Tree ensembles including random forest (RF) [47], adaptive boosting (AdaBoost) [47], and extreme gradient boosting (XGBoost) [48] are also popular. Some of these methods, like those based on decision trees, can also output numeric predictions, though linear regression is the most common regression model. These traditional machine learning models often have the advantage over deep learning models in health research given the small datasets and need for interpretability [43]. While deep transfer learning models have overcome this challenge for certain modalities like text and audio [49], [50], time series are more common in passive sensor data. Recurrent Neural Networks (RNN) are deep learning models for sequential data like time series [51], with popular variants being Long Short-Term

Memory networks (LSTMs) [52] and Gated Recurrent Units (GRUs) [53]. While some studies have investigated interpretability among deep learning models operating on time series[54], [55], such methods are relatively new and less popular.

While predictive models are required for the diagnosis related objectives, exploratory models can provide important information to guide the development of the predictive models. Exploratory modeling is particularly common for healthcare datasets which do not always contain sufficient data instances for predictive models. For example, linear models describe the relationship between predictors and an outcome and generally report a correlation (*r)* between individual predictors and the outcome; linear mixed effects models, also known as hierarchical models, combine fixed effects (e.g., age, race) which impact all data points consistently with random effects, which account for individual variations not explained by fixed effects [56]. Linear mixed models are especially useful in analyzing data from multivariate or longitudinal studies, allowing for the exploration of both general trends and unique individual or group variations. While exploratory models can be valuable to guide research, they do not share the same translational value as predictive models, which could be implemented in clinical settings.

Background on Passive Sensing

The digital health community has been exploring the ability to replace or augment traditional mental health disorder screening surveys with machine learning models that use a variety of screening modalities. The goal of such research is typically to improve the ability to screen for mental health disorders over more traditional screening approaches, to increase the rates of mental health disorder screening by lowering the participant burden, and predict long and short term outcomes. Digital phenotype data and digital biomarker data are of particular interest as screening modalities. Digital phenotyping [10] is defined as the use of digital sensors and data, especially from smartphones and wearables to understand individual experiences and behaviors as they relate to disease. Uncovered via digital phenotyping, *digital biomarkers* comprise those objective behavioral and physiologic data with particular capacity to detect, predict, or understand important mental health related outcomes.

In this narrative review, we focus on the digital phenotype data that is produced from mobile and wearable passive sensors. Modern smartphones have many embedded sensors including microphones, accelerometers, gyroscopes, and ambient light sensors. Further digital phenotype data can be extracted from smartphones such as communication logs, location logs, Bluetooth logs, WiFi connectivity, app usage, and phone usage.

There are a number of wearable sensors that can provide similar and complementary data to smartphones. Fitness trackers in particular have reached a certain level of popularity, with products ranging from the well known Fitbit watches [57] to the newer Oura rings [58]. Unlike phone sensors, these fitness trackers can also capture digital biomarker data like heart rate and pulse oximetry. While worn like watches, Actigraphs are specifically designed to monitor movement during sleep and produce actigraphy data [59].

Within digital mental health research, microphones are more commonly used to record voice during an activity that requires active participant engagement [60]–[63]. While there are privacy concerns with passively collecting audio, passive sensing research has employed two innovative solutions. The first, deployed by the StudentLife app [64], is to detect when there is conversation and store the duration of the conversation instead of the raw audio. Di Matteo et al. [65] alternatively repeatedly recorded short durations of environmental audio, a method pioneered by Mehl [66].

There are also other passive screening modalities with privacy concerns [67]. For example, most modern phones are GPS enabled, but storing location history can make it easy to locate someone. Likewise, the message content in SMS logs may not only reveal identity but also private information. While the metadata from the SMS logs can be extracted without the content and the phone numbers one-way-hashed [68], not all

modalities can be anonymized as easily. Sometimes a modality with less privacy concerns can be used to approximate another modality, such as determining location through WiFi instead of GPS [69]. Alternatively, features can be extracted for collection instead of the raw data [70], but this limits feature engineering research as the features have to be decided on prior to the collection. While a major concern for research, privacy is less of a concern for implementation as a screening model could be deployed on-device [71].

While traditional screening instruments work well in clinical practices [37], these passive sensors are important for increasing screening rates. Passive sensor data can simultaneously screen for many different mental health disorders, whereas there are separate traditional screening instruments for each disorder. These traditional screening instruments often ask direct questions that can be perceived as invasive [72], rely on patient recall, and are subject to conscious and unconscious bias [73]. Further, mental health disorder symptoms may manifest in a way that is not recognized or perceived as a physical health concern by patients [72]. Certain symptoms can also interfere with help seeking behavior [74]. Thus, by the time patients seek help and are screened by traditional screening instruments, the mental health disorder symptoms may be quite advanced.

Treatments are more effective when symptoms are less advanced [75], and addressing symptoms early can prevent some of the more serious social, financial, and health consequences associated with mental health disorders. Screening for mental health with passive sensor data can solve many of the challenges associated with traditional screening instruments, both catching symptoms earlier through continuous monitoring and removing the need for active participation in the screening process. Passive sensing technologies for mental health disorder screening are generally considered to be acceptable once privacy concerns are addressed [67]. Thus, with appropriate translational validation of current digital phenotyping research, passive sensing holds promise to complement and extend current clinical practice.

Background on Retrospective Versus Prospective Studies

There are two different ways in which digital phenotype data can be collected. *Prospective* collections, like that of Reality Commons' Social Evolution Dataset [76], enrolls a participant and monitors them over the course of the study. The Mood Capable Assessment Framework (Moodable) [77] introduced the idea of a *retrospective* collection in 2020, where a participant is enrolled in a study and their stored data is scraped. In comparison to prospective collections, retrospective collections are currently relatively rare [62], [77], [78].

Each of these collection approaches naturally have positives and negatives, both with regards to research collections and clinical translational value. As prospective collections are longitudinal by nature, the data can be labeled throughout the collection process. Meanwhile, for retrospective collections, the data labels are collected concurrently with the log scraping, a process that can be completed within five minutes, afterwhich the app can be uninstalled. Thus, retrospective collections do not tax the phone battery like prospective collections [76], which require an app to be installed and running for the collection duration. Due to the time commitment, it is more challenging to recruit for prospective collections and participant dropout is an issue [65]. Furthermore, participants in prospective collections know they are being monitored which can impact their behavior. However, participants in retrospective collections could remove logs prior to participation in the study, thus greatly reducing the predictive quality of their data.

Alternatively, retrospective and prospective have been used to describe the temporal placement of the screening survey in relation to the digital phenotype data. In this case, *retrospective placement* would refer to a collection where the screening survey was administered upon enrollment and then data was *prospectively collected*. Meanwhile, *prospective placement* would involve *prospectively collecting* the data and then

administering the screening survey. A *retrospective collection* can only have a *prospective placement* of the screening instrument as the data already exists at the time the participant enrolls in the study. In contrast, a *prospective collection* could have both retrospective placement and prospective placement of the screening surveys, as is modeled by StudentLife [64]. While a retrospective placement can be translationally useful in conjunction with a prospective placement, retrospective placement alone holds limited translational value.

Translational Paper Identification Methodology

As this is a narrative review, we do not report on every study that falls within the broad umbrella of passive sensing, but focus on those that provide translational value. We loosely define those studies with translational value to be those that: (1) have mental health disorder labels, (2) obtained the mental health disorder labels through clinical assessment or a validated screening survey, (3) predict the mental health labels with passive sensing data, (4) collected the passive sensing data on the phone, and (5) have a sufficient number of participants for the results to reflect on the potential translational value of the predictive methodology.

There is a large body of related literature that falls outside of the scope of our narrative review. For example, substance abuse disorders and neurodevelopmental disorders have a high co-occurrence with mental health disorders [79], [80]], and there is research that uses passive sensing to screen for substance abuse and neurodevelopmental symptoms [81], [82]. There is also related passive sensing research [76], [83] where participants report feeling stressed or depressed. Similarly, there are also studies where the labels rely on participants' self-disclosure of diagnoses [84]. These data labeling strategies do not provide the same translational value of studies that leverage data labeled through clinical assessment or a validated screening survey. Likewise, while studies that use exploratory analysis without predictive models provide important information to guide predictive work, they lack the direct translational value of the studies that leverage predictive modeling.

We also limit the scope of our narrative review to studies that collect passive sensing data from the phones, rather than through third-party apps on the phones. For example, studies that consider the amount of time using third-party apps would be included whereas studies that extract activity from the third-party apps would not be included. In particular, this criteria regards the disclusion of social media studies, which is a broadly enough used screening modality to require a separate review [85]. Furthermore, most social media users are lurkers in that they consume content rather than create content [86], and therefore the translational value of such studies is limited.

Lastly, the digital health phenotype research space is known for small datasets, reflecting the difficulty in recruiting participants for such studies. This number can change based on the evaluation method, but it outright excludes the many studies with less than 50 participants. While these studies provide value in introducing new screening modalities or features, they lack the translational value of research that models such modalities or features collected from a greater number of participants. Additionally, the evaluation strategy employed by the study can influence transitional insights. For example, a study that leverages leave-one-out cross-validation does not need as many participants as a study that leverages a single held-out test set.

We do not intentionally limit the temporal scope of the papers in this narrative review. However, it is bound by the development of passive sensing technology and the date that we completed our literature review, which was September 2023. As we present a narrative literature review rather than a comprehensive literature review, we did not employ any formal search criteria to identify papers. Rather, we conducted a thorough search for passive sensing studies that provide translational value.

In addition to focusing on the translational value of the studies, our narrative review is unique in the broadness of scope in relation to types of sensors and mental health disorders. Many related reviews limit themselves to a single screening modality [85], [87] and/or a single mental health disorder [87]–[89]. However,

this approach obfuscates connections between studies that leverage different modalities or screen for different mental health disorders. Further, due to the plethora of studies that focus on depression, many of these reviews focus solely on depression, which limits the understanding of how the passive sensors are used across different mental health disorders. Thus, our narrative review offers a unique translational perspective on the breadth and depth of research studies that screen for mental health disorders with passive sensing data.

Depression

Major depressive disorder is a highly prevalent DSM-5 defined disorder [6]. In 2021, an estimated 21 million U.S adults, or 8.3% of the country's population, had experienced at least one major depressive episode over the year [7]. Depression represents a significant burden to public health, ranking as one of the highest causes for disability globally, representing its prevalence worldwide [90]. Further, MDD is associated with a higher risk of cardiovascular disorders and diabetes [28]. Despite this, only 40.6% of U.S. adults with depression received treatment in 2021, a process which begins with efficient and effective screening and diagnosis of the disorder [7].

The current DSM-5 criteria for MDD includes a period of two or more weeks with consistent depressed mood or loss of pleasure and a number of additional symptoms, including problems with sleep, appetite, concentration, energy, and self-worth [6]. This disorder is therefore highly heterogeneous; two patients could each present with an entirely different set of symptoms under the same diagnosis [91]. Therefore, it remains important to continue further study and classification of different phenotypes of depression to expedite the potential for patients to both be clinically assessed and treated in a more personalized manner. Personalized treatment holds promise for expediting care, and can include considerations of which medications and therapies will most benefit a particular patient given their symptom profile [92].

Major depressive disorder still primarily relies on subjective self-report and clinician assessment, however, recent trends have pushed for precision psychiatry to offer a more objective means of quantifying depression onset, prognosis, and severity, including passive sensing and neurological advancements. Growing evidence does support knowledge of the neural substrates of depression, but such study remains far from practical for clinical translation given that many assessment methods rely on posthumous analysis or costly medical tests [93]. Another precision approach aims to target specific genes thought to be responsible for the disorder, given that the genetic contribution is approximately 35%, and even higher in those with rare mutations [94]. However, MDD is heterogeneous not only in symptom presentation, but also underlying cause with environmental risk factors, including prenatal complications; sexual, physical or emotional abuse; and traumatic or stressful life events also contributing significantly to one's risk of developing the disorder [95]. Passive sensing holds great promise for MDD assessment given its emphasis on personalized assessment, objective data, and reduction in patient burden. Passive sensing is also promising for depression more specifically given the heterogeneity in symptom profile and ways in which these symptoms can be masked, contributing to a cycle of undiagnosed and untreated mental illness [48]. For example, common symptoms of MDD include fatigue, irritability, and trouble concentrating [6]. An individual experiencing a major depressive episode may not link these symptoms together and may ascribe them to menstruation, a tumultuous time in their personal life, or related to a pre-existing health condition.

Many passive sensing studies of depressed samples collect ground truth data through regular assessment of mood utilizing validated questionnaires like the PHQ-9 [97], QIDS [98], and BDI-II [99]. The PHQ-9 has been validated in a number of settings and populations, and is the most commonly deployed depression screener in the United States [100]. It asks users to reflect on the previous two weeks using nine questions [38]. However, some studies make use of the Quick Inventory of Depressive Symptomatology (QIDS) instead. The QIDS is a 16-item questionnaire available in both self-report and clinician-rated formats

[101]. One study [98], referenced in Table 1, highlights the use-case of QIDS over PHQ-9 for passive sensing purposes given it is a more detailed questionnaire that is able to differentiate between decreased appetite and increased appetite. In comparison, the PHQ-9 asks about poor appetite and overeating within the same item. There are a number of other differences between these different depression screeners which influence which screener is utilized. Ultimately, ground truth data is necessary for supervised predictive models, so understanding and utilizing the optimal depression screener is important to both manage the need for quality assessment data and reduce patient burden.

Many exploratory studies have highlighted important associations between passively collected data and known symptoms of depression, revealing potential digital biomarkers which help inform the design of further study. GPS has proven to be a reliable and important biomarker for depression assessment in study samples [102]–[104]. For example, Asare et al. found that patients with depression demonstrated less mobility, more sleep time, less physical activity, and low mood utilizing a wearable health ring and smartphone-based data [102]. Understanding broad digital biomarkers for depression begins the process for adapting data regarding depressed populations to personalized understandings between depressed patients, incorporating environmental risk factors and demographic information. Predictive modeling allows for the leveraging of demographic and socioeconomic variables, which can improve model performance in some studies [99], [105]. For example, Razavi et al. compared balanced accuracies including (BA = 0.768) and not including (BA = 0.811) age and gender demographic information in their predictive models [99]. Xu et al. created a new method to extract features from passive sensing data that leverage knowledge of larger contexts, instead of considering features individually [106]. For example, this might combine average sleep with location data to generate an understanding of average sleep during the night when students are off-campus and have high activity levels. This modeling approach relates to the larger task of identifying digital depression phenotypes by exploring differences between depressed and non-depressed groups that either (1) share similar contexts with different behavior or (2) having contexts that are common in one class and uncommon in the other. In these ways, exploratory and predictive studies can build an emphasis on charting and then applying a personalized understanding of population and individual depression.

Many multimodal passive sensing studies (Table 1) have utilized both wearable and smartphone-based data to predict MDD with moderate to high performance [98], [107], [108]. Wearable devices can add complementary data streams and often do so with high accuracy [109]. For example, compared to only using smartphone data (F1=0.67), one study achieved better results when incorporating smartphone and smartwatch data together (F1=0.77) [98]. However, combining all features did not always generate the best results, which could be explained by the additional noise that multiple data types can incorporate into a predictive model, leading to reductions in performance. For Chikersal et al., the best results for detecting depression included Bluetooth, Call, Phone Usage, and Sleep features, and so they only used a subset of total features collected [107]. Horwitz et al. found that incorporating smartwatch features to mood features had no appreciable effect, as mood features alone (AUC = 0.749) and mood features with smartwatch features (AUC = 0.750) yielded similar evaluation metrics [30].

While passive sensing offers seemingly limitless opportunities to collect and model data, results like these indicate that features and modalities should be carefully selected for predictive applications for several reasons. One reason is to minimize patient burden, as carrying and maintaining an additional device for long-term, daily use and, in some cases, filling out regular mood assessments for ground truth data, can create additional work for a patient population that already faces symptoms which impact one's ability to carry out daily life. Another reason is that this evidence reveals that additional features do not always improve model performance, but sometimes contribute noise to the model. A third reason is that more features require a more complex predictive model which requires more computational power, which is associated with more expenses and waste when such systems are scaled [111].

In order to position passive sensing as a more equitable and accessible option for depressed patients, it is important that models can utilize data from a larger amount of devices. To this aim, there is a large body of predictive and exploratory passive sensing research within the domain of MDD allowing for unique machine learning methods to build upon prior work, and approach solutions for disorder-spanning passive sensing limitations [98], [106], [107], [112]. For example, Lu et al. jointly modeled both Android and iPhone data to improve depression prediction accuracy [98]. The methods that different sensors use differ substantially in how they collect data, so extraction of features that are relevant to the depression and data preprocessing are important. Therefore, modeling approaches that can jointly model both Android and iOS data can overcome this technical limitation of passive sensing for mental health assessment, given that operating systems sample data differently. For example, Android phones collect location data every 10 minutes and iPhones utilize an event-triggering mechanism [98]. As previously discussed, a large number of individuals experience barriers to traditional depression assessment and treatment, whether these are financial, geographical, or related to stigma. For passive sensing to bridge these barriers, it is important that data can be collected from a larger number of devices given what is known about the demographic and socioeconomic differences between Android and iPhone smartphone owners [113].

It is important that predictive models generalize to large at-risk populations. One of the most promising aspects of passive sensing for MDD patients is its potential to open access to care for patients whose disorder causes burdensome symptoms, which means that it is ideal that passive sensing require little responsibility on the part of the patient. One area of improvement for passive sensing is that predictive models need to be able to generalize to unseen data with high performance to reduce the patient workload and increase compliance, comfort, and use of such technology. Xu et al. made steps towards this end by evaluating cross-dataset generalizability using longitudinal behavior models to detect depression, as well as designing the benchmark platform, GLOBEM [112]. The GLOBEM dataset combined multiple longitudinal datasets from two different academic institutions over a two-year span, leading to over four institute-year datasets with over 500 participants. Their findings indicate that a number of machine learning algorithms do not generalize well on their datasets, which highlights the importance of considering the role of individual differences in this challenge of model generalization.

The episodic nature of this disorder encourages an *objective and temporal* understanding of how symptoms change over time as an important tool for proper diagnosis and treatment. As mentioned, this positions passive sensing as a unique opportunity to obtain unobtrusive, temporal data and to reduce bias related to traditional screening methods. A number of studies have aimed to identify clinically relevant PHQ-9 scores, and then predict *future* depression severity as a means to allow for intervention [107], [108], [110]. Bai et al. [108] monitored the mood status and stability of patients with MDD, characterizing patients as Steady or Swing, with subsets being Steady-Remission, Steady-Depressed, Swing-Moderate, and Swing-Drastic. This presents an opportunity for translational impact, in which ‘swing’ patients might need more frequent clinical assessment than ‘steady’ patients. As well, Horwitz et al. [110] were able to detect depression at day 92 after only 14 days of data collection in a large sample of first-year medical interns. This presents the promise that passive sensing data can be utilized to build predictive frameworks that can track changes in mood utilizing objective data, rather than just subjective recall, and employ just-in-time-adaptive-interventions. Future research could provide more insights into how less obtrusive data could be modeled to understand changes in depression without relying on mood assessment.

A number of studies have utilized unimodal text and call logs (Table 2) to assess depression with moderate performance [24], [70], [78], [114]–[118]. The majority of these studies [24], [78], [114]–[116] utilize the communication log metadata rather than the message content for depression assessment analysis due to the privacy concerns of participants. Two of them considered the number of messages, message duration, and amount of unique contacts of the incoming and outgoing call and SMS text messages [24], [115]. Using time

series features of retrospectively collected logs, outgoing text logs (F1=0.72) proved more useful for screening than incoming call logs (F1=0.65). Time series were also used by the study that introduced the retrospectively collected DepreST-CAT (Call and Text log) dataset [78], the largest dataset of smartphone logs with mental health disorder labels, specifically for depression and anxiety. Separately, Tlachac et al. [116] also studied reply latency, a feature of direct messages that records the time passed between receiving a message from a contact and then returning a message. Using just 49 DepreST-CAT participants, a depression screening model reached a balanced accuracy of 0.7, but incorporating an out-of-distributions sample (46 Moodable/EMU participants) yielded a balanced accuracy of 0.66 [114].

The message content of the SMS logs have also been modeled by a few studies [70], [117], [118]. By considering content, it introduces the potential to identify the quality of the interactions rather than just the quantity, potentially revealing the presence of strained relationships which are known to reduce access to emotional support [70]. In 2020, Tlachac et al. [117] first revealed the value of SMS text content for depression screening purposes, extracting a variety of text features including pre-existing lexical categories. When analyzing content, it is important to consider both the privacy concerns and the colloquial nature of text messages. To tackle the first, Liu et al. [70] extracted pre-existing lexical categories from participant phones instead of raw data, which achieved an AUC of 0.72 in unimodal depression screening models. However, it is important to process text in a manner that optimizes their colloquial nature, as they contain slang and abbreviations that are absent from other more formal texts that are often used in the training of natural language processing tools [119]. Tlachac et al. [118] proposed a strategy to automatically construct alternative lexicons that contain this more relevant language. They discovered that vacation was the most important lexical category in their best performing lexicon (F1=0.79) [118]. Additionally, there are three other studies with less translational value that leverage SMS text content that have depression labels; they extract features from received texts [120], perform message-level screening with the raw text [121], and fit a linear mixed model to identify differences in depressed language usage between contact subsets [122].

Internet and Bluetooth data, including traffic/usage and mobility using Wifi signals, have also yielded moderate to high performance [69], [97], [123]. By collecting the amount of nearby Bluetooth devices, this data can serve as a proxy for understanding an individual's social connection. In an association analysis, Zhang et al. found that changes in nearby Bluetooth device count were inversely related to worsening in depression symptoms [123]. Within their study, they utilized a subset (n=316) of the total RADAR-MDD sample (N=623). The Remote Assessment of Disease and Relapse- Major Depressive Disorder (RADAR-MDD) study is a multicenter, prospective observational cohort study and a part of the RADAR-CNS program [124]. The study is the largest multimodal passive sensing study in the field of mental health, with over 623 participants from the United Kingdom, Spain, and the Netherlands [125]. The study utilized mood questionnaires and wearable device data including ambient light and noise, Bluetooth connections, and GPS to predict depression relapse. This work builds on the work of the StudentLife [64], Ware et al. [69], and Saeb et al [103] , whose prior work has laid foundational knowledge for using passive sensing for depression assessment. By utilizing such a large diverse sample, the RADAR-MDD study holds the potential to further validate passive sensing for translational use by tackling a number of gaps/limitations in the research, including the use of small sample sizes (n<50) which lack generalizability.

In addition to StudentLife [64], there are a number of other studies that have a small sample size and/or employ an exploratory analysis. Like Saeb et al [103], Gerych et al. [126] also presents an innovative approach to working with the GPS logs released as part of the original StudentLife dataset [64]. Additionally, there are other foundational digital phenotype studies that leverage GPS logs in datasets with less than 50 participants, such as Saeb et al. [127], Canzian et al. [128], and Wahle et al. [129]. Additionally, there are more recent studies that have more participants, but opt for a more exploratory analysis, like correlation analysis [65], [68],

[130]. Even the RADAR-MDD, which has sufficient participants for predictive modeling, opted to start with exploratory modeling [131]–[133] to gain insights about the data.

Passive sensing has been able to serve as a tool for depression assessment using a variety of modalities, both unimodal and multimodal, with moderate to high performance in most cases. As passive sensing becomes a more reliable and trusted tool for assessment of mental health, this presents an opportunity for already reliable sources of data to become integrated to strengthen the tool. For example, integration with patient electronic health records gains the ability to leverage information regarding close/emergency contacts, home and work address, and medical history. As well, for depressed patients, this again functions as an opportunity to reduce front-end workload and increase the ease of using unobtrusive mental health assessment. In conclusion, passive sensing holds promise to open avenues for personalized assessment and treatment in a way that reduces patient burden for depressed populations using objective and unobtrusive data, which can bring treatment to the millions of individuals who go undiagnosed and untreated.

Bipolar Disorder

Bipolar disorders are chronic mood disorders distinguished by discrete periods of persistent and disabling *elevated* or *irritable* mood coupled with symptoms such as impulsivity, grandiosity, decreased need for sleep, and increased activity [6]. Such discrete episodes are termed mania or hypomania and may occur alternately or mixed with intermittent depressive episodes. Periods of normal mood, termed euthymia, characterize the middle ground between depressive and manic and hypomanic states. The DSM-5 characterizes hypomanic states as lasting for at least 4 consecutive days, whereas a manic state must last for at least 1 week [6]. Manic episodes are generally more severe than hypomanic, with psychotic symptoms being common for acute manic episodes. Given this, hypomanic symptoms are often underrecognized, which presents additional challenges for patient care when later experiencing depressive or manic episodes [134]. This highlights a particular need for acuity in detection of hypomanic episodes for more effective diagnosis of such a heterogenous mental disorder. Further, the onset of a major depressive episode by itself may represent a unipolar or bipolar disorder, which complicates effective accurate diagnosis. . The average delay between illness onset and diagnosis is 5-10 years [135].

There are no clear distinctions between depressive states in MDD and BD, according to the criteria outlined in the DSM-5, so it can be difficult to provide the proper diagnosis at the onset of the first depressive episode. This can be dangerous as administering antidepressants to BD patients may trigger mania or “rapid cycling” in which a patient experiences manic and depressive episodes in quicker succession [136]. Activity metrics acquired via passive sensing hold the promise of distinguishing between depressive states in these two psychiatric disorders. One exploratory study attempts this type of differential diagnosis using actigraphy data, between BD and MDD patients in depressive states [137]. In that study, Tanaka et al. discovered that participants with BD showcased higher activity during specific nighttime hours and lower activity levels during those same nighttime hours in comparison to MDD patients.

In contrast, there are also a small number of exploratory studies, provided in Table 3, that have distinguished between BD patients and healthy controls (HC) [138], [139]. They utilize location data, step count, actigraphy data, sleep data, and call and text metadata to explore relationships between these modalities and bipolar symptoms. Building off of these relationships, Faurholt-Jepsen et al. used predictive modeling of location data to detect bipolar symptoms among BD patients and HC [140]. Their results indicated high model performance for accurately detecting bipolar disorder among BD and HC participants (AUC = 0.82), but also high performance between *euthymic* BD patients and HC (AUC = 0.82), between depressive state BD patients and HC (AUC=0.83), and between manic state BD patients vs HC (AUC=0.84). This represents an important area for translational impact, in which passive sensing needs to have high performance across

separate affective states, especially during depressive or manic states when assessment and treatment become vital for patient outcomes. Similarly, being able to distinguish between euthymic BD patients and HC is important to identify those who might need more monitoring and resources for risk management while between states compared to those without the disorder.

Using predictive modeling, Bennett et al. combined activity data from accelerometers and typing dynamics to serve as a proxy for clinical data for modeling bipolar symptom trajectory, using a large (n=291) naturalistic dataset. They found that this data was capable of predicting clinically-relevant changes in PHQ-8 scores (AUC = 0.9442) [141]. PHQ-8 is a shortened form of the PHQ-9 screener which lacks item 9 which queries thoughts of death and self-harm [142]. However, unlike most other digital phenotyping studies which often face the inverse, one limitation of their study was that their integrated mobile app only worked on the iPhone Platform, excluding Android users. . Given known demographic and socioeconomic differences between iPhone and Android individuals as well as the personal burdens and disabilities that bipolar disorders can cause an individual, it is likely that technology with such exclusions would limit those that might benefit from predictive technology [113].

Anxiety Disorders

Anxiety disorders, including generalized anxiety disorder (GAD), social anxiety disorder (SAD), and panic disorder (PD), are one of the most commonly diagnosed categories of chronic mental disorders, with approximately 30% of adults in the U.S. experience some type of anxiety disorder during their lives [143]. While each disorder is unique in specific symptom profiles [6], all anxiety disorders share the central characteristics of fear, worry, or excessive apprehension [144]. While often debilitating and distressing, once diagnosed, effective treatments, including psychotherapy and medications, exist for anxiety disorders [145].

While chronic in nature, anxiety disorders are dynamic with symptoms that can fluctuate, sometimes rapidly, over time and depending on context [146], [147]. Due to the dynamic nature of anxiety disorder symptoms, understanding temporally nuanced aspects of the disorder can aid in assessment and treatment. Given that effective treatments exist, early diagnosis and symptom tracking is paramount. Passive sensors have parallel benefits and modalities as previously established in mood disorders. For anxiety disorders more specifically, passive sensing provides an opportunity to capture physiologic disturbances (such as increased heart rate and breathing levels during a panic attack) in order to associate those with the relevant contextual cues or triggers. This could highlight a leverage point for which self-management of symptoms or a clinician's guidance could become particularly useful. Although a number of studies show promising results leveraging such data as included in Table 4, many studies in this domain have small sample sizes limiting their inclusion in this narrative review given the guidelines for sample size established.

GAD is one of the most common anxiety disorders, impacting approximately 3.7% of people worldwide at some point in their lifetime [148]. The core feature of GAD is persistent uncontrolled worry, associated with some several psychiatric and physical symptoms, including restlessness, fatigue, muscle pain, and sleep disturbance [6]. GAD is commonly assessed using self report screening measures, such as the GAD-7 [149], or as part of structured interviews, such as the Structured Clinical Interview for the DSM-5 (SCID-5) [150]. Passive movement data has successfully been used in a moderate-sized study (N>250) to both detect persons with elevated GAD (AUC=0.89) and predict symptom severity across a continuum (r=0.511)[19]. In this study, correlations were highest between passive movement-based predicted risk scores and self-reported restlessness, irritability, difficulty controlling worry, and difficulty putting worry out of mind. Results suggest that signal in passive movement data alone may have clinical utility in both identification of persons at risk for GAD and in tracking symptom severity after GAD is diagnosed. Given the low burden and limited privacy concerns

around passive movement data, it is likely that these methods would find acceptability among general and clinical populations.

SAD, formerly known as *Social Phobia* in the DSM-IV, is characterized by marked anxiety in one or more social situations driven by an intense fear of social rejection, judgment, or negative evaluation [6]. Approximately 12.1% of adults in the United States experience SAD over their lifetime [151]. SAD is an area in which the use of passive sensing can be especially useful for assessment and diagnosis. Due to the presentation of the disorder, it is not uncommon for those with SAD to avoid pursuing assessment and treatment [152]. The DemonicSalmon study [153] explored the relationship between passive sensors, including accelerometry, GPS, and communication logs and social anxiety scores measured by the self reported Social Interaction Anxiety Scale (SIAS) in a college student population utilizing Android over 2 weeks. While Boukhechba et al. performed only exploratory statistical evaluation, finding moderate correlation ($|r| < 0.6$) between individual features and SIAS scores, Jacobson et al. analyzed these data in a predictive modeling framework, finding a moderate ($r=0.702$) correlation between predicted and observed SIAS scores [152].

Using similar features and the same modalities to Boukhechba et al., Gong et al., also assessed an undergraduate sample using the SIAS prior to the data collection period. Gong et al. explored how actigraphy data differs between those with lower levels of Social Anxiety compared to the fidgeting and other “fine movements” displayed by those with comparatively high levels of social anxiety. These “fine movements” were also shown to vary based on GPS location, allowing Gong et al. to determine that, not surprisingly, social anxiety levels increase in public settings [154]. Such fusion of passive sensor types, exemplified here, allows for real time symptom contextualization, not possible with traditional screening instruments. Further, a study conducted by Jacobson and Bhattacharya (2022), suggested that passive sensing data and personalized models can also be used to determine the fluctuation of one’s symptoms of SAD and GAD throughout a given day [155]. This presents with translational impact as this technology can offer a better understanding of which environments and times of day lead to increased anxiety, measure the effectiveness of certain therapies or coping mechanisms, and employ the use of just in time adaptations given the real time symptom contextualization which can help personalize the adaptation to the specific type of environment or stressor.

Panic attacks are marked by an intense and discrete period of fear and anxiety, associated with physical symptoms including chest pain, nausea, difficulty breathing, and trembling [156]. The DSM-5 diagnostic criteria for panic disorder requires multiple attacks followed by either prolonged worry or change in behavior related to the panic attack [157], which separates panic disorder from the experience of a one of more singular panic attacks, which are common among the general population in the United States [158]. McGinnis et al. utilized data collected from the Apple Watches of participants who experience regular panic attacks, with results indicating that variation in resting heart rate, heart rate variability, respiratory rate, and ambient noise levels can predict a panic attack on the following day [158]. These findings combined with the ease of data collection through a widely used consumer device creates pathways for increased clinical intervention before a panic attack occurs, potentially lessening the discomfort and severity of attacks. Knowledge of increased panic attack risk could then be coupled with just in time adaptive interventions (JITAIs) to promote behavioral change at points when the participant is optimally receptive [159]. Future benefits of this type of prediction might also include reducing burden on emergency services for those who misinterpret panic attack symptoms with those of a heart attack or other medical emergency.

Trauma and Stressor Related Disorders

Posttraumatic stress disorder (PTSD), acute distress disorder, reactive attachment disorder, disinhibited social engagement disorder, and adjustment disorders comprise the DSM-5 defined trauma and stressor-related disorders [6], which are all characterized by a precipitating exposure to one or more traumatic events [6], with

subsequent alterations to cognitions, mood, or physiologic arousal [160]. PTSD, the most well-known of the aforementioned trauma and stressor-related disorders, has a lifetime prevalence of 8.3% in the general population [161], with substantially higher rates in United States military veterans [162], [163]. PTSD is most commonly associated with symptoms of intrusion, avoidance, negative alterations in cognition, increased arousal, and can follow a chronic course if left untreated [6], [164]. Resultantly, clinical presentation of PTSD is highly variable in both presence and degree of symptoms [165], with chronic untreated symptoms leading to an increase in additional physical health problems such as low engagement in physical activity [166], hypertension [167], obesity [29], and stroke [159], which parallels the heterogeneity and associated risks observed in MDD. Furthermore, PTSD is highly comorbid with other mental health disorders [169], highlighting the difficulty in identifying PTSD in primary care settings, with detection rates consistently less than 50% [170], [171]. Low detection rates may be due, in part, to the reliance of structured questionnaires in a clinical setting as means of acquiring patient information [172], [173], which is further confounded by the distress reported in a high proportion of individuals with PTSD when required to discuss traumatic events in a structured clinical interview [174]. Given these barriers to adequate PTSD diagnosis and screening, a more effective approach to screening and monitoring individuals with PTSD may be with the use of naturalistic, passively-collected sensor data via smartphones and wearable devices which incorporates mechanisms for assessment without adding additional burden or distress on the part of the patient.. Specifically, passive sensors may be able to better capture the distinct digital phenotypes related to the aforementioned symptom domains of PTSD, thus improving efforts with disorder screening, monitoring, and treatment.

Many trauma-related passive sensing studies to date have investigated HRV, as provided in Table 5, likely related to the arousal subdomain in PTSD, which specifically considers alteration in arousal, such as irritability and aggression [6]. Consistent with the RDoC framework [8], HRV serves as a biomarker of arousal, thereby serving as a strong example of how passive sensing via smartphone and wearable sensors can best capture the digital phenotypes of a symptom domain. Circadian rhythms, were evaluated by Ulmer et al. (2016) who found that individuals in a young adult cohort with high PTSD symptom severity had less high-frequency HRV [175] during periods of sleep; however, there were no differences during rest or activity during the high and low sleep symptom severity groups [176]. Furthermore, continuous passive collection of HRV has shown efficacy in detecting PTSD triggers, with an association between PTSD trigger onset and heart rate increase [177] and prediction of post-deployment PTSD [178], both of which highlight the utility of passive sensing modalities as a means of monitoring individuals with PTSD to allow for symptom mitigation. As previously noted, multimodal approaches allow for distinct but complementary signals to improve model performance. To complement any signal captured by HRV, accelerometers are also a suitable sensor for detecting alterations in arousal, and have been shown to be a more reliable method of capturing an individual's movement behavior compared to self-report, including in primary mood disorder populations [179], [180]. Indeed, Sasangohar et al. (2022) highlighted the utility of this sensor pairing by using a machine learning approach to predict (Accuracy 83%, AUC 0.70) hyperarousal events among military veterans. Further, model introspection via SHapley Additive exPlanations found the model's predictions to be most correlated with average heart rate, minimum heart rate, and acceleration [181], thus further establishing the basis for incorporating passive sensing, specifically those sensors with clear relationship to symptom domains of interest, in studies related to PTSD outcomes.

Psychotic Disorders

Psychotic disorders, such as schizophrenia, schizoaffective disorder, and delusional disorder, are characterized by a *disturbance* and *disorganization* in thought processes, often involving distortions of perceived reality. Such disturbances are termed *positive* symptoms and may include delusions, paranoia, and

hallucinations. Some psychotic disorders, notably schizophrenia and schizoaffective disorder, also involve *negative* symptoms, comprising depressed mood, reduced interest and motivational drive, often leading to social withdrawal, anhedonia, blunted affect, and cognitive symptoms. Left unchecked, both positive and negative symptoms of psychotic disorders contribute to a severe and often irreversible functional decline associated with the onset of the disorder. Psychotic disorders, particularly schizophrenia, are often observed to be preceded by a clinically high risk, prodromal phase, lasting 4-7 years [182], during which nonspecific symptoms, such as mood and sleep disturbances, anxiety, and suicidal ideations may appear, prior to the onset of full psychosis [183]. The risk for psychosis is modulated by both immutable baseline and *modifiable dynamic risk factors*, cannabis-use an example of the latter, showing a dose positive dependent association with psychoses [184]. The prodromal, clinically high risk period provides a time window where disease trajectory modification is possible. Indeed, early intervention in schizophrenia during the prodromal phase has been associated with reduced conversion to full psychosis and improved functional outcomes [182]. Trajectories are manifold and relapse and remission common in persons with psychotic disorders.

Considering the nature of psychotic disorders, several important areas of mobile passive sensing involvement emerge. First, we might imagine the development of biomarkers and prediction of early risk factors for psychotic disorders in clinically high risk persons. This has historically been challenging due to the nonspecific nature of prodromal symptoms, combined with the potential harms, including stigma, associated with a false positive psychotic disorder label. Passive methodologies could be employed to both identify high risk phenotypes and identify high risk behaviors among persons possessing those phenotypes. To our knowledge, no mobile passive indicators have been developed for identifying the phenotype of the psychotic disorder prodrome. Such studies, if available, could translate and scale with ubiquitous sensors to unobtrusively screen large populations for psychotic disorder risk.

Another potential area for involvement is in predicting symptom trajectory once clinical symptoms have emerged. Several studies, referenced in Table 6, to date have explored the use of passive sensing to predict symptom change over time [44], [185], [186]. Predicting symptom change overtime is often done using ecological momentary assessment and is more challenging, given that it requires repeat intra-individual outcome measurements. The ability to predict symptom trajectories using passive sensing data could provide valuable prognostic information for patients, families, and their mental health treatment teams, reducing the uncertainty associated with a highly heterogeneous disorder.

An important area for ubiquitous sensing using consumer grade wearables in psychotic disorders is *relapse prediction*. Relapse is a common and serious outcome in affected patients [187], with early identification potentially allowing for prevention using just in time adaptive interventions (JITAIs). Multimodal passive data from the CrossCheck study (2016) have been used in multiple analyses, aimed at examining personalized factors associated with psychotic disorder relapse [185]. Building on the pioneering work of StudentLife [64], CrossCheck deployed a mobile app for Android combined with a cloud based platform for secure data storage and processing [21]. The CrossCheck App collected an extensive suite of multimodal passive sensing data, including accelerometry, GPS, ambient sound, call and text logs, smartphone usage metrics, such as app usage and unlock events [21]. Such features are translationally valuable as they present the opportunity for scale, given that they can be collected through ubiquitous mobile devices and they are highly unobtrusive, allowing for collection without participant burden.

Multiple CrossCheck analyses included up to 62 participants with psychotic disorders recruited from an outpatient treatment program in New York followed over 12 months. Outcomes varied across analyses with some including objective, electronic health record (EHR)- and clinician-validated markers of relapse [186], [188], [189] and others limited to self-report outcomes derived from EMA [44], [185]. Objectively-validated relapse outcomes, such as EHR-based hospitalization or clinician diagnosis confer high translational value, given their direct and proximal position in relation to real outcomes. Wang et al. (2017) further highlighted an

important translational use-case, which we term *human-in-the-loop targeted intervention*. [186]. In this translational use-case, high risk cases, as predicted by machine learning models, were sent to humans to inform their clinical decisions. As such, patients at high risk for relapse could be identified early and provided with preventative measures to reduce their risk of relapse. The clinician administered 7-item Brief Psychiatric Rating Scale (BPRS) could be predicted with 30 days of passively collected data with an MAE=1.80, representing 4.3% of the BPRS score range. With a larger sample (N=60), CrossCheck further demonstrated moderate predictive capacity for objectively assessed relapse outcomes (Sens=0.25; Spec=0.88).

Outside of CrossCheck, we find several other moderately-sized (100<N<150) studies which explore smartphone and wearable based sensors for psychotic disorder detection, prediction, or characterization. Both GPS and actigraphy were common modalities across studies, either used by themselves [190], [191] or in a multimodal framework in conjunction with other sensors [25]. GPS-derived metrics demonstrated between group (schizophrenia vs control) differences, with strong, statistically significant effect sizes (0.70<Cohen's D <0.90), showing generally lower travel and more time at home among persons with schizophrenia [190]. Strauss, et al. [191] utilized both smartphone and smartband accelerometry measures to explore negative symptoms in schizophrenia, finding modest statistically significant negative correlations *only* between the smartband accelerometry scores and clinical negative symptom scales. Despite the absence of a predictive framework in Depp, et al. (2019) and Strauss et al. (2022), as both studies utilize an exploratory statistical approach without a held-out test set, we include these studies as two of the largest to highlight promising passive sensing modalities for future research. Further, we consider accelerometers and GPS to be highly favorable for translational purposes, given their ease of de-identification and their low participant burden.

We further include mention of the publicly available PSYKOSE [192] dataset to highlight an important consideration in participant sampling for *controlled* assessment studies. PSYKOSE collected actigraphy from inpatients with schizophrenia with controls comprising non-inpatient hospital employees and students, among others. While the recruitment strategy no doubt aimed to enlist participants with more severe schizophrenia phenotypes, the choice to utilize non-inpatient controls almost certainly confounded the outcome (diagnostic status: control vs schizophrenia); that is, persons hospitalized would be expected to move less regardless of their mental health pathology type. Thus, to maximize translational value in controlled protocols, it would be important to match hospitalization status between groups.

Suicidal Ideation and Behavior

Suicide[1] is considered in most cases to be the most serious adverse outcome of all psychiatric disorders, and thus we include suicide and suicidal ideation in our narrative review as a transdiagnostic severe adverse outcome. Suicide is a serious public health concern, with a yearly average of over 800,000 people completing suicide worldwide [193], as well as being the leading cause of death in youth in the United States [194]. Worse still, the high incidence does not capture the full societal impact, with prevalence rates likely much higher due to both death certificate misclassification [195], and non-fatal suicidal-related behaviors estimated as high as twenty times more prevalent than suicide completion [196]. Furthermore, comorbid psychopathology is associated with higher odds of both suicidal ideation and suicide attempts in preadolescent children [197] and adults [147], highlighting that certain populations are particularly at risk for suicide behaviors.

Unfortunately, there are significant barriers to suitable intervention and prevention as self-stigma and reluctance to seek professional help is prominent, particularly in those individuals with previous suicidal

[1] It should be noted that suicide is included in the in the DSM-5 as suicidal behavior disorder under "Conditions for Further Study"; however, these criteria have not been intended for clinical use, but rather to guide further research and considerations of the disorder for future iterations of the DSM. Our treatment of suicide in this review will be as a serious transdiagnostic outcome.

behavior [199]. Low sensitivity in detecting suicide or suicidal ideation from survey instruments or clinical diagnoses [200], [201] suggests the need for improved methods of detection and treatment for individuals with or susceptible to suicidal ideation. Passive sensing may offer a scalable and effective solution to this issue as early efforts suggest there is opportunity for identification of suicidal identification and prevention of suicidal behavior [202], and these efforts could benefit from monitoring efforts that may avoid the need for frequent, clinical visits.

Research to date has investigated the prediction of suicidal ideation and suicidal behavior utilizing machine learning methods. Regarding predicted outcomes, the papers we examined (Table 7) primarily utilized either the ninth item of the PHQ-9 (self-reported), “Thoughts that you would be better off dead, or of hurting yourself” or some version of the Columbia-Suicide Severity Rating Scale (C-SSRS), either self reported or clinician administered. Only a minority of studies used a clinician administered outcome measure or objectively validated the outcome measure, such as by electronic medical record review[203]–[205]. Most studies performed their analyses using traditional machine learning approaches, as opposed to deep learning algorithms, and reported performance metrics as AUC or F1, with AUC generally in the range of 0.6-0.85.

As in other mental disorder categories, most studies were limited by small to moderately small samples (100<N<400 participants), with the largest study examining suicidal ideation among medical interns reaching N=2459. The majority of studies used devices running the Android platform and the most common wearable used was Fitbit. A subset of studies, which we excluded from Table 7, were limited in their ecological validity and translational value by having been conducted in a laboratory setting [206]–[208], requiring prompted responses from participants.

Furthermore, we can build on behaviors already established as related to positive outcomes to better tailor studies to leveraging passive sensing modalities. For example, help-seeking behavior is associated with individuals who attribute suicide to isolation [209], thus monitoring location via passively-collected GPS information, along with voice analysis to check for proximity to other individuals, serves as one approach that suggests how passive sensing may provide unique insight into how to predict suicidal ideation for individuals.

Discussion

Overview

At the confluence of unprecedented computing power, advances in mobile sensors, and the growing prevalence of smartphones and wearables, passive sensing technologies paired with predictive supervised models hold enormous translational potential in the clinical behavioral sciences. With a rapidly growing body of methodologically diverse passive sensing research in mental health, regular synthesis and summary of the evidence is paramount. The degree of direct translational capacity is a very important dimension of such research, and as such, regular review and appraisal of translational value is warranted both to guide future translational research and inform which particular findings may be ripe for adaptation to real clinical populations. In this narrative review, we examine current passive sensing research in the context of its value to behavioral clinical practice, with particular focus on whether it could be implemented with real patients.

We overall find the most predictive passive sensing studies in the areas of depressive disorders. This is not surprising, given its high prevalence [210], coupled with the relative ease of recruitment compared to more functionally impaired populations, such as those with psychotic disorders. Methodologically, across mental disorder domains, we find a mixture of unimodal and multimodal sensor approaches. Commonly used sensor types are actigraphy, accelerometry, GPS, heart rate, ambient light and sound, wifi and bluetooth connectivity. In particular, we find numerous studies, for example [78], [114], [211], utilizing call and text metadata to predict anxiety, depression, and suicidal ideation. Among PTSD studies, we find many using HRV. Multimodal data

fusion is an important consideration in translational studies, as it can not only augment signal by providing additional predictive information, but can also “fill in gaps” from missing data in one modality; consider, for instance, the complementarity of Wifi and GPS for providing user location information [104].

Limitations of Current Research

We note several limitations that span across disorder categories, which offer actionable information with which this field can utilize to advance its goals. Key limitations for depressive passive sensing studies include the use of homogeneous populations that do not generalize well to larger populations, sample sizes that are too small to model, and the dependence on frequent mood assessments as ground truth which can limit compliance and continued use of such technology by depressed patients. Similar limitations are observed across bipolar disorder focused passive sensing studies, in which many utilize self-assessment of mood to forecast manic or depressive episode symptomatology without incorporating passive sensing modalities to perform predictive modeling or small samples are used that reduce the generalizability of the results [212], [213]. Additionally, many studies, which we considered out of scope for our purposes in this review, make use of speech and voice recording to assess bipolar disorder diagnosis and state recognition, which can lack some of the benefits of passive sensing–including less patient burden and naturalistic collection–which prompt passive sensing as translationally valuable as the collection is often an active process on the part of the participant [214], [215].

Many anxiety disorder-focused passive sensing studies make use of smaller sample sizes, which limit their ability to use machine learning or other predictive models and limit their translational impact, which is partially responsible for the limited number of anxiety studies included in Table 4. Additionally, many anxiety-related studies use homogenous samples of undergraduate students or samples which potentially have some self-selection bias such as crowdsourced participants. Consider, for instance, Boukhechba et al. and Jacobson et al. who study a college student sample and Tlachac et al. who study a crowdsourced MTurk population. Additionally, many studies focus on anxiety related symptomatology by evaluating stress or mental health more broadly without applying more rigid diagnostic labels.

For trauma-related disorders, the primary limitation is sample size, with the majority of the identified studies analyzing samples of less than 200 participants. While this can likely be attributed to the unique participant populations being considered (i.e., active military or veterans) it does limit the generalizability to the respective sensor's utility in the modeling task. Furthermore, of the identified studies, adults were the primary study sample age demographic. There are apparent considerations related to sensitivity of discussing and studying trauma and trauma-related events in a clinical setting; however, there is little to no research or literature on pre-teen and teenage age groups related to PTSD diagnosis [216]. It is likely that long-term, naturalistic monitoring of these age groups may help identify novel, longitudinal digital biomarkers related to PTSD, and offer opportunities for earlier diagnosis and intervention. Lastly, there was a small representation of deep learning approaches among PTSD studies. Deep learning is well-suited to explore dense time-series data collected from smartphone or wearable sensors, and may offer a unique methodological approach, allowing for improved performance for disorder-related metrics related to detection, monitoring, or treatment.

Among the passive sensor studies in psychotic disorders, we find several important nearly universal limitations, which we assert in order to guide further translational research. First, we find overall the studies in the psychotic disorder domain have modest sample sizes (N<150), which would limit their generalizability; related to small sample size, we find over-representation, often nearly 50-50 splits of psychotic disorders compared to control participants in detection studies, such as Jaboksen et al. (2020), which had 41% (22/54) of individuals with psychotic disorders [192]. This would potentially limit the capacity of such models to be used for screening on general population or clinical samples, where the prevalence of schizophrenia is considerably lower [6]. Third, in line with other pathologies, we find mostly Android-dependent studies, likely to limit generalizability to users on other mobile OS platforms. Last, we note that some studies in schizophrenia utilize

either more burdensome or more invasive modalities which may not be widely accepted in clinical or general populations. For example, one study incorporated passive ambient sound [217], which may pose privacy concerns.

For suicidal ideation and behavior passive sensing studies, it is important to consider current limitations in the literature to allow for continued improvement in detection and monitoring. Firstly, ethical and liability considerations result in many individuals with suicidal ideation will be excluded from studies [218] limiting the opportunity to monitor these individuals and allow for prevention interventions. It follows that suicide is a rare outcome, and thus difficult to deploy and evaluate machine learning models with suitable training data [219]. Resultantly, many studies consider imperfect proxies for suicidal behavior, such as self-report and evidenced based clinician administered questionnaires, such as the Columbia Suicide Severity Risk Scale (C-SSRS), or physiological signals, such as increases in autonomic arousal [220], which can rely on proprietary algorithms that may limit inter-study utility and generalizability. Furthermore, many studies which explored suicidal ideation did so among depressed persons. This could potentially mean that these studies are unintentionally utilizing suicidal thoughts as an indirect indicator of intense depression, thereby capturing evidence of severe depression rather than specifically of suicidal ideation. To address this limitation, it is crucial for future research to explore signal for suicidal ideation across a broader spectrum of psychiatric conditions, including disorders such as bipolar disorder and borderline personality disorder, where suicidal ideations are also present.

Therefore, two main limitations become evident across pathologies–the use of small or homogenous sample sizes that do not generalize to the larger population and the almost exclusive use of the Android operating system. Further consideration of these trends follows.

Study Sample: Size, Makeup, and Research Aims

With limited exception, e.g. [26], [78], [99], [110], across all mental disorders, we find a trend toward small to moderate sample size, limiting the generalizability of the associated findings. This is likely due in part to the challenging and expensive task of recruiting participants for clinical research [221]. In addition to generally limited sample sizes, we observed that many studies utilized convenience samples, such as crowdsourced cohorts and college students, as seen in examples including StudentLife (college students) [64], DemonicSalmon (college students) [153], and DeprestCAT (crowdsourced) [78]. These sample choices suggest important translational considerations. For instance, it is established that students and crowdsourced participants differ from the general population in important ways, including rates of mental illness, smartphone usage, and willingness to share data. Crowdsourced participants, for example, have been shown to have higher rates of depression compared to the general population [222] and college students differ from nonstudents in important psychosocial ways [223]. Furthermore, research to date suggests that even among similar convenience samples (e.g., multiple related student samples) results may not generalize [224]. The observed limitations in generalizability, particularly among related convenience samples, imply that a universal model capable of accurately detecting signals in any human cohort may not be feasible. Rather, it may be essential to train and develop distinct models tailored to particular populations. Therefore, caution should be used before presuming that a model's applicability will extend to new populations. This underscores the importance of thorough validation processes when models are applied beyond the initial sample group in which they were developed.

A further consideration regarding result generalizability are the aims and protocols of the study from which participants were drawn. Some studies, such as [220] employed passive sensing while delivering a mental health intervention. Obviously, results obtained from intervention studies may not be generalizable to populations where such an intervention is not delivered. Even EMA studies without an explicitly defined intervention may inadvertently alter participants' naturalistic course; EMA itself has been suggested in some

studies to positively influence behavioral and health outcomes [225], [226], perhaps by bolstering self reflection and monitoring.

Mobile Operating Systems and Device Types

We observed that the large majority of studies leverage the Android platform, either by supplying participants with an Android phone or requiring the use of their personal Android device for study participation. This heavy dependence on Android might be attributed to its enhanced flexibility and more permissive environment for apps to access inbuilt sensors, in contrast to iOS [15]. However, this creates challenges when generalizing findings to broader populations and clinical samples, with mixed smartphone platform usage. Firstly, data collection methodologies might vary across platforms, complicating the translation from Android to iPhone; note that some passive sensing methodologies may lack feasibility all together on non-Android devices. Secondly, while empirical evidence varies [189]–[191], there may be inherent differences between Android and iPhone users, biasing the predictive model performance by mobile phone use type. Offering an Android phone to participants irrespective of their preferred phone type could counteract this disparity. Yet, this approach is imperfect, as it can produce its own biases. Specifically, usage patterns are likely to differ between a personal device and one provided for a study.

From a device perspective, we also find that some studies[108], [110] [bai 2021, Horwitz 2022] make use of device-specific proprietary algorithms, Horwitz et al. [110], for instance, utilizes features derived from the Fitbit. While this approach limits the need for extensive preprocessing and saving large data stores, it poses translational challenges, given the likely brandwise differences between devices and the oftentimes inscrutable nature of such on-device proprietary algorithms. Additionally mobile phone type may impact the quality of data able to be obtained from passive sensors, with Boonstra et al. [227] showing lower completed sensing scans on iOS compared to Android. Further, consideration must be given to the trade off between the frequency sampling and battery life [227], which are inversely related and impact both participant experience and capacity to gather data.

Predicted Outcomes

Regarding predicted outcomes, most studies we investigated across disorder domains utilized self reported outcomes–for instance, the PHQ-9 or the GAD-7, administered either as EMA or self report, requiring recall of weeks to months. While subjective self report is and will remain a necessary component in assessment of psychological and behavioral disorders [228], training models to predict subjective self report raises important considerations. Self report instruments, such as the PHQ-9 and GAD-7, are, themselves imperfect proxies and subject to recall biases [229]; stigmatized symptoms, such as suicidal ideations, may also be underreported. While EMA partially mitigates the problem of recall bias, it has higher participant burden; further it is possible that EMA, by encouraging frequent self reflection and self monitoring, may perturb the natural symptom trajectory [226], limiting generalizability to populations not using the EMA component.

We found a minority of passive sensing studies, which used objectively validated clinical outcomes, notably overrepresented among the studies in persons with psychosis. This finding is likely due to the commonality of more severe clinical outcomes, such as hospitalization which have clear, objective markers in the EHR. It may also be due to necessity, given the fact that persons with psychosis often have more impairing symptoms and less insight into their symptoms, compared to non psychotic disorders. Importantly, even among suicidal ideation and behavior, most of studies we surveyed utilized self reported outcomes, most often the ninth item of the PHQ-9 (Highlighted in Table 7). While self disclosed suicidal ideation is a risk factor for completed suicide [230], it is also an imperfect proxy. The choice for utilizing self reported instruments raises important translational considerations, including the need to realize that when self report screeners are used

for the outcome, trained models will be limited to disorder screening and the conclusions learned from inference on such models have no more consequence than the participant self report on the associated self report screening instrument.

Modeling Approaches: Predictive Framework vs Exploratory Analysis

For those studies which utilized a predictive framework, most used traditional machine learning models, such as KNN and RF operating on derived passive sensing features; consider for instance average sleep duration derived from actigraphy. This is in contrast to the fewer studies we found, which used deep learning methods, capable of ingesting minimally processed time series data. Deep learning methods, utilized by an increasing number of passive studies, partially address the challenge of feature selection, though come with their own cost. Deep learning models are plagued by lack of interpretability, limiting their clinical usefulness to patients and clinicians. Many of the passive sensing studies we examined attempted prediction of their respective outcomes using either traditional machine learning methods, namely gradient boosted regression trees [185], [186], or deep learning methods [188]. Modeling decisions bring up important considerations for translational research: traditional machine learning methods, such as logistic regression or tree-based classifiers, are less computationally intensive and better suited for on-device mobile deployment, however more preprocessing and specialized domain knowledge is required in the feature extraction and selection. In contrast, deep learning models, such as that used by [cite deep learning studies], require far less data preprocessing and have the capacity to ingest near-raw time series. They require less feature engineering and less domain knowledge-based decisions upfront. However, their cost comes with their increased computational processing requirements, in many cases precluding their use on-mobile-device, and requiring large and often expensive computing resources. Further, given the magnitude of their parameters and connections, they are plagued by a relative lack of interpretability[231] [231]. Raw data can not be shared, thus limiting the ability to use deep learning models. Further, deep learning models typically require more participants than are in the existing datasets. Notably, in our tables, we only report on the most successful models so studies may have included deep learning models that were simply not as successful as the traditional machine learning models, which is likely when working with smaller dataset sizes. The exceptions to this would be the use of GRU for modeling the DepreST-CAT logs.

Aside from deep learning and traditional machine learning predictive frameworks, we find many valuable exploratory studies, which examine statistical associations between passive sensing features and mental health outcomes [cite studies], but do not utilize a predictive modeling framework with a held-out test or validation set. Although exploratory studies utilizing basic statistical methods do not have immediate clinical translational value, they play a crucial role in guiding empirical feature selection for predictive modeling. Careful and informed feature selection is paramount in traditional machine learning models. Relying on an indiscriminate inclusion of variables — the "throw everything at the wall to see what sticks" approach — can lead to undesirable outcomes. This includes unnecessary burdens on participants from collecting excessive data and a potential decline in model performance. This decline can result from overfitting or the introduction of noise into the model. Indeed, there is empirical evidence indicating that adding information can degrade model performance, as in [107], [110], [220].

COVID and College Populations

A number of studies [24], [107], [232] focus on collecting passive sensing data from undergraduate student samples, benefiting from factors that can make recruiting and studying this population easier than crowdsourced or clinical populations, ergo many refer to student samples as “convenience samples” [224]. For one, student samples are often more receptive towards research opportunities, either as a factor of higher trust

in research as a discipline or as a function of the monetary or academic compensation. While this trend towards undergraduate students is not specific to passive sensing or mental health related study, students serve as an especially useful population for passive sensing of anxiety and depressive screening studies given that 27% experience either clinically significant anxiety or depression [233], a number which has likely increased since the significant disruption of the COVID-19 pandemic. Additionally, it is important to note that other disorders such as schizophrenia have a typical onset age of late teens to early twenties, meaning sampling from an undergraduate population could provide an optimal opportunity to study the prodromal phase and begin early intervention and treatment [234].

During a critical period for social and neurological development, many students were isolated from social networks and displaced from campus [235], with mental health repercussions compounded by increased anxiety levels due to other pandemic related stressors [236]. While the acute effects of the pandemic have largely wound down, its lasting impact has underscored the need for increased mental health care services across academic institutions as the supply of mental health professionals remains unable to meet the demand of students [237]. Passive sensing offers institutions an unobtrusive and easily widespread mechanism to keep a pulse on the mental health of their students, many of whom are already utilizing smartphone and smartwatch devices. Many institutions have opted to offer students access to third-party online therapy services to help increase access to care, however, only a subset of students reach out to these resources [238]. This highlights the need for an unobtrusive tool which can assist students and administrators with determining individual need for mental health care before a crisis might occur, perhaps allowing providers to selectively contact students who display a greater need for care. However, this larger trend of a lack of mental health services alongside a greater need is not limited to college students, but rather a larger trend in mental health care since the onset of the pandemic, which has served as a green light for many innovative approaches to initiate and service mental health care [239].

Translational Implementations of Passive Sensing Research

There is tremendous potential benefit in disseminating research toward implementing unobtrusive predictive mental health models. Building automated predictive models allows for scaling state of the science evidenced-based assessment methods, which would allow for the provision of care to people and areas previously underserved, potentially minimizing existing inequities [240]. However, it is also an important consideration that translational methods which rely on expensive or emerging novel technologies may risk perpetuating existing inequities in the healthcare system, given their lack of reach to persons without these technologies. Even smartphones at present have dramatically unequal distribution globally [241]. Thus, the sociodemographic and geographic reach of technologies required to implement new mental health sensing methods is an important consideration in translational research which seeks to alleviate healthcare disparities.

In addition to allowing for the scaling of mental disorder screening, as well as the early prediction of high risk states, continuous passive sensing which track symptom trajectory in real time would make possible the delivery of just in time adaptive interventions [159], which can provide a tailored intervention when an individual is both in need of the intervention and expected to be most receptive.

Concluding Thoughts

Assessing mental health disorders remains a significant challenge with considerable public health implications, and passive sensing technologies, powered by advances in computing and material sciences, alongside the proliferation of mobile devices, offer a promising solution. Passive sensing has the potential to transform screening and assessment in mental healthcare, allowing for timely and effective care to be provided at scale. However, a prerequisite to such implementation is the critical appraisal of the current research, particularly the

translational potential and limitations. Understanding such factors can help to identify research products ripe for clinical implementation and also help to guide future studies in passive sensing. Our findings highlight key considerations for translation and generalizability, including a trend towards smaller, specialized sample populations, a predominant use of Android platform applications and a reliance on self-reported measures as proxies for mental health outcomes.

Data Availability Statement

The present work represents a narrative review, and thus the manuscript does include any primary data. All studies described in the review are cited for individual reference.

Conflict of Interest Statement

NCJ has received a grant from Boehringer-Ingelheim. NCJ has edited a book through Academic Press and receives book royalties, and NCJ also receives speaking fees related to his research.

References

[1] D. Arias, S. Saxena, and S. Verguet, “Quantifying the global burden of mental disorders and their economic value,” *eClinicalMedicine*, vol. 54, Dec. 2022, doi: 10.1016/j.eclinm.2022.101675.

[2] W. T. Carpenter and B. Kirkpatrick, “The Heterogeneity of the Long-Term Course of Schizophrenia,” *Schizophr. Bull.*, vol. 14, no. 4, pp. 645–652, Jan. 1988, doi: 10.1093/schbul/14.4.645.

[3] D. Goldberg, “The heterogeneity of ‘major depression,’” *World Psychiatry*, vol. 10, no. 3, pp. 226–228, Oct. 2011, doi: 10.1002/j.2051-5545.2011.tb00061.x.

[4] G. Y. Toh and M. W. Vasey, “Heterogeneity in Autonomic Arousal Level in Perseverative Worry: The Role of Cognitive Control and Verbal Thought,” *Front. Hum. Neurosci.*, vol. 11, Mar. 2017, doi: 10.3389/fnhum.2017.00108.

[5] M. V. Heinz, N. X. Thomas, N. D. Nguyen, T. Z. Griffin, and N. C. Jacobson, “Technological Advances in Clinical Assessment,” in *Reference Module in Neuroscience and Biobehavioral Psychology*, Elsevier, 2021. doi: 10.1016/B978-0-12-818697-8.00171-0.

[6] American Psychiatric Association, *Diagnostic and statistical manual of mental disorders: DSM-5.*, 5th ed. Arlington, VA: American Psychiatric Association, 2013.

[7] SAMHSA, “Key substance use and mental health indicators in the United States: Results from the 2021 National Survey on Drug Use and Health,” Center for Behavioral Health Statistics and Quality, HHS Publication No. PEP22-07-01-005, 2022. [Online]. Available: . https://www.samhsa.gov/data/report/2021-nsduh-annual-national-report

[8] T. Insel *et al.*, “Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders,” *Am. J. Psychiatry*, vol. 167, no. 7, pp. 748–751, Jul. 2010, doi: 10.1176/appi.ajp.2010.09091379.

[9] R. Kotov *et al.*, “The Hierarchical Taxonomy of Psychopathology (HiTOP): A dimensional alternative to traditional nosologies,” *J. Abnorm. Psychol.*, vol. 126, no. 4, pp. 454–477, 2017, doi: 10.1037/abn0000258.

[10] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, “New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research,” *JMIR Ment. Health*, vol. 3, no. 2, p. e5165, May 2016, doi: 10.2196/mental.5165.

[11] Pew Research Center, S. 800 Washington, and D. 20036 U.-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, “Mobile Fact Sheet,” Pew Research Center: Internet, Science & Tech. Accessed: Apr. 18, 2022. [Online]. Available: https://www.pewresearch.org/internet/fact-sheet/mobile/

[12] E. a Vogels, “About one-in-five Americans use a smart watch or fitness tracker,” Pew Research Center. Accessed: Oct. 01, 2023. [Online]. Available: https://www.pewresearch.org/short-reads/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/

[13] V. J. Reddi, H. Yoon, and A. Knies, “Two Billion Devices and Counting,” *IEEE Micro*, vol. 38, no. 1, pp. 6–21, Jan. 2018, doi: 10.1109/MM.2018.011441560.

[14] J. Shalf, “The future of computing beyond Moore’s Law,” *Philos. Trans. R. Soc. Math. Phys. Eng. Sci.*, vol. 378, no. 2166, p. 20190061, Jan. 2020, doi: 10.1098/rsta.2019.0061.

[15] A. Trifan, M. Oliveira, and J. L. Oliveira, “Passive Sensing of Health Outcomes Through Smartphones: Systematic Review of Current Solutions and Possible Limitations,” *JMIR MHealth UHealth*, vol. 7, no. 8, p. e12649, Aug. 2019, doi: 10.2196/12649.

[16] M. Sheikh, M. Qassem, and P. A. Kyriacou, “Wearable, Environmental, and Smartphone-Based Passive Sensing for Mental Health Monitoring,” *Front. Digit. Health*, vol. 3, 2021, Accessed: Jul. 13, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdgth.2021.662811

[17] I. Moura, A. Teles, D. Viana, J. Marques, L. Coutinho, and F. Silva, “Digital Phenotyping of Mental Health using multimodal sensing of multiple situations of interest: A Systematic Literature Review,” *J. Biomed. Inform.*, vol. 138, p. 104278, Feb. 2023, doi: 10.1016/j.jbi.2022.104278.

[18] A. Abd-alrazaq *et al.*, “Wearable Artificial Intelligence for Anxiety and Depression: Scoping Review,” *J. Med. Internet Res.*, vol. 25, no. 1, p. e42672, Jan. 2023, doi: 10.2196/42672.

[19] N. C. Jacobson and B. Feng, “Digital phenotyping of generalized anxiety disorder: using artificial intelligence to accurately predict symptom severity using wearable sensors in daily life,” *Transl. Psychiatry*, vol. 12, no. 1, Art. no. 1, Aug. 2022, doi: 10.1038/s41398-022-02038-1.

[20] D. Lekkas and N. C. Jacobson, “Using artificial intelligence and longitudinal location data to differentiate persons who develop posttraumatic stress disorder following childhood trauma,” *Sci. Rep.*, vol. 11, no. 1, Art. no. 1, May 2021, doi: 10.1038/s41598-021-89768-2.
[21] D. Ben-Zeev *et al.*, “CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse.,” *Psychiatr. Rehabil. J.*, vol. 40, no. 3, pp. 266–275, Sep. 2017, doi: 10.1037/prj0000243.
[22] E. K. Czyz, C. A. King, N. Al-Dajani, L. Zimmermann, V. Hong, and I. Nahum-Shani, “Ecological Momentary Assessments and Passive Sensing in the Prediction of Short-Term Suicidal Ideation in Young Adults,” *JAMA Netw. Open*, vol. 6, no. 8, p. e2328005, Aug. 2023, doi: 10.1001/jamanetworkopen.2023.28005.
[23] S. D. Dlima, S. Shevade, S. R. Menezes, and A. Ganju, “Digital Phenotyping in Health Using Machine Learning Approaches: Scoping Review,” *JMIR Bioinforma. Biotechnol.*, vol. 3, no. 1, p. e39618, Jul. 2022, doi: 10.2196/39618.
[24] S. Ware *et al.*, “Automatic depression screening using social interaction data on smartphones,” *Smart Health*, vol. 26, p. 100356, Dec. 2022, doi: 10.1016/j.smhl.2022.100356.
[25] S. M. Narkhede *et al.*, “Machine Learning Identifies Digital Phenotyping Measures Most Relevant to Negative Symptoms in Psychotic Disorders: Implications for Clinical Trials,” *Schizophr. Bull.*, vol. 48, no. 2, pp. 425–436, Mar. 2022, doi: 10.1093/schbul/sbab134.
[26] A. S. Cakmak *et al.*, “Classification and Prediction of Post-Trauma Outcomes Related to PTSD Using Circadian Rhythm Changes Measured via Wrist-Worn Research Watch in a Large Longitudinal Cohort,” *IEEE J. Biomed. Health Inform.*, vol. 25, no. 8, pp. 2866–2876, Aug. 2021, doi: 10.1109/JBHI.2021.3053909.
[27] P. S. Wang, P. A. Berglund, M. Olfson, and R. C. Kessler, “Delays in Initial Treatment Contact after First Onset of a Mental Disorder,” *Health Serv. Res.*, vol. 39, no. 2, pp. 393–416, 2004, doi: 10.1111/j.1475-6773.2004.00234.x.
[28] M. A. Whooley and J. M. Wong, “Depression and Cardiovascular Disorders,” *Annu. Rev. Clin. Psychol.*, vol. 9, no. 1, pp. 327–354, 2013, doi: 10.1146/annurev-clinpsy-050212-185526.
[29] O. M. Farr *et al.*, “Posttraumatic stress disorder, alone or additively with early life adversity, is associated with obesity and cardiometabolic risk,” *Nutr. Metab. Cardiovasc. Dis.*, vol. 25, no. 5, pp. 479–488, May 2015, doi: 10.1016/j.numecd.2015.01.007.
[30] null The Lancet Global Health, “Mental health matters,” *Lancet Glob. Health*, vol. 8, no. 11, p. e1352, Nov. 2020, doi: 10.1016/S2214-109X(20)30432-0.
[31] Health Resources & Services Administration, “Health Professional Shortage Areas,” Health Workforce Shortage Areas. Accessed: Sep. 04, 2023. [Online]. Available: https://data.hrsa.gov/topics/health-workforce/shortage-areas
[32] D. Vigo, G. Thornicroft, and R. Atun, “Estimating the true global burden of mental illness,” *Lancet Psychiatry*, vol. 3, no. 2, pp. 171–178, Feb. 2016, doi: 10.1016/S2215-0366(15)00505-2.
[33] B. Druss and E. Walker, “Mental Disorders and Medical Comorbidity,” *Synth. Proj. Res. Synth. Rep.*, pp. 1–26, Feb. 2011.
[34] E. T. Isometsä, “Psychological autopsy studies – a review,” *Eur. Psychiatry*, vol. 16, no. 7, pp. 379–385, Nov. 2001, doi: 10.1016/S0924-9338(01)00594-6.
[35] A. D. Moreland and J. E. Dumas, “Categorical and dimensional approaches to the measurement of disruptive behavior in the preschool years: A meta-analysis,” *Clin. Psychol. Rev.*, vol. 28, no. 6, pp. 1059–1070, Jul. 2008, doi: 10.1016/j.cpr.2008.03.001.
[36] World Health Organization, “ICD-11.” Accessed: Sep. 02, 2023. [Online]. Available: https://icd.who.int/en
[37] M. L. Savoy and D. T. O’Gurek, “Screening Your Adult Patients for Depression,” *Fam. Pract. Manag.*, vol. 23, no. 2, pp. 16–20, 2016.
[38] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, “The PHQ-9,” *J. Gen. Intern. Med.*, vol. 16, no. 9, pp. 606–613, 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.
[39] D. Colombo *et al.*, “Affect Recall Bias: Being Resilient by Distorting Reality,” *Cogn. Ther. Res.*, vol. 44, no. 5, pp. 906–918, Oct. 2020, doi: 10.1007/s10608-020-10122-3.
[40] S. D. Targum, C. Sauder, M. Evans, J. N. Saber, and P. D. Harvey, “Ecological momentary assessment as a measurement tool in depression trials,” *J. Psychiatr. Res.*, vol. 136, pp. 256–264, Apr. 2021, doi:

10.1016/j.jpsychires.2021.02.012.
[41] S. Shiffman, A. Stone, and M. Hufford, “Ecolocial Momentary Assessment,” *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, Feb. 2008, doi: 10.1146/annurev.clinpsy.3.022806.091415.
[42] M. D. Nemesure *et al.*, “Depressive Symptoms as a Heterogeneous and Constantly Evolving Dynamical System: Idiographic Depressive Symptom Networks of Rapid Symptom Changes among Persons with Major Depressive Disorder.” PsyArXiv, Oct. 27, 2022. doi: 10.31234/osf.io/pf4kc.
[43] D. B. Dwyer, P. Falkai, and N. Koutsouleris, “Machine Learning Approaches for Clinical Psychology and Psychiatry,” *Annu. Rev. Clin. Psychol.*, vol. 14, no. 1, pp. 91–118, May 2018, doi: 10.1146/annurev-clinpsy-032816-045037.
[44] B. Buck *et al.*, “Capturing behavioral indicators of persecutory ideation using mobile technology,” *J. Psychiatr. Res.*, vol. 116, pp. 112–117, Sep. 2019, doi: 10.1016/j.jpsychires.2019.06.002.
[45] “The Balanced Accuracy and Its Posterior Distribution | IEEE Conference Publication | IEEE Xplore.” Accessed: Nov. 05, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5597285?casa_token=XytYNuJq_e8AAAAA:j0h1o2NmgQKwrWiWKJjzIB2YBTY7rlwo8qvk0xxyPDoB5Spy_U7hmKv_fJLQ_bVaYvcaTc7n
[46] D. Hand and P. Christen, “A note on using the F-measure for evaluating record linkage algorithms,” *Stat. Comput.*, vol. 28, no. 3, pp. 539–547, 2017, doi: 10.1007/s11222-017-9746-6.
[47] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.
[48] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[49] M. Tlachac *et al.*, “StudentSADD: Rapid Mobile Depression and Suicidal Ideation Screening of College Students during the Coronavirus Pandemic,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–32, Jul. 2022, doi: 10.1145/3534604.
[50] E. Toto, M. Tlachac, and E. A. Rundensteiner, “AudiBERT: A Deep Transfer Learning Multimodal Classification Framework for Depression Screening,” in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Virtual Event Queensland Australia: ACM, Oct. 2021, pp. 4145–4154. doi: 10.1145/3459637.3481895.
[51] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. in Adaptive computation and machine learning. Cambridge, Massachusetts: The MIT Press, 2016.
[52] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[53] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, 2014, pp. 1724–1734. doi: 10.3115/v1/D14-1179.
[54] M. V. Heinz *et al.*, “Association of Selective Serotonin Reuptake Inhibitor Use With Abnormal Physical Movement Patterns as Detected Using a Piezoelectric Accelerometer and Deep Learning in a Nationally Representative Sample of Noninstitutionalized Persons in the US,” *JAMA Netw. Open*, vol. 5, no. 4, p. e225403, Apr. 2022, doi: 10.1001/jamanetworkopen.2022.5403.
[55] G. Price, M. V. Heinz, A. C. Collins, and N. C. Jacobson, “Detecting Major Depressive Disorder Presence Using Passively-Collected Wearable Movement Data in a Nationally-Representative Sample.” PsyArXiv, Jul. 03, 2023. doi: 10.31234/osf.io/9p4xr.
[56] S. G. Luke, “Evaluating significance in linear mixed-effects models in R,” *Behav. Res. Methods*, vol. 49, no. 4, pp. 1494–1502, Aug. 2017, doi: 10.3758/s13428-016-0809-y.
[57] “Fitbit Official Site for Activity Trackers & More.” Accessed: Nov. 01, 2023. [Online]. Available: https://www.fitbit.com/global/us/home
[58] “Oura Ring. Smart Ring for Fitness, Stress, Sleep & Health.,” Oura Ring. Accessed: Nov. 01, 2023. [Online]. Available: https://ouraring.com
[59] C. Acebo and M. K. LeBourgeois, “Actigraphy,” *Respir. Care Clin. N. Am.*, vol. 12, no. 1, pp. 23–30, viii, Mar. 2006, doi: 10.1016/j.rcc.2005.11.010.
[60] Z. Huang, J. Epps, D. Joachim, and M. Chen, “Depression Detection from Short Utterances via Diverse

Smartphones in Natural Environmental Conditions," in *Interspeech 2018*, ISCA, Sep. 2018, pp. 3393–3397. doi: 10.21437/Interspeech.2018-1743.

[61] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Commun.*, vol. 71, pp. 10–49, Jul. 2015, doi: 10.1016/j.specom.2015.03.004.

[62] M. L. Tlachac, R. Flores, E. Toto, and E. Rundensteiner, "Early Mental Health Uncovering with Short Scripted and Unscripted Voice Recordings," in *Deep Learning Applications, Volume 4*, vol. 1434, M. A. Wani and V. Palade, Eds., in Advances in Intelligent Systems and Computing, vol. 1434. , Singapore: Springer Nature Singapore, 2023, pp. 79–110. doi: 10.1007/978-981-19-6153-3_4.

[63] E. W. McGinnis *et al.*, "Giving Voice to Vulnerable Children: Machine Learning Analysis of Speech Detects Anxiety and Depression in Early Childhood," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2294–2301, Nov. 2019, doi: 10.1109/JBHI.2019.2913590.

[64] R. Wang *et al.*, "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, in UbiComp '14. New York, NY, USA: Association for Computing Machinery, Sep. 2014, pp. 3–14. doi: 10.1145/2632048.2632054.

[65] D. Di Matteo *et al.*, "The Relationship Between Smartphone-Recorded Environmental Audio and Symptomatology of Anxiety and Depression: Exploratory Study," *JMIR Form. Res.*, vol. 4, no. 8, p. e18751, Aug. 2020, doi: 10.2196/18751.

[66] M. R. Mehl, J. W. Pennebaker, D. M. Crow, J. Dabbs, and J. H. Price, "The Electronically Activated Recorder (EAR): a device for sampling naturalistic daily activities and conversations," *Behav. Res. Methods Instrum. Comput. J. Psychon. Soc. Inc*, vol. 33, no. 4, pp. 517–523, Nov. 2001, doi: 10.3758/bf03195410.

[67] J. Rooksby, A. Morrison, and D. Murray-Rust, "Student Perspectives on Digital Phenotyping: The Acceptability of Using Smartphone Data to Assess Mental Health," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow Scotland Uk: ACM, May 2019, pp. 1–14. doi: 10.1145/3290605.3300655.

[68] M. Boukhechba, A. R. Daros, K. Fua, P. I. Chow, B. A. Teachman, and L. E. Barnes, "DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones," *Smart Health*, vol. 9–10, pp. 192–203, Dec. 2018, doi: 10.1016/j.smhl.2018.07.005.

[69] S. Ware *et al.*, "Large-scale Automatic Depression Screening Using Meta-data from WiFi Infrastructure," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 4, p. 195:1-195:27, Dec. 2018, doi: 10.1145/3287073.

[70] T. Liu *et al.*, "The relationship between text message sentiment and self-reported depression," *J. Affect. Disord.*, vol. 302, pp. 7–14, Apr. 2022, doi: 10.1016/j.jad.2021.12.048.

[71] S. S. Ogden and T. Guo, "Layercake: Efficient Inference Serving with Cloud and Mobile Resources," in *2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*, Bangalore, India: IEEE, May 2023, pp. 191–202. doi: 10.1109/CCGrid57682.2023.00027.

[72] R. M. Epstein *et al.*, "'I Didn't Know What Was Wrong:' How People With Undiagnosed Depression Recognize, Name and Explain Their Distress," *J. Gen. Intern. Med.*, vol. 25, no. 9, pp. 954–961, Sep. 2010, doi: 10.1007/s11606-010-1367-0.

[73] M. L. Tlachac, M. Reisch, B. Lewis, R. Flores, L. Harrison, and E. Rundensteiner, "Impact assessment of stereotype threat on mobile depression screening using Bayesian estimation," *Healthc. Anal.*, vol. 2, p. 100088, Nov. 2022, doi: 10.1016/j.health.2022.100088.

[74] K. Demyttenaere, A. Bonnewyn, R. Bruffaerts, T. Brugha, R. De Graaf, and J. Alonso, "Comorbid painful physical symptoms and depression: Prevalence, work loss, and help seeking," *J. Affect. Disord.*, vol. 92, no. 2–3, pp. 185–193, Jun. 2006, doi: 10.1016/j.jad.2006.01.007.

[75] A. Halfin, "Depression: the benefits of early and appropriate treatment," *Am. J. Manag. Care*, vol. 13, no. 4 Suppl, pp. S92-97, Nov. 2007.

[76] A. Madan, M. Cebrian, S. Moturu, K. Farrahi, and A. "Sandy" Pentland, "Sensing the 'Health State' of a Community," *IEEE Pervasive Comput.*, vol. 11, no. 4, pp. 36–45, Oct. 2012, doi: 10.1109/MPRV.2011.79.

[77] A. Dogrucu *et al.*, "Moodable: On feasibility of instantaneous depression assessment using machine

learning on voice samples with retrospectively harvested smartphone and social media data," *Smart Health*, vol. 17, p. 100118, Jul. 2020, doi: 10.1016/j.smhl.2020.100118.

[78] M. Tlachac, R. Flores, M. Reisch, K. Houskeeper, and E. A. Rundensteiner, "DepreST-CAT: Retrospective Smartphone Call and Text Logs Collected during the COVID-19 Pandemic to Screen for Mental Illnesses," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 2, p. 75:1-75:32, Jul. 2022, doi: 10.1145/3534596.

[79] C. M. Jones and E. F. McCance-Katz, "Co-occurring substance use and mental disorders among adults with opioid use disorder," *Drug Alcohol Depend.*, vol. 197, pp. 78–82, Apr. 2019, doi: 10.1016/j.drugalcdep.2018.12.030.

[80] D. Comer-HaGans, B. E. Weller, C. Story, and J. Holton, "Developmental stages and estimated prevalence of coexisting mental health and neurodevelopmental conditions and service use in youth with intellectual disabilities, 2011–2012," *J. Intellect. Disabil. Res.*, vol. 64, no. 3, pp. 185–196, 2020, doi: 10.1111/jir.12708.

[81] L. A. Marsch *et al.*, "The application of digital health to the assessment and treatment of substance use disorders: The past, current, and future role of the National Drug Abuse Treatment Clinical Trials Network," *J. Subst. Abuse Treat.*, vol. 112S, pp. 4–11, Mar. 2020, doi: 10.1016/j.jsat.2020.02.005.

[82] D. Campolo, F. Taffoni, G. Schiavone, C. Laschi, F. Keller, and E. Guglielmelli, "A novel technological approach towards the early diagnosis of neurodevelopmental disorders," *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf.*, vol. 2008, pp. 4875–4878, 2008, doi: 10.1109/IEMBS.2008.4650306.

[83] A. Sano *et al.*, "Identifying Objective Physiological Markers and Modifiable Behaviors for Self-Reported Stress and Mental Health Status Using Wearable Sensors and Mobile Phones: Observational Study," *J. Med. Internet Res.*, vol. 20, no. 6, p. e210, Jun. 2018, doi: 10.2196/jmir.9410.

[84] S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Curr. Opin. Behav. Sci.*, vol. 18, pp. 43–49, Dec. 2017, doi: 10.1016/j.cobeha.2017.07.005.

[85] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *Npj Digit. Med.*, vol. 3, no. 1, p. 43, Mar. 2020, doi: 10.1038/s41746-020-0233-7.

[86] M. Alkhathlan, M. L. Tlachac, L. Harrison, and E. Rundensteiner, "'Honestly I Never Really Thought About Adding a Description': Why Highly Engaged Tweets Are Inaccessible," in *Human-Computer Interaction – INTERACT 2021*, vol. 12932, C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, and K. Inkpen, Eds., in Lecture Notes in Computer Science, vol. 12932. , Cham: Springer International Publishing, 2021, pp. 373–395. doi: 10.1007/978-3-030-85623-6_23.

[87] J. Shin and S. M. Bae, "A Systematic Review of Location Data for Depression Prediction," *Int. J. Environ. Res. Public. Health*, vol. 20, no. 11, p. 5984, May 2023, doi: 10.3390/ijerph20115984.

[88] W. F. Heckler, J. V. De Carvalho, and J. L. V. Barbosa, "Machine learning for suicidal ideation identification: A systematic literature review," *Comput. Hum. Behav.*, vol. 128, p. 107095, Mar. 2022, doi: 10.1016/j.chb.2021.107095.

[89] D. Highland and G. Zhou, "A review of detection techniques for depression and bipolar disorder," *Smart Health*, vol. 24, p. 100282, Jun. 2022, doi: 10.1016/j.smhl.2022.100282.

[90] G. S. Malhi and J. J. Mann, "Depression," *Lancet Lond. Engl.*, vol. 392, no. 10161, pp. 2299–2312, Nov. 2018, doi: 10.1016/S0140-6736(18)31948-2.

[91] E. I. Fried and R. M. Nesse, "Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR*D study," *J. Affect. Disord.*, vol. 172, pp. 96–102, Feb. 2015, doi: 10.1016/j.jad.2014.10.010.

[92] P. Cuijpers, C. F. Reynolds III, T. Donker, J. Li, G. Andersson, and A. Beekman, "Personalized Treatment of Adult Depression: Medication, Psychotherapy, or Both? A Systematic Review," *Depress. Anxiety*, vol. 29, no. 10, pp. 855–864, 2012, doi: 10.1002/da.21985.

[93] A. M. Buch and C. Liston, "Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics," *Neuropsychopharmacology*, vol. 46, no. 1, Art. no. 1, Jan. 2021, doi: 10.1038/s41386-020-00789-3.

[94] C. Otte *et al.*, "Major depressive disorder," *Nat. Rev. Dis. Primer*, vol. 2, no. 1, Art. no. 1, Sep. 2016, doi:

10.1038/nrdp.2016.65.
[95] J. D. Tubbs, J. Ding, L. Baum, and P. C. Sham, “Systemic neuro-dysregulation in depression: Evidence from genome-wide association,” *Eur. Neuropsychopharmacol.*, vol. 39, pp. 1–18, Oct. 2020, doi: 10.1016/j.euroneuro.2020.08.007.
[96] R. Z. Fisch and G. Nesher, “Masked depression,” *Postgrad. Med.*, vol. 80, no. 3, pp. 165–169, Sep. 1986, doi: 10.1080/00325481.1986.11699519.
[97] C. Yue *et al.*, “Automatic Depression Prediction Using Internet Traffic Characteristics on Smartphones,” *Smart Health Amst. Neth.*, vol. 18, p. 100137, Nov. 2020, doi: 10.1016/j.smhl.2020.100137.
[98] J. Lu *et al.*, “Joint Modeling of Heterogeneous Sensing Data for Depression Assessment via Multi-task Learning,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 2, no. 1, p. 21:1-21:21, Mar. 2018, doi: 10.1145/3191753.
[99] R. Razavi, A. Gharipour, and M. Gharipour, “Depression screening using mobile phone usage metadata: a machine learning approach,” *J. Am. Med. Inform. Assoc. JAMIA*, vol. 27, no. 4, pp. 522–530, Apr. 2020, doi: 10.1093/jamia/ocz221.
[100] E. O’Connor *et al.*, *Screening for Depression in Adults: An Updated Systematic Evidence Review for the U.S. Preventive Services Task Force*. in U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews. Rockville (MD): Agency for Healthcare Research and Quality (US), 2016. Accessed: Aug. 30, 2023. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK349027/
[101] A. J. Rush *et al.*, “The 16-Item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression,” *Biol. Psychiatry*, vol. 54, no. 5, pp. 573–583, Sep. 2003, doi: 10.1016/S0006-3223(02)01866-8.
[102] K. Opoku Asare *et al.*, “Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis,” *Pervasive Mob. Comput.*, vol. 83, p. 101621, Jul. 2022, doi: 10.1016/j.pmcj.2022.101621.
[103] S. Saeb, E. G. Lattie, S. M. Schueller, K. P. Kording, and D. C. Mohr, “The relationship between mobile phone location sensor data and depressive symptom severity,” *PeerJ*, vol. 4, p. e2537, Sep. 2016, doi: 10.7717/peerj.2537.
[104] C. Yue *et al.*, “Fusing Location Data for Depression Prediction,” *IEEE Trans. Big Data*, vol. 7, no. 2, pp. 355–370, Jun. 2021, doi: 10.1109/TBDATA.2018.2872569.
[105] A. Pratap *et al.*, “The accuracy of passive phone sensors in predicting daily mood,” *Depress. Anxiety*, vol. 36, no. 1, pp. 72–81, 2019, doi: 10.1002/da.22822.
[106] X. Xu *et al.*, “Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, no. 3, p. 116:1-116:33, Sep. 2019, doi: 10.1145/3351274.
[107] P. Chikersal *et al.*, “Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection,” *ACM Trans. Comput.-Hum. Interact.*, vol. 28, no. 1, pp. 1–41, Feb. 2021, doi: 10.1145/3422821.
[108] R. Bai *et al.*, “Tracking and Monitoring Mood Stability of Patients With Major Depressive Disorder by Machine Learning Models Using Passive Digital Data: Prospective Naturalistic Multicenter Study,” *JMIR MHealth UHealth*, vol. 9, no. 3, p. e24365, Mar. 2021, doi: 10.2196/24365.
[109] B. W. Nelson, C. A. Low, N. Jacobson, P. Areán, J. Torous, and N. B. Allen, “Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research,” *NPJ Digit. Med.*, vol. 3, p. 90, 2020, doi: 10.1038/s41746-020-0297-4.
[110] A. G. Horwitz *et al.*, “Using machine learning with intensive longitudinal data to predict depression and suicidal ideation among medical interns over time,” *Psychol. Med.*, pp. 1–8, Sep. 2022, doi: 10.1017/S0033291722003014.
[111] E. Strubell, A. Ganesh, and A. McCallum, “Energy and Policy Considerations for Deep Learning in NLP,” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3645–3650. doi: 10.18653/v1/P19-1355.
[112] X. Xu *et al.*, “GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 6, no. 4, p. 190:1-190:34, Jan. 2023, doi:

10.1145/3569485.
[113] M. Jamalova and C. Milán, “The Comparative Study of the Relationship Between Smartphone Choice and Socio-Economic Indicators,” *Int. J. Mark. Stud.*, vol. 11, no. 3, p. 11, Jul. 2019, doi: 10.5539/ijms.v11n3p11.
[114] M. Tlachac and S. S. Ogden, “Left on Read: Reply Latency for Anxiety & Depression Screening,” in *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers*, in UbiComp/ISWC ’22 Adjunct. New York, NY, USA: Association for Computing Machinery, Apr. 2023, pp. 500–502. doi: 10.1145/3544793.3563429.
[115] M. Tlachac, V. Melican, M. Reisch, and E. Rundensteiner, “Mobile Depression Screening with Time Series of Text Logs and Call Logs,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Jul. 2021, pp. 1–4. doi: 10.1109/BHI50953.2021.9508582.
[116] M. L. Tlachac and E. A. Rundensteiner, “Depression Screening from Text Message Reply Latency,” in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Jul. 2020, pp. 5490–5493. doi: 10.1109/EMBC44109.2020.9175690.
[117] M. Tlachac and E. Rundensteiner, “Screening For Depression With Retrospectively Harvested Private Versus Public Text,” *IEEE J. Biomed. Health Inform.*, vol. 24, no. 11, pp. 3326–3332, Nov. 2020, doi: 10.1109/JBHI.2020.2983035.
[118] M. Tlachac, A. Shrestha, M. Shah, B. Litterer, and E. A. Rundensteiner, “Automated Construction of Lexicons to Improve Depression Screening With Text Messages,” *IEEE J. Biomed. Health Inform.*, vol. 27, no. 6, pp. 2751–2759, Jun. 2023, doi: 10.1109/JBHI.2022.3203345.
[119] T. Ek, C. Kirkegaard, H. Jonsson, and P. Nugues, “Named Entity Recognition for Short Text Messages,” *Procedia - Soc. Behav. Sci.*, vol. 27, pp. 178–187, 2011, doi: 10.1016/j.sbspro.2011.10.596.
[120] M. Tlachac, E. Toto, and E. Rundensteiner, “You’re Making Me Depressed: Leveraging Texts from Contact Subsets to Predict Depression,” in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, Chicago, IL, USA: IEEE, May 2019, pp. 1–4. doi: 10.1109/BHI.2019.8834481.
[121] M. Tlachac *et al.*, “Text Generation to Aid Depression Detection: A Comparative Study of Conditional Sequence Generative Adversarial Networks,” in *2022 IEEE International Conference on Big Data (Big Data)*, Osaka, Japan: IEEE, Dec. 2022, pp. 2804–2813. doi: 10.1109/BigData55660.2022.10020224.
[122] J. Meyerhoff *et al.*, “Analyzing text message linguistic features: Do people with depression communicate differently with their close and non-close contacts?,” *Behav. Res. Ther.*, vol. 166, p. 104342, Jul. 2023, doi: 10.1016/j.brat.2023.104342.
[123] Y. Zhang *et al.*, “Predicting Depressive Symptom Severity Through Individuals’ Nearby Bluetooth Device Count Data Collected by Mobile Phones: Preliminary Longitudinal Study,” *JMIR MHealth UHealth*, vol. 9, no. 7, p. e29840, Jul. 2021, doi: 10.2196/29840.
[124] F. Matcham *et al.*, “Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study,” *BMC Psychiatry*, vol. 22, no. 1, p. 136, Feb. 2022, doi: 10.1186/s12888-022-03753-1.
[125] C. Oetzmann *et al.*, “Lessons learned from recruiting into a longitudinal remote measurement study in major depressive disorder,” *Npj Digit. Med.*, vol. 5, no. 1, Art. no. 1, Sep. 2022, doi: 10.1038/s41746-022-00680-z.
[126] W. Gerych, E. Agu, and E. Rundensteiner, “Classifying Depression in Imbalanced Datasets Using an Autoencoder- Based Anomaly Detection Approach,” in *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, Newport Beach, CA, USA: IEEE, Jan. 2019, pp. 124–127. doi: 10.1109/ICOSC.2019.8665535.
[127] S. Saeb *et al.*, “Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study,” *J. Med. Internet Res.*, vol. 17, no. 7, p. e175, Jul. 2015, doi: 10.2196/jmir.4273.
[128] L. Canzian and M. Musolesi, “Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis,” in *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, Osaka Japan: ACM, Sep. 2015, pp. 1293–1304. doi: 10.1145/2750858.2805845.
[129] F. Wahle, T. Kowatsch, E. Fleisch, M. Rufer, and S. Weidt, “Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild,” *JMIR MHealth UHealth*, vol. 4, no. 3, p. e111, Sep. 2016, doi:

10.2196/mhealth.5960.
[130] J. Meyerhoff *et al.*, “Evaluation of Changes in Depression, Anxiety, and Social Anxiety Using Smartphone Sensor Features: Longitudinal Cohort Study,” *J. Med. Internet Res.*, vol. 23, no. 9, p. e22844, Sep. 2021, doi: 10.2196/22844.
[131] P. Laiou *et al.*, “The Association Between Home Stay and Symptom Severity in Major Depressive Disorder: Preliminary Findings From a Multicenter Observational Study Using Geolocation Data From Smartphones,” *JMIR MHealth UHealth*, vol. 10, no. 1, p. e28095, Jan. 2022, doi: 10.2196/28095.
[132] Y. Zhang *et al.*, “Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study,” *JMIR Ment. Health*, vol. 9, no. 3, p. e34898, Mar. 2022, doi: 10.2196/34898.
[133] Y. Zhang *et al.*, “Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study,” *JMIR MHealth UHealth*, vol. 9, no. 4, p. e24604, Apr. 2021, doi: 10.2196/24604.
[134] I. Grande, M. Berk, B. Birmaher, and E. Vieta, “Bipolar disorder,” *The Lancet*, vol. 387, no. 10027, pp. 1561–1572, Apr. 2016, doi: 10.1016/S0140-6736(15)00241-X.
[135] M. Berk *et al.*, “History of illness prior to a diagnosis of bipolar disorder or schizoaffective disorder,” *J. Affect. Disord.*, vol. 103, no. 1–3, pp. 181–186, Nov. 2007, doi: 10.1016/j.jad.2007.01.027.
[136] J. R. Calabrese, M. D. Shelton, D. J. Rapport, M. Kujawa, S. E. Kimmel, and S. Caban, “Current research on rapid cycling bipolar disorder and its treatment,” *J. Affect. Disord.*, vol. 67, no. 1, pp. 241–255, Dec. 2001, doi: 10.1016/S0165-0327(98)00161-X.
[137] T. Tanaka, K. Kokubo, K. Iwasa, K. Sawa, N. Yamada, and M. Komori, “Intraday Activity Levels May Better Reflect the Differences Between Major Depressive Disorder and Bipolar Disorder Than Average Daily Activity Levels,” *Front. Psychol.*, vol. 9, p. 2314, Dec. 2018, doi: 10.3389/fpsyg.2018.02314.
[138] S. Melbye *et al.*, “Automatically Generated Smartphone Data in Young Patients With Newly Diagnosed Bipolar Disorder and Healthy Controls,” *Front. Psychiatry*, vol. 12, 2021, Accessed: Jul. 31, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyt.2021.559954
[139] C. N. Kaufmann, A. Gershon, C. A. Depp, S. Miller, J. M. Zeitzer, and T. A. Ketter, “Daytime midpoint as a digital biomarker for chronotype in bipolar disorder,” *J. Affect. Disord.*, vol. 241, pp. 586–591, Dec. 2018, doi: 10.1016/j.jad.2018.08.032.
[140] M. Faurholt-Jepsen *et al.*, “Daily mobility patterns in patients with bipolar disorder and healthy individuals,” *J. Affect. Disord.*, vol. 278, pp. 413–422, Jan. 2021, doi: 10.1016/j.jad.2020.09.087.
[141] C. C. Bennett, M. K. Ross, E. Baek, D. Kim, and A. D. Leow, “Smartphone accelerometer data as a proxy for clinical data in modeling of bipolar disorder symptom trajectory,” *Npj Digit. Med.*, vol. 5, no. 1, Art. no. 1, Dec. 2022, doi: 10.1038/s41746-022-00741-3.
[142] Y. Wu *et al.*, “Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: A systematic review and individual participant data meta-analysis,” *Psychol. Med.*, vol. 50, no. 8, pp. 1368–1380, Jun. 2020, doi: 10.1017/S0033291719001314.
[143] *American Psychiatric Association. Anxiety Disorders. In: Diagnostic and Statistical Manual of Mental Disorders.*, 5th ed. Text Revision. American Psychiatric Association, 2022.
[144] “Anxiety Disorders.” [Online]. Available: https://dictionary.apa.org/anxiety-disorder
[145] B. Bandelow, M. Reitt, C. Röver, S. Michaelis, Y. Görlich, and D. Wedekind, “Efficacy of treatments for anxiety disorders: a meta-analysis,” *Int. Clin. Psychopharmacol.*, vol. 30, no. 4, pp. 183–192, Jul. 2015, doi: 10.1097/YIC.0000000000000078.
[146] R. B. Weisberg, “Overview of Generalized Anxiety Disorder: Epidemiology, Presentation, and Course,” *J Clin Psychiatry*.
[147] K. L. Szuhany and N. M. Simon, “Anxiety Disorders: A Review,” *JAMA*, vol. 328, no. 24, pp. 2431–2445, Dec. 2022, doi: 10.1001/jama.2022.22744.
[148] K. Leonard and A. Abramovitch, “Cognitive functions in young adults with generalized anxiety disorder,” *Eur. Psychiatry*, vol. 56, pp. 1–7, Feb. 2019, doi: 10.1016/j.eurpsy.2018.10.008.
[149] Y. Kim *et al.*, “Screening Tool for Anxiety Disorders: Development and Validation of the Korean Anxiety Screening Assessment,” *Psychiatry Investig.*, vol. 15, no. 11, pp. 1053–1063, Nov. 2018, doi: 10.30773/pi.2018.09.27.2.
[150] M. B. First, “Structured Clinical Interview for the DSM (SCID),” in *The Encyclopedia of Clinical*

Psychology, John Wiley & Sons, Ltd, 2015, pp. 1–6. doi: 10.1002/9781118625392.wbecp351.
[151] "Social Anxiety Disorder." [Online]. Available: https://www.nimh.nih.gov/health/statistics/social-anxiety-disorder
[152] N. C. Jacobson, B. Summers, and S. Wilhelm, "Digital Biomarkers of Social Anxiety Severity: Digital Phenotyping Using Passive Smartphone Sensors," *J. Med. Internet Res.*, vol. 22, no. 5, p. e16875, May 2020, doi: 10.2196/16875.
[153] M. Boukhechba, A. R. Daros, K. Fua, P. I. Chow, B. A. Teachman, and L. E. Barnes, "DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones," *Smart Health*, vol. 9–10, pp. 192–203, Dec. 2018, doi: 10.1016/j.smhl.2018.07.005.
[154] J. Gong *et al.*, "Understanding behavioral dynamics of social anxiety among college students through smartphone sensors," *Inf. Fusion*, vol. 49, pp. 57–68, Sep. 2019, doi: 10.1016/j.inffus.2018.09.002.
[155] N. C. Jacobson and S. Bhattacharya, "Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments," *Behav. Res. Ther.*, vol. 149, p. 104013, Feb. 2022, doi: 10.1016/j.brat.2021.104013.
[156] M. G. Craske *et al.*, "Panic disorder: a review of DSM-IV panic disorder and proposals for DSM-V," *Depress. Anxiety*, vol. 27, no. 2, pp. 93–112, 2010, doi: 10.1002/da.20654.
[157] S. A. and M. H. S. Administration, "Table 3.10, Panic Disorder and Agoraphobia Criteria Changes from DSM-IV to DSM-5." Accessed: Aug. 24, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK519704/table/ch3.t10/
[158] E. W. McGinnis *et al.*, "Discovering Digital Biomarkers of Panic Attack Risk in Consumer Wearables Data." medRxiv, p. 2023.03.01.23286647, Mar. 06, 2023. doi: 10.1101/2023.03.01.23286647.
[159] I. Nahum-Shani *et al.*, "Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support," *Ann. Behav. Med.*, vol. 52, no. 6, pp. 446–462, May 2018, doi: 10.1007/s12160-016-9830-8.
[160] D. J. Stein, M. A. Craske, M. J. Friedman, and K. A. Phillips, "Anxiety Disorders, Obsessive-Compulsive and Related Disorders, Trauma- and Stressor-Related Disorders, and Dissociative Disorders in DSM-5," *Am. J. Psychiatry*, vol. 171, no. 6, pp. 611–613, Jun. 2014, doi: 10.1176/appi.ajp.2014.14010003.
[161] D. G. Kilpatrick, H. S. Resnick, M. E. Milanak, M. W. Miller, K. M. Keyes, and M. J. Friedman, "National Estimates of Exposure to Traumatic Events and PTSD Prevalence Using *DSM-IV* and *DSM-5* Criteria: *DSM-5* PTSD Prevalence," *J. Trauma. Stress*, vol. 26, no. 5, pp. 537–547, Oct. 2013, doi: 10.1002/jts.21848.
[162] R. H. Pietrzak, R. B. Goldstein, S. M. Southwick, and B. F. Grant, "Prevalence and axis I comorbidity of full and partial posttraumatic stress disorder in the United States: Results from Wave 2 of the National Epidemiologic Survey on Alcohol and Related Conditions," *J. Anxiety Disord.*, vol. 25, no. 3, pp. 456–465, Apr. 2011, doi: 10.1016/j.janxdis.2010.11.010.
[163] K. H. Seal, T. J. Metzler, K. S. Gima, D. Bertenthal, S. Maguen, and C. R. Marmar, "Trends and risk factors for mental health diagnoses among Iraq and Afghanistan veterans using Department of Veterans Affairs Health Care, 2002–2008," *Am. J. Public Health*, vol. 99, no. 9, pp. 1651–1658, Sep. 2009, doi: 10.2105/AJPH.2008.150284.
[164] J. Sareen, "Posttraumatic stress disorder in adults: Impact, comorbidity, risk factors, and treatment," *Can. J. Psychiatry*, vol. 59, no. 9, pp. 460–467, Sep. 2014, doi: 10.1177/070674371405900902.
[165] I. R. Galatzer-Levy and R. A. Bryant, "636,120 ways to have posttraumatic stress disorder," *Perspect. Psychol. Sci.*, vol. 8, no. 6, pp. 651–662, Nov. 2013, doi: 10.1177/1745691613504115.
[166] L. D. Kubzansky *et al.*, "The weight of traumatic stress: A prospective study of posttraumatic stress disorder symptoms and weight status in women," *JAMA Psychiatry*, vol. 71, no. 1, p. 44, Jan. 2014, doi: 10.1001/jamapsychiatry.2013.2798.
[167] E. J. Paulus, T. R. Argo, and J. A. Egge, "The Impact of Posttraumatic Stress Disorder on Blood Pressure and Heart Rate in a Veteran Population: Effect of PTSD on Blood Pressure and Heart Rate," *J. Trauma. Stress*, vol. 26, no. 1, pp. 169–172, Feb. 2013, doi: 10.1002/jts.21785.
[168] M.-H. Chen *et al.*, "Risk of stroke among patients with post-traumatic stress disorder: nationwide longitudinal study," *Br. J. Psychiatry*, vol. 206, no. 4, pp. 302–307, Apr. 2015, doi: 10.1192/bjp.bp.113.143610.

[169] Y. Neria *et al.*, “Long-term course of probable PTSD after the 9/11 attacks: A study in urban primary care,” *J. Trauma. Stress*, vol. 23, no. 4, pp. 474–482, Aug. 2010, doi: 10.1002/jts.20544.
[170] K. M. Magruder *et al.*, “Prevalence of posttraumatic stress disorder in Veterans Affairs primary care clinics,” *Gen. Hosp. Psychiatry*, vol. 27, no. 3, pp. 169–179, May 2005, doi: 10.1016/j.genhosppsych.2004.11.001.
[171] R. Kimerling *et al.*, “Brief report: Utility of a short screening scale for DSM-IV PTSD in primary care,” *J. Gen. Intern. Med.*, vol. 21, no. 1, pp. 65–67, Jan. 2006, doi: 10.1111/j.1525-1497.2005.00292.x.
[172] A. Elklit and M. Shevlin, “The structure of PTSD symptoms: A test of alternative models using confirmatory factor analysis,” *Br. J. Clin. Psychol.*, vol. 46, no. 3, pp. 299–313, Sep. 2007, doi: 10.1348/014466506X171540.
[173] C. P. Sullivan, A. J. Smith, M. Lewis, and R. T. Jones, “Network analysis of PTSD symptoms following mass violence,” *Psychol. Trauma Theory Res. Pract. Policy*, vol. 10, no. 1, pp. 58–66, Jan. 2018, doi: 10.1037/tra0000237.
[174] R. A. Parslow, A. F. Jorm, B. I. O’Toole, R. P. Marshall, and D. A. Grayson, “Distress experienced by participants during an epidemiological survey of posttraumatic stress disorder,” *J. Trauma. Stress*, vol. 13, no. 3, pp. 465–471, Jul. 2000, doi: 10.1023/A:1007785308422.
[175] S. Akselrod, D. Gordon, F. A. Ubel, D. C. Shannon, A. C. Berger, and R. J. Cohen, “Power spectrum analysis of heart rate fluctuation: a quantitative probe of beat-to-beat cardiovascular control,” *Science*, vol. 213, no. 4504, pp. 220–222, Jul. 1981, doi: 10.1126/science.6166045.
[176] M. B. Rissling *et al.*, “Circadian Contrasts in Heart Rate Variability Associated With Posttraumatic Stress Disorder Symptoms in a Young Adult Cohort,” *J. Trauma. Stress*, vol. 29, no. 5, pp. 415–421, Oct. 2016, doi: 10.1002/jts.22125.
[177] A. D. McDonald, F. Sasangohar, A. Jatav, and A. H. Rao, “Continuous monitoring and detection of post-traumatic stress disorder (PTSD) triggers among veterans: A supervised machine learning approach,” *IISE Trans. Healthc. Syst. Eng.*, vol. 9, no. 3, pp. 201–211, Jul. 2019, doi: 10.1080/24725579.2019.1583703.
[178] A. Minassian *et al.*, “Association of Predeployment Heart Rate Variability With Risk of Postdeployment Posttraumatic Stress Disorder in Active-Duty Marines,” *JAMA Psychiatry*, vol. 72, no. 10, pp. 979–986, Oct. 2015, doi: 10.1001/jamapsychiatry.2015.0922.
[179] D. J. Biddle, R. Robillard, D. F. Hermens, I. B. Hickie, and N. Glozier, “Accuracy of self-reported sleep parameters compared with actigraphy in young people with mental ill-health,” *Sleep Health*, vol. 1, no. 3, pp. 214–220, Sep. 2015, doi: 10.1016/j.sleh.2015.07.006.
[180] S. M. Patterson, D. S. Krantz, L. C. Montgomery, P. A. Deuster, S. M. Hedges, and L. E. Nebel, “Automated physical activity monitoring: Validation and comparison with physiological and self-report measures,” *Psychophysiology*, vol. 30, no. 3, pp. 296–305, May 1993, doi: 10.1111/j.1469-8986.1993.tb03356.x.
[181] M. Sadeghi, A. D. McDonald, and F. Sasangohar, “Posttraumatic stress disorder hyperarousal event detection using smartwatch physiological and activity data,” *PLOS ONE*, vol. 17, no. 5, p. e0267749, May 2022, doi: 10.1371/journal.pone.0267749.
[182] S. C. Cheng and K. G. Schepp, “Early Intervention in Schizophrenia: A Literature Review,” *Arch. Psychiatr. Nurs.*, vol. 30, no. 6, pp. 774–781, Dec. 2016, doi: 10.1016/j.apnu.2016.02.009.
[183] M. George, S. Maheshwari, S. Chandran, J. S. Manohar, and T. S. Sathyanarayana Rao, “Understanding the schizophrenia prodrome,” *Indian J. Psychiatry*, vol. 59, no. 4, pp. 505–509, 2017, doi: 10.4103/psychiatry.IndianJPsychiatry_464_17.
[184] A. Marconi, M. Di Forti, C. M. Lewis, R. M. Murray, and E. Vassos, “Meta-analysis of the Association Between the Level of Cannabis Use and Risk of Psychosis,” *Schizophr. Bull.*, vol. 42, no. 5, pp. 1262–1269, Sep. 2016, doi: 10.1093/schbul/sbw003.
[185] R. Wang *et al.*, “CrossCheck: toward passive sensing and detection of mental health changes in people with schizophrenia,” in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, in UbiComp ’16. New York, NY, USA: Association for Computing Machinery, Sep. 2016, pp. 886–897. doi: 10.1145/2971648.2971740.
[186] R. Wang *et al.*, “Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–24, Sep. 2017, doi: 10.1145/3130976.

[187] R. Emsley, B. Chiliza, L. Asmal, and B. H. Harvey, “The nature of relapse in schizophrenia,” *BMC Psychiatry*, vol. 13, no. 1, p. 50, Feb. 2013, doi: 10.1186/1471-244X-13-50.
[188] D. A. Adler *et al.*, “Predicting Early Warning Signs of Psychotic Relapse From Passive Sensing Data: An Approach Using Encoder-Decoder Neural Networks,” *JMIR MHealth UHealth*, vol. 8, no. 8, p. e19962, Aug. 2020, doi: 10.2196/19962.
[189] B. Buck *et al.*, “Relationships between smartphone social behavior and relapse in schizophrenia: A preliminary report,” *Schizophr. Res.*, vol. 208, pp. 167–172, Jun. 2019, doi: 10.1016/j.schres.2019.03.014.
[190] C. A. Depp *et al.*, “GPS mobility as a digital biomarker of negative symptoms in schizophrenia: a case control study,” *Npj Digit. Med.*, vol. 2, no. 1, Art. no. 1, Nov. 2019, doi: 10.1038/s41746-019-0182-1.
[191] G. P. Strauss *et al.*, “Validation of accelerometry as a digital phenotyping measure of negative symptoms in schizophrenia,” *Schizophrenia*, vol. 8, no. 1, Art. no. 1, Apr. 2022, doi: 10.1038/s41537-022-00241-z.
[192] P. Jakobsen *et al.*, “PSYKOSE: A Motor Activity Database of Patients with Schizophrenia,” in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, Rochester, MN, USA: IEEE, Jul. 2020, pp. 303–308. doi: 10.1109/CBMS49503.2020.00064.
[193] World Health Organization, “Mental health: suicide prevention.” [Online]. Available: http://www.who.int/mental_health/suicide-prevention/en/
[194] H. Hedegaard, S. C. Curtin, and M. Warner, “Increase in Suicide Mortality in the United States, 1999-2018,” *NCHS Data Brief*, no. 362, pp. 1–8, Apr. 2020.
[195] C. Katz, J. Bolton, and J. Sareen, “The prevalence rates of suicide are likely underestimated worldwide: why it matters,” *Soc. Psychiatry Psychiatr. Epidemiol.*, vol. 51, no. 1, pp. 125–127, Jan. 2016, doi: 10.1007/s00127-015-1158-3.
[196] World Health Organization, “World suicide prevention day media release: suicide prevention.” [Online]. Available: http://www.who.int/mental_health/prevention/suicide/suicideprevent/en
[197] H. R. Lawrence *et al.*, “Prevalence and correlates of suicidal ideation and suicide attempts in preadolescent children: A US population-based study,” *Transl. Psychiatry*, vol. 11, no. 1, p. 489, Sep. 2021, doi: 10.1038/s41398-021-01593-3.
[198] G. Milos, A. Spindler, U. Hepp, and U. Schnyder, “Suicide attempts and suicidal ideation: links with psychiatric comorbidity in eating disorder subjects,” *Gen. Hosp. Psychiatry*, vol. 26, no. 2, pp. 129–135, Mar. 2004, doi: 10.1016/j.genhosppsych.2003.10.005.
[199] A. Reynders, A. J. F. M. Kerkhof, G. Molenberghs, and C. Van Audenhove, “Help-seeking, stigma and attitudes of people with and without a suicidal past. A comparison between a low and a high suicide rate country,” *J. Affect. Disord.*, vol. 178, pp. 5–11, Jun. 2015, doi: 10.1016/j.jad.2015.02.013.
[200] J. T. Walkup, L. Townsend, S. Crystal, and M. Olfson, “A systematic review of validated methods for identifying suicide or suicidal ideation using administrative or claims data: METHODS FOR IDENTIFYING SUICIDE USING CLAIMS DATA,” *Pharmacoepidemiol. Drug Saf.*, vol. 21, pp. 174–182, Jan. 2012, doi: 10.1002/pds.2335.
[201] J.-I. Lee *et al.*, “Prevalence of Suicidal Ideation and Associated Risk Factors in the General Population,” *J. Formos. Med. Assoc.*, vol. 109, no. 2, pp. 138–147, Feb. 2010, doi: 10.1016/S0929-6646(10)60034-4.
[202] K. Szanto, A. Gildengers, B. H. Mulsant, G. Brown, G. S. Alexopoulos, and C. F. Reynolds, “Identification of Suicidal Ideation and Prevention of Suicidal Behaviour in the Elderly:,” *Drugs Aging*, vol. 19, no. 1, pp. 11–24, 2002, doi: 10.2165/00002512-200219010-00002.
[203] A. Haines-Delmont *et al.*, “Testing Suicide Risk Prediction Algorithms Using Phone Measurements With Patients in Acute Mental Health Settings: Feasibility Study,” *JMIR MHealth UHealth*, vol. 8, no. 6, p. e15901, Jun. 2020, doi: 10.2196/15901.
[204] P. Moreno-Muñoz, L. Romero-Medrano, Á. Moreno, J. Herrera-López, E. Baca-García, and A. Artés-Rodríguez, “Passive detection of behavioral shifts for suicide attempt prevention.” arXiv, Nov. 14, 2020. doi: 10.48550/arXiv.2011.09848.
[205] M. L. Barrigon *et al.*, “One-Week Suicide Risk Prediction Using Real-Time Smartphone Monitoring: Prospective Cohort Study,” *J. Med. Internet Res.*, vol. 25, no. 1, p. e43719, Sep. 2023, doi: 10.2196/43719.
[206] J. Rottenberg, F. H. Wilhelm, J. J. Gross, and I. H. Gotlib, “Respiratory sinus arrhythmia as a predictor of

outcome in major depressive disorder," *J. Affect. Disord.*, vol. 71, no. 1–3, pp. 265–272, Sep. 2002, doi: 10.1016/s0165-0327(01)00406-2.
[207] D. Adolph, T. Teismann, T. Forkmann, A. Wannemüller, and J. Margraf, "High frequency heart rate variability: Evidence for a transdiagnostic association with suicide ideation," *Biol. Psychol.*, vol. 138, pp. 165–171, Oct. 2018, doi: 10.1016/j.biopsycho.2018.09.006.
[208] S. T. Wilson *et al.*, "Heart rate variability and suicidal behavior," *Psychiatry Res.*, vol. 240, pp. 241–247, Jun. 2016, doi: 10.1016/j.psychres.2016.04.033.
[209] A. L. Calear, P. J. Batterham, and H. Christensen, "Predictors of help-seeking for suicidal ideation in the community: Risks and opportunities for public suicide prevention campaigns," *Psychiatry Res.*, vol. 219, no. 3, pp. 525–530, Nov. 2014, doi: 10.1016/j.psychres.2014.06.027.
[210] L. Gutiérrez-Rojas, A. Porras-Segovia, H. Dunne, N. Andrade-González, and J. A. Cervilla, "Prevalence and correlates of major depressive disorder: a systematic review," *Braz. J. Psychiatry*, vol. 42, pp. 657–672, Aug. 2020, doi: 10.1590/1516-4446-2020-0650.
[211] J. Hong *et al.*, "Depressive Symptoms Feature-Based Machine Learning Approach to Predicting Depression Using Smartphone," *Healthcare*, vol. 10, no. 7, Art. no. 7, Jul. 2022, doi: 10.3390/healthcare10071189.
[212] J. Busk, M. Faurholt-Jepsen, M. Frost, J. E. Bardram, L. V. Kessing, and O. Winther, "Forecasting Mood in Bipolar Disorder From Smartphone Self-assessments: Hierarchical Bayesian Approach," *JMIR MHealth UHealth*, vol. 8, no. 4, p. e15028, Apr. 2020, doi: 10.2196/15028.
[213] C.-H. Cho, T. Lee, M.-G. Kim, H. P. In, L. Kim, and H.-J. Lee, "Mood Prediction of Patients With Mood Disorders by Machine Learning Using Passive Digital Phenotypes Based on the Circadian Rhythm: Prospective Observational Cohort Study," *J. Med. Internet Res.*, vol. 21, no. 4, p. e11029, Apr. 2019, doi: 10.2196/11029.
[214] J. Gideon, E. M. Provost, and M. McInnis, "MOOD STATE PREDICTION FROM SPEECH OF VARYING ACOUSTIC QUALITY FOR INDIVIDUALS WITH BIPOLAR DISORDER," *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. Spons. Inst. Electr. Electron. Eng. Signal Process. Soc. ICASSP Conf.*, vol. 2016, pp. 2359–2363, Mar. 2016, doi: 10.1109/ICASSP.2016.7472099.
[215] N. Vanello *et al.*, "Speech analysis for mood state characterization in bipolar patients," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2012, pp. 2104–2107. doi: 10.1109/EMBC.2012.6346375.
[216] M. S. Scheeringa, "PTSD in Children Younger Than the Age of 13: Toward Developmentally Sensitive Assessment and Management," *J. Child Adolesc. Trauma*, vol. 41, no. 3, pp. 181–197, Sep. 2011, doi: 10.1080/19361521.2011.597079.
[217] I. M. Raugh *et al.*, "DIGITAL PHENOTYPING ADHERENCE, FEASIBILITY, AND TOLERABILITY IN OUTPATIENTS WITH SCHIZOPHRENIA," *J. Psychiatr. Res.*, vol. 138, pp. 436–443, Jun. 2021, doi: 10.1016/j.jpsychires.2021.04.022.
[218] J. Meyerhoff, K. P. Kruzan, K.-Y. A. Kim, K. Van Orden, and D. C. Mohr, "Exploring the Safety of a General Digital Mental Health Intervention to Effect Symptom Reduction among Individuals with and without Suicidal Ideation: A Secondary Analysis," *Arch. Suicide Res.*, vol. 27, no. 3, pp. 966–983, Jul. 2023, doi: 10.1080/13811118.2022.2096520.
[219] C. G. Walsh *et al.*, "Prospective Validation of an Electronic Health Record–Based, Real-Time Suicide Risk Model," *JAMA Netw. Open*, vol. 4, no. 3, p. e211428, Mar. 2021, doi: 10.1001/jamanetworkopen.2021.1428.
[220] E. M. Kleiman *et al.*, "Can passive measurement of physiological distress help better predict suicidal thinking?," *Transl. Psychiatry*, vol. 11, no. 1, p. 611, Dec. 2021, doi: 10.1038/s41398-021-01730-y.
[221] L. K. Berger, A. L. Begun, and L. L. Otto-Salaj, "Participant recruitment in intervention research: scientific integrity and cost-effective strategies," *Int. J. Soc. Res. Methodol.*, vol. 12, no. 1, pp. 79–92, Feb. 2009, doi: 10.1080/13645570701606077.
[222] M. Tlachac, E. Toto, J. Lovering, R. Kayastha, N. Taurich, and E. Rundensteiner, "EMU: Early Mental Health Uncovering Framework and Dataset," in *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec. 2021, pp. 1311–1318. doi: 10.1109/ICMLA52953.2021.00213.
[223] J. Henrich, S. J. Heine, and A. Norenzayan, "The weirdest people in the world?," *Behav. Brain Sci.*, vol. 33, no. 2–3, pp. 61–83, Jun. 2010, doi: 10.1017/S0140525X0999152X.

[224] R. A. Peterson and D. R. Merunka, “Convenience samples of college students and research reproducibility,” *J. Bus. Res.*, vol. 67, no. 5, pp. 1035–1041, May 2014, doi: 10.1016/j.jbusres.2013.08.010.
[225] J. D. Runyan, T. A. Steenbergh, C. Bainbridge, D. A. Daugherty, L. Oke, and B. N. Fry, “A Smartphone Ecological Momentary Assessment/Intervention ‘App’ for Collecting Real-Time Data and Promoting Self-Awareness,” *PLOS ONE*, vol. 8, no. 8, p. e71325, Aug. 2013, doi: 10.1371/journal.pone.0071325.
[226] C. J. Reback, D. Rünger, J. B. Fletcher, and D. Swendeman, “Ecological Momentary Assessments for Self-Monitoring and Counseling to Optimize Methamphetamine Treatment and Sexual Risk Reduction Outcomes among Gay and Bisexual Men,” *J. Subst. Abuse Treat.*, vol. 92, pp. 17–26, Sep. 2018, doi: 10.1016/j.jsat.2018.06.005.
[227] T. W. Boonstra, J. Nicholas, Q. J. Wong, F. Shaw, S. Townsend, and H. Christensen, “Using Mobile Phone Sensor Technology for Mental Health Research: Integrated Analysis to Identify Hidden Challenges and Potential Solutions,” *J. Med. Internet Res.*, vol. 20, no. 7, p. e10131, Jul. 2018, doi: 10.2196/10131.
[228] A. P. A. American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5)*. Arlington, Virginia: American Psychiatric Association, 2013.
[229] A. G. Horwitz, Z. Zhao, and S. Sen, “Peak-end bias in retrospective recall of depressive symptoms on the PHQ-9,” *Psychol. Assess.*, vol. 35, no. 4, pp. 378–381, Apr. 2023, doi: 10.1037/pas0001219.
[230] D. J. Hallford, D. Rusanov, B. Winestone, R. Kaplan, M. Fuller-Tyszkiewicz, and G. Melvin, “Disclosure of suicidal ideation and behaviours: A systematic review and meta-analysis of prevalence,” *Clin. Psychol. Rev.*, vol. 101, p. 102272, Apr. 2023, doi: 10.1016/j.cpr.2023.102272.
[231] H. Li *et al.*, “Modern deep learning in bioinformatics,” *J. Mol. Cell Biol.*, vol. 12, no. 11, pp. 823–827, Nov. 2020, doi: 10.1093/jmcb/mjaa030.
[232] S. Nepal *et al.*, “COVID Student Study: A Year in the Life of College Students during the COVID-19 Pandemic Through the Lens of Mobile Phone Sensing,” in *CHI Conference on Human Factors in Computing Systems*, New Orleans LA USA: ACM, Apr. 2022, pp. 1–19. doi: 10.1145/3491102.3502043.
[233] R. C. Kessler *et al.*, “Anxious and non-anxious major depressive disorder in the World Health Organization World Mental Health Surveys,” *Epidemiol. Psychiatr. Sci.*, vol. 24, no. 3, pp. 210–226, Jun. 2015, doi: 10.1017/S2045796015000189.
[234] M. K. Larson, E. F. Walker, and M. T. Compton, “Early signs, diagnosis and therapeutics of the prodromal phase of schizophrenia and related psychotic disorders,” *Expert Rev. Neurother.*, vol. 10, no. 8, pp. 1347–1359, Aug. 2010, doi: 10.1586/ern.10.93.
[235] W. E. Copeland *et al.*, “Impact of COVID-19 Pandemic on College Student Mental Health and Wellness,” *J. Am. Acad. Child Adolesc. Psychiatry*, vol. 60, no. 1, pp. 134-141.e2, Jan. 2021, doi: 10.1016/j.jaac.2020.08.466.
[236] L. T. Hoyt, A. K. Cohen, B. Dull, E. Maker Castro, and N. Yazdani, “‘Constant Stress Has Become the New Normal’: Stress and Anxiety Inequalities Among U.S. College Students in the Time of COVID-19,” *J. Adolesc. Health*, vol. 68, no. 2, pp. 270–276, Feb. 2021, doi: 10.1016/j.jadohealth.2020.10.030.
[237] J. A. Elharake, F. Akbar, A. A. Malik, W. Gilliam, and S. B. Omer, “Mental Health Impact of COVID-19 among Children and College Students: A Systematic Review,” *Child Psychiatry Hum. Dev.*, vol. 54, no. 3, pp. 913–925, Jun. 2023, doi: 10.1007/s10578-021-01297-1.
[238] M. Carrasco, “Colleges Seek Virtual Mental Health Services,” Inside Higher Ed. Accessed: Oct. 14, 2023. [Online]. Available: https://www.insidehighered.com/news/2021/09/20/colleges-expand-mental-health-services-students
[239] H. Kobayashi, R. Saenz-Escarcega, A. Fulk, and F. B. Agusto, “Understanding mental health trends during COVID-19 pandemic in the United States using network analysis,” *PLOS ONE*, vol. 18, no. 6, p. e0286857, Jun. 2023, doi: 10.1371/journal.pone.0286857.
[240] S. Yu, “Uncovering the hidden impacts of inequality on mental health: a global study,” *Transl. Psychiatry*, vol. 8, no. 1, Art. no. 1, May 2018, doi: 10.1038/s41398-018-0148-0.
[241] 1615 L. St NW, S. 800 Washington, and D. 20036 U.-419-4300 | M.-857-8562 | F.-419-4372 | M. Inquiries, “Smartphone ownership in advanced economies higher than in emerging,” Pew Research Center’s Global Attitudes Project. Accessed: Oct. 19, 2023. [Online]. Available: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-

world-but-not-always-equally/pg_global-technology-use-2018_2019-02-05_0-01/
[242] R. S. McIntyre *et al.*, “Ecological momentary assessment of depressive symptoms using the mind.me application: Convergence with the Patient Health Questionnaire-9 (PHQ-9),” *J. Psychiatr. Res.*, vol. 135, pp. 311–317, Mar. 2021, doi: 10.1016/j.jpsychires.2021.01.012.
[243] X. Xu *et al.*, “Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–27, Mar. 2021, doi: 10.1145/3448107.
[244] A. A. Farhan *et al.*, “Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data,” in *2016 IEEE Wireless Health (WH)*, Oct. 2016, pp. 1–8. doi: 10.1109/WH.2016.7764553.
[245] N. C. Jacobson, D. Lekkas, R. Huang, and N. Thomas, “Deep learning paired with wearable passive sensing data predicts deterioration in anxiety disorder symptoms across 17–18 years,” *J. Affect. Disord.*, vol. 282, pp. 104–111, Mar. 2021, doi: 10.1016/j.jad.2020.12.086.
[246] M. Tahmasian *et al.*, “Differentiation chronic post traumatic stress disorder patients from healthy subjects using objective and subjective sleep-related parameters,” *Neurosci. Lett.*, vol. 650, pp. 174–179, May 2017, doi: 10.1016/j.neulet.2017.04.042.
[247] M. M. Misgar and M. Bhatia, “Utilizing deep convolutional neural architecture with attention mechanism for objective diagnosis of schizophrenia using wearable IoMT devices,” *Multimed. Tools Appl.*, Oct. 2023, doi: 10.1007/s11042-023-17119-6.
[248] R. Wang *et al.*, “On Predicting Relapse in Schizophrenia using Mobile Sensing in a Randomized Control Trial,” in *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, Austin, TX, USA: IEEE, Mar. 2020, pp. 1–8. doi: 10.1109/PerCom45495.2020.9127365.
[249] M. Tlachac, M. Reisch, and M. Heinz, “Mobile Communication Log Time Series to Detect Depressive Symptoms”.
[250] M. Tlachac, K. Dixon-Gordon, and E. Rundensteiner, “Screening for Suicidal Ideation with Text Messages,” in *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, Athens, Greece: IEEE, Jul. 2021, pp. 1–4. doi: 10.1109/BHI50953.2021.9508486.

Table 1. Predictive Passive Sensing Studies in Depression - Multimodal

Ref.	Year	DATA	OS	Duration	N	Pop.	Sensor / Features	Model	Outcome	Eval
Hong et al. [211]	2022		AND	4 w	106	Psychiatric Patients w/ MDD	ACC, GYRO, GPS, screen usage, call and text, activity, Wifi, Bluetooth	RF	PHQ-9≥10 (SR)	ACC=74.07%
Horwitz et al. [110]	2022	IHS	Watch	2 w	2459	PGY-1's	Sleep, activity	ENR	PHQ-9≥10 (SR)	AUC=0.750
Xu et al. [112]	2022	GLO	AND; iOS	10 w	534	UG	Location, phone usage, activity, sleep, Bluetooth, Wifi, call logs	Reorder	PHQ-4 > 2, BDI-II > 13 (SR)	ROC AUC=0.616
Bai et al. [108]	2021		AND	12 w	261	OP w/ MDD	Sleep, step count, HR, call and text data, app usage, GPS	RF	PHQ-9≥10; Mood State Tracking (SR)	ACC=84.27%
Chikersal et al. [107]	2021		AND; iOS	10 w	138	UG	Bluetooth, location, phone usage, call data, step count, sleep	AdaBoost	BDI-II ≥ 14 (SR)	F1=0.82
McIntyre et al. [242]	2021		AND	90 d	200	CR	GPS, social network activity, call and text logs	XGBoost	PHQ-9 ≥ 5 (SR)	F1=0.91
Razavi et al. [99]	2020		AND		412	CR	Mobile Usage Patterns: log counts and usage duration	RF	BDI-II ≥ 14 (SR)	BA=0.768
Xu et al. [106]	2019		AND;iOS	106 d	138	UG	Bluetooth, call logs, screen status, location, sleep, step count: contextually-filtered [243]	AdaBoost	BDI-II ≥ 14 (SR)	F1=0.881
Lu et al. [98]	2018		AND;iOS;Watch	10 w	103	UG	Location, physical activity, heart rate, sleep	MTL w/ Four Tasks	QIDS ≥ 6 (OA)	F1=0.77
Pratap et al. [105]	2018	BRI	BRI	12 w	271	CR	Call and text data, GPS	RF	PHQ-2 ≥ 3 (SR)	AUC>0.50 for 80.6%

ACC, accelerometry; AND, Android; BRI, Brighten; CR, Crowdsourced Adults; ENR, Elastic Net Regression; GLO, Globem; GYRO, gyroscope; HR, Heart Rate; IHS, The Intern Health Study; MTL; Multi-Task Learning; OA, Objectively Assessed, OP, Outpatient; PGY-1; Post-Graduate Year-One/First Year Training Physician; RF, Random Forest; SR, Subjectively Reported; UG, Undergraduates

Table 2. Predictive Passive Sensing Studies in Depression - Unimodal

Ref.	Year	DATA	OS	Duration	N	Pop.	Sensor/Features	Model	Outcome	Eval
Tlachac et al. [118]	2023	MO, EMU	AND	2 w	88	CR	Text message content: lexical categories [117]	LR	PHQ-9≥10 (SA)	F1=0.79
Tlachac & Ogden [114]	2023	MO, EMU, CAT	AND	2 w	95	CR	Text log metadata: reply latency [116]	RF	PHQ-9≥10 (SA)	BA=0.66
Tlachac et al. [78]	2022	CAT	AND	2-16 w	369	CR	Text and call log metadata	GRU	PHQ-9≥10 (SA)	F1=0.68
Ware et al. [24]	2022		AND		59	UG	Text and call	SVM	PHQ-9≥10 (OA)	F1=0.82
Liu et al. [70]	2021		AND	16 w	219	CR	Text message content: lexical categories	LR w/ L2 regularization	PHQ-8 ≥10 (SR)	AUC=0.72
Tlachac et al. [115]	2021	MO, EMU	AND	2 w	312	CR	Text and call log metadata: time series	LR (text); RF(call)	PHQ-9 ≥10 (SR)	F1=0.72 (outgoing text logs); F1=0.65 (incoming call logs)
Zhang et al. [133]	2021	R-MDD	AND		316	Multisite	Bluetooth	HBLR	PHQ-8 ≥10 (SR)	RMSE=3.891
Yue et al. [104]	2021		AND; iOS		79	UG and GR students	GPS and Wifi	SVM w/ RBFK	PHQ-9≥10 (OA)	F1=0.66 (AND); F1=0.76 (iOS)
Yue et al. [97]	2020		AND; iOS	8 m	79	UG	Internet traffic and usage	SVM	PHQ-9≥10 (OA)	F1=0.80
Ware et al. [69]	2018		AND; iOS	11 m	103	UG	Wifi metadata	SVM-RFE	PHQ-9≥10 (OA)	F1=0.85
Farhan et al. [244]	2016		AND; iOS		79	UG and GR students	GPS, physical activity	SVM	PHQ-9≥10 (OA)	F1=0.82 (AND); F1=0.81 (iOS)

AND, Android;CR, Crowdsourced Adults; CAT, DepreST-CAT; EMU, Early Mental Health Uncovering Framework and DatasetGR, Graduate; GRU, Gated Recurrent Unit; HBLR, Hierarchical Bayesian Linear Regression; LR, Logistic Regression; MO, Moodable; OA, Objectively Assessed; R-MDD, RADAR-MDD; RBFK, Radial Basis Function Kernel; RF, Random Forest; SR, Subjectively Reported; SVM, Support Vector Machine; SVM-RFE, SVM-Recursive Feature Elimination; UG-Undergraduate

Table 3. Predictive Passive Sensing Studies in Bipolar Disorder

Ref.	Year	DATA	OS	Duration	N	Pop.	Sensor / Features	Model	Outcome	Eval
Bennett et al. [141]	2022		iOS	3 m	291	Crowdsourced BD or likely BD patients	Typing dynamics, accelerometry	Random Forest	PHQ-8 change ≥4 points one week before occurring (SR)	AUC=0.9442
Faurholt-Jepsen et al. [140]	2021		AND	4 w - 9m	77	BD patients	location	Decision Trees	BD Detection and State Differentiation (OA)	AUC=0.82 (BD vs HC); AUC=0.82 (EU vs HC); AUC=0.83 (DEP vs HC)

AND, Android; BD, Bipolar Disorder; EU, Euthymic Patients; HC, Healthy Controls; OA, Objectively assessed; SR, Subjectively Reported

Table 4. Predictive Passive Sensing Studies in Anxiety

Ref.	Year	DATA	OS	Duration	N	Pop.	Sensor / Features	Model	Outcome	Eval
Tlachac et al. [78]	2022	CAT	AND	2-16 w	369	CR	Text and call log metadata	GRU	GAD-7 ≥ 10 (SR)	F1=0.58
Jacobson and Feng [19]	2022	NHANES	Actigraph Wearable	7 d	264	CR	Movement time series data	Ensemble Learning Model, XGBoost	GAD symptom severity based on CIDI score (SR)	AUC=0.89
Jacobson et al. [245]	2021	MIDUS	Actigraph Wearable	7 d	265	CR	sleep time, wake time, total activity, average activity, maximum activity	Ensemble Learning, XGBoost	GAD/PAD symptom severity prediction over 17 y based on CIDI score (SR)	AUC = 0.696
Jacobson et al. [152]	2020		Android	2 w	59	UG	ACL, incoming and outgoing call and text logs	XGBoost	SIAS ≥ 34 (SR)	r = 0.70 (between predicted and observed)

ACL, accelerometry; CAT, DepreST-CAT; CR, Crowdsourced; GAD, Generalized Anxiety Disorder; NHANES, National Health and Nutrition Examination Survey; PAD, Panic Disorder; UG, Undergraduates

Table 5. Predictive Passive Sensing Studies in Trauma and Stressor Related Disorders

Ref.	Year	DATA	OS	Duration	N	Pop.	Sensor / Features	Model	Outcome	Eval
Sadeghi, et al. [181]	2022		Watch	7 d	99	100% OP PD	ABA, average heart rate, min. and max. heart rate, heart rate standard deviation, heart rate range	XGBoost	Hyperarousal Event (SR, OA)	AUC=0.70
Cakmak, et al. [26]	2021		WD	8 w	1,618	100% IP PD	Interdaily and intradaily variability, movement, circadian rhythm strength, rest start index, CR, NNI, spectral content, acceleration and deceleration capacity of the heart, total power, low/high frequency ratio, sample entropy, approximate entropy of heart rate, signal quality index	LR	PCL-5/PROM-Pain4a; PTSD, Pain Interference (SR)	AUC=0.70
Lekkas, et al. [20]	2021		SP	7 d	228	100% OP PD	Daily minutes spent away from home, maximum daily radius around home	xgbDART, PLS, LGL, SVM, CIRF	PTSD Diagnosis (OA)	AUC=0.82
McDonald, et al. [177]	2019		AND	3-7 d	100	100% OP PD	FDC, energy ratio, change quantiles, aggregated linear trend	SVM	PTSD Trigger Event (SR)	AUC=0.67
Tahmasian, et al. [246]	2017		AEEG	1 d	64	50% OP PD	Total sleep time, sleep efficiency, , awakenings count, awake duration, sleep stage durations	SVM	PTSD Diagnosis (OA)	Accuracy=65%

ABA, Average Body Acceleration; AEEG, Ambulatory Electroencephalography; CIRF, Conditional Inference Random Forest; CR, Cosinor-based Rhytmometry; FDC, Fourier Decomposition Coefficients; IP, inpatient; LR, Logistic Regression; LGL, Lasso-regularized Generalized Linear Model; NNI, Normal-to-Normal Intervals; OP, outpatient; PLS, Partial Least Squares; RSI, Rest Start Index; SP, Smartphone; SVM, Support Vector Machine; xgbDART, extreme gradient boosting machine with Deep Neural Net Dropout Techniques; WD, Wrist-worn Device

Table 6. Predictive Passive Sensing Studies in Psychotic Disorders

Ref.	Year	DATA	OS	Duration	N	Pop.	Sensor / Features	Model	Outcome	Eval
Misgar, et al. [247]	2023	PK	Watch	24 h	54	41% IP PD	ACT	CNN	PD Dx (OA)	Accuracy=94.0
Narkhede et al. [25]	2022		AND	6 d	92	48% OP PD	GPS, ACL ACL and GPS derived features	RF, KNN	PD Dx; Negative Symptom Presence (OA)	AUC=0.86 (PD); AUC=0.82-0.92 (Negative Symptom)
Adler, et al. [188]	2020	CC	AND	12 m	60	100% OP PD	ACL, App Use, communication log, Microphone, GPS, Screen Activity Hourly Aggregates	NN Autoencoder	PD Relapse (OA)	Sens=0.25; Spec=0.88
Jakobsen, et al. [192]	2020	PK	Watch	12.7 d	54	41% IP PD	ACT Mean, SD, 0 Proportion	LR	PD Dx (OA)	AUC=0.92
Wang, et al. [248]	2020	CC	AND	12 m	61	100% OP PD	ACL, App Use, CL, Microphone, GPS, Screen Activity Hourly Aggregates	RF	PD Relapse (OA)	F1=0.223

ACL, accelerometry; ACT, Actigraphy; AND, Android; BPRS, Brief Psychotic Rating Scale; CC, CrossCheck; CNN, Convolutional neural network; Dx, Diagnosis; EDA, electrodermal activity; GBRT, gradient boosted regression trees; GPS, Global Positioning System; IP, inpatient; KNN, K Nearest Neighbor; LSTM, long- short-term memory network; NN, neural network; OA, objectively assessed; OP, outpatient; PD, psychotic disorder; PK, Psykose; RF, random forest; SD, standard deviation; SR, self reported

Table 7. Predictive Passive Sensing Studies in Suicidal Ideation & Behavior

Ref.	Year	DATA	OS	Duration	N	Pop.	Sensor / Features	Model	Outcome	Eval
Barrigon, et al. [205]	2023		iOS	6 m	225	100% OP PD	Distance traveled, step count, time at home, app usage	HMM	Suicide Attempt (OA)	AUC=0.78
Czyz et al. [22]	2023		FB	8 w	102	100% ER	Resting heart rate, heart rate variability, sleep duration, step count	CART	Modified-C-SSR (SR)	AUC=0.84
Tlachac et al. [249]	2023	MO	AND	2 w	312	54% OP PD	Call and text logs: Count, Length, Contact	LR	PHQ-9 Item 9 (SR)	BA=0.67
Horwitz, et al. [110]	2022		EDA	9 d	2459	100% RMS	Total sleep, sleep duration, sleep efficiency, active minutes, resting heart rate	ENR	PHQ-9 Item 9 (SR)	AUC=0.69
Tlachac, et al. [250]	2021	MO, EMU	AND	1-8 w	66	CR	Empath lexical categories, parts of speech tags, Sentiment	SVC	PHQ-9 Item 9 (SR)	AUC=0.88
Haines-Delmo, et al. [203]	2020		FB, iOS	≤ 7 d	66	100% IP PD	Step count, step frequency, Sleep, phone use metadata	KNN	Binarized C-SSRS (OA)	Mean Accuracy=0.68
Moreno-Muñoz, et al. [204]	2020		AB, AND	346 d	301	100% OP PD	Step count, distance traveled, phone use metadata, location traces, time at home	CPDM	C-SSRS (OA)	AUC=0.71
Dogrucu, et al. [77]	2020	MO	AND	≥ 2 w	335	CR	Voice samples, phone use metadata	RF	PHQ-9 Item 9 (SR)	F1=0.85

AB, Armband; AND, Android; BA, Balanced Accuracy; CART, Classification And Regression Trees; CPDM, Change Point Detection Model; CR, Crowdsourced Adults; C-SSRS, Columbia Suicide Severity Rating Scale; EDA, electrodermal activity; EMU, Early Mental Health Uncovering Framework and Dataset; ER, Emergency Room; FB, FitBit; HMM, Heterogeneous Mixture Model; IP, inpatient; KNN, K Nearest Neighbor; LR, Logistic Regression; MO, Moodable; OP, outpatient; PD, psychotic disorder; RF, random forest; RMS, Rising Medical Students; SVC, Support Vector Classifier