# *Beyond Detection: Towards Actionable Sensing Research in Clinical Mental Healthcare*

# Beyond Detection: Towards Actionable Sensing Research in Clinical Mental Healthcare

DANIEL A. ADLER, Cornell Tech, Cornell University, USA
YUEWEN YANG, Cornell Tech, Cornell University, USA
THALIA VIRANDA, Cornell Tech, Cornell University, USA
XUHAI XU, Columbia University, USA
DAVID C. MOHR, Northwestern University Feinberg School of Medicine, USA
ANNA R. VAN METER, New York University Grossman School of Medicine, USA
JULIA C. TARTAGLIA, Weill Cornell Medicine, USA
NICHOLAS C. JACOBSON, Geisel School of Medicine at Dartmouth College, USA
FEI WANG, Weill Cornell Medicine, USA
DEBORAH ESTRIN, Cornell Tech, Cornell University, USA
TANZEEM CHOUDHURY, Cornell Tech, Cornell University, USA

Researchers in ubiquitous computing have long promised that passive sensing will revolutionize mental health measurement by detecting individuals in a population experiencing a mental health disorder or specific symptoms. Recent work suggests that detection tools do not generalize well when trained and tested in more heterogeneous samples. In this work, we contribute a narrative review and findings from two studies with 41 mental health clinicians to understand these generalization challenges. Our findings motivate research on actionable sensing, as an alternative to detection research, studying how passive sensing can augment traditional mental health measures to support actions in clinical care. Specifically, we identify how passive sensing can support clinical actions by revealing patients' presenting problems for treatment and identifying targets for behavior change and symptom reduction, but passive data requires additional contextual information to be appropriately interpreted and used in care. We conclude by suggesting research at the intersection of actionable sensing and mental healthcare, to align technical research in ubiquitous computing with clinical actions and needs.

CCS Concepts: • **Human-centered computing** → **Empirical studies in ubiquitous and mobile computing**; **Empirical studies in HCI**; • **Computing methodologies** → *Machine learning*; • **Applied computing** → Health informatics.

Additional Key Words and Phrases: passive sensing, mHealth, behavioral health, mental health, clinical decision support, user-centered design, human-computer interaction, digital phenotyping, digital biomarkers, personal sensing, context-awareness

**ACM Reference Format:**
Daniel A. Adler, Yuewen Yang, Thalia Viranda, Xuhai Xu, David C. Mohr, Anna R. Van Meter, Julia C. Tartaglia, Nicholas C. Jacobson, Fei Wang, Deborah Estrin, and Tanzeem Choudhury. 2024. Beyond Detection: Towards Actionable Sensing Research in Clinical Mental Healthcare. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 4, Article 160 (December 2024), 33 pages. https://doi.org/10.1145/3699755

Authors' Contact Information: Daniel A. Adler, daa243@cornell.edu, Cornell Tech, Cornell University, USA; Yuewen Yang, Cornell Tech, Cornell University, USA; Thalia Viranda, Cornell Tech, Cornell University, USA; Xuhai Xu, Columbia University, USA; David C. Mohr, Northwestern University Feinberg School of Medicine, USA; Anna R. Van Meter, New York University Grossman School of Medicine, USA; Julia C. Tartaglia, Weill Cornell Medicine, USA; Nicholas C. Jacobson, Geisel School of Medicine at Dartmouth College, USA; Fei Wang, Weill Cornell Medicine, USA; Deborah Estrin, Cornell Tech, Cornell University, USA; Tanzeem Choudhury, Cornell Tech, Cornell University, USA.


ACM 2474-9567/2024/12-ART160
https://doi.org/10.1145/3699755

## 1 Introduction

Ubiquitous computing technologies continue to permeate our everyday lives. Sensors embedded in these technologies collect fine-grained data on location, sound, light, and movement [15] that can approximate behavior and physiology [2, 27, 119]. Researchers have repurposed this data to monitor the health and well-being of individuals and populations by uncovering associations between digital data, clinical measurements, disease risk [41, 49, 102], and well-being [21]. From a clinical perspective, ubiquitous technologies are appealing because they gather *passive sensing data*: real-time data automatically collected from patients' everyday lives with little-or-no user participation, enabling the collection of clinically-relevant information both inside and outside the clinic, revealing insights that are difficult to discern during infrequent, point-in-time clinical encounters [123].

Almost 15 years of interdisciplinary work in ubiquitous computing, psychology, and psychiatry has focused on using passive sensing data to train machine learning tools that detect mental health status [17, 31, 170]. In this work, we use the term **detection** to refer to research on machine learning models that process passive sensing data to detect individuals in a population experiencing a mental health disorder or specific symptoms. Passive sensing-detection research intersects with areas of computational psychiatry interested in the real-time quantification of mental health from passive data, called *digital phenotyping* [160], *digital biomarkers* [8], *personal sensing* [93, 114], or *behavioral sensing* [113]. Passive sensing data have been applied to detect mental health disorders including depression [6, 177], schizophrenia [18, 168], anxiety [78, 79], and bipolar disorder [16, 56]. The stated "value" of detection research is to create a low-burden method, using passive sensing, to continuously measure mental health outside of the clinic, and identify and manage emergent symptoms early-on [8, 77].

Despite this research and promise, recent work suggests that passive sensing-mental health detection tools do not generalize well when they are trained and tested in more heterogeneous samples. By *generalization*, we refer to the performance of task-specific (eg, depression prediction) detection models on data unseen during model training [131]. By *heterogeneity*, we focus on inter-individual differences, with data collected from similar points in time, devices, and data collection methods. We focus on generalization across individuals because it has been the focus of recent detection research [4, 105, 106, 178]. For example, Table 1 summarizes studies that have trained machine learning models with smartphone passive data to classify individuals experiencing clinically significant depression. These studies suggest that depression classification tools are more accurate in homogeneous samples, for example a sample of collocated undergraduates (AUC=0.81) [170], compared to heterogeneous samples (eg, a larger U.S.-wide sample, AUC=0.55) [4]. Studies using wearable sensing [129], detecting continuous depression symptoms [110, 150, 182], major psychiatric events in schizophrenia [3, 22, 124], and individuals living with bipolar disorder [57] have also struggled to find strong, generalizable associations between passive data and mental health. We recognize these studies are examples, and a systematic review is needed to conclude that detection models do not generalize well. Yet, these findings question the effectiveness of detection, and motivate alternative opportunities for passive sensing research in mental healthcare.

In contrast to detection, a growing line of work in ubiquitous computing and digital mental health has explored how passive sensing can complement traditional mental health measures (eg, patient symptom self-reports, clinician-rated scales) to support clinical actions and decision making [53, 123, 146]. In this study, we use the term **actionable sensing** to refer to research studying how passive sensing data can augment traditional mental health measures to support actions in clinical mental healthcare. Actionable sensing offers a contrasting and alternative research direction to detection work (Figure 1). Detection research focuses on validating how passive sensing can approximate traditional mental health measures, and assumes these approximations generalize across individuals to some extent. In contrast, actionable sensing research does not assume that passive sensing offers a generalizable approximation, and instead envisions passive and traditional mental health data working together to drive more effective care [123].

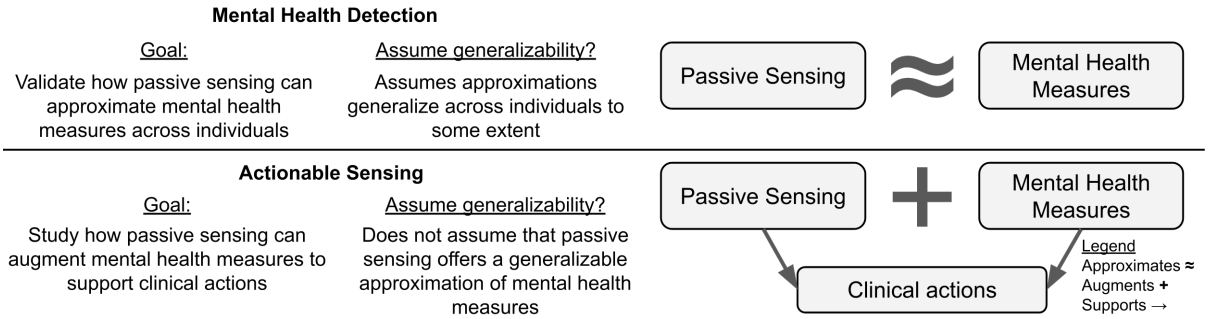


Fig. 1. **Contrasting the goals of mental health detection and actionable sensing research.** Passive sensing refers to passively collected data on observable behavior and physiology, like sleep duration or heart rate variability. Mental health measures include measures of mental health disorders (eg, major depressive disorder) or specific symptoms (eg, loss of interest or pleasure), collected from patient self-reports or clinician-rated scales.

| | Study | N | Sample Description | Depression Criteria | Performance |
|---|---|---|---|---|---|
| Lower heterogeneity | [139] | 18 | Individuals living across the United States | PHQ-9 ≥ 5 | BA = 0.87 |
| | [116] | 57 | Students at a university in the northeastern United States | ICD-10 criteria for major depressive disorder | AUC = 0.82 |
| | [170] | 83 | Students at a university in the northeastern United States | PHQ-4 ≥ 3 | AUC = 0.81 |
| Higher heterogeneity | [178] | 534 | Students at two universities in the United States | PHQ-4 ≥ 3 | AUC = 0.58<br>BA = 0.55 |
| | [4] | 650 | Individuals living across the United States | PHQ-8 ≥ 10 | AUC = 0.55 |
| | [105] | 678 | Students from universities across 8 different countries | Positive versus negative mood score | AUC = 0.52 |
| | [116] | 5,262 | Individuals living across the United States | ICD-10 criteria for major depressive disorder | AUC = 0.57 |

Table 1. Example performance differences across studies using smartphone passive sensing to detect clinically significant depression. Heterogeneity refers to the number of participants ("N") in each study and/or the diversity of participants in the study sample (see "Sample Description"). PHQ-x are validated self-reported depression symptom measures [90–92]. The reported performance for all studies except for [170] were computed by using cross-validation partitioning data by participants. [116] explored depression detection models in both small (N=57) and large (N=5,262) samples. [105, 178] focused on generalization across new datasets. [178] describes the validation accuracy from the model trained using three out of four collected datasets, validated on the fourth dataset. [105] describes the validation accuracy of the model trained and validated on all available data. AUC = the area under the received operating curve; BA = the balanced accuracy.

In this work, we explored through literature review and user studies with mental health experts both (1) detection generalization challenges, and (2) alternative opportunities for actionable sensing research in mental healthcare. We explored these two objectives through the following research questions:

**RQ1:** From literature, why do passive sensing-mental health detection tools not generalize well?
**RQ2:** How do mental health experts interpret these generalization challenges?
**RQ3:** What opportunities do mental health experts identify for actionable sensing research in clinical care?
**RQ4:** Based upon the opportunities identified from RQ3, how could passive sensing data impact care?

We explored RQ1 by conducting a narrative review of ubiquitous computing, psychiatry, and psychology literature (Narrative Review: Section 3). We then conducted a formative interview study with 21 mental health experts, specifically practicing clinicians, to answer RQ2 and RQ3 (Study 1: Sections 4 and 5). Finally, to answer RQ4, we conducted a design probe study where 20 additional mental health clinicians were shown collected passive sensing data, and we explored how this data could impact care (Study 2: Sections 6 and 7). Our contributions are:

- Findings from a narrative review and formative interviews explaining why generalizable detection is challenging: inter-individual differences in mental health symptom presentation and reporting result in patient-specific passive signals associated with symptom changes. These findings imply that generalization challenges will be difficult to solve with ubiquitous computing research because they stem from inherent reliability challenges in mental health measurement.
- Formative interview findings supporting alternative opportunities for actionable sensing research in clinical care. These opportunities focus on using individual-level passive sensing data to identify patient-specific presenting problems; generate insight on the relationships between observed behavior, physiology and symptoms; identify treatment targets that could be modified to reduce symptoms; and monitor patients' treatment response.
- Findings from a second, design probe study that refine these opportunities for actionable sensing research. Our findings chart a path towards future research exploring the use of contextual information that illuminate individual-level, causal associations between passive sensing data and symptoms. These associations could support behavioral interventions and a more effective mental healthcare.

## 2 Related Work

We first provide an overview of the types of data relevant to this work (Section 2.1). We then give an overview of passive sensing-mental health detection research and perspectives on generalizability (Section 2.2). Finally, we conclude with a review of actionable sensing research in mental healthcare (Section 2.3).

### 2.1 Data in Mental Healthcare

There are a variety of data used in mental healthcare most relevant to this work. Traditional data for measuring mental health symptoms include validated *self-reports*, like the PHQ-9 [90] for depression, or the GAD-7 for anxiety [149]. These self-reports assess persistent, 2-week symptoms, while other self-reports can assess in-the-moment symptoms. For example, clinicians may ask patients to self-report their sleep quality each morning, a practice called *experience sampling* [152]. Self-reports, broadly, are considered *active sensing* data because they require users' active participation for data collection [123]. Outside of active sensing, clinicians can assess symptoms by conducting clinical interviews [59] and quantify symptom severity on a clinician-rated symptom scale [173]. Clinician-rated symptom scales may be more frequently used for conditions, like schizophrenia, where impairment affects patients' ability to self-report [133]. Scales and other data collected during clinical interviews are used to determine whether a patient meets criteria to be diagnosed with a mental health disorder (eg, major depressive disorder, schizoaffective disorder) [59, 90].

Clinicians may also wish to collect and review data on patients' observed behavior and physiology in their everyday environments. These signals can be estimated using *passive sensing* data: data collected from digital devices with little-to-no user effort [123]. Passive sensing data can be collected through smartphones or wearables [17], online platforms, such as social media [140] or search engines [23]. Raw data from these devices and platforms may include GPS location data [137], phone usage, voice, light sensor, and accelerometer data [168], text message data [111], and the frequency and/or content of digital interactions [22]. Sleep patterns, like onset/wake times, can be estimated by tracking phone usage [2], or with physiological and behavioral signals collected by many consumer wearable devices [60]. Passive sensing offers a low-burden method to collect data related to mental health that is less arduous to administer than self-reports [123]. But, while clinical guidelines exist for using self-reports, like the PHQ-9, in care [145], similar guidelines do not currently exist for other types of active or passive sensing data, and their use tends to be case-by-case [123].

## 2.2 Passive Sensing-Mental Health Detection and Generalization

Researchers in ubiquitous computing have explored using passive sensing to detect different types of mental health outcomes. These studies build *detection models* using supervised machine learning. These models process passive sensing data to approximate a *ground truth* mental health measure, for example, a self-reported active sensing measure such as the PHQ-9 [97, 136, 170, 178], or in-the-moment symptoms [7, 33, 105]. Studies have also used clinician-rated symptom scales as detection outcomes [169]. Symptom severity scores can be thresholded to produce classification outcomes associated with mental health disorder diagnosis [90]. Researchers have also developed passive sensing tools to detect diagnostic status using outcomes from "gold-standard" structured clinical interviews as ground truths [59, 80].

The performance of detection models are measured using the *generalization accuracy*, estimating how well the model approximates a ground truth mental health measure within data not included in model training [138]. Researchers debate what level of generalization is necessary before introducing models into clinical settings. Studies show that clinical machine learning models are unlikely to generalize across heterogeneous samples [175], and it may be more practical to train and validate models in targeted populations [68, 148]. However, developing models for smaller populations can lead to bias, where models are underparameterized, do not capture the true data distribution [131], and overestimate performance [165]. In addition, there are no best practices to identify target populations with similar passive sensing-mental health relationships, and these relationships are heterogeneous within demographic and socioeconomic subgroups [4, 7]. In this work, it is not our intention to suggest what level of generalization is necessary, but instead explore what generalization challenges reveal about our ability to detect mental health with passive sensing.

## 2.3 Actionable Sensing Research

The term "actionable sensing" comes from sensing and AI research describing actions with data-driven tools [24–26, 98, 180]. Similar to actionable sensing, human-AI interaction researchers have studied how AI can enable actions in clinical care, for example, as an aid to clinical decision-making. These researchers have studied how AI outputs, often predicting disease risk, can be viewed alongside clinical data to support care [101, 147, 179]. In this work, we situate actionable sensing research analogously, to study how passive sensing can augment traditional mental health data to support actions in clinical care.

Researchers in ubiquitous computing, human-computer interaction, and digital mental health have explored opportunities for actionable sensing research in mental healthcare. For example, personal informatics researchers study how passive data can inform conversations between caregivers [36, 51, 53, 67]. Passive data can also support interventions, for example, by measuring the treatment effect in N-of-1 trials [44], or enabling just-in-time adaptive interventions (JITAIs) [89, 118, 130]. Interviews with mental health clinicians have described specific

opportunities for passive data to enable actions in care including reflection and behavior change [46, 123, 146]. These findings align with principles of behavior therapy [37, 172] and the U.S. National Institute of Mental Health's (NIMH) call for focused research on *treatment targets* – eg, disruptive behaviors – over "symptoms", as targets are the "mechanism of action by which the intervention might ultimately modify...symptom(s) of interest" [76]. Contrasting detection research, actionable sensing does not assume that passive sensing data can approximate mental health across a population. Instead, actionable sensing research suggests that there are specific care contexts – for example, patient-specific behavior change in psychotherapy – where relationships between passive sensing and mental health exist and are useful to monitor [123].

Informed by these ideas, we conducted a narrative review followed by two studies with mental health experts, specifically practicing clinicians. We wished to first explore mental health-passive sensing detection challenges, and then offer alternative research directions on actionable sensing in mental healthcare.

## 3 Narrative Review (RQ1)

Motivated by RQ1, we looked to generate a hypothesis suggesting why passive sensing-mental health detection tools do not generalize well. We conducted a narrative review, which surveys literature to support hypothesis generation [127]. Our narrative review suggested the following hypothesis, summarized in Figure 2:

> **Hypothesis**: Passive sensing-mental health detection tools do not generalize well when trained and tested in more heterogeneous samples due to inter-individual differences in mental health symptom presentation and reporting.

To conduct the narrative review, we followed the process of literature accumulation and synthesis described in [127]. First, we accumulated and synthesized recent literature suggesting researchers have struggled to find generalizable passive features that detect mental health (Section 3.1). We then synthesized relevant literature in psychiatry and psychology to explore why these generalization challenges arise (Section 3.2).

### 3.1 Inter-individual Differences in Predictive Passive Sensing Features

*3.1.1 Personalization Improves Model Performance.* Detection models must assume that associations between passive sensing features and mental health symptoms generalize across individuals [131]. Literature suggests this is not the case, and that personalized machine learning models tend to perform better than models trained on an entire sample. Taylor et al. and Tseng et al. showed that models trained within specific groups improved symptom severity prediction [156, 162]. [105, 168] used a different personalization approach to improve performance by adding a small amount of data for each participant to models during training, and testing model performance on participants' remaining data. Outside of mobile sensor data, personalization has improved mental health detection models trained with social media data [141].

*3.1.2 Predictive Passive Features are Heterogeneous.* Studies have also explicitly modeled individual and group (eg, demographic) differences in predictive passive features. For example, papers studying group-personalized models have found that passive features predictive of mental health differ across sample subgroups [156, 162]. [105] found that relationships between smartphone-sensed behavior and mood differ across individuals living within different countries, and that these differences impact detection model performance. Extending these ideas to a single-country sample, a recent study found that a depression detection model had variable performance across demographic and socioeconomic subgroups, and associations between passively-sensed behavior and depression differed across subgroups [4]. Other work has qualitatively compared individuals to identify differences across passive features predicting mental health. In [117], researchers found that wearable sensing measures of self-reported and physiological stress were highly heterogeneous across study participants. [3] conducted a similar analysis and identified that smartphone-sensed behaviors predictive of psychotic relapse were different across individuals.

**a. Differences in symptom presentation**

Patients diagnosed with depression (A and B) can have different passive sensing features predictive of depression because they experience non-overlapping symptoms.

**b. Differences in reporting**

Two patients (C and D) can experience the same underlying behavior change, but only one patient interprets these changes as clinically significant symptoms.

| Patient | 2 Week Passive Sensing Trends | | DSM-5 Major Depressive Disorder Criteria<br>Five or more symptoms must be present nearly everyday in the past two weeks, and at least one symptom is either depressed mood or loss of interest or pleasure.<br>A filled in box indicates a symptom is present and clinically significant. | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Daily Sleep Duration | Daily Step Count | Depressed Mood | Loss of interest or pleasure | Significant weight or appetite changes | Sleep disturbances | Psychomotor changes | Loss of energy or fatigue | Worthlessness or guilt | Impaired concentration or indecisiveness | Suicidal ideation |
| A | + 2 hours | – 2,000 steps | ■ | | ■ | | ■ | ■ | ■ | | |
| B | – 2 hours | + 2,000 steps | | ■ | ■ | ■ | | | | ■ | ■ |
| C | + 3 hours | – 3,000 steps | | ■ | | ■ | ■ | ■ | | ■ | |
| D | + 3 hours | – 3,000 steps | | | | | | | | | |

Fig. 2. **Why do passive sensing-mental health detection tools not generalize well when trained and tested in more heterogeneous samples?** A visual description of our hypothesis from Section 3 suggesting that detection tools do not generalize well due to inter-individual differences in **a.** mental health symptom presentation and **b.** reporting. The table shows data for four patients, including 2-week passive sensing feature trends, as well as depression symptoms patients report experiencing (the black squares). Symptoms represent DSM-5 diagnostic criteria for major depressive disorder [9, 163].

*3.1.3 Challenges with Personalization.* Taken together, these studies suggest that (1) personalized models may improve detection and (2) inter-individual differences in predictive passive sensing features, amplified in more heterogeneous samples, may explain why models need to be personalized to improve detection accuracy. While these findings could imply that personalized models "are the answer" to improving detection tools, they also show that personalization is difficult in practice. As Nagaraj et al. state, future work should embrace "the individuality of stress; no two people experience stress exactly the same way" [117]. [178] describe that individual-level personalization is difficult because it requires fine tuning multiple machine learning models on small amounts of participant data. In addition, training and testing personalized models on small datasets can lead to bias [131, 165]. In the following section, we further explore: why are predictive passive features different across individuals? Are there inter-individual differences in mental health symptoms that explain these findings?

## 3.2 Inter-individual Differences in Symptom Presentation and Reporting

The prior section suggests that it is difficult to identify passive features that predict mental health reliably across individuals. We turned to literature in psychology and psychiatry to better understand why mental health detection with passive sensing is difficult.

*3.2.1 Inter-individual Symptom Differences.* First, we found evidence that individuals can be diagnosed with the same mental health disorder, but experience vastly different symptoms (*inter-individual symptom differences*). For example, Fried and Nesse found that among 3,703 individuals diagnosed with major depressive disorder, there were over 1,030 unique symptom profiles, and the most common profile was found in only 1.8% of the sample [66]. Network analyses of symptom profiles also show that symptoms are shared across disorders, and boundaries between disorders are unclear [28, 42]. These papers suggest that mental health disorders are ill-defined and are not a single construct, and Figure 2a describes why symptom differences can affect passive sensing mental

health-detection tools. Two patients meeting criteria for major depressive disorder (A and B) can have different passive sensing features predictive of depression because they experience non-overlapping symptoms.

*3.2.2 Inter-individual Reporting Differences.* That said, differences across individuals discussed in Section 3.1 extend beyond disorder detection, and were also found in symptom detection (eg, mood, stress) models. Literature suggests this could be due to symptom *reporting differences*. For example, mental health questionnaires intentionally ask individuals to report persistent (eg, 2 week) symptoms, but studies find that some individuals are more likely to report recent or briefly elevated symptoms [11, 74]. For psychotic disorders, like schizophrenia, self-reports are also affected by *patients' insight* – their awareness of symptoms — and self-reported symptoms change as patients gain insight [20]. Similar associations between insight and symptom severity have also been identified for individuals diagnosed with depressive disorders [181]. This work may explain why mental health self-reports are only intended for within-patient monitoring and screening [47, 62, 65]: self-reporting behaviors are likely more consistent within-patients, self-reports often overestimate disease prevalence [95, 158], and thresholds to distinguish clinically significant symptoms from continuous symptom scores are unclear [65]. Figure 2b shows why reporting differences can affect passive sensing-mental health detection tools. Two patients (C and D) can experience the same underlying behavior change, but only one patient interprets these changes as symptoms.

*3.2.3 Are Challenges Due to Self-report?* These findings could suggest that detection challenges are caused by patients' self-reporting behaviors, and clinician-rated scales offer a potentially more reliable detection outcome. It is difficult to prove that clinician-rated assessments are more reliable than self-reports. Field trails of diagnostic criteria show that among 23 mental health disorders, the test-retest and inter-rater reliability (agreement by two clinicians diagnosing independently across clinical visits) was rated as questionable or unacceptable for 9 disorders, potentially because many of these disorders were highly comorbid and difficult to distinguish [40, 132]. Furthermore, studies show that clinicians often miss patient deterioration [72]. Researchers contest that clinician-rated and self-reported scales are synonymous, suggesting self-reports may be more predictive of treatment response and both types of scales complement each other in clinical care [43, 164]. This is why comprehensive neuropsychological evaluations, often used for diagnosis, are a multi-step process that draw from symptom scales, cognitive tests, structured interviews, observation, and collateral information from family and friends [32, 104], though full evaluations are impractical in large-scale, longitudinal detection research.

### 3.3 Narrative Review Conclusion

Taken together, this review hypothesizes that (1) mental health detection tools do not generalize well because passive features predictive of mental health vary across individuals, and (2) this variation may be explained by inter-individual differences in symptom presentation and reporting. This hypothesis suggests that mental health measurement challenges, which affect the assessments used as ground truths to train detection models, make it difficult to identify reliable associations between passive sensing data and mental health symptoms. Solving these challenges may be beyond the scope of ubiquitous computing research in the absence of better ground truths. In the rest of this work, we present findings from user studies with mental health experts that further explore these challenges. In addition, we looked to suggest alternative opportunities for actionable sensing research that are not predicated upon generalizable associations between passive sensing data and mental health measures.

## 4 Study 1 Methods: Formative Interviews

We conducted interviews with mental health experts to interpret the detection model challenges hypothesized in the literature review (RQ2) and identify alternative opportunities for actionable sensing research (RQ3). We recruited mental health clinicians as participants to clarify our reading of the clinical literature with experts

who observe the heterogeneity of mental illness in their practice. In addition, we wished to identify with our participants specific clinical needs that could be supported with actionable sensing.

Methodologically, we drew from work in user-centered design [84, 122, 123] to explore mental health clinicians' current use of data – defined broadly (see Section 2.1) – and how passive data, specifically, may be used within future models of care. In this section, we detail the formative study procedures, including participant recruitment (Section 4.1), and how data was collected and analyzed (Section 4.2). All study procedures were approved by the coauthors' institutional review boards (IRBs).

### 4.1 Participant Recruitment

We enrolled as participants mental health clinicians, including psychiatrists, clinical psychologists, licensed clinical social workers (LCSWs), and licensed mental health counselors (LMHCs). We intentionally recruited providers from these different clinical orientations to better understand how data is used in multiple aspects of treatment, from medication management, managed by psychiatrists, to psychotherapy, often delivered by clinical psychologists and social workers [108]. Participants were recruited via a combination of convenience, purposive, and snowball sampling [52, 70]. Specifically, the first author sent a recruitment email and flier to staff working in psychiatry departments (where multiple types of mental health clinicians practice) across the United States. Interviewed participants often forwarded the recruitment message to their colleagues or to trade organization that have mental health clinicians as their members. Potential participants were asked to provide informed consent after being provided information about the study procedures.

### 4.2 Data Collection and Analysis

Interviews were held via Zoom over two 1-hour sessions, attended by the first three authors, and participants were reimbursed $30 per hour for their time. In the first session, participants described how they used data as a part of their current clinical practice. In the second session, participants described how data could be used within future models of care. Specific interview questions were broad to allow for on-the-spot adaption and probing [19], and included: (in session 1) *what types of data do you use to inform your clinical practice? are there certain types of patients you use this data with? how does this data inform care?*; (in session 2) *what data would you use in the future to inform care outcomes? to improve care quality?*

Interviews were recorded with participants' permission, transcribed by a professional service, and de-identified by the first author. Transcripts were analyzed using a reflexive thematic analysis approach adopted from [29], combining both inductive and deductive elements: codes and themes arose from the data, but were guided by our specific research questions and literature synthesized in the narrative review [30]. The first author qualitatively coded transcripts to develop an initial codebook. Codes were refined, a final codebook was developed, and all transcripts were recoded using the final codebook. Themes were developed from the codes by the first author. The second and third authors, who participated in the interviews, validated that the themes and quotes accurately represented participants' views.

## 5 Study 1 Findings: Formative Interviews

Interviews occurred from October 2023 through January 2024. 21 mental health clinicians participated in the formative interviews (see Table 2). Out of these participants, 19 participants completed the full 2-hour study. One participant (C27) completed only 1 hour of the study, and another participant (C36) was only available for a 30 minute interview. Findings were derived using data from all 21 participants. Due to the small sample size and to protect anonymity, we chose to not collect or present more specific demographic information, including intersectional identities, and we did not interpret how participants' identities influenced our findings. Participants are quoted using study IDs (eg, C45) to retain anonymity.

| Study 1 participants | N=21 mental health clinicians |
|---|---|
| Median (IQR) years of clinical experience | 9 (5 – 15) |
| Clinical training | 11 Clinical Psychology<br>5 Psychiatry<br>4 Clinical Social Work<br>1 Mental Health Counseling |
| Practice setting | 16 Academic Medical Center<br>8 Private Practice<br>2 Community Mental Health Center |
| Geographic location (in the USA) | 18 Northeast<br>2 Southeast<br>1 West Coast |

Table 2. Background information of formative interview study participants. Participants could list multiple practice settings. IQR = interquartile range

## 5.1 Clarifying Why Inter-individual Differences Make Detection Difficult (RQ2)

*5.1.1 Inter-individual Differences in Symptom Presentation.* Participants supported that symptom presentation within conditions differ across individuals and specific sections of the population. For example, C32 highlighted how there are "more specialized depression scales for older adults, or more specialized depression scales for postpartum women because there's just so many symptoms." Some participants preferred to use detailed scales with more specific symptoms because symptom nuances were important for treatment. One participant, C25, mentioned that "the PHQ [a short depression self-report] is usually a screener and it won't account for things like delusions, guilt or taking care of yourself, and it's only nine questions, but the HAM-D [a clinician-rated depression scale] is a lot more nuanced, so I wouldn't use it to screen a patient, but I would use it when I'm monitoring someone's symptoms." C23 agreed, stating "you do have to know a little bit more than just the number. You actually need to see what the answers are on individual [symptom] questions."

Participants also mentioned how patients with different symptoms may present with different problematic behaviors in treatment. As one participant stated, treating depression for some patients might look like "I'd be able to get out of bed and get to work on time", but for other patients it may look like "I'd be much more present with my children" (C37). This is important for detection because it implies that behaviors associated with symptom change, even for the same disorder, differ across patients (Figure 2a). Another participant believed that, in some cases, it made more sense to focus on behaviors over symptoms in treatment, stating how they start monitoring patients "with symptom measures, assuming that was the reason that someone came in. And then I look at their goals and goal progress. Does that line up with the symptoms or do those things seem to be discrepant in some way? Because some goals may be that 'I want to get along better with my partner' and that's not linked with depression or anxiety as closely as other things might be" (C43).

*5.1.2 Inter-individual Differences in Symptom Reporting.* Participants also explained that inter-individual symptom reporting differences could be explained by recall bias, when patients have difficulty recollecting symptoms, or choose to not report specific symptoms. For example, though validated symptom measures, like the PHQ-9 and GAD-7, ask patients to recall persistent symptoms, one participant noted how patients often report transient symptoms, because "some people are not insightful enough to think about, 'how have I felt on average over the last two weeks?' You catch a 12-year-old on a bad day and they're going to answer the assessment [with all high

scores] and make you think that you need to send them to an intensive program" (C38). Another participant noted how the context surrounding scale administration can affect recall, stating that "sometimes the patient does not feel comfortable fully disclosing their answers. We will see a discrepancy sometimes between how they fill out questionnaires with their doctor and how they fill it out with a mental health professional" (C37). One participant gave an example of a patient "who did not endorse any suicidality to the research assistant but did with me in our session together, and the measure was the same" (C39).

Given these concerns, we asked participants how reporting differences would influence our ability to interpret symptom scores across patients and identify generalizable passive features. C43 believed that "people are reporting as accurately as possible, but their understanding of what these measures mean in their everyday life and how these numbers map onto their experience shifts over time." Another participant explained how "every patient fills in their questionnaires differently, but it's relative to how they usually function in life" (C42). If reported symptoms are relative to baseline functioning, individuals can experience similar life changes, and only some individuals will experience and interpret these changes as symptoms, explaining why passive feature changes map to mental health symptoms in some individuals, but not others (Figure 2b). A participant agreed, stating that "what's a 10 to me [on a scale] might not be a 10 to you", but symptom scores are still useful because they are "data driven in nature, trackable, and are a jumping off point for more conversations" (C38).

## 5.2 Identifying Alternative Opportunities for Actionable Sensing Research (RQ3)

In addition to clarifying our hypothesis, we wished to identify with participants alternative opportunities for actionable sensing research (RQ3). These opportunities are summarized in Figure 3 and described in the following sections.

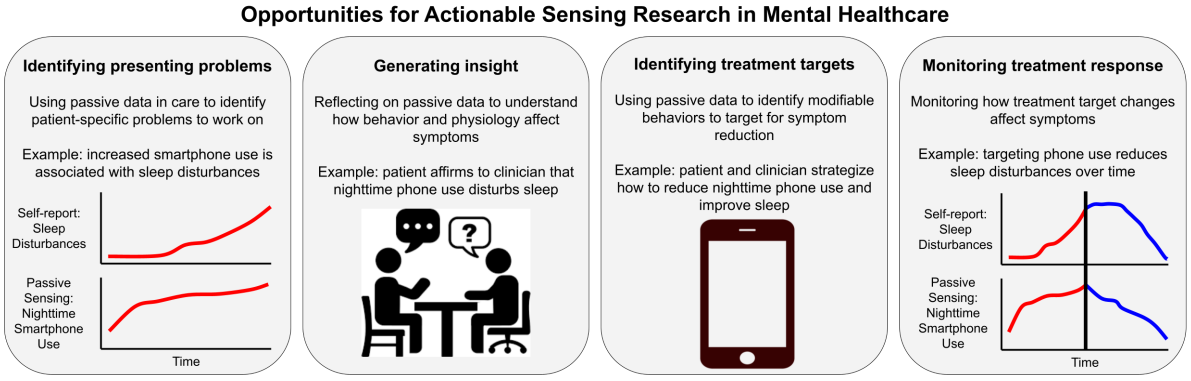


Fig. 3. **Opportunities for actionable sensing research in mental healthcare**, identified in Section 5.2.

*5.2.1 Identifying Presenting Problems.* Participants described how data highlighted patient-specific presenting problems. For example if a patient self-reported "symptoms, let's say sleep disturbance" the data would "inform a deeper assessment what's going on at night" and help clinicians better understand "how does the patient prepare for bed? What's getting in the way?" (C37). Participants believed that passive sensing data could help them more efficiently and effectively identify presenting problems. One participant mentioned how "traditionally you just deal with whatever a person is bringing in" and "try to do a detailed inquiry, not just take everything at face value, but really probe", and passive sensing could "bring evidence, tracking the patient in their daily life" (C37). Another participant described how passive data could inspire questions to further understand behavior,

specifically "is my patient getting up 20 times a night because they have an enlarged prostate and need to pee every five seconds, or because they're super anxious?" (C24).

*5.2.2 Generating Insight.* Participants described how data helped patients generate insight: a greater understanding of themselves and their symptoms. One participant mentioned, in the context of patients with suicidal ideation, that gaining insight, specifically realizing the "connection between negative emotions and suicidality" was a first step towards managing symptoms (C39). This participant was interested in how passive data could improve insight by shedding light on how behaviors impact symptoms, mentioning that patients often "think they're sleeping better than they probably are" and "the objective [passive] data could suggest that your body and brain are not actually resting as much as they can" or "challenge the distorted view that I have poor sleep" (C39). Another participant noted how patients "have difficulty with objectively monitoring their own experiences" and thus saw "for things like substance use, sleep, behavioral stuff that we can monitor easily, I see a lot of utility in that [passive sensing]" (C34).

*5.2.3 Identifying Treatment Targets.* Participants described how they used data to manage patients' specific presenting problems by identifying treatment targets – modifiable behavioral or physiological signals that mediate downstream mental health symptoms. One participant, a child and adolescent psychologist, described that getting their patients to "school is a target" as well as "their ability to engage in exposures" (C30), which were everyday, fearful, situations that their patients, who were living with obsessive-compulsive disorder (OCD), typically avoided. Our participants were interested in how passive data could be used to identify and monitor patient-specific treatment targets. For example, one participant stated that for bipolar disorder, where "the main treatment that you do is social rhythm therapy" where patients aim to "get to sleep at a similar time" and you "actually have them track what time they go to sleep" and "literal sleep data would be amazing for that" (C42). Another participant gave a specific example on how passive sensing data from a wrist-worn actigraphy sensor informed treatment. The participant was treating "a seven year old who was insistent that she was staying in bed all night, but her mom said she was tired all the time." The participant gave the patient an actigraph, and upon reviewing the data, found that the patient "was awake and playing with the dogs", and that she asked the patient's mom "to take the dogs out of the room, and two weeks later, the patient was sleeping fine" (C35).

*5.2.4 Monitoring Treatment Response.* Participants also used data to understand patients' specific treatment response, including what treatments patients respond to, and how patients respond to treatment. As C43 described, data gave "more insight into how they respond, giving me a sense of like, is treatment working? Are we on the right track? Or are we off track?" Participants highlighted the ways in which passive data could give them tangible, longitudinal data to show patients' treatment response. For example, C23 mentioned how "people who are depressed are usually less energetic, they're less motivated. So if you see people are walking more, or moving more, that's something else that you can think about" or "if the patient tells me that she's been walking twice as much as usual and that her mood is better, then I'll take it. That's good news" (C31).

## 5.3 Study 1 Conclusion

Our formative interview findings highlight both the complexities of detection while inspiring alternative opportunities for actionable sensing research. These opportunities focus on how patients and their clinicians could identify passive sensing data relevant for care and act on this data to reduce symptoms. In the rest of this paper, we describe findings from a second study that refine these opportunities and inform future research on passive sensing in clinical mental healthcare.

## 6 Study 2 Methods: Design Probe (RQ4)

We conducted a follow-up study to explore how the specific opportunities for actionable sensing research identified in study 1 could impact care (RQ4). Specifically, we conducted a mixed-methods design probe study, where we showed mental health clinicians real, de-identified passive sensing and self-reported mental health data, and asked them qualitative and quantitative questions to explore using passive data in care. In this section, we detail our study methodology. Section 6.1 details the visualization tool we created to display passive and self-reported data. We then detail the study design in Section 6.2, and participant recruitment, data, and analysis procedures in Section 6.3. Study procedures were approved by the coauthors' IRB.

### 6.1 Visualization Tool

We built a design probe using the Streamlit Python library [153]. Our study goal *was not* to validate the utility of this specific probe, or to identify broadly generalizable conclusions about the use of passive data in care. Instead, in the design probe tradition [5, 69, 166], we wished to deeply study the impact of using passive data to support clinical actions. The de-identified data the tool displayed was collected during a U.S.-based NIMH-funded study to identify associations between smartphone passive sensing data and depression symptoms (see [4, 110, 150] for data collection methods). A screenshot of the tool can be found in Figure 4. Note that we use the word "patient" to refer to individuals whose data was displayed on the design probe. The probe had the following affordances:

(1) **Individual-level data**: Findings in Section 5 suggest that clinicians primarily use data at an individual-level. We thus displayed data on the tool for one patient at a time.
(2) **Longitudinal data**: Findings in Sections 5.2 suggest that clinicians were interested in using passive data to identify patient-specific presenting problems and treatment response. We displayed longitudinal total depression symptom severity scores, self-reported using the PHQ-8 [92], responses to specific PHQ-8 questions as symptoms, and passive sensing data to suggest presenting problems (if symptoms worsen) or treatment response (if symptoms improve). *PHQ-8 Total* scores range from 0 to 24 and higher scores indicate greater symptom severity. Symptom scores range from 0 to 3, where higher scores represent more frequent symptoms. We displayed two passive sensing-symptom pairs. 1) The first PHQ-8 question, measuring patients' *(1) Loss of Interest/Pleasure*, which was paired with passive sensing data measuring the daily *Percentage Time at Home*: the percentage of time a patient spent at home throughout a day. Increased time at home can suggest social isolation, a sign of loss of interest or pleasure [171], and improving real-life social contact can reduce depression symptoms [109]. 2) The third question on the PHQ-8 measuring *(3) Sleep Disturbances* paired with daily *Nighttime Screen On Events*: screen on events recorded between 12-6 AM. Screen on events can indicate sleep disturbances [54, 134], and reducing late night phone use can reduce sleep disturbances [100].
(3) **Associated slopes**: Findings in Section 5.2 suggest how clinicians perceive using passive data to improve patient insight by identifying changes in behavior and physiology associated with mental health symptoms. We thus displayed data suggesting associated slopes between mental health self-reports and passively sensed-behaviors. In addition, our formative findings describe how passive data could be used to identify and monitor treatment targets (reduce phone usage to decrease sleep disturbances; reduce time at home to decrease loss of interest or pleasure). We hoped the data would inspire conversations to explore associations in data as treatment targets.

### 6.2 Study 2 Design

Figure 5 summarizes the study design. One participant attended each study session, and we conducted the study over Zoom. After consenting, participants provided information about their clinical background. We then introduced participants to the design probe, showing data from an "educational patient" to teach participants

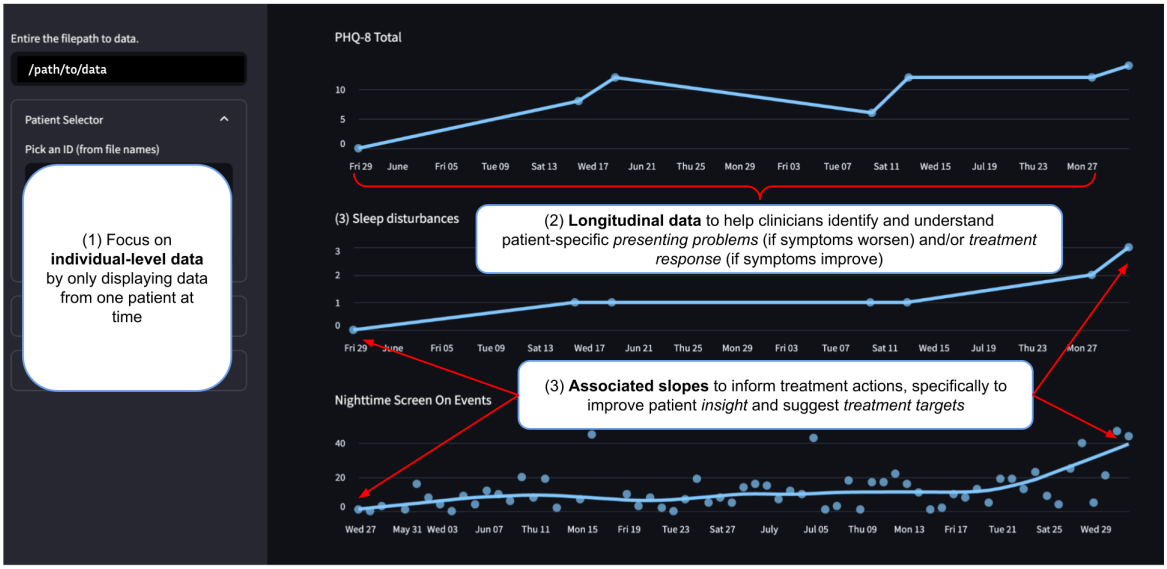


Fig. 4. **The design probe**. Labeled affordances are further described in Section 6.1, and are motivated by our findings in Section 5.2. Dots represent data points, and lines connect data points in the PHQ-8 Total and symptom (sleep disturbance) graphs, but indicate average trends (a LOESS line) for passive data, in this case, "Nighttime Screen On Events".

about the different data types the probe displayed. After learning about the data, participants rated their baseline familiarity with each data type: *from your prior experience, how familiar are you with each type of data (PHQ-8 total, symptom, passive sensing) we have reviewed? 1 (no familiarity) to 5 (lots of familiarity).*

*6.2.1 Patient Cases.* We then showed participants data from four different patients (*patient cases*), one patient at a time. The order of the patient cases was randomized across participants to reduce order bias. During each case, participants were first shown a patient's longitudinal PHQ-8 total score. We then added symptom (sleep disturbances OR loss of interest/pleasure) data to the screen, followed by passive data (nighttime phone use OR percentage of time at home). Similar to [147], for each patient, we kept the data type presentation order fixed, adding data types one at a time so participants could gradually acquaint themselves to the patient case.

The four patient cases were inspired by prior work investigating AI clinical decision support tools [101, 147]. In this work, clinicians were shown AI-generated recommendations that were both in concordance with and subverted their expectations. Analogously, we showed participants two patient cases that contained expected behavior-depression relationships from literature [110, 139, 159] (**positive, + cases**), and two cases that subverted these expectations (**negative, – cases**) to understand how passive data impacts care when it confirms and challenges assumptions about how behavior impacts mental health. Four patient cases with apparent and sound associations were chosen from collected data in collaboration with a non-participant practicing clinician who co-authored this work. The cases (described in Figure 5a) were:

**Phone+** PHQ-8 total score, sleep disturbances, and nighttime screen on events *all increase.*
**Phone–** PHQ-8 total score and sleep disturbances *decrease*, but nighttime screen on events *increases.*
**Home+** PHQ-8 total score, loss of interest/pleasure, and percentage time at home *all increase.*
**Home–** PHQ-8 total score and loss of interest/pleasure *decrease*, but percentage time at home *increases.*

*6.2.2 Question Blocks.* We asked participants quantitative and qualitative questions as we showed them patient data. These questions were motivated by concepts including working alliance, specifically how well clinicians understand patients' treatment goals [71, 99], technology acceptance [38, 45, 161] and perceived usefulness [101, 147], which may influence clinicians' use of passive data. Within each patient case, as we added a data type to the screen (eg, the PHQ-8 total score, the symptom, and then passive sensing data), we asked participants *what observations could you make about this patient, using this data?*; *how you would approach treating this patient, once*

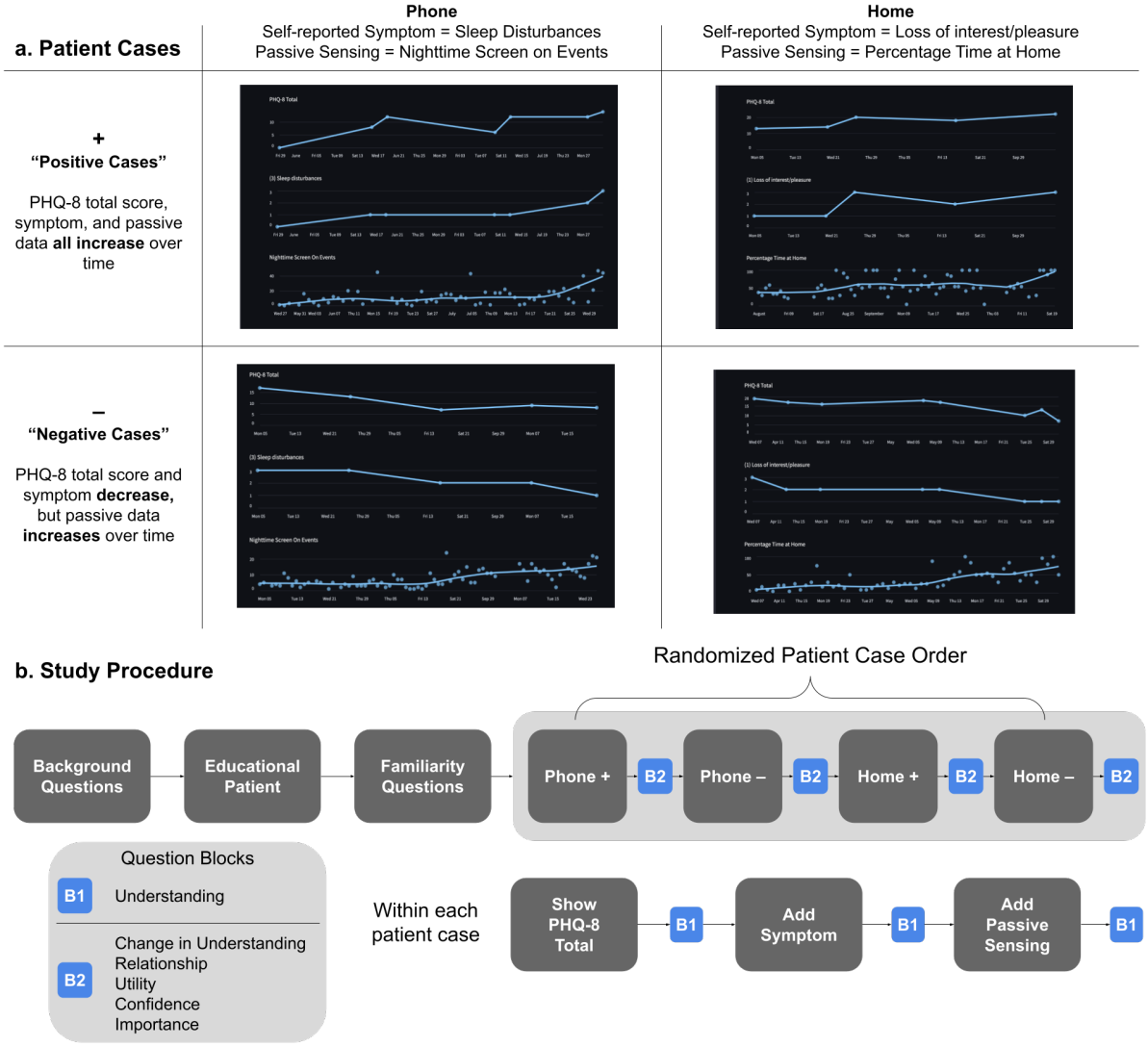


Fig. 5. **Study 2 design. a** describes the four patient cases participants viewed in the study. **b.** describes the study procedure. We gradually showed participants data (the PHQ-8 total, symptom, passive sensing) within each case, and intermixed question blocks to understand how different data types impacted participants' understanding of patients, and interest in using passive data to inform treatment actions. See Section 6.2 for more details.

*they came in for their visit, using this data?*; and quantitatively to rate *on a scale of 1 (low) to 5 (high), based upon the data on the screen, how much do you feel you understand this patient?* This question block is labeled as **B1** in Figure 5b, and we called this quantitative question the **Understanding** question in Figure 5 and Section 7.

After we finished showing participants data for a single patient, we asked the following quantitative questions (**B2** in Figure 5b): **Change in Understanding**, *did the smartphone [passive sensing] data change your understanding of this patient? 1 (little-to-no change) to 5 (great change)*; **Relationship**, *did you see a relationship between the smartphone data and the patient's self-reported symptoms? (no relationship; as the smartphone data increased, the symptom increased; as the smartphone data increased, the symptom decreased)*; **Utility**, *how useful was the smartphone data to help you understand this patient? 1 (not useful) to 5 (very useful)*; **Confidence**, *if a patient shared this smartphone data with you, how confident would you feel using this data as a part of treatment? 1 (not confident) to 5 (very confident)*; and **Importance**, *rank order the importance of each type of data (PHQ-8 total, symptom, and smartphone) for understanding and treating this patient. (1 = highest rank, 3 = lowest rank).* We probed participants to collect qualitative data that further explained numerical responses.

### 6.3 Recruitment, Data Collection, and Analysis

We recruited and reimbursed mental health clinicians as participants using the same methods described in Section 4.1. Recruited participants did not participate in study 1. The first two authors attended the study sessions. Interviews were recorded with participants' consent, transcribed, and de-identified by the first author. The first two authors analyzed the transcripts following the same reflexive coding procedure described in Section 4.2.

## 7 Study 2 Findings: Design Probe (RQ4)

20 mental health clinicians completed study 2 from January through March 2024 (see Table 3). We refrained from conducting significance tests of the quantitative results because the small sample size (n=20) would result in low statistical power and undermine any conclusions [34]. Instead, we contribute a descriptive analysis of our findings. Participants are quoted using study IDs (eg, U10) to retain anonymity.

| Study 2 participants | N=20 mental health clinicians |
|---|---|
| Median (IQR) years of clinical experience | 16.5 (10 – 23.5) |
| Clinical training | 7 Clinical Psychology<br>7 Clinical Social Work<br>6 Psychiatry |
| Practice setting | 17 Private Practice<br>16 Academic Medical Center<br>14 Community Mental Health Center<br>10 Non-academic Health System |
| Geographic location (in the USA) | 16 Northeast<br>2 Midwest<br>2 Southeast |

Table 3. Background information of mental health clinicians participating in study 2. These participants did not take part in study 1. Participants could list multiple practice settings. IQR = interquartile range

The majority of participants reported a high baseline familiarity (Figure 6, left) with the PHQ-8 total score (1 = no familiarity, 5 = high familiarity, n=18, 90% responses $\geq$ 4), loss of interest/pleasure (n=20, 100% $\geq$ 4), and sleep

disturbance (n=19, 95% ≥ 4) symptoms, but low familiarity with the nighttime screen on events (n=14, 70% ≤ 2) and percentage of time at home (n=19, 95% ≤ 2) passive sensor data. Participants often described that while they had never used passive data in clinical practice, they often asked about "phone use, especially at night, especially if there's sleep disturbances" (U21) and that they had experiences using and tracking "how much time patients spend on their phone at nighttime as part of sleep hygiene" (U14). In contrast, participants were slightly less familiar with the percentage of time at home measure, but participants could see the relationship between time at home, "isolating and things like that" (U07) and that "if you're depressed, you might be spending a lot of time at home" (U13). Most participants were also able to identify the intended associations (Figure 6, right) between the passive sensor and mental health symptom data slopes for each patient case (identification rates: Phone+ n=19, 95%; Phone− n=20, 100%; Home+ n=17, 85%; Home− n=20, 100%).

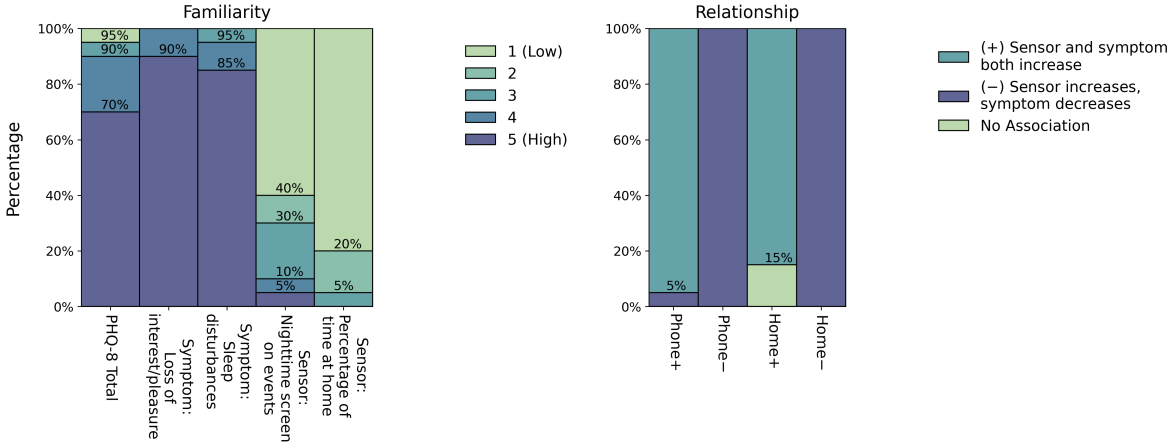


Fig. 6. **Baseline familiarity and validating the relationships in the data. Left.** Histograms show response distributions to the question: *from your prior experience, how familiar are you with each type of data we have reviewed? 1 (no familiarity) to 5 (lots of familiarity)*. The x-axis describes the data types participants were shown. **Right.** We validated if participants noticed the intended relationships between the passive and mental health symptom data slopes within each patient case. Bars are specific to the patient cases described in Figure 5, and the + and − on the x-axis ticks identifies the intended associations. On all histograms, numbers above the bars indicate the cumulative distribution (eg, 90% of values ≥ 3).

## 7.1 Refining Opportunities for Actionable Sensing

*7.1.1 Identifying Presenting Problems.* Participants felt the passive data increased their understanding of the patient when it more clearly identified meaningful signal on patients' presenting problems. For example, 12 (60%) participants stated that the nighttime phone use data increased their understanding (score change ≥ 1, see Figure 7, top-left) when phone usage increased with sleep disturbances over time (Phone+ case). Participants stated that the Phone+ case was in "line with what one might expect" and that they were "starting to have a more coherent view [of the patient], starting to develop some more clear hypotheses of what might be happening" (U08). Another participant mentioned how the passive data gave "validity to what's actually happening with that patient" (U12). Participants, in general, were fairly confident using passive sensing data when they could more easily reason through the data to identify presenting problems. For example, 18 (90%) participants were reasonably confident (score ≥ 4, on a scale of 1, low confidence, to 5, very confident, Figure 7, middle-right) using the passive data in the Phone+ case because "there's a clearer correlation with sleep" (U01).

*7.1.2 Challenging Assumptions.* Participants were also interested in passive data when it challenged their assumptions about how behavior affects mental health. For example, 9 (45%) participants ranked that the passive data led to a greater change in understanding when it subverted their expectations (higher score for – versus + in both Phone and Home cases, see Figure 7, top-right). In the Phone– case, where nighttime phone use increased but symptoms decreased, U07 stated that the data gave them "a different perspective", and another participant interpreted that the data could indicate that the patient "works a night shift, maybe they're asleep from 9:00 AM to 4:00 PM, so then they're on their phone at night" (U06).

*7.1.3 Insight Generation.* Participants were interested in reflecting on the data with their patients to improve insight. For example, U16 mentioned how in the Home– case they thought that the patient could be "anxious and depressed, and if they stay home, the anxiety drops out and they feel a little bit better" and they could ask their patient "at sweet spots in your life do you like being at home?" to understand if time at home was "an avoidance strategy". Another participant (U02) mentioned how time at home data could help patients gain insight because "the biggest issue you have with anyone who's struggling with anxiety or depression is that they view everything through what they're feeling in that moment" and "they're not always the most accurate reporters".

*7.1.4 Identifying Treatment Targets.* Participants stated how the passive data in the – cases may indicate treatment targets: behaviors that lead to symptom reduction. U20 mentioned how the Phone– case, where phone use increased but symptoms decreased, the patient could "wake up in the night, and they have to put on their hypnosis or meditation app". U03 mentioned in the Home– cases that the passive data "is giving me some context to now we [the patient] spend more time at home and we feel better." In contrast, some participants (7, 35%) participants believed that the nighttime phone usage data was more useful (higher utility score, Figure 7, middle-left) in the Phone+ case than the Phone– case, because in this case, they felt the treatment target was clearer. As one participant stated, in the Phone+ case they could explore with patients "strategies to minimize the use of the phone overnight and improve sleep" (U08).

## 7.2 Uncovering Challenges

*7.2.1 The Need for Context.* Many participants believed that the data would not be useful without additional contextual information to clarify associations between passive sensing and symptoms. Otherwise, their interpretation of the data "would all just kind of be shots in the dark" (U06). As U10 stated, "they couldn't just use that data on its own" and "they would have to follow up" with the patient. U22 mentioned how in the Home+ case, that they did not "want to just assume that if you're leaving the home you're healthy and if you're home, you're not healthy", stating that patients "could be avoiding something at home". These quotes suggest that, in general, participants were more confident using the passive data to inspire follow up questions for patients, and not use the passive data, alone, to inform treatment actions. As U14 stated "numbers can mean anything", and only felt confident using the data "if they [the patient] were sharing it and I would have the opportunity for context." These findings explain why participants rarely ranked the passive sensing as the most important data type (Figure 7, bottom). As U15 stated that they "don't think the passive data could be a sole intervention or treatment", and U05 believed that "it was the combined data that enhanced the smartphone data value". U14, agreed, stating that: "I would want to understand what their baseline for sleep is, to understand if it's actually a sleep disturbance. So if they are an executive and they're used to going on five hours of sleep, and now, they're getting eight hours of sleep, for them, that might feel like they're sleeping too much, but that might actually be healthy."

The need for context was prominent when participants perceived passive sensing-symptom relationships as unexpected or ambiguous. For example, only 6 (30%) participants reported an increased understanding of the patient (score change $\geq 1$, Figure 7, top-left) when nighttime phone use increased, but sleep disturbances decreased

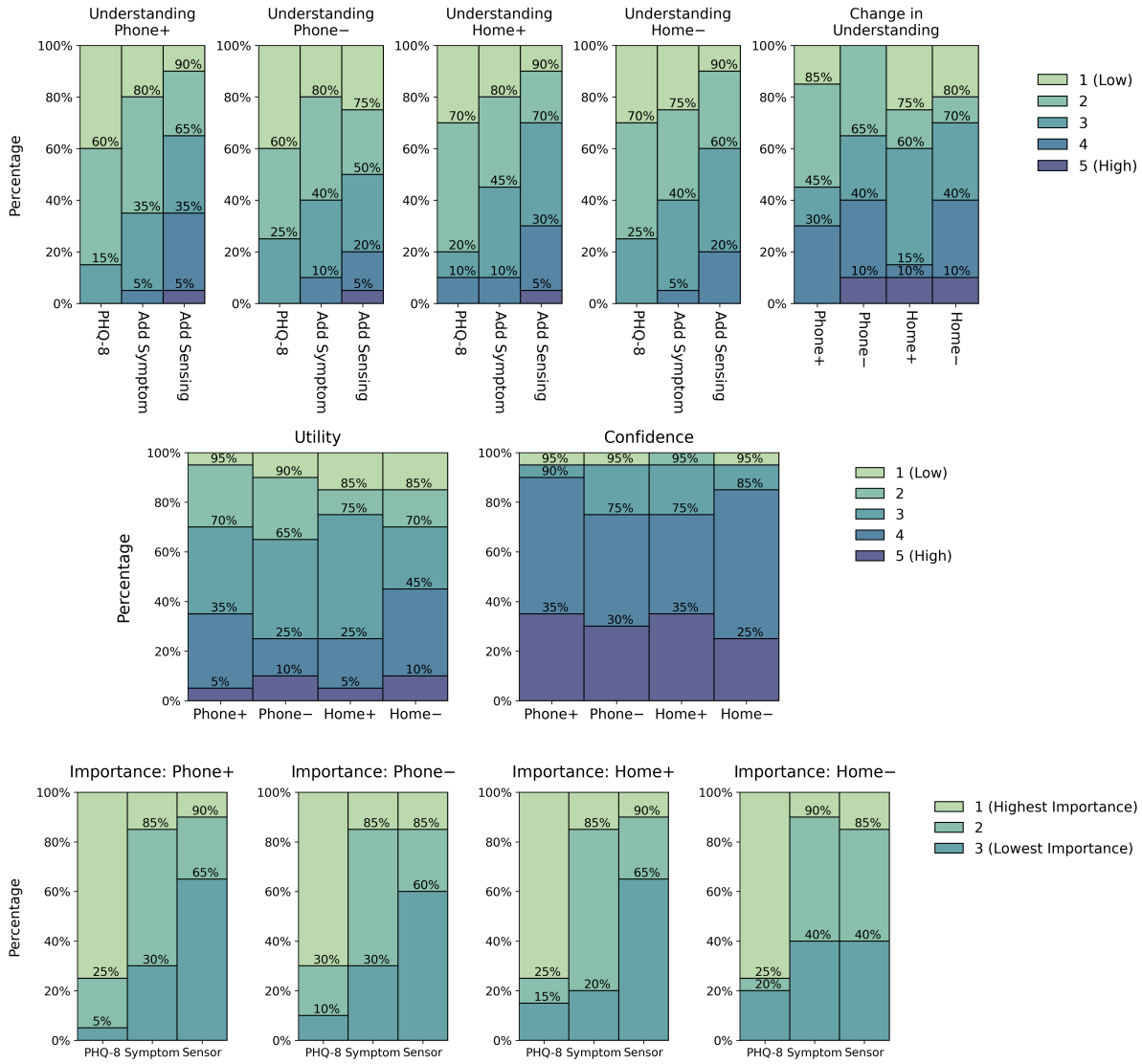


Fig. 7. **Study 2 quantitative response distributions. Top, Left.** *On a scale of 1 (low) to 5 (high), based upon the data on the screen, how much do you feel you understand this patient?* Each histogram is specific to the patient cases described in Figure 5. Participants were always first shown patients' PHQ-8 total scores, then symptom, and finally smartphone sensing (x-axes). **Top, Right.** *Did the smartphone data change your understanding of this patient? 1 (little-to-no change) to 5 (great change).* **Middle, left.** *How useful was the smartphone data to help you understand this patient? 1 (not useful) to 5 (very useful).* **Middle, right.** *if a patient shared this smartphone data with you, how confident would you feel using this data as a part of treatment? 1 (not confident) to 5 (very confident).* Each bar is specific to the patient cases on the x-axis described in Figure 5. **Bottom.** *Rank order the importance of each type of data for understanding and treating this patient.* Each histogram is specific to the patient cases described in Figure 5. The x-axis describes the data types shown to participants. On all histograms, numbers above the bars indicate the cumulative distribution (eg, 90% of values $\geq 3$)

(Phone− case). In fact, 2 (5%) participants lowered their understanding scores for this case. U05 mentioned how they were "surprised that sleep has gotten better, but there's more time on the phone". Similarly, participants were split on whether the time at home data, in both cases, increased their understanding (Home+ n=10, 50%; Home− n=9, 45%). U22 found the implications of the time at home data ambiguous, and wanted "to look more at the research" to understand "are people who spend more time at home more depressed or less depressed?"

*7.2.2 Reckoning with Personal Beliefs.* Some participants reported that data types changed their understanding only when the associations confirmed their personal beliefs on how behavior affects mental health. For example, 6 (30%) participants in the − case, and 8 (40%) participants in the + case ranked that nighttime phone use led to a greater change (higher score, Figure 7, top-right) in understanding than the time at home data. U18 explained, the phone usage data"tells me more information" and "I wouldn't expect someone to use a phone that much during the night" while they doubted that less time at home should improve mental health, stating that "I'm much happier at home". Comparatively, only 5 (25%) participants in both +/− cases reported a higher change in understanding for the time at home compared to the phone use cases. This included U16, who stated: "for me, I find that if I don't leave the house even during a single day, it's a very different kind of day than when I go out and walk around the block".

## 8 Discussion

Taken together, our findings present actionable sensing as an alternative research direction to detection while surfacing both opportunities and challenges for using passive sensing to support actions in clinical care. Specifically, our findings suggest that actionable sensing research can bring individual-level passive measures of behavior and physiology into care that highlight patients' presenting problems, treatment response, and motivate inquiry on how behaviors and treatment impact mental health. That said, our findings also suggest that clinicians share different beliefs on how behavior impacts mental health, and require additional contextual information to interpret and integrate passive data in care. In this discussion, we consider these findings with the literature to clarify the future development of actionable sensing technologies for clinical mental healthcare.

### 8.1 Mining Contextualized, Longitudinal Multimodal Passive and Clinical Data

Our findings support designing technologies that help clinicians and their patients better navigate contextualized, longitudinal multimodal passive and mental health data. These technologies can assist with identifying presenting problems, treatment response, and allow clinicians and their patients to reflect on data to increase insight. These ideas intersect with personal informatics research [39, 51], recent work studying clinician-patient preferences for visualizing passive and self-reported data [36, 143], and "context-aware" computing applications [48].

One open research question is how to best identify relevant passive and clinical data highlighting presenting problems or treatment response, and integrate insights from this data into clinical workflows. Otherwise, clinicians may experience "information overload", which could perpetuate burnout [135], and choose to not use passive data in treatment [123]. Many papers have used simple statistical approaches to identify passive sensing features associated with symptom changes [3, 105] that may be relevant for treatment, but our findings suggest that it would be difficult to validate if these associations are meaningful without additional contextual information. For example, recent work has explored how to create "context-aware" interventions that combine passive sensing and treatment data with patients' specific goals and preferences [121]. Contextualized goals and preferences could be summarized with clinical and passive data by interactive, intelligent agents using AI [50, 120], and integrated into visualizations to guide clinicians and patients towards data relevant for care. Outside of building technology, passive data use in clinical care requires workforce and compensation changes, for example, hiring "digital navigators" who assist with data use [174], and developing reimbursement mechanisms for time spent reviewing passive data [154].

Another challenge is how to operationalize data sharing preferences. Patients can have varied preferences about sharing personal data with clinicians [115], and clinicians may be wary to receive and view passive data if they perceive data usage as a violation of patients' privacy [123]. From a technical perspective, this calls for developing flexible methods of "tiered passive data access", analogous to tiered access in research data sharing [13, 83], that allow patients and clinicians to seamlessly move from less (eg, percentage of time at home) to more (eg, fine-grained location) sensitive passive data as sharing preferences evolve in treatment. From a sociotechnical perspective, theories such as contextual integrity [125, 126] motivate creating norms and policies surrounding when and how passive data could be shared within clinical contexts.

## 8.2 Identifying and Monitoring Treatment Targets

In addition to supporting patient-clinician reflection, our findings support actionable sensing research on identifying and monitoring treatment targets. The concept of a "treatment target" persists across different types of behavioral therapy. For example, in behavioral activation, patients undergo a therapeutic process of focused behavior change to improve mood [73, 81]; in exposure therapy, patients expose themselves to fearful stimuli in order to reduce their fear [61]; in social rhythm therapy, patients regularize their routines to reduce sleep-wake cycle dysregulation and improve functioning [63]. In these therapies, the "target" of treatment is engagement in a specific task (behavior change, exposures, or routine regulation), and task engagement over the course of treatment reduces mental health symptoms [61, 157].

One question for actionable sensing research is: can passive sensing be used as a target to monitor task engagement? Recent work has found that clinicians are interested in using passive data to monitor engagement in psychotherapy [53]. But, using passive sensing as a treatment target assumes causality, for example, that decreased phone use should reduce insomnia. Establishing causality is challenging, and studies that identify individual-level, causal associations, specifically with small data, must reduce the risk of bias in their results. For example, [117] used a causal discovery algorithm to identify individual-level wearable sensor indicators causally associated with mental health. This work established static causal indicators, while passive data is longitudinal and causal associations may change over time [8]. The dynamic nature of passive data calls for adaptable, time series causal discovery methods, for example from [14, 85] for treatment target identification, as well as causal approaches to predict the effects of modifying treatment targets on mental health [58].

A prospective method for treatment target identification would be to modulate passive sensing variables by having patients take specific actions during treatment (eg, decrease nighttime phone), and tracking the effects of these actions. In this setting, treatment delivery must be randomized to establish causality, for example, borrowing methods from micro-randomized trials [88], or approaches specifically designed for small data, like N-of-1 trials [44]. Researchers have already studied mental health applications that create "micro-interventions" using passive sensing to track engagement in behavior change [64, 107]. However, a critique of this direction is that it reduces mental health to modifiable behaviors trackable through passive data [151], and our participants noticed the flaws in this assumption. For example, nighttime phone use may not cause poor sleep: a person may not sleep because they live in a loud urban environment. We intend for future actionable sensing research to develop technologies and design interactions that account for these external factors [128], as well as other social factors that influence device ownership, passive data collection, and capabilities for behavior change [167].

## 8.3 Monitoring Treatment Response to Improve Care Quality

Section 8.1 focused on methods to identify treatment response in the context of individual patient care. From a health systems and public health perspective, there are other motivations to identify indicators of treatment response. Global healthcare costs have increased, and governments are testing programs that hold health systems accountable for delivering treatments that improve patient outcomes and reduce cost [103, 176]. In mental health,

there is no consensus for what measures should assess treatment outcomes and the quality of care delivery [55]. Existing mental health quality measures focus on process, not outcomes, for example, measuring how often clinicians screen for symptoms [86]. These quality measures may not reflect the complexities of mental illness in everyday life [12, 87], are often distal from patients' and clinicians' care needs [155], and are difficult to translate into concrete actions that improve patient care [35]. In addition, our findings and the literature show that mental health providers are inclined to rely on assessments and treatments informed by their clinical judgement over evidence and research [82, 96].

This creates opportunities in actionable sensing to study how passive data change throughout the course of treatment, and how practitioners can use this information to improve care quality. Studies have linked passive measures of activity with chronic disease and mortality [94, 102, 142], and have also identified associations between passive data and symptom reduction [1]. Yet, our findings motivate research into how passive sensing can measure treatment outcomes beyond symptom reduction. For example, treatment engagement is a proximal outcome of psychotherapy, and studies show that passive sensing can measure engagement in treatment [10, 53, 144]. Researchers have also used AI to identify effective patient-clinician interactions in psychotherapy [112], and insights can be fed back to clinicians to improve the quality of patient encounters [75]. We envision a world where passive sensing data promotes evidence-informed interventions, and insights from passive data are shared with clinicians and health officials to improve mental health services.

### 8.4 Limitations and Future Work

Our findings reflect our interpretation of the literature joined with the perspectives of 41 mental health clinicians, and should not be interpreted to reflect mental health clinicians as a whole. All participants in this study were based in the United States, and these findings are biased towards mental healthcare in the U.S. In addition, our design probe methodology allowed for in-depth engagement with shown data, but these probes did not explore all possible passive sensing-mental health relationships. We chose to conduct a narrative review instead of a scoping or systematic reviews, which give a more complete summary and/or synthesis of the literature. A more rigorous review may reach different conclusions. We interviewed mental health experts, specifically practicing clinicians, to understand how passive sensing can align with existing data practices and clinical needs. We plan to include patients' perspectives, exploring use cases in both self-management and clinical care as future work. Finally, we hope that the research directions outlined in this discussion motivate exciting future work for ubiquitous computing researchers at the intersection of actionable sensing and clinical mental healthcare.

### 8.5 Conclusion

In this work, we call for more focused research on actionable sensing as a vision to bridge promising technical research in passive sensing with clinical actions and needs in mental healthcare. We present actionable sensing as an alternative to but not a replacement for detection research, and we hope that research in mental health detection using passive sensing data improves screening and assessment. Simultaneously, actionable sensing research can develop and evaluate technologies that bring passive sensing data on behavior and physiology into clinical encounters and improve care. We are excited to engage and collaborate with the community on designing innovative technologies that support the needs of patients, clinicians, and health systems, improve treatment outcomes, and mental health service delivery.

### 8.6 Positionality

The first, second, and third authors are graduate students in computer and information science. These authors recruited participants, collected, and analyzed all of the data. Two authors are both clinical researchers and practicing mental health clinicians who worked with the first author on the study protocols, including creating

the design probe. All other authors were either clinical researchers, or researchers in computing and information science who contributed to the final manuscript. All authors were based in the United States.

## Acknowledgments

D.A. is supported by a National Science Foundation Graduate Research Fellowship under Grant No. DGE-2139899, a Digital Life Initiative Doctoral Fellowship, and a Siegel PiTech PhD Impact Fellowship. The data used in the design probes was collected through a National Institute of Mental Health Grant No. R01MH111610 awarded to D.C.M. Computing costs were funded by a Microsoft Azure Cloud Computing Grant through the Cornell Center for Data Science for Enterprise and Society, awarded to T.C. Transcription and participant reimbursement costs were supported by a multi-investigator seed grant through the Cornell Academic Integration Program, awarded to T.C. and F.W. Publication costs were funded by a National Science Foundation Grant No. 2212351 awarded to T.C. N.C.J was partially funded by the National Institute of Mental Health and the National Institute of General Medical Sciences under grant R01MH123482-01. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not reflect the views of the funders. Thank you to Marianne Aubin Le Quéré for providing feedback on this paper.

## References

[1] Anzar Abbas, Colin Sauder, Vijay Yadav, Vidya Koesmahargyo, Allison Aghjayan, Serena Marecki, Miriam Evans, and Isaac R. Galatzer-Levy. 2021. Remote Digital Measurement of Facial and Vocal Markers of Major Depressive Disorder Severity and Treatment Response: A Pilot Study. *Frontiers in Digital Health* 3 (March 2021), 610006. https://doi.org/10.3389/fdgth.2021.610006

[2] Saeed Abdullah, Mark Matthews, Elizabeth L. Murnane, Geri Gay, and Tanzeem Choudhury. 2014. Towards circadian computing: "early to bed and early to rise" makes some of us unhealthy and sleep deprived. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Seattle Washington, 673–684. https://doi.org/10.1145/2632048.2632100

[3] Daniel A Adler, Dror Ben-Zeev, Vincent W-S Tseng, John M Kane, Rachel Brian, Andrew T Campbell, Marta Hauser, Emily A Scherer, and Tanzeem Choudhury. 2020. Predicting Early Warning Signs of Psychotic Relapse From Passive Sensing Data: An Approach Using Encoder-Decoder Neural Networks. *JMIR mHealth and uHealth* 8, 8 (Aug. 2020), e19962. https://doi.org/10.2196/19962

[4] Daniel A. Adler, Caitlin A. Stamatis, Jonah Meyerhoff, David C. Mohr, Fei Wang, Gabriel J. Aranovich, Srijan Sen, and Tanzeem Choudhury. 2024. Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. *npj Mental Health Research* 3, 1 (April 2024), 1–11. https://doi.org/10.1038/s44184-024-00057-y Publisher: Nature Publishing Group.

[5] Daniel A. Adler, Emily Tseng, Khatiya C. Moon, John Q. Young, John M. Kane, Emanuel Moss, David C. Mohr, and Tanzeem Choudhury. 2022. Burnout and the Quantified Workplace: Tensions around Personal Sensing Interventions for Stress in Resident Physicians. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 430:1–430:48. https://doi.org/10.1145/3555531

[6] Daniel A. Adler, Vincent W.-S. Tseng, Gengmo Qi, Joseph Scarpa, Srijan Sen, and Tanzeem Choudhury. 2021. Identifying Mobile Sensing Indicators of Stress-Resilience. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 2 (June 2021), 51:1–51:32. https://doi.org/10.1145/3463528

[7] Daniel A. Adler, Fei Wang, David C. Mohr, and Tanzeem Choudhury. 2022. Machine learning for passive mental health symptom prediction: Generalization across different longitudinal mobile sensing studies. *PLOS ONE* 17, 4 (April 2022), e0266516. https://doi.org/10.1371/journal.pone.0266516 Publisher: Public Library of Science.

[8] Daniel A. Adler, Fei Wang, David C. Mohr, Deborah Estrin, Cecilia Livesey, and Tanzeem Choudhury. 2022. A call for open data to develop mental health digital biomarkers. *BJPsych Open* 8, 2 (March 2022). https://doi.org/10.1192/bjo.2022.28 Publisher: Cambridge University Press.

[9] Substance Abuse and Mental Health Services Administration. 2016. DSM-5 Child Mental Disorder Classification. In *DSM-5 Changes: Implications for Child Serious Emotional Disturbance [Internet]*. Substance Abuse and Mental Health Services Administration (US). https://www.ncbi.nlm.nih.gov/books/NBK519712/

[10] Elena Agapie, Patricia A. Areán, Gary Hsieh, and Sean A. Munson. 2022. A Longitudinal Goal Setting Model for Addressing Complex Personal Problems in Mental Health. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–28. https://doi.org/10.1145/3555160

[11] Adrian Aguilera, Stephen Schueller, and Yan Leykin. 2015. Daily Mood Ratings via Text Message as a Proxy for Clinic Based Depression Assessment. *Journal of affective disorders* 175 (April 2015), 471–474. https://doi.org/10.1016/j.jad.2015.01.033

[12] Margarita Alegría, Richard G. Frank, Helena B. Hansen, Joshua M. Sharfstein, Ruth S. Shim, and Matt Tierney. 2021. Transforming Mental Health And Addiction Services. *Health Affairs* 40, 2 (Feb. 2021), 226–234. https://doi.org/10.1377/hlthaff.2020.01472 Publisher: Health Affairs.

[13] All of Us. 2024. Data Access Tiers – All of Us Research Hub. https://www.researchallofus.org/data-tools/data-access/

[14] Charles K. Assaad, Emilie Devijver, and Eric Gaussier. 2022. Survey and Evaluation of Causal Discovery Methods for Time Series. *Journal of Artificial Intelligence Research* 73 (Feb. 2022), 767–819. https://doi.org/10.1613/jair.1.13428

[15] Jakob E. Bardram. 2022. From Sensing to Acting—Can Pervasive Computing Change the World? *IEEE Pervasive Computing* (2022), 1–7. https://doi.org/10.1109/MPRV.2022.3182489 Conference Name: IEEE Pervasive Computing.

[16] Jakob E. Bardram, Mads Frost, Károly Szántó, Maria Faurholt-Jepsen, Maj Vinberg, and Lars Vedel Kessing. 2013. Designing mobile health technology for bipolar disorder: a field trial of the monarca system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, Paris, France, 2627–2636. https://doi.org/10.1145/2470654.2481364

[17] Jakob E. Bardram and Aleksandar Matic. 2020. A Decade of Ubiquitous Computing Research in Mental Health. *IEEE Pervasive Computing* 19, 1 (Jan. 2020), 62–72. https://doi.org/10.1109/MPRV.2019.2925338 Conference Name: IEEE Pervasive Computing.

[18] Ian Barnett, John Torous, Patrick Staples, Luis Sandoval, Matcheri Keshavan, and Jukka-Pekka Onnela. 2018. Relapse prediction in schizophrenia through digital phenotyping: a pilot study. *Neuropsychopharmacology* 43, 8 (July 2018), 1660–1666. https://doi.org/10.1038/s41386-018-0030-z

[19] K. L. Barriball and A. While. 1994. Collecting data using a semi-structured interview: a discussion paper. *Journal of Advanced Nursing* 19, 2 (Feb. 1994), 328–335. https://doi.org/10.1111/j.1365-2648.1994.tb01088.x

[20] Morris Bell, Joanna Fiszdon, Randall Richardson, Paul Lysaker, and Gary Bryson. 2007. Are self-reports valid for schizophrenia patients with poor insight? Relationship of unawareness of illness to psychological self-report instruments. *Psychiatry Research* 151, 1 (May 2007), 37–46. https://doi.org/10.1016/j.psychres.2006.04.012

[21] Ethan M. Berke, Tanzeem Choudhury, Shahid Ali, and Mashfiqui Rabbi. 2011. Objective Measurement of Sociability and Activity: Mobile Sensing in the Community. *The Annals of Family Medicine* 9, 4 (July 2011), 344–350. https://doi.org/10.1370/afm.1266 Publisher: The Annals of Family Medicine Section: Methodology.

[22] M. L. Birnbaum, S. K. Ernala, A. F. Rizvi, E. Arenare, A. R. Van Meter, M. De Choudhury, and J. M. Kane. 2019. Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. *npj Schizophrenia* 5, 1 (Oct. 2019), 1–9. https://doi.org/10.1038/s41537-019-0085-9

[23] Michael L. Birnbaum, Hongyi Wen, Anna Van Meter, Sindhu K. Ernala, Asra F. Rizvi, Elizabeth Arenare, Deborah Estrin, Munmun De Choudhury, and John M. Kane. 2020. Identifying emerging mental illness utilizing search engine activity: A feasibility study. *PLOS ONE* 15, 10 (Oct. 2020), e0240820. https://doi.org/10.1371/journal.pone.0240820 Publisher: Public Library of Science.

[24] Erik Blasch, Robert Cruise, Alexander Aved, Uttam Majumder, and Todd Rovito. 2019. Methods of AI for Multimodal Sensing and Action for Complex Situations. *AI Magazine* 40, 4 (Dec. 2019), 50–65. https://doi.org/10.1609/aimag.v40i4.4813 Number: 4.

[25] Erik Blasch, Zheng Liu, and Yufeng Zheng. 2020. Image fusion for context-aided automatic target recognition. In *Automatic Target Recognition XXX*, Vol. 11394. SPIE, 163–174. https://doi.org/10.1117/12.2564876

[26] Erik P. Blasch, Uttam Majumder, Todd Rovito, and Ali K. Raz. 2019. Artificial Intelligence in Use by Multimodal Fusion. In *2019 22th International Conference on Information Fusion (FUSION)*. 1–8. https://doi.org/10.23919/FUSION43075.2019.9011267

[27] Alexander J. Boe, Lori L. McGee Koch, Megan K. O'Brien, Nicholas Shawen, John A. Rogers, Richard L. Lieber, Kathryn J. Reid, Phyllis C. Zee, and Arun Jayaraman. 2019. Automating sleep stage classification using wireless, wearable sensors. *npj Digital Medicine* 2, 1 (Dec. 2019), 131. https://doi.org/10.1038/s41746-019-0210-1

[28] Lynn Boschloo, Claudia D. van Borkulo, Mijke Rhemtulla, Katherine M. Keyes, Denny Borsboom, and Robert A. Schoevers. 2015. The Network Structure of Symptoms of the Diagnostic and Statistical Manual of Mental Disorders. *PLOS ONE* 10, 9 (Sept. 2015), e0137621. https://doi.org/10.1371/journal.pone.0137621 Publisher: Public Library of Science.

[29] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa Publisher: Routledge _eprint: https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa.

[30] Virginia Braun and Victoria Clarke. 2021. One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology* 18, 3 (July 2021), 328–352. https://doi.org/10.1080/14780887.2020.1769238 Publisher: Routledge _eprint: https://doi.org/10.1080/14780887.2020.1769238.

[31] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J Karr, Emily Giangrande, and David C Mohr. 2011. Harnessing Context Sensing to Develop a Mobile Intervention for Depression. *Journal of Medical Internet Research* 13, 3 (Aug. 2011), e55. https://doi.org/10.2196/jmir.1838

[32] Shane S. Bush and Noah K. Kaufman. 2023. Diagnostic Testing: Rating Scales and Psychological and Neuropsychological Tests. In *Atlas of Psychiatry*, Waguih William IsHak (Ed.). Springer International Publishing, Cham, 201–226. https://doi.org/10.1007/978-3-031-15401-0_7

[33] Jonas Busk, Maria Faurholt-Jepsen, Mads Frost, Jakob E. Bardram, Lars Vedel Kessing, and Ole Winther. 2020. Forecasting Mood in Bipolar Disorder From Smartphone Self-assessments: Hierarchical Bayesian Approach. *JMIR mHealth and uHealth* 8, 4 (2020), e15028. https://doi.org/10.2196/15028 Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[34] Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan Flint, Emma S. J. Robinson, and Marcus R. Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14, 5 (May 2013), 365–376. https://doi.org/10.1038/nrn3475 Publisher: Nature Publishing Group.

[35] Rachel Campbell, Angela Ju, Madeleine T. King, and Claudia Rutherford. 2022. Perceived benefits and limitations of using patient-reported outcome measures in clinical practice with individual patients: a systematic review of qualitative studies. *Quality of Life Research* 31, 6 (June 2022), 1597–1620. https://doi.org/10.1007/s11136-021-03003-z

[36] Sarah Chang, Lucy Gray, Noy Alon, and John Torous. 2023. Patient and Clinician Experiences with Sharing Data Visualizations Integrated into Mental Health Treatment. *Social Sciences* 12, 12 (Dec. 2023), 648. https://doi.org/10.3390/socsci12120648 Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[37] Alexander L. Chapman. 2006. Dialectical behavior therapy: current indications and unique elements. *Psychiatry (Edgmont (Pa.: Township))* 3, 9 (Sept. 2006), 62–68.

[38] Patrick Y. K. Chau. 1996. An Empirical Assessment of a Modified Technology Acceptance Model. *Journal of Management Information Systems* 13, 2 (1996), 185–204. https://www.jstor.org/stable/40398221 Publisher: Taylor & Francis, Ltd..

[39] Chia-Fang Chung, Kristin Dew, Allison Cole, Jasmine Zia, James Fogarty, Julie A. Kientz, and Sean A. Munson. 2016. Boundary Negotiating Artifacts in Personal Informatics: Patient-Provider Collaboration with Patient-Generated Data. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing (CSCW '16)*. Association for Computing Machinery, New York, NY, USA, 770–786. https://doi.org/10.1145/2818048.2819926

[40] Diana E. Clarke, William E. Narrow, Darrel A. Regier, S. Janet Kuramoto, David J. Kupfer, Emily A. Kuhl, Lisa Greiner, and Helena C. Kraemer. 2013. DSM-5 Field Trials in the United States and Canada, Part I: Study Design, Sampling Strategy, Implementation, and Analytic Approaches. *American Journal of Psychiatry* 170, 1 (Jan. 2013), 43–58. https://doi.org/10.1176/appi.ajp.2012.12070998 Publisher: American Psychiatric Publishing.

[41] Andrea Coravos, Sean Khozin, and Kenneth D. Mandl. 2019. Developing and adopting safe and effective digital biomarkers to improve patient outcomes. *npj Digital Medicine* 2, 1 (March 2019), 1–5. https://doi.org/10.1038/s41746-019-0090-4 Number: 1 Publisher: Nature Publishing Group.

[42] Angélique O. J. Cramer, Lourens J. Waldorp, Han L. J. van der Maas, and Denny Borsboom. 2010. Comorbidity: A network perspective. *Behavioral and Brain Sciences* 33, 2-3 (June 2010), 137–150. https://doi.org/10.1017/S0140525X09991567

[43] Pim Cuijpers, Juan Li, Stefan G. Hofmann, and Gerhard Andersson. 2010. Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: A meta-analysis. *Clinical Psychology Review* 30, 6 (Aug. 2010), 768–778. https://doi.org/10.1016/j.cpr.2010.06.001

[44] Karina W. Davidson, Michael Silverstein, Ken Cheung, Rocco A. Paluch, and Leonard H. Epstein. 2021. Personalized (N-of-1) Trials: A Primer. *JAMA pediatrics* 175, 4 (April 2021), 404–409. https://doi.org/10.1001/jamapediatrics.2020.5801

[45] Fred D. Davis. 1989. Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly* 13, 3 (1989), 319–340. https://doi.org/10.2307/249008 Publisher: Management Information Systems Research Center, University of Minnesota.

[46] Valeria de Angel, Serena Lewis, Katie M White, Faith Matcham, and Matthew Hotopf. 2022. Clinical Targets and Attitudes Toward Implementing Digital Health Tools for Remote Measurement in Treatment for Depression: Focus Groups With Patients and Clinicians. *JMIR Mental Health* 9, 8 (Aug. 2022), e38934. https://doi.org/10.2196/38934

[47] Janine DeSimone and Bryan R. Hansen. 2023. The Impact of Measurement-Based Care in Psychiatry: An Integrative Review. *Journal of the American Psychiatric Nurses Association* (June 2023), 10783903231177707. https://doi.org/10.1177/10783903231177707 Publisher: SAGE Publications Inc STM.

[48] Anind K. Dey, Gregory D. Abowd, and Daniel Salber. 2001. A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications. *Human–Computer Interaction* 16, 2-4 (Dec. 2001), 97–166. https://doi.org/10.1207/S15327051HCI16234_02 Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/S15327051HCI16234_02.

[49] Jessilyn Dunn, Lukasz Kidzinski, Ryan Runge, Daniel Witt, Jennifer L. Hicks, Sophia Miryam Schüssler-Fiorenza Rose, Xiao Li, Amir Bahmani, Scott L. Delp, Trevor Hastie, and Michael P. Snyder. 2021. Wearable sensors enable personalized predictions of clinical laboratory measurements. *Nature Medicine* 27, 6 (June 2021), 1105–1112. https://doi.org/10.1038/s41591-021-01339-0 Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Laboratory techniques and procedures;Translational research Subject_term_id: laboratory-techniques-and-procedures;translational-research.

[50] Zachary Englhardt, Chengqian Ma, Margaret E. Morris, Xuhai "Orson" Xu, Chun-Cheng Chang, Lianhui Qin, Daniel McDuff, Xin Liu, Shwetak Patel, and Vikram Iyer. 2023. From Classification to Clinical Insights: Towards Analyzing and Reasoning About Mobile and Behavioral Health Data With Large Language Models. https://doi.org/10.48550/arXiv.2311.13063 arXiv:2311.13063 [cs].

[51] Daniel A. Epstein, Clara Caldeira, Mayara Costa Figueiredo, Xi Lu, Lucas M. Silva, Lucretia Williams, Jong Ho Lee, Qingyang Li, Simran Ahuja, Qiuer Chen, Payam Dowlatyari, Craig Hilby, Sazeda Sultana, Elizabeth V. Eikey, and Yunan Chen. 2020. Mapping and Taking Stock of the Personal Informatics Literature. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 1–38. https://doi.org/10.1145/3432231

[52] Ilker Etikan, Sulaiman Abubakar Musa, and Rukayya Sunusi Alkassim. 2015. Comparison of Convenience Sampling and Purposive Sampling. *American Journal of Theoretical and Applied Statistics* 5, 1 (Dec. 2015), 1. https://doi.org/10.11648/j.ajtas.20160501.11 Number: 1 Publisher: Science Publishing Group.

[53] Hayley I. Evans, Myeonghan Ryu, Theresa Hsieh, Jiawei Zhou, Kefan Xu, Kenneth W. Akers, Andrew M. Sherrill, and Rosa I. Arriaga. 2024. Using Sensor-Captured Patient-Generated Data to Support Clinical Decision-making in PTSD Therapy. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 1–28. https://doi.org/10.1145/3637426

[54] Liese Exelmans and Jan Van den Bulck. 2016. Bedtime mobile phone use and sleep in adults. *Social Science & Medicine* 148 (Jan. 2016), 93–101. https://doi.org/10.1016/j.socscimed.2015.11.037

[55] Gregory K. Farber, Suzanne Gage, and Danielle Kemmer. 2023. A Collaborative Effort to Establish Common Metrics for Use in Mental Health. *JAMA Psychiatry* 80, 10 (Oct. 2023), 981–982. https://doi.org/10.1001/jamapsychiatry.2023.2282

[56] Maria Faurholt-Jepsen, Jonas Busk, Darius Adam Rohani, Mads Frost, Morten Lindberg Tønning, Jakob Eyvind Bardram, and Lars Vedel Kessing. 2022. Differences in mobility patterns according to machine learning models in patients with bipolar disorder and patients with unipolar disorder. *Journal of Affective Disorders* 306 (June 2022), 246–253. https://doi.org/10.1016/j.jad.2022.03.054

[57] Maria Faurholt-Jepsen, Darius Adam Rohani, Jonas Busk, Morten Lindberg Tønning, Mads Frost, Jakob Eyvind Bardram, and Lars Vedel Kessing. 2024. Using digital phenotyping to classify bipolar disorder and unipolar disorder – exploratory findings using machine learning models. *European Neuropsychopharmacology* 81 (April 2024), 12–19. https://doi.org/10.1016/j.euroneuro.2024.01.003

[58] Stefan Feuerriegel, Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Konstantin Hess, Alicia Curth, Stefan Bauer, Niki Kilbertus, Isaac S. Kohane, and Mihaela van der Schaar. 2024. Causal machine learning for predicting treatment outcomes. *Nature Medicine* 30, 4 (April 2024), 958–968. https://doi.org/10.1038/s41591-024-02902-1 Publisher: Nature Publishing Group.

[59] Michael B. First. 2015. Structured Clinical Interview for the DSM (SCID). In *The Encyclopedia of Clinical Psychology*. American Cancer Society, 1–6. https://doi.org/10.1002/9781118625392.wbecp351

[60] Fitbit. 2023. What should I know about Fitbit sleep stages? https://help.fitbit.com/articles/en_US/Help_article/2163.htm

[61] Edna B. Foa and Carmen P. McLean. 2016. The Efficacy of Exposure Therapy for Anxiety-Related Disorders and Its Underlying Mechanisms: The Case of OCD and PTSD. *Annual Review of Clinical Psychology* 12, Volume 12, 2016 (March 2016), 1–28. https://doi.org/10.1146/annurev-clinpsy-021815-093533 Publisher: Annual Reviews.

[62] John C. Fortney, Jürgen Unützer, Glenda Wrenn, Jeffrey M. Pyne, G. Richard Smith, Michael Schoenbaum, and Henry T. Harbin. 2017. A Tipping Point for Measurement-Based Care. *Psychiatric Services* 68, 2 (Feb. 2017), 179–188. https://doi.org/10.1176/appi.ps.201500439 Publisher: American Psychiatric Publishing.

[63] Ellen Frank, Holly A. Swartz, and David J. Kupfer. 2001. Interpersonal and Social Rhythm Therapy: Managing the Chaos of Bipolar Disorder. In *Bipolar Disorder*. Routledge. Num Pages: 12.

[64] Ellen Frank, Meredith L. Wallace, Mark L. Matthews, Jeremy Kendrick, Jeremy Leach, Tara Moore, Gabriel Aranovich, Tanzeem Choudhury, Nirav R. Shah, Zeenia Framroze, Greg Posey, Samuel Burgess, and David J. Kupfer. 2022. Personalized digital intervention for depression based on social rhythm principles adds significantly to outpatient treatment. *Frontiers in Digital Health* 4 (2022). https://www.frontiersin.org/articles/10.3389/fdgth.2022.870522

[65] Eiko I. Fried, Jessica K. Flake, and Donald J. Robinaugh. 2022. Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology* (April 2022), 1–11. https://doi.org/10.1038/s44159-022-00050-2 Publisher: Nature Publishing Group.

[66] Eiko I. Fried and Randolph M. Nesse. 2015. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders* 172 (Feb. 2015), 96–102. https://doi.org/10.1016/j.jad.2014.10.010

[67] Mads Frost, Afsaneh Doryab, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2013. Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (UbiComp '13)*. Association for Computing Machinery, Zurich, Switzerland, 133–142. https://doi.org/10.1145/2493432.2493507

[68] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* 2, 9 (Sept. 2020), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2 Publisher: Elsevier.

[69] William Gaver. 2014. Science and Design: The Implications of Different Forms of Accountability. In *Ways of Knowing in HCI*. Springer New York, New York, NY, 143–165. https://doi.org/10.1007/978-1-4939-0378-8_7

[70] Leo A. Goodman. 1961. Snowball Sampling. *The Annals of Mathematical Statistics* 32, 1 (1961), 148–170. https://www.jstor.org/stable/2237615 Publisher: Institute of Mathematical Statistics.

[71] Robert L. Hatcher and J. Arthur Gillaspy. 2006. Development and validation of a revised short version of the working alliance inventory. *Psychotherapy Research* 16, 1 (Jan. 2006), 12–25. https://doi.org/10.1080/10503300500352500 Publisher: Routledge _eprint: https://doi.org/10.1080/10503300500352500.

[72] Derek Hatfield, Lynn McCullough, Shelby H. B. Frantz, and Kenin Krieger. 2010. Do we know when our clients get worse? an investigation of therapists' ability to detect negative client change. *Clinical Psychology & Psychotherapy* 17, 1 (2010), 25–32. https://doi.org/10.1002/cpp.656 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpp.656.

[73] Derek R. Hopko, C. W. Lejuez, and Sandra D. Hopko. 2004. Behavioral Activation as an Intervention for Coexistent Depressive and Anxiety Symptoms. *Clinical Case Studies* 3, 1 (Jan. 2004), 37–48. https://doi.org/10.1177/1534650103258969 Publisher: SAGE Publications.

[74] Adam G. Horwitz, Zhuo Zhao, and Srijan Sen. 2023. Peak-end bias in retrospective recall of depressive symptoms on the PHQ-9. *Psychological Assessment* 35, 4 (April 2023), 378–381. https://doi.org/10.1037/pas0001219

[75] Zac E. Imel, Brian T. Pace, Christina S. Soma, Michael Tanana, Tad Hirsch, James Gibson, Panayiotis Georgiou, Shrikanth Narayanan, and David C. Atkins. 2019. Design feasibility of an automated, machine-learning based feedback system for motivational interviewing. *Psychotherapy (Chicago, Ill.)* 56, 2 (June 2019), 318–328. https://doi.org/10.1037/pst0000221

[76] Thomas R. Insel. 2014. The NIMH Research Domain Criteria (RDoC) Project: precision medicine for psychiatry. *The American Journal of Psychiatry* 171, 4 (April 2014), 395–397. https://doi.org/10.1176/appi.ajp.2014.14020138

[77] Thomas R. Insel. 2017. Digital Phenotyping: Technology for a New Science of Behavior. *JAMA* 318, 13 (Oct. 2017), 1215–1216. https://doi.org/10.1001/jama.2017.11295 Publisher: American Medical Association.

[78] Nicholas C. Jacobson and Sukanya Bhattacharya. 2022. Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. *Behaviour Research and Therapy* 149 (Feb. 2022), 104013. https://doi.org/10.1016/j.brat.2021.104013

[79] Nicholas C. Jacobson, Berta Summers, and Sabine Wilhelm. 2020. Digital Biomarkers of Social Anxiety Severity: Digital Phenotyping Using Passive Smartphone Sensors. *Journal of Medical Internet Research* 22, 5 (May 2020), e16875. https://doi.org/10.2196/16875 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[80] Nicholas C. Jacobson, Hilary Weingarden, and Sabine Wilhelm. 2019. Digital biomarkers of mood disorders and symptom change. *npj Digital Medicine* 2, 1 (Feb. 2019), 1–3. https://doi.org/10.1038/s41746-019-0078-0 Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Diagnostic markers;Human behaviour;Predictive markers Subject_term_id: diagnostic-markers;human-behaviour;predictive-markers.

[81] Neil S. Jacobson, Christopher R. Martell, and Sona Dimidjian. 2001. Behavioral Activation Treatment for Depression: Returning to Contextual Roots. *Clinical Psychology: Science and Practice* 8, 3 (2001), 255–270. https://doi.org/10.1093/clipsy.8.3.255 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1093/clipsy.8.3.255.

[82] Amanda Jensen-Doss and Kristin M. Hawley. 2010. Understanding barriers to evidence-based assessment: Clinician attitudes toward standardized assessment tools. *Journal of clinical child and adolescent psychology* 39, 6 (Nov. 2010), 885–896. https://doi.org/10.1080/15374416.2010.517169

[83] Yann Joly, Stephanie O.M. Dyke, Bartha M. Knoppers, and Tomi Pastinen. 2016. Are Data Sharing and Privacy Protection Mutually Exclusive? *Cell* 167, 5 (Nov. 2016), 1150–1154. https://doi.org/10.1016/j.cell.2016.11.004

[84] Kazi Sinthia Kabir, Stacey A. Kenfield, Erin L. Van Blarigan, June M. Chan, and Jason Wiese. 2022. Ask the Users: A Case Study of Leveraging User-Centered Design for Designing Just-in-Time Adaptive Interventions (JITAIs). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (July 2022), 59:1–59:21. https://doi.org/10.1145/3534612

[85] Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. 2022. Causal Machine Learning: A Survey and Open Problems. (2022).

[86] Ira R. Katz, Edward P. Liebmann, Sandra G. Resnick, Rani A. Hoff, and Eric M. Schmidt. 2022. Progress toward a performance measure for mental health based on a generic patient- reported outcome measure: Findings from the Veterans Outcome Assessment survey. *Psychiatry Research* 317 (Nov. 2022), 114797. https://doi.org/10.1016/j.psychres.2022.114797

[87] Amy M. Kilbourne, Kathryn Beck, Brigitta Spaeth-Rublee, Parashar Ramanuj, Robert W. O'Brien, Naomi Tomoyasu, and Harold Alan Pincus. 2018. Measuring and improving the quality of mental health care: a global perspective. *World Psychiatry* 17, 1 (2018), 30–38. https://doi.org/10.1002/wps.20482 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/wps.20482.

[88] Predrag Klasnja, Eric B. Hekler, Saul Shiffman, Audrey Boruvka, Daniel Almirall, Ambuj Tewari, and Susan A. Murphy. 2015. Micro-Randomized Trials: An Experimental Design for Developing Just-in-Time Adaptive Interventions. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association* 34, 0 (Dec. 2015), 1220–1228. https://doi.org/10.1037/hea0000305

[89] Predrag Klasnja, Shawna Smith, Nicholas J Seewald, Andy Lee, Kelly Hall, Brook Luers, Eric B Hekler, and Susan A Murphy. 2019. Efficacy of Contextually Tailored Suggestions for Physical Activity: A Micro-randomized Optimization Trial of HeartSteps. *Annals of Behavioral Medicine* 53, 6 (May 2019), 573–582. https://doi.org/10.1093/abm/kay067

[90] Kurt Kroenke, Robert L Spitzer, and Janet B W Williams. 2001. The PHQ-9. *Journal of General Internal Medicine* 16, 9 (Sept. 2001), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[91] Kurt Kroenke, Robert L. Spitzer, Janet B. W. Williams, and Bernd Löwe. 2009. An Ultra-Brief Screening Scale for Anxiety and Depression: The PHQ–4. *Psychosomatics* 50, 6 (Nov. 2009), 613–621. https://doi.org/10.1016/S0033-3182(09)70864-3

[92] Kurt Kroenke, Tara W. Strine, Robert L. Spitzer, Janet B. W. Williams, Joyce T. Berry, and Ali H. Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders* 114, 1 (April 2009), 163–173. https://doi.org/10.1016/j.jad.2008.06.026

[93] Nicholas Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (Sept. 2010), 140–150. https://doi.org/10.1109/MCOM.2010.5560598

[94] I-Min Lee, Eric J. Shiroma, Masamitsu Kamada, David R. Bassett, Charles E. Matthews, and Julie E. Buring. 2019. Association of Step Volume and Intensity With All-Cause Mortality in Older Women. *JAMA Internal Medicine* 179, 8 (Aug. 2019), 1105–1112. https://doi.org/10.1001/jamainternmed.2019.0899

[95] Brooke Levis, Andrea Benedetti, John P. A. Ioannidis, Ying Sun, Zelalem Negeri, Chen He, Yin Wu, Ankur Krishnan, Parash Mani Bhandari, Dipika Neupane, Mahrukh Imran, Danielle B. Rice, Kira E. Riehm, Nazanin Saadat, Marleine Azar, Jill Boruff, Pim Cuijpers, Simon Gilbody, Lorie A. Kloda, Dean McMillan, Scott B. Patten, Ian Shrier, Roy C. Ziegelstein, Sultan H. Alamri, Dagmar Amtmann, Liat Ayalon, Hamid R. Baradaran, Anna Beraldi, Charles N. Bernstein, Arvin Bhana, Charles H. Bombardier, Gregory Carter, Marcos H. Chagas, Dixon Chibanda, Kerrie Clover, Yeates Conwell, Crisanto Diez-Quevedo, Jesse R. Fann, Felix H. Fischer, Leila Gholizadeh, Lorna J. Gibson, Eric P. Green, Catherine G. Greeno, Brian J. Hall, Emily E. Haroz, Khalida Ismail, Nathalie Jetté, Mohammad E. Khamseh, Yunxin Kwan, Maria Asunción Lara, Shen-Ing Liu, Sonia R. Loureiro, Bernd Löwe, Ruth Ann Marrie, Laura Marsh, Anthony McGuire, Kumiko Muramatsu, Laura Navarrete, Flávia L. Osório, Inge Petersen, Angelo Picardi, Stephanie L. Pugh, Terence J. Quinn, Alasdair G. Rooney, Eileen H. Shinn, Abbey Sidebottom, Lena Spangenberg, Pei Lin Lynnette Tan, Martin Taylor-Rowan, Alyna Turner, Henk C. van Weert, Paul A. Vöhringer, Lynne I. Wagner, Jennifer White, Kirsty Winkley, and Brett D. Thombs. 2020. Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis. *Journal of Clinical Epidemiology* 122 (June 2020), 115–128.e1. https://doi.org/10.1016/j.jclinepi.2020.02.002

[96] Scott O. Lilienfeld, Lorie A. Ritschel, Steven Jay Lynn, Robin L. Cautin, and Robert D. Latzman. 2013. Why many clinical psychologists are resistant to evidence-based practice: Root causes and constructive remedies. *Clinical Psychology Review* 33, 7 (Nov. 2013), 883–900. https://doi.org/10.1016/j.cpr.2012.09.008

[97] Tony Liu, Jonah Meyerhoff, Johannes C. Eichstaedt, Chris J. Karr, Susan M. Kaiser, Konrad P. Kording, David C. Mohr, and Lyle H. Ungar. 2022. The relationship between text message sentiment and self-reported depression. *Journal of Affective Disorders* 302 (April 2022), 7–14. https://doi.org/10.1016/j.jad.2021.12.048

[98] Kevin J. Major, Jasbinder S. Sanghera, Ishwar D. Aggarwal, Mikella E. Farrell, Ellen L. Holthoff, Paul M. Pellegrino, and Kenneth J. Ewing. 2019. Demonstration of a Human Color Vision Mimic in the Infrared. *Analytical Chemistry* 91, 21 (Nov. 2019), 14058–14065. https://doi.org/10.1021/acs.analchem.9b03749 Publisher: American Chemical Society.

[99] Brent Mallinckrodt and Yacob T. Tekie. 2016. Item response theory analysis of Working Alliance Inventory, revised response format, and new Brief Alliance Inventory. *Psychotherapy Research* 26, 6 (Nov. 2016), 694–718. https://doi.org/10.1080/10503307.2015.1061718

[100] Susan Kohl Malone. 2011. Early to Bed, Early to Rise?: An Exploration of Adolescent Sleep Hygiene Practices. *The Journal of School Nursing* 27, 5 (Oct. 2011), 348–354. https://doi.org/10.1177/1059840511410434 Publisher: SAGE Publications.

[101] Marta M. Maslej, Stefan Kloiber, Marzyeh Ghassemi, Joanna Yu, and Sean L. Hill. 2023. Out with AI, in with the psychiatrist: a preference for human-derived clinical decision support in depression care. *Translational Psychiatry* 13, 1 (June 2023), 1–9. https://doi.org/10.1038/s41398-023-02509-z Number: 1 Publisher: Nature Publishing Group.

[102] Hiral Master, Jeffrey Annis, Shi Huang, Joshua A. Beckman, Francis Ratsimbazafy, Kayla Marginean, Robert Carroll, Karthik Natarajan, Frank E. Harrell, Dan M. Roden, Paul Harris, and Evan L. Brittain. 2022. Association of step counts over time with the risk of chronic disease in the All of Us Research Program. *Nature Medicine* (Oct. 2022), 1–8. https://doi.org/10.1038/s41591-022-02012-w Publisher: Nature Publishing Group.

[103] Mark McClellan, Krishna Udayakumar, Andrea Thoumi, Jonathan Gonzalez-Smith, Kushal Kadakia, Natalia Kurek, Mariam Abdulmalik, and Ara W. Darzi. 2017. Improving Care And Lowering Costs: Evidence And Lessons From A Global Analysis Of Accountable Care Reforms. *Health Affairs* 36, 11 (Nov. 2017), 1920–1927. https://doi.org/10.1377/hlthaff.2017.0535 Publisher: Health Affairs.

[104] Shawn M. McClintock, Lex Minto, David A. Denney, K. Chase Bailey, C. Munro Cullum, and Vonetta M. Dotson. 2021. Clinical Neuropsychological Evaluation in Older Adults With Major Depressive Disorder. *Current Psychiatry Reports* 23, 9 (July 2021), 55. https://doi.org/10.1007/s11920-021-01267-3

[105] Lakmal Meegahapola, William Droz, Peter Kun, Amalia de Götzen, Chaitanya Nutakki, Shyam Diwakar, Salvador Ruiz Correa, Donglei Song, Hao Xu, Miriam Bidoglia, George Gaskell, Altangerel Chagnaa, Amarsanaa Ganbold, Tsolmon Zundui, Carlo Caprini, Daniele Miorandi, Alethia Hume, Jose Luis Zarza, Luca Cernuzzi, Ivano Bison, Marcelo Rodas Britez, Matteo Busso, Ronald Chenu-Abente, Can Günel, Fausto Giunchiglia, Laura Schelenz, and Daniel Gatica-Perez. 2023. Generalization and Personalization of Mobile Sensing-Based Mood Inference Models: An Analysis of College Students in Eight Countries. *Proceedings of the ACM on Interactive, Mobile, Wearable*

*and Ubiquitous Technologies* 6, 4 (Jan. 2023), 176:1–176:32. https://doi.org/10.1145/3569483

[106] Lakmal Meegahapola, Hamza Hassoune, and Daniel Gatica-Perez. 2024. M3BAT: Unsupervised Domain Adaptation for Multimodal Mobile Sensing with Multi-Branch Adversarial Training. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (May 2024), 1–30. https://doi.org/10.1145/3659591

[107] Jingbo Meng, Syed Ali Hussain, David C. Mohr, Mary Czerwinski, and Mi Zhang. 2018. Exploring User Needs for a Mobile Behavioral-Sensing Technology for Depression Management: Qualitative Study. *Journal of Medical Internet Research* 20, 7 (July 2018), e10139. https://doi.org/10.2196/10139 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[108] Mental Health America. 2024. Types of Mental Health Professionals. https://mhanational.org/types-mental-health-professionals

[109] Dar Meshi and Morgan E. Ellithorpe. 2021. Problematic social media use and social support received in real-life versus on social media: Associations with depression, anxiety and social isolation. *Addictive Behaviors* 119 (Aug. 2021), 106949. https://doi.org/10.1016/j.addbeh.2021.106949

[110] Jonah Meyerhoff, Tony Liu, Konrad P. Kording, Lyle H. Ungar, Susan M. Kaiser, Chris J. Karr, and David C. Mohr. 2021. Evaluation of Changes in Depression, Anxiety, and Social Anxiety Using Smartphone Sensor Features: Longitudinal Cohort Study. *Journal of Medical Internet Research* 23, 9 (Sept. 2021), e22844. https://doi.org/10.2196/22844 Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

[111] Jonah Meyerhoff, Tingting Liu, Caitlin A. Stamatis, Tony Liu, Harry Wang, Yixuan Meng, Brenda Curtis, Chris J. Karr, Garrick Sherman, Lyle H. Ungar, and David C. Mohr. 2023. Analyzing text message linguistic features: Do people with depression communicate differently with their close and non-close contacts? *Behaviour Research and Therapy* 166 (July 2023), 104342. https://doi.org/10.1016/j.brat.2023.104342

[112] Adam S. Miner, Scott L. Fleming, Albert Haque, Jason A. Fries, Tim Althoff, Denise E. Wilfley, W. Stewart Agras, Arnold Milstein, Jeff Hancock, Steven M. Asch, Shannon Wiltsey Stirman, Bruce A. Arnow, and Nigam H. Shah. 2022. A computational approach to measure the linguistic characteristics of psychotherapy timing, responsiveness, and consistency. *npj Mental Health Research* 1, 1 (Dec. 2022), 1–12. https://doi.org/10.1038/s44184-022-00020-9 Number: 1 Publisher: Nature Publishing Group.

[113] David C. Mohr, Katie Shilton, and Matthew Hotopf. 2020. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. *npj Digital Medicine* 3, 1 (March 2020), 1–2. https://doi.org/10.1038/s41746-020-0251-5 Number: 1 Publisher: Nature Publishing Group.

[114] David C. Mohr, Mi Zhang, and Stephen M. Schueller. 2017. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annual review of clinical psychology* 13 (May 2017), 23–47. https://doi.org/10.1146/annurev-clinpsy-032816-044949

[115] Elizabeth L. Murnane, Tara G. Walker, Beck Tench, Stephen Voida, and Jaime Snyder. 2018. Personal Informatics in Interpersonal Contexts: Towards the Design of Technology that Supports the Social Ecologies of Long-Term Mental Health Management. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (Nov. 2018), 127:1–127:27. https://doi.org/10.1145/3274396

[116] Sandrine R. Müller, Xi (Leslie) Chen, Heinrich Peters, Augustin Chaintreau, and Sandra C. Matz. 2021. Depression predictions from GPS-based mobility do not generalize well to large demographically heterogeneous samples. *Scientific Reports* 11 (July 2021), 14007. https://doi.org/10.1038/s41598-021-93087-x

[117] Sujay Nagaraj, Sarah Goodday, Thomas Hartvigsen, Adrien Boch, Kopal Garg, Sindhu Gowda, Luca Foschini, Marzyeh Ghassemi, Stephen Friend, and Anna Goldenberg. 2023. Dissecting the heterogeneity of "in the wild" stress from multimodal sensor data. *npj Digital Medicine* 6, 1 (Dec. 2023), 1–9. https://doi.org/10.1038/s41746-023-00975-9 Number: 1 Publisher: Nature Publishing Group.

[118] Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. 2017. Just-in-Time Adaptive Interventions (JITAIs) in Mobile Health: Key Components and Design Principles for Ongoing Health Behavior Support. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine* 52, 6 (Dec. 2017), 446–462. https://doi.org/10.1007/s12160-016-9830-8

[119] Benjamin W. Nelson and Nicholas B. Allen. 2019. Accuracy of Consumer Wearable Heart Rate Measurement During an Ecologically Valid 24-Hour Period: Intraindividual Validation Study. *JMIR mHealth and uHealth* 7, 3 (2019), e10828. https://doi.org/10.2196/10828 Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[120] Subigya Nepal, Arvind Pillai, William Campbell, Talie Massachi, Eunsol Soul Choi, Orson Xu, Joanna Kuc, Jeremy Huckins, Jason Holden, Colin Depp, Nicholas Jacobson, Mary Czerwinski, Eric Granholm, and Andrew T. Campbell. 2024. Contextual AI Journaling: Integrating LLM and Time Series Behavioral Sensing Technology to Promote Self-Reflection and Well-being using the MindScape App. https://doi.org/10.1145/3613905.3650767 arXiv:2404.00487 [cs].

[121] Subigya Nepal, Arvind Pillai, Emma M. Parrish, Jason Holden, Colin Depp, Andrew T. Campbell, and Eric L. Granholm. 2024. Social Isolation and Serious Mental Illness: The Role of Context-Aware Mobile Interventions. *IEEE Pervasive Computing* (2024), 1–11. https://doi.org/10.1109/MPRV.2024.3377200 Conference Name: IEEE Pervasive Computing.

[122] Ada Ng, Rachel Kornfield, Stephen M. Schueller, Alyson K. Zalta, Michael Brennan, and Madhu Reddy. 2019. Provider Perspectives on Integrating Sensor-Captured Patient-Generated Data in Mental Health Care. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 115:1–115:25. https://doi.org/10.1145/3359217

[123] Jodie Nghiem, Daniel A. Adler, Deborah Estrin, Cecilia Livesey, and Tanzeem Choudhury. 2023. Understanding Mental Health Clinicians' Perceptions and Concerns Regarding Using Passive Patient-Generated Health Data for Clinical Decision-Making: Qualitative Semistructured Interview Study. *JMIR formative research* 7 (Aug. 2023), e47380. https://doi.org/10.2196/47380

[124] Viet Cuong Nguyen, Nathaniel Lu, John M. Kane, Michael L. Birnbaum, and Munmun De Choudhury. 2022. Cross-Platform Detection of Psychiatric Hospitalization via Social Media Data: Comparison Study. *JMIR Mental Health* 9, 12 (Dec. 2022), e39747. https://doi.org/10.2196/39747 Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

[125] Jennifer Nicholas, Katie Shilton, Stephen M. Schueller, Elizabeth L. Gray, Mary J. Kwasny, and David C. Mohr. 2019. The Role of Data Type and Recipient in Individuals' Perspectives on Sharing Passively Collected Smartphone Data for Mental Health: Cross-Sectional Questionnaire Study. *JMIR mHealth and uHealth* 7, 4 (April 2019), e12578. https://doi.org/10.2196/12578

[126] Helen Nissenbaum. 2004. PRIVACY AS CONTEXTUAL INTEGRITY. *Washington Law Review* 79 (2004), 39.

[127] Guy Paré and Spyros Kitsiou. 2017. Chapter 9 Methods for Literature Reviews. In *Handbook of eHealth Evaluation: An Evidence-based Approach [Internet]*. University of Victoria. https://www.ncbi.nlm.nih.gov/books/NBK481583/

[128] Sachin R Pendse, Daniel Nkemelu, Nicola J Bidwell, Sushrut Jadhav, Soumitra Pathare, Munmun De Choudhury, and Neha Kumar. 2022. From Treatment to Healing:Envisioning a Decolonial Digital Mental Health. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–23. https://doi.org/10.1145/3491102.3501982

[129] George D. Price, Michael V. Heinz, Amanda C. Collins, and Nicholas C. Jacobson. 2024. Detecting major depressive disorder presence using passively-collected wearable movement data in a nationally-representative sample. *Psychiatry Research* 332 (Feb. 2024), 115693. https://doi.org/10.1016/j.psychres.2023.115693

[130] Mashfiqui Rabbi, Min Hane Aung, Mi Zhang, and Tanzeem Choudhury. 2015. MyBehavior: automatic personalized health feedback from user behaviors and preferences using smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, Osaka, Japan, 707–718. https://doi.org/10.1145/2750858.2805840

[131] Sebastian Raschka. 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. (2018).

[132] Darrel A. Regier, William E. Narrow, Diana E. Clarke, Helena C. Kraemer, S. Janet Kuramoto, Emily A. Kuhl, and David J. Kupfer. 2013. DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *American Journal of Psychiatry* 170, 1 (Jan. 2013), 59–70. https://doi.org/10.1176/appi.ajp.2012.12070999 Publisher: American Psychiatric Publishing.

[133] Paola Rocca, Claudio Brasso, Cristiana Montemagni, Silvio Bellino, Alessandro Rossi, Alessandro Bertolino, Dino Gibertoni, Eugenio Aguglia, Mario Amore, Ileana Andriola, Antonello Bellomo, Paola Bucci, Antonino Buzzanca, Bernardo Carpiniello, Alessandro Cuomo, Liliana Dell'Osso, Angela Favaro, Giulia Maria Giordano, Carlo Marchesi, Palmiero Monteleone, Lucio Oldani, Maurizio Pompili, Rita Roncone, Rodolfo Rossi, Alberto Siracusano, Antonio Vita, Patrizia Zeppegno, Silvana Galderisi, and Mario Maj. 2021. Accuracy of self-assessment of real-life functioning in schizophrenia. *NPJ Schizophrenia* 7 (Feb. 2021), 11. https://doi.org/10.1038/s41537-021-00140-9

[134] Naja Hulvej Rod, Agnete Skovlund Dissing, Alice Clark, Thomas Alexander Gerds, and Rikke Lund. 2018. Overnight smartphone use: A new public health challenge? A novel study design based on high-resolution smartphone data. *PLOS ONE* 13, 10 (Oct. 2018), e0204811. https://doi.org/10.1371/journal.pone.0204811 Publisher: Public Library of Science.

[135] Lisa S. Rotenstein, A. Jay Holmgren, Michael J. Healey, Daniel M. Horn, David Y. Ting, Stuart Lipsitz, Hojjat Salmasian, Richard Gitomer, and David W. Bates. 2022. Association Between Electronic Health Record Time and Quality of Care Metrics in Primary Care. *JAMA Network Open* 5, 10 (Oct. 2022), e2237086. https://doi.org/10.1001/jamanetworkopen.2022.37086

[136] Tomasz Rutowski, Elizabeth Shriberg, Amir Harati, Yang Lu, Piotr Chlebek, and Ricardo Oliveira. 2020. Depression and Anxiety Prediction Using Deep Language Models and Transfer Learning. In *2020 7th International Conference on Behavioural and Social Computing (BESC)*. 1–6. https://doi.org/10.1109/BESC51023.2020.9348290

[137] Sohrab Saeb, Emily G. Lattie, Stephen M. Schueller, Konrad P. Kording, and David C. Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (Sept. 2016). https://doi.org/10.7717/peerj.2537

[138] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C. Mohr, and Konrad P. Kording. 2017. The need to approximate the use-case in clinical machine learning. *GigaScience* 6, gix019 (May 2017). https://doi.org/10.1093/gigascience/gix019

[139] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *Journal of Medical Internet Research* 17, 7 (July 2015). https://doi.org/10.2196/jmir.4273

[140] Koustuv Saha, Ayse E. Bayraktaroglu, Andrew T. Campbell, Nitesh V. Chawla, Munmun De Choudhury, Sidney K. D'Mello, Anind K. Dey, Ge Gao, Julie M. Gregg, Krithika Jagannath, Gloria Mark, Gonzalo J. Martinez, Stephen M. Mattingly, Edward Moskal, Anusha Sirigiri, Aaron Striegel, and Dong Whi Yoo. 2019. Social Media as a Passive Sensor in Longitudinal Studies of Human Behavior and Wellbeing. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for

Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3290607.3299065

[141] Koustuv Saha, Ted Grover, Stephen M. Mattingly, Vedant Das swain, Pranshu Gupta, Gonzalo J. Martinez, Pablo Robles-Granda, Gloria Mark, Aaron Striegel, and Munmun De Choudhury. 2021. Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (March 2021), 1–32. https://doi.org/10.1145/3448117

[142] Pedro F. Saint-Maurice, Richard P. Troiano, David R. Bassett, Jr, Barry I. Graubard, Susan A. Carlson, Eric J. Shiroma, Janet E. Fulton, and Charles E. Matthews. 2020. Association of Daily Step Count and Step Intensity With Mortality Among US Adults. *JAMA* 323, 12 (March 2020), 1151–1160. https://doi.org/10.1001/jama.2020.1382

[143] Luke Scheuer and John Torous. 2022. Usable Data Visualization for Digital Biomarkers: An Analysis of Usability, Data Sharing, and Clinician Contact. *Digital Biomarkers* 6, 3 (Sept. 2022), 98–106. https://doi.org/10.1159/000525888

[144] Jessica Schroeder, Jina Suh, Chelsey Wilks, Mary Czerwinski, Sean A. Munson, James Fogarty, and Tim Althoff. 2020. Data-Driven Implications for Translating Evidence-Based Psychotherapies into Technology-Delivered Interventions. *International Conference on Pervasive Computing Technologies for Healthcare : [proceedings]. International Conference on Pervasive Computing Technologies for Healthcare* 2020 (May 2020), 274–287. https://doi.org/10.1145/3421937.3421975

[145] Gregory E. Simon, Nathalie Moise, and David C. Mohr. 2024. Management of Depression in Adults: A Review. *JAMA* (June 2024). https://doi.org/10.1001/jama.2024.5756

[146] Simone Schmidt and Simon D'Alfonso. 2023. Clinician perspectives on how digital phenotyping can inform client treatment. *Acta Psychologica* 235 (May 2023), 103886. https://doi.org/10.1016/j.actpsy.2023.103886 Publisher: North-Holland.

[147] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–18. https://doi.org/10.1145/3544548.3581075

[148] Matthew Sperrin, Richard D. Riley, Gary S. Collins, and Glen P. Martin. 2022. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagnostic and Prognostic Research* 6, 1 (Dec. 2022), 24. https://doi.org/10.1186/s41512-022-00136-8

[149] Robert L. Spitzer, Kurt Kroenke, Janet B. W. Williams, and Bernd Löwe. 2006. A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. *Archives of Internal Medicine* 166, 10 (May 2006), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092 Publisher: American Medical Association.

[150] Caitlin A. Stamatis, Jonah Meyerhoff, Yixuan Meng, Zhi Chong Chris Lin, Young Min Cho, Tony Liu, Chris J. Karr, Tingting Liu, Brenda L. Curtis, Lyle H. Ungar, and David C. Mohr. 2024. Differential temporal utility of passively sensed smartphone features for depression and anxiety symptom prediction: a longitudinal cohort study. *npj Mental Health Research* 3, 1 (Jan. 2024), 1–8. https://doi.org/10.1038/s44184-023-00041-y Number: 1 Publisher: Nature Publishing Group.

[151] Luke Stark. 2018. Algorithmic psychometrics and the scalable subject. *Social Studies of Science* 48, 2 (April 2018), 204–231. https://doi.org/10.1177/0306312718772094 Publisher: SAGE Publications Ltd.

[152] Arthur A. Stone and Saul Shiffman. 2002. Capturing momentary, self-report data: A proposal for reporting guidelines. *Annals of Behavioral Medicine* 24, 3 (Aug. 2002), 236–243. https://doi.org/10.1207/S15324796ABM2403_09

[153] Streamlit. 2021. Streamlit • A faster way to build and share data apps. https://streamlit.io/

[154] Mitchell Tang, Carter H. Nakamoto, Ariel D. Stern, and Ateev Mehrotra. 2022. Trends in Remote Patient Monitoring Use in Traditional Medicare. *JAMA Internal Medicine* 182, 9 (Sept. 2022), 1005–1006. https://doi.org/10.1001/jamainternmed.2022.3043

[155] Justin S. Tauscher, Eliza B. Cohn, Tascha R. Johnson, Kaylie D. Diteman, Richard K. Ries, David C. Atkins, and Kevin A. Hallgren. 2021. What do clinicians want? Understanding frontline addiction treatment clinicians' preferences and priorities to improve the design of measurement-based care technology. *Addiction Science & Clinical Practice* 16 (2021), 38. https://doi.org/10.1186/s13722-021-00247-5

[156] Sara Ann Taylor, Natasha Jaques, Ehimwenma Nosakhare, Akane Sano, and Rosalind Picard. 2017. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE Transactions on Affective Computing* (2017), 1–1. https://doi.org/10.1109/TAFFC.2017.2784832

[157] Thomas Insel. 2013. NIMH's New Focus in Clinical Trials - National Institute of Mental Health (NIMH). https://www.nimh.nih.gov/funding/grant-writing-and-application-process/concept-clearances/2013/nimhs-new-focus-in-clinical-trials

[158] Brett D. Thombs, Linda Kwakkenbos, Alexander W. Levis, and Andrea Benedetti. 2018. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. *CMAJ* 190, 2 (Jan. 2018), E44–E49. https://doi.org/10.1503/cmaj.170691 Publisher: CMAJ Section: Analysis.

[159] Sara Thomée, Annika Härenstam, and Mats Hagberg. 2011. Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults - a prospective cohort study. *BMC Public Health* 11, 1 (Jan. 2011), 66. https://doi.org/10.1186/1471-2458-11-66

[160] John Torous, Mathew V. Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. 2016. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health* 3, 2 (May 2016), e5165. https://doi.org/10.2196/mental.5165 Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health Publisher: JMIR Publications Inc., Toronto, Canada.

[161] Harry C. Triandis. 1982. Dimensions of Cultural Variation as Parameters of Organizational Theories. *International Studies of Management & Organization* 12, 4 (1982), 139–169. https://www.jstor.org/stable/40396948 Publisher: Taylor & Francis, Ltd..

[162] Vincent W.-S. Tseng, Akane Sano, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Marta Hauser, John M. Kane, Emily A. Scherer, Rui Wang, Weichen Wang, Hongyi Wen, and Tanzeem Choudhury. 2020. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. *Scientific Reports* 10, 1 (Sept. 2020), 15100. https://doi.org/10.1038/s41598-020-71689-1 Number: 1 Publisher: Nature Publishing Group.

[163] Rudolf Uher, Jennifer L. Payne, Barbara Pavlova, and Roy H. Perlis. 2014. Major Depressive Disorder in Dsm-5: Implications for Clinical Practice and Research of Changes from Dsm-Iv. *Depression and Anxiety* 31, 6 (2014), 459–471. https://doi.org/10.1002/da.22217 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.22217.

[164] Rudolf Uher, Roy H. Perlis, Anna Placentino, Mojca Zvezdana Dernovšek, Neven Henigsberg, Ole Mors, Wolfgang Maier, Peter McGuffin, and Anne Farmer. 2012. Self-Report and Clinician-Rated Measures of Depression Severity: Can One Replace the Other? *Depression and Anxiety* 29, 12 (2012), 1043–1049. https://doi.org/10.1002/da.21993 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/da.21993.

[165] Andrius Vabalas, Emma Gowen, Ellen Poliakoff, and Alexander J. Casson. 2019. Machine learning algorithm validation with a limited sample size. *PLOS ONE* 14, 11 (Nov. 2019), e0224365. https://doi.org/10.1371/journal.pone.0224365 Publisher: Public Library of Science.

[166] Jayne Wallace, John McCarthy, Peter C. Wright, and Patrick Olivier. 2013. Making design probes work. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 3441–3450. https://doi.org/10.1145/2470654.2466473

[167] Monica L. Wang, Cristina M. Gago, and Kate Rodriguez. 2024. Digital Redlining—The Invisible Structural Determinant of Health. *JAMA* 331, 15 (April 2024), 1267–1268. https://doi.org/10.1001/jama.2024.1628

[168] Rui Wang, Emily A. Scherer, Vincent W. S. Tseng, Dror Ben-Zeev, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, and Michael Merrill. 2016. 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 - UbiComp '16*. ACM Press, Heidelberg, Germany, 886–897. https://doi.org/10.1145/2971648.2971740

[169] Rui Wang, Weichen Wang, Min S. H. Aung, Dror Ben-Zeev, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A. Scherer, and Megan Walsh. 2017. Predicting Symptom Trajectories of Schizophrenia Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3 (Sept. 2017), 110:1–110:24. https://doi.org/10.1145/3130976

[170] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (March 2018), 43:1–43:26. https://doi.org/10.1145/3191775

[171] Rebecca Watson, Kate Harvey, Ciara McCabe, and Shirley Reynolds. 2020. Understanding anhedonia: a qualitative study exploring loss of interest and pleasure in adolescent depression. *European Child & Adolescent Psychiatry* 29, 4 (April 2020), 489–499. https://doi.org/10.1007/s00787-019-01364-y

[172] Amy Wenzel. 2017. Basic Strategies of Cognitive Behavioral Therapy. *Psychiatric Clinics* 40, 4 (Dec. 2017), 597–609. https://doi.org/10.1016/j.psc.2017.07.001 Publisher: Elsevier.

[173] Janet B. W. Williams. 1988. A Structured Interview Guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry* 45, 8 (Aug. 1988), 742–747. https://doi.org/10.1001/archpsyc.1988.01800320058007

[174] H. Wisniewski and J. Torous. 2020. Digital navigators to implement smartphone and digital tools in care. *Acta Psychiatrica Scandinavica* 141, 4 (2020), 350–355. https://doi.org/10.1111/acps.13149 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/acps.13149.

[175] Andrew Wong, Erkin Otles, John P. Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, Muhammad Ghous, and Karandeep Singh. 2021. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. *JAMA Internal Medicine* (June 2021). https://doi.org/10.1001/jamainternmed.2021.2626

[176] World Economic Forum. 2023. *The Moment of Truth for Healthcare Spending: How Payment Models can Transform Healthcare Systems*. Technical Report. World Economic Forum. https://www3.weforum.org/docs/WEF_The_Moment_of_Truth_for_Healthcare_Spending_2023.pdf

[177] Wanwan Xu, Chang Su, Yan Li, Steven Rogers, Fei Wang, Kun Chen, and Robert Aseltine. 2021. Improving suicide risk prediction via targeted data fusion: proof of concept using medical claims data. *Journal of the American Medical Informatics Association* (Nov. 2021), ocab209. https://doi.org/10.1093/jamia/ocab209

[178] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (Jan. 2023), 190:1–190:34. https://doi.org/10.1145/3569485

[179] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–11. https://doi.org/10.1145/3290605.3300468

[180] Y Yao, E Tung, and B Glisic. 2014. High-resolution sensing sheet for damage detection based on large area electronics. In *Bridge Maintenance, Safety, Management and Life Extension*. CRC Press, 2706–2713. https://doi.org/10.1201/b17063-405

[181] Cheng-Fang Yen, Cheng-Chung Chen, Yu Lee, Tze-Chun Tang, Chih-Hung Ko, and Ju-Yu Yen. 2005. Insight and correlates among outpatients with depressive disorders. *Comprehensive Psychiatry* 46, 5 (Sept. 2005), 384–389. https://doi.org/10.1016/j.comppsych.2004.11.004

[182] Yuezhou Zhang, Amos A. Folarin, Shaoxiong Sun, Nicholas Cummins, Yatharth Ranjan, Zulqarnain Rashid, Pauline Conde, Callum Stewart, Petroula Laiou, Faith Matcham, Carolin Oetzmann, Femke Lamers, Sara Siddi, Sara Simblett, Aki Rintala, David C. Mohr, Inez Myin-Germeys, Til Wykes, Josep Maria Haro, Brenda W. J. H. Penninx, Vaibhav A. Narayan, Peter Annas, Matthew Hotopf, Richard J. B. Dobson, and Radar-Cns Consortium. 2021. Predicting Depressive Symptom Severity Through Individuals' Nearby Bluetooth Device Count Data Collected by Mobile Phones: Preliminary Longitudinal Study. *JMIR mHealth and uHealth* 9, 7 (July 2021), e29840. https://doi.org/10.2196/29840 Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.