Seamful Design Considerations for Human-in-the-Loop Digital Phenotyping of Mental Health

VEDANT DAS SWAIN, New York University, USA
TUNWA TONGTAWEE, Northeastern University, USA
OLIVIA WANG, Northeastern University, USA
JOYCE HSU, Northeastern University, USA
NICHOLAS C. JACOBSON, Dartmouth College, USA
VARUN MISHRA, Northeastern University, USA

Digital Phenotyping of Mental Health (DPMH) through passive sensing is a promising approach for personal health informatics and digital wellbeing. Its appeal lies in unobtrusiveness, making it appear seamless. However, this very quality leads users to find it impersonal, untrustworthy, and disengaging. To counteract challenges of seamlessness, researchers propose seamful design to deliberately engage users. Yet, it remains unclear how this principle can be incorporated into digital phenotyping. To address this, we conducted a formative study by developing DYMOND. It is a technology probe that estimates depression, explains estimates, reveals discrepancies, and provides user control over the underlying model. In a 6-week deployment, 22 individuals with moderate–severe depression monitored their state with DYMOND. They interviewed every two weeks with researchers to collaboratively reconfigure the model and co-design new interfaces. Our analysis of 57 sessions revealed (i) seams—friction points—across data, modeling, and output, and (ii) design requirements helping users evaluate and mitigate seams. These findings inform the design requirements for human-in-the-loop DPMH to support agency, transparency, and collaborative reflection. This study provides insight into theoretical re-conceptualization for passive sensing, opportunities to integrate large language models and human-AI interaction for better interfaces for digital mental health, and pathways to involve expert stakeholders in DPMH.

CCS Concepts: • **Human-centered computing → Collaborative and social computing systems and tools**; **Empirical studies in HCI**; • **Applied computing** → *Psychology*; **Health informatics**.

Additional Key Words and Phrases: Digital phenotyping, Mental health, Digital health, Psychiatry, Human-computer interaction, Personal health informatics, Seamful design, Human-in-the-loop, Passive sensing, Depression

ACM Reference Format:
Vedant Das Swain, Tunwa Tongtawee, Olivia Wang, Joyce Hsu, Nicholas C. Jacobson, and Varun Mishra. 2026. Seamful Design Considerations for Human-in-the-Loop Digital Phenotyping of Mental Health. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3, Article 154 (September 2026), 33 pages. https://doi.org/10.1145/3832021

1 INTRODUCTION

For people facing behavioral mental health challenges, recalling their experiences during clinical visits (or therapy) can be limited and daily journaling can be burdensome. A compelling means to complement these

Authors' Contact Information: Vedant Das Swain (corresponding author), New York University, New York, New York, USA, v.das.swain@nyu.edu; Tunwa Tongtawee, Northeastern University, Boston, Massachusetts, USA, tongtawee.t@northeastern.edu; Olivia Wang, Northeastern University, Boston, Massachusetts, USA, owang388@gmail.com; Joyce Hsu, Northeastern University, Boston, Massachusetts, USA, joyce.hsu028@gmail.com; Nicholas C. Jacobson, Dartmouth College, Hanover, New Hampshire, USA, Nicholas.C.Jacobson@dartmouth.edu; Varun Mishra, Northeastern University, Boston, Massachusetts, USA, v.mishra@northeastern.edu.

ACM 2474-9567/2026/9-ART154
https://doi.org/10.1145/3832021

pieces of information is by passively sensing human behaviors through their digital technology, known as *digital phenotyping* [44]. So far, ***digital phenotyping of mental health*** **(DPMH)** has shown promise in fostering continuous, unobtrusive, and automatic monitoring of mental health outcomes like anxiety, stress, depression, and substance use disorders [10, 23, 47, 105]. Large volumes of multimodal data (e.g., sleep, mobility, exercise, screen time) can be monitored over long periods to infer a user's mental state [26, 70, 76, 108]. By design, DPMH systems are intended to be *seamless*. These quietly gather high-fidelity naturalistic data as users continue their daily routines. Although algorithms continue to advance, even when trained with hundreds of data-subjects over several years and with sophisticated models, DPMH struggles to perform [66, 108] and generalize [70]. Despite over a decade of research in DPMH, the needle has barely moved [44, 66]. Unlike passive sensing of physical phenomena (e.g., sleep tracking and glucose monitors), DPMH systems indirectly approximate their target—user's mental state—based on subjective and complex ground truth [23, 98]. For instance, one depressed individual may sleep more when depressed, while another has trouble sleeping. Similarly, a depressed individual may want more rest but their work schedule forces them to stay active. When deployed, the imperceptible monitoring and unintelligible modeling of DPMH can lead to lack of trust, and lack of motivation [79], which in turn leads to reduced adoption, compliance, and adherence [11, 37, 101, 102]. DPMH users are often left with little clarity and stifled in self-determining care [1, 31]. As a result, existing seamless DPMH systems are rarely in the position to engage individuals enough for personalization.

By contrast, personalization, customization, and configurability are known to improve engagement in personal health informatics tools [48, 97]. We posit that the seamlessness needs to be complemented by *seamful* design [12]. This principle aims to reveal *seams*, which are logical points to demarcate the scope of a system and in turn guide users to appropriate the system meaningfully [43]. Users will encounter friction even in well refined systems, such as a high-performing DPMH model, because one-size-fits-all systems seldom exist. Consider DPMH as an infinite fabric woven into a user's daily routine. *Seams* are the stitching lines — they reveal where the fabric might fray, but also where a user can tailor it to fit.

In machine learning applications, an analog of seamful design is human-in-the-loop (HitL) user flows that enable users to customize systems based on personal intent even after deployment [9]. Unlike traditional participatory design, which involves stakeholders at the prototyping phase, these approaches propose *interactive machine learning* where users can "proactively teach" everyday intelligent systems about their (non-)preference during regular use [55]. However, these methods are not directly transferrable to DPMH because of several limitations in prior work. First, seamful design and HitL literature focus on systems with explicit friction points (e.g., poor network coverage zones [12]), which are handled by experts. Users of DPMH systems are neither mental health clinicians nor data scientists, due to which the friction is far less legible. Second, user interaction is only investigated in task-specific ML applications (e.g., tumor detection) [9]. DPMH systems work continuously, are task agnostic, and often infer something about the user and their identity, as opposed to helping them complete a task. Last, the evidence of friction in everyday intelligent tools is based on retrospective accounts [55]. Most challenges for DPMH are situated within a user's daily routine that involves a naturalistic loop of data collection and inference. Even though calls for centering human needs in DPMH continue to grow [1, 25, 101, 113], we still witness a gap in practically implementing interfaces to support user control.

In traditional care, patients periodically reflect and renegotiate how they are being diagnosed and treated. How might this happen in DPMH systems? Elegantly integrating seamfulness can provide users an opportunity for similar recalibration of DPMH tools. However, training a DPMH system is expensive, takes a long period of time, and involves significant human resources. This reality makes it impractical—and possibly, unethical—to deploy an interactive DPMH tool at scale without foundational knowledge of the design space, user intentions, and user burdens. To mitigate this barrier to design, we need to identify **what types of friction users in DPMH encounter, and how they can configure DPMH to overcome this friction with minimal burden.** For this, we designed and deployed a DPMH system for daily depression estimation as a *technology probe* [42]. Probes are not designed for comparative experiments, and are artifacts deployed in natural user settings to gather *formative* design insight through qualitative synthesis that is needed to design systems better. The method has strong precedent in Ubicomp research [30, 57] and has again gained renewed prominence in IMWUT studies focused on personal health informatics [16, 67]. Our system, DYMOND (DYnamic MONitoring for Depression), lets users scrutinize how the estimation model is using their passive sensing data to estimate their mental state, and also gives them an interface to adjust aspects of the underlying model. As a technology probe, DYMOND is neither intended to accurately forecast depression, nor does it promise seamful design. Instead, **the probe surfaces conditions to elicit situated user feedback about their experience with a DPMH.** This probe was directed to **gain a foundational understanding of seamful designs in DPMH** by answering the following research questions: **[RQ I]—**What are the *seams* in DPMH for estimating depression? **[RQ II]—**How do users want to navigate the *seams* of DPMH?

Much like conventional DPMH systems, DYMOND collated multimodal data from users' smartphones and wearables to algorithmically model a user's depression level. We focused on depression as it is co-morbid with other concerns like anxiety and substance use [33, 64] and has drawn significant attention in DPMH research [46, 105, 108], but has shown little evidence of involving end-user interaction. Furthermore, depression manifests in heterogeneous ways [32] that vary by unique lived experiences. Uniquely, our system situated users deeper into the DPMH loop by (i) explicitly disclosing discrepancies between the model's estimate of depression and what the user reported; and (ii) providing users an opportunity to modify the features being used by the underlying algorithm. We deployed DYMOND with 22 individuals, who screened positive for moderate–severe depression. Each participant used our probe for six weeks, and regularly assessed DYMOND for discrepancies. Every two weeks, we conducted collaborative evaluation sessions to make sense of the system's limitations and guide them in configuring DYMOND's features. Through our investigation, we make the following key contributions to design human-in-the-loop DPMH:

- **Theoretical—Identifies Seams in DPMH (RQI, §4)**: *Seams* in DPMH exist throughout its pipeline, encompassing the input (*data seams*), the processing (*modeling seams*), and the output (*outcome seams*). The naturalistic probe helped highlight and subvert the assumptions underlying these systems. These seams equip stakeholders with shared vocabulary to describe the friction in the system as epistemic and not merely technical. Our taxonomy promotes researchers to rethink the roles and paradigms in passive sensing (§6.1).
- **Practical—Designs to Navigate Seams in DPMH (RQII, §5)**: Users were motivated to identify errors and appropriate these systems, but face barriers of cognitive burden, limited controllability, and lack of expertise. To mitigate these challenges, we derived six design requirements (**DR**s) from co-design artifacts. Furthermore, we discuss how these DRs guide user-centered development of DPMH that synergizes its seamless and seamful aspects (§6.2) for improved personalization with minimal data-work.

Our technology probe, DYMOND, helped longitudinally inspect user experiences with a DPMH system that not only collects their data, but actually estimates their mental state. This led us to produce one of the first papers to amplify and anticipate how users want to leverage DPMH tools in their everyday life.

2 BACKGROUND

DPMH has been described as "moment-by-moment quantification of the individual-level human phenotype in situ using data from personal digital devices" [78]. This umbrella term describes approaches to passively leverage behavioral traces from ubiquitous everyday technology (e.g., smartphones and wearables) to understand social, emotional, and cognitive states in free-living conditions[1]. DPMH presents an important step to patient-centered

[1]Throughout our paper we use the term "DPMH" to refer to passive sensing–enabled DPMH. Other forms may leverage data sources that are not passively collected, e.g., electronic health records.

care [93] by providing users with unique insight into their health. DPMH has shown promise in supporting anxiety-related disorders [47], depression [105], and substance use [10]. The value proposition of DPMH is that users can get continuous digital insights with negligible effort compared to daily journaling and frequently responding to extensive surveys [89].

2.1 Perils of Seamless DPMH

Passive sensing is automatic, portable and can remain always-on [21]. After all, gaining insights for low user effort implies greater adoption [104]. Patients in clinical settings have expressed willingness to use DPMH to enhance their care [89]. However, this assumes that passive sensing is a *seamless* method that "just works" in the background. In actual deployment, typical approaches to passive sensing struggle to retain users [37, 102]. Meta-analyses of depression applications indicate a dropout rate of 47.8% and daily engagement of only 4% [100]. The systems can be unobtrusive to a fault, and be perceived as less personal, less intelligible, and more invasive [2]. Therefore, despite the best intentions, the core design feature of DPMH—passive collection and modeling of data—also gives rise to its core design limitation.

2.1.1 DPMH is Distinct from Other Forms of Passive Tracking of Behavior. Since the conception of the *quantified self* movement [53], digital tools for personal tracking have become commonplace (e.g., step counters and glucose monitors). DPMH, however, often goes beyond simple counting or measurement. This process collects different kinds of data—often representing indirect relationships to the target of interest (e.g., sleep to infer stress)—and algorithmically models these data to estimate a complex psychosocial states that are often non-trivial to label for training [98]. The data pipeline for DPMH involves a sequence of abstraction which can make it ambiguous for users to understand how the estimate was derived from the initial data [23]. The barriers to readily adopt DPMH is not purely about functionality, but related to nebulous semantics. Unlike standard activity trackers, users struggle to just setup and forget [25, 55, 100]. While modeling continues to improve [66, 70], by overlooking these user experiences we risk building monitoring that lack practical meaning and become difficult to verify. Research on HitL machine learning inspires our study [9] but does not sufficiently support DPMH. These systems involve very different data streams (activity tracking), different life-cycle of deployment (continuously in the background), and different users (non-experts who are also data-subjects). To address this, we studied experiences with DPMH systems through a longitudinal deployment of DYMOND because it situates user perspectives in naturalistic practice of evaluating algorithmic estimates of one's mental state.

2.1.2 Current DPMH Lacks Methods for Users' Configurations. Mohr et al. note the challenges in DPMH stem from its "top-down" design, where clinicians' and developers' assertions supersede the patient's preferences [73]. Consequently, users foster anxieties of uncertain self-presentation through the DPMH model [1]. Not only does the system fail to represent users' lived experience in their data, but the system also fails to provide an opportunity to remediate. Users have an underlying urge to self-determine their care and this supports their wellbeing [31]. However, their ability to actually exercise control on their health is often influenced by the technology that tracks it because it represents "socio-material arrangements, which offer different capacities for actions" [77]. Participatory design to involve users in the development of the system [101] is a step towards capturing user perspectives, but it still leads to a rigid design. Predefined health tracking approaches are known to hinder agency, but instead, supporting individual preferences in an interactive manner improves mindfulness and meaningful engagement [3]. It also respects users' continually changing behavioral dynamics [28]. Our technology probe, DYMOND, was designed to allow fluid customization that accommodates users' continually evolving everyday experiences.

2.1.3 Control in DPMH is Fundamental to Engagement. We know from traditional care and treatment settings that patients are more likely to discontinue treatment when they experience lack of trust, connection, and mutual respect between themselves and clinicians [4]. These issues have been shown to create hesitation in seeking and continuing treatment for depression [60] and substance use disorder [34]. Considering that DPMH systems are also positioned as instruments for care, the lack of control a user has on their data can demotivate continuous usage. Existing DPMH paradigms neither provide any feedback on how user data will be modeled, nor do they give users a medium to give feedback to the underlying model. To mitigate this problem, Grönvall and Verdezoto suggest moving the user closer to the locus of control for their data [35]. Conceptually, this is a sound recommendation, but how might systems practically facilitate such an interaction? DPMH systems emerged from the need to reduce user burden, and expecting excessive user control can make mental health monitoring exhausting. Hence, as a technology probe, DYMOND created conditions for users to envision interaction paradigms for DPMH by giving them an opportunity for configuration.

Recent studies on DPMH show increased attention to its inherent challenges [6, 49, 112, 114] and have started to grapple with solutions to these perils through an aesthetic lens [7] and through refined model engineering [69]. **Our research extends prior efforts by considering the perils of seamlessness as a design opportunity itself.** This study distinguishes itself by investigating the conception and design of an alternative DPMH system by considering its interactive elements, underlying mechanisms, and naturalistic context.

2.2 Seamfulness in Passive Systems

Instead of trying to make DPMH work seamlessly, say, by trying to improve algorithmic accuracy, our study embraces the contrasting concept of *seamfulness*. Chalmers describes *seams* as points that account for the "finite and physical nature of digital media" [12]. Their research showed that despite making many algorithmic improvements to a spatial database for cell network coverage, the users had to adjust to many errors and uncertainties in the system. Moreover, users learned to navigate the bounds of the technology (e.g., choosing not to make phone calls in areas they know have weak cellphone coverage). The seamful approach accounts for the seams of a digital system by "deliberately revealing" them to the user so that they can be used flexibly [12]. Since the conception of seams, Inman and Ribes have argued that *seamlessness* and *seamfulness* are complementary, not contradictory ideas [43]. According to them, *seams* indicate some point of variation in how a user uses a system and how they change their lives outside the digital context visible to the system. Driven by these studies, **we posit that DPMH tools are also appropriated, but to make such systems robust to adaptation, we need to understand how users adapt**. An unobtrusive system is hard for users to adjust to because its discrepancies are unclear. Carter and Mankoff stated that "implicit sensing can be too unobtrusive" and users of such tools will "desire hooks allowing them to control" the sensors monitoring their behaviors [11]. Identifying the seams of DPMH and opportunities to leverage them could help users find a balance between passive sensing and flexible control. Thus, we designed DYMOND to illuminate the uncertainty in DPMH to then learn how users adjust their interaction with the system.

The *seams* of DPMH can also be viewed as *agential cuts* [114]. According to this concept, users have limited *agency* on the data they produce as it is constantly entangled with their ecology. This leads to *cuts* or logical points that separate real-world factors from the target outcome (e.g., depression). For instance, in menstruation tracking apps the (ir)regularity of users' cycles is a type of *agential cut* that explains the misestimate of menstruation cycles [114]. Similarly, the heterogeneity of user goals for migraine tracking presents another such separation between a user's experience and the capacities of the system [86]. These studies call for the need to take stock of varying experiences and goals by supporting user appropriation to adjust to uncertainty. Seamful design is a way to do this [12, 43], but we have little to no knowledge on the seams in DPMH. Instead, we look towards other studies of uncertainty and appropriation in passive sensing to formulate our investigation. Yang and Newman's study of smart thermostats shows that for users to adapt to a system's limits, the system should support *exception flagging* (users should notice and articulate errors) and *constrained engagement* (users can input without being overwhelmed) [109]. We followed this recommendation **to study DPMH in practice through a system that probes for seams to lay the foundation for seamful design.**

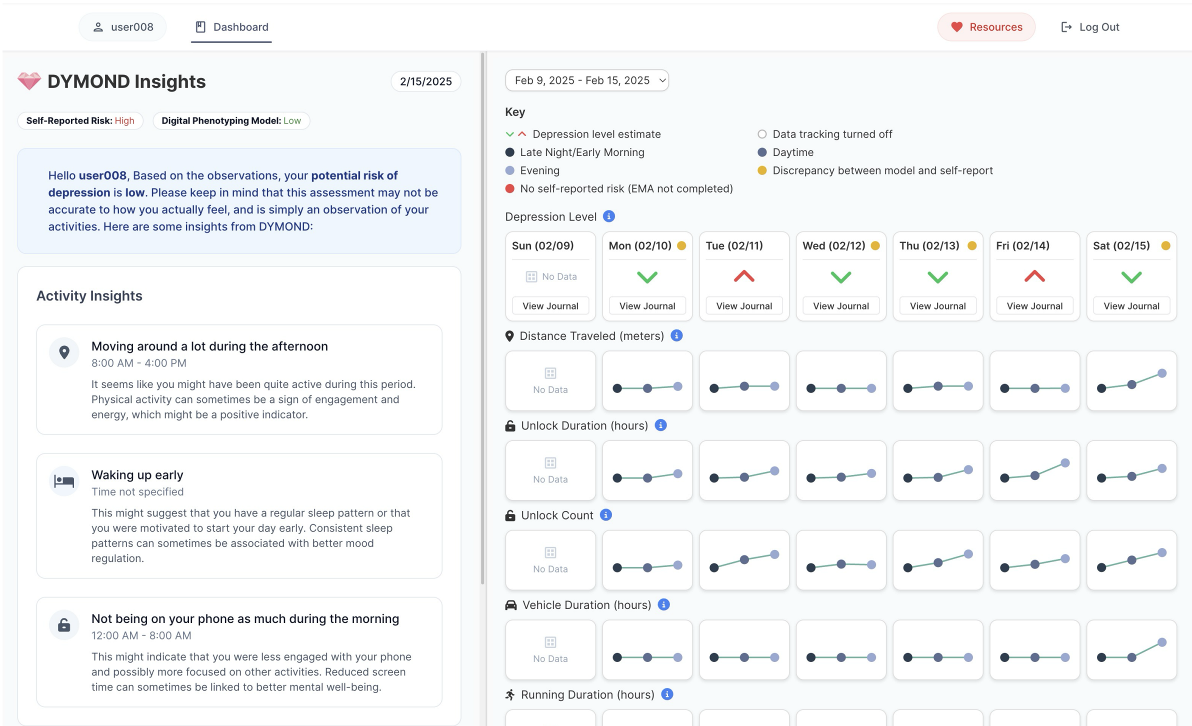


Fig. 1. DYMOND allows users to retroactively evaluate how DPMH models estimate their mental state. (Right pane) A minimalist visualization showing how DYMOND estimated their depression and how they behaved on those days. (Left pane) For the latest day, DYMOND explains the rationale behind its estimate. When DYMOND's estimate did not match what was self-reported by the user it is denoted by a yellow dot.

3 DYMOND: A TECHNOLOGY PROBE FOR SEAMFUL DPMH

Users' lived experiences vary, and therefore, monolithic DPMH systems can be incompatible with diverse situational contexts. One way to involve users is through Participatory design (PD) [62]—iterative development of a system through several workshops. Relying on PD alone would still produce a fixed artifact that does not allow user involvement after deployment. Alternatively, a seamful design can be represented through HitL approaches [29, 82], which accommodate new user experiences by allowing them to debug and update the system when needed. However, to build such a system assumes users have sufficient knowledge of the design space. For DPMH, this knowledge does not yet exist — we do not know where users encounter friction, what they are capable of configuring, or how much burden they can sustain. Thus, to design such a system, we need to first identify what users need to adapt to (RQI) and how they try to adapt (RQII). Following Inman and Ribes [43], to get a deeper understanding of seams, we designed DYMOND as a *technology probe* [42] to make explicit certain embodied or implicit assumptions of DPMH.

Table 1. DYMOND Features with directional contributions by epoch. Epochs: Late Night/Early Morning (00:00–08:00), Daytime (08:00–16:00), Evening (16:00–24:00). (+) adds to depression score, (−) subtracts from score.

Category	**Features**
Physical Activity	Running Duration: Late Night/Early Morning (−), Daytime (−), Evening (−) Walking Duration: Late Night/Early Morning (+), Daytime (−), Evening (−) Stationary Duration: Late Night/Early Morning (−), Daytime (+), Evening (+)
Commute Behavior	Vehicle Duration: Late Night/Early Morning (−), Daytime (−), Evening (−) Bicycle Duration: Late Night/Early Morning (−), Daytime (−), Evening (−)
Location Behavior	Distance Traveled: Late Night/Early Morning (+), Daytime (−), Evening (−)
Sleep Patterns	Sleep Duration (+), Sleep Start Time (+), Sleep End Time (+)
Phone Usage	Unlock Duration: Late Night/Early Morning (+), Daytime (−), Evening (+) Unlock Count: Late Night/Early Morning (+), Daytime (+), Evening (+)

Technology Probe: DYMOND is a *technology probe* [42]—a working tool to reveal user experiences through real use. DYMOND takes inspiration from probes like *Pearl* [51] (a tool to reflect on automatically logged work behavior data) and *MigraineTracker* [86] (a user-configured self-tracking tool). DYMOND distinguishes itself by focusing on algorithmic estimation of everyday mental-health outcomes (depression) with passively collected data. Some recent studies have used similar approaches in mental health, but they have either focused on probes to study explanations [56] or interventions [40, 58]. By contrast, DYMOND is the first probe centered around interacting with the model itself.

System Outline: DYMOND augments traditional DPMH by exposing uncertainty in the system and allowing users to correct it. At its core, it leverages a passive sensing platform to gather and interpret behavioral data. Additionally, it includes a dashboard for users (i) to view their data, (ii) to learn DYMOND's interpretation of their mental health, and (iii) to configure the data they provide to DYMOND. This section elaborates on the major technical components of our prototype, the underlying passive sensing platform, and our investigation protocol.

3.1 Passive Sensing Platform

The core of every DPMH system is a passive sensing platform that captures multiple streams of activity data and processes them into behavioral features. We leveraged the data-collection platform developed by Mishra et al., which has been deployed in a variety of DPMH studies [59, 63, 72, 95] collecting over 100,000 person-days of passive sensing data. This smartphone application passively recorded users' physical activity, locations, and phone usage. Additionally, participants received a Garmin wearable to help DYMOND retrieve users' sleep patterns (through the accelerometer and heart rate). Our choice of features was informed by prominent DPMH studies focused on depression monitoring [105]. The iOS application used the SwiftUI kit for the UI and also incorporated the Garmin Health SDK for real-time data access to the Garmin wearable. The backend data management server was built using Flask and MongoDB. The server also maintained the custom preferences for each unique user. Thus, we could customize the notification schedule for completing daily surveys and reviewing the dashboard based on participants' sleep schedule. At regular intervals, the app syncs data back to the server over WiFi and flushes it from the local storage of the iPhone.

3.2 Depression Estimation Model

Ideally, health monitoring in free-living should be deployed with pre-trained machine learning models. However, standard DPMH models rarely generalize between cohorts and even require retraining within similar cohorts [70].

An alternative would be to first accumulate training data for several weeks to personalize to the user [56]. Yet, this time constraint leads to a *cold-start* problem where users have no interaction opportunity with our probe at early stages of participation, risking disengagement. Moreover, DYMOND needs to help users assess depression estimates, but standard machine learning approaches often require post-hoc interpretability (e.g., *SHAP* or *LIME*). Even if we aggregate enough representative data to train such models, explaining these models rarely provides actionable insight to non-technical end-users [20]. Instead, although linear models sacrifice performance for transparency [68], their explanations do not involve "complicated decision pathways", making them more usable [83]. Given the lack of appropriate standardized models to adapt, we opted to design a domain-grounded heuristically weighted model. This helps balance the practical constraints of depression monitoring with the aims of the technology probe. Howe et al. used a similar method to probe stress monitoring and interventions in free-living [40]. We determined the features (and initial weights) for DYMOND to mirror standard features from prior DPMH studies that approximate nomothetic correlations of clinical symptoms of depression described in the *DSM-5* [46, 75, 105]. For instance, phone unlock at night contributes additively towards the depression score. Alternatively, healthy sleep or walking in the morning reduces the score. Table 1 lists all our features and their valence contribution. For initial deployment, features were standardized based on the sample distributions in the *Dartmouth* dataset [76]. After collecting 14 days of data, DYMOND would scale the features as per the users' distributions. DYMOND's feature weights were refined to get the best results on the *Dartmouth* dataset. Our initial model was more confident in estimating low depression level than high (Negative Prediction Value of 76.8%). However, overall, the prediction accuracy was only slightly better than chance (55.3%). While it may seem that sophisticated models can be more accurate, recent analyses of multiple DPMH algorithms in the GLOBEM datasets show that the ceiling of model performance is not notably better than random (maximum accuracy rarely exceeds 0.57) [108]. Therefore, the uncertainty of our probe is comparable to the state-of-the-art. The inaccuracy also favors our intention to examine friction in DPMH, as "purposeful malfunction" [54] is a valid method to steer users to deeply evaluate systems. Thus, our probe was robust despite the performance, because DYMOND only estimates depression for past days, reveals inaccuracies, and translates feature heuristics to the users.

$$\mathrm{DYMOND}_d = \sum_{i=1}^{n} w_i \cdot z_{i,d} \tag{1}$$

$$\text{where } z_{i,d} = \frac{f_{i,d} - \mu_i}{\sigma_i} \tag{2}$$

$$\mu_i, \sigma_i = \begin{cases} \text{Dartmouth dataset} & \text{if } d \leq 14 \\ \text{User's data} & \text{if } d > 14 \end{cases} \tag{3}$$

3.3 Dashboard Components

To pursue seamfulness, Inman and Ribes recommend making certain *seams* in a system visible to foster discovery of new seams [43]. In other words, we designed DYMOND such that implicit and overlooked aspects of DPMH were intentionally more pronounced to the user, and to learn how they would adapt, we designed affordances for configuring the model to align with their needs. It contained three primary components [2]:

(1) **Behavior visualization:** First, we wanted to make users' own behaviors and deviations in captured data more legible. Recognizing our users are not DPMH experts, we considered a minimal visualization approach derived from previous studies on behavioral data exploration [15, 27, 96], particularly emulating Das Swain

[2]Built on Vue.js and Node.js

et al.'s multimodal data-walkthrough visualization [27]. The dashboard's central section graphically described users' behaviors, DYMOND's depression estimation, and user self-reported estimates (right pane in Fig. 1). The top row indicated DYMOND's depression level estimates: high (*red upward arrow*) or low (*green downward arrow*), available only for past days with collected data. A *yellow dot* signified when DYMOND's estimate mismatched users' self-reports. Subsequent rows visualized major features using dot strings representing each activity across three daily epochs ("late night/early morning": 12am–8am; "daytime": 8am–4pm; "evening": 4pm–12am). Horizontally spaced dots in different blue shades had vertical positions based on deviation from the model's mean (above-mean values placed higher).

(2) **Insights and journaling:** Second, we wanted to surface the scope of the underlying model so that users can rationalize how and why the estimates diverge. We drew from Kim et al.'s *MindScope* study that showed detailed explanations were the most preferred way to understand models [56]. Explanations nudge users to recall past events and encourage visual data exploration [15]. A panel of insights accompanied the daily visualizations. We leveraged GPT-4o [41] to articulate, in natural language, the top features that contributed to the estimated depression level (left pane in Fig. 1). DYMOND captured users' in-the-moment reflections through a journaling text-input (Fig. 2) by prompting them to consider how the system might be (mis)interpreting their mental state.

(3) **Configuration panel:** Finally, we wanted to inspect the feasibility of providing users with agency and learn how to design for the actions they would like to take on encountering friction. After evaluating DYMOND's efficacy, users had the option to change which data the model should consider going forward. End-user selection to track personal health data has been typically restricted to self-reported data [86]. We extended this notion to passively collected data for mental health monitoring. As an initial seed into seamfulness, users could toggle certain features (at the epoch level) for DYMOND to exclude from, or include in, its estimation model. To prevent unintended and uninformed tinkering with the model, we restricted this panel under an administrative lock. Therefore, the configuration panel could only be accessed under the supervision of a researcher.

Expert Iteration and Ethical Design Considerations. To refine the design of our probe, we conducted interviews with four domain experts who have extensively published research in DPMH broadly and their expertise covered sub-domains including health informatics design, end-user data evaluation, health data interaction, passive sensing, and data privacy. Their early feedback drove us to design an interface that was less "data-dense" and more human-centered. For instance, early versions showed discrepancies through an exclamation symbol, but experts steered us to a less alarming notation. Through their input, we not only made DYMOND more user-friendly as a probe but also established information guardrails to ensure we consider ethical needs. We included tooltips in the interface to provide deeper information on features. We modified our GPT-4o prompt to be easily comprehensible and concise without being patronizing. Experts also urged us to represent the system as a memorable entity to remind users that it is a "co-performing agent". Akin to Yoo et al.'s pie-like mascot [113] and Kim et al.'s color-changing bubble [56], throughout the design, DYMOND was represented by a vibrant heart-shaped diamond icon. Importantly, we embedded a "resources" page with information for emergency mental health facilities and provisions that were readily accessible through DYMOND. The overall design was finalized after a two-week internal pilot study.

3.4 Study Design and Participants

Consistent with technology probe methodology [16, 30, 42, 51, 57, 67, 86], this study does not employ a control condition — formative probes are designed to generate design knowledge, not to evaluate comparative efficacy. DPMH studies can vary in length and can span anything between a few weeks [40, 56, 71], to several months [13, 26, 84, 105], to more recently even years [76, 108]. Most of these studies are completely passive. By contrast, we

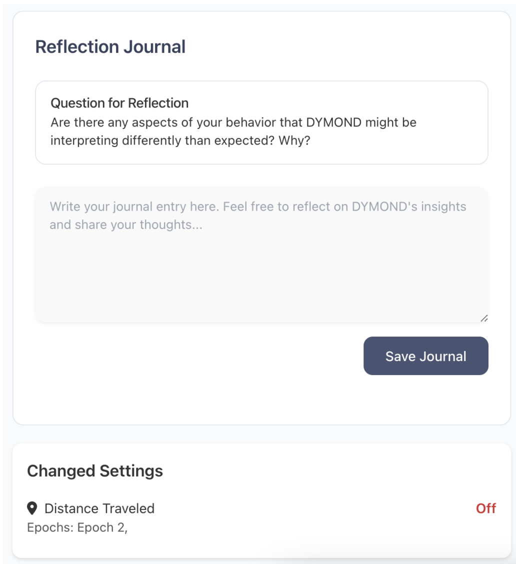


Fig. 2. DYMOND purposefully invokes the user to frequently note and express where they disagree with the model or how the model might be incomplete.

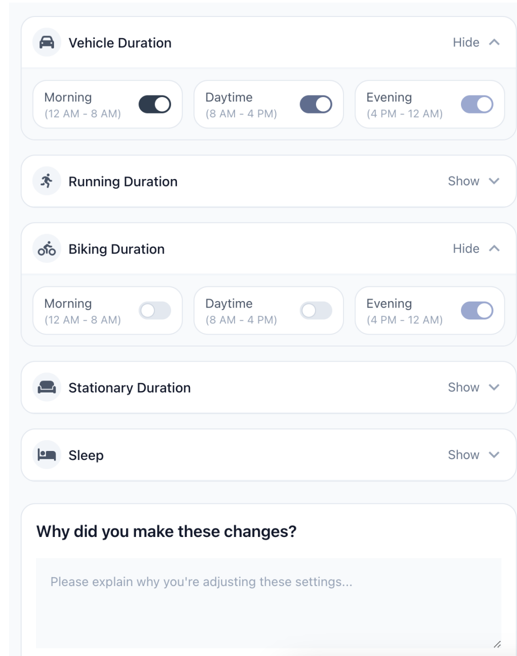


Fig. 3. DYMOND provides affordances for users to personalize the model by toggling certain features.

expected that DYMOND required a lot of active engagement from the users to elicit rich formative insights. This form of data-work would be difficult to sustain over a long period of time, which is why we designed a 6-week study for our probe. This section elaborates on the study protocol, recruitment, and analysis methods.

3.4.1 Study Protocol. The study was divided into three two week *periods*. Three periods allowed personalization of passive data within each [56], while enabling us to observe whether users maintain or iterate configurations across check-ins. Fig. 4 outlines the participant expectations through different aspects of our study: **(a) Full study (passive):** Continued their daily routines as DYMOND recorded their activity (§ 3.1). **(b) End-of-day (active):** Completed an Ecological Momentary Assessment via the phone app to report their depression (PHQ-8 [75]) and affect (PANAS [99]). Each daily item of PHQ-8 was modified to begin with "In the last day..." to capture the inter-day variability in symptoms as has been shown in prior work [75]. Similarly, the PANAS was administered with instructions to reflect on "today". **(c) Morning (active):** Participants reviewed DYMOND's estimate (for past days), their behavior patterns, and journaled their reflections on how DYMOND correctly (or incorrectly) inferred their depression level. **(d) End-of-period (collaborative):** Interviewed with researchers through a check-in at the end of two weeks to reflect on their journals and configured DYMOND for the next period.

3.4.2 Collaboratively Evaluating and Defining Seams. An emerging clinical practice in mental health care is collaborative documentation of visit notes [94]. This method involves both patients and clinicians in preparing the written artifacts, thus, promising increased transparency, improved patient trust, and better treatment adherence. We applied this same metaphor to conduct check-ins with participants so that DYMOND's future behavior was informed by users' own experience along with consultation with domain-experts (us, researchers). The check-in comprised two segments:

Understanding user needs and personalizing configurations: Using a semi-structured interview script, we triangulated insights from the DYMOND's dashboard, the user's journal notes, and the user's recalled experience. This sensemaking encouraged users to share additional points of friction—apart from model accuracy. At the end of this segment, we guided users to define configurations by toggling features.

Co-designing new affordances for configurations: Arguably, our probe only allowed rudimentary means to explicitly appropriate DYMOND. To illuminate new approaches, we conducted a co-design exercise. Each check-in focused on co-designing one of the three components of DYMOND (§3.3). We had prepared some initial *seed* designs—static prototypes showing alternative interfaces—to trigger brainstorming (Fig. 5)[3]. First, we presented participants with the concepts behind the seeds. Then, participants were tasked to come up with 2-3 designs through rapid prototyping. They were encouraged to use any material or method available to them (written notes, doodles, or digital annotations). Researchers did not interfere at this stage. Finally, they presented their designs back to us, and we discussed modifications and combinations of these designs. Over the three check-ins, we were able to ensure each participant had an opportunity to redesign each component.

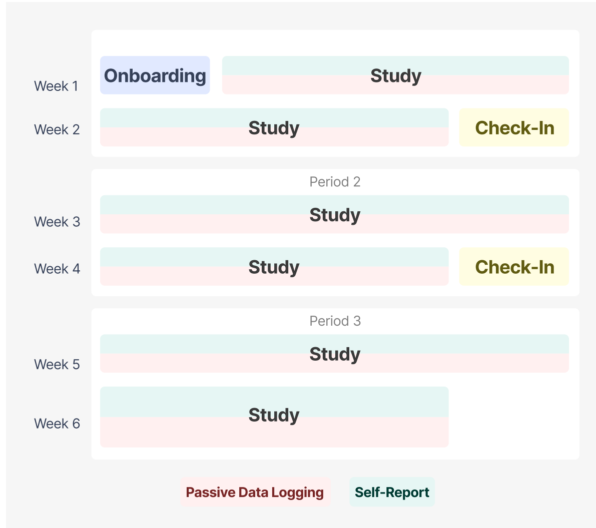
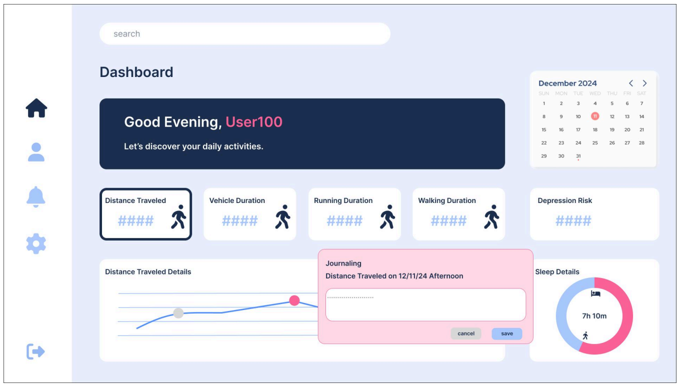


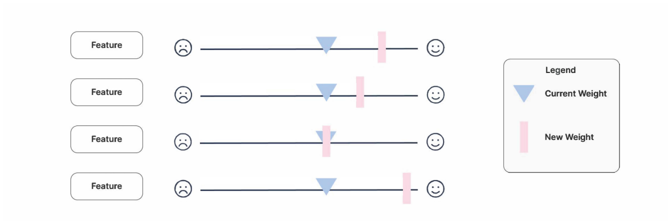
Fig. 4. Participant commitment to the study involved passive data logging and active reflection.

3.4.3 Recruitment and Participants. We recruited individuals in the U.S. through Meta ads, between October 2024 and April 2025. The initial screening had two major criteria: (i) participants were required to possess an iPhone and have access to a laptop or desktop; (ii) participants must have moderate–severe depressive symptoms over the two weeks preceding the enrollment ($\geq$ 10 on PHQ-9 [61]). We excluded participants at risk of suicidality. Those who scored $\geq$ 0 for item-9 (about self-harm) of the PHQ-9 were directed to complete the *Columbia-Suicide Severity Rating Scale Screener* (C-SSRS [80]). They were excluded if they scored moderate or high risk based on the C-SSRS[4]. A phone interview verified demographic eligibility and fluency in English, during which we administered the MINI [90] to confirm a current depression episode. The final screener was an address verification via post. The 22 participants that successfully onboarded are described in Table 2, along with a full demographic overview, including age, gender, race, and PHQ-9 scores at enrollment. Fig. 7 shows how their daily scores varied across the spectrum. Admittedly, the participant pool was dominated by the female gender (81.8%), but these proportions correspond to other studies on depression monitoring [75]. Participants were compensated through a flexible and progressive incentive structure where they could earn up to $150 as a gift card. The sensing platform was instrumented both in-person and remotely. During onboarding, participants were informed of the data and corresponding features for depression monitoring. Their task was to scrutinize how the model represented their depression. All check-in sessions were conducted and recorded over Zoom. The entire study was approved by the IRB of the first author's institute. Our safety protocol monitored survey responses, journal entries, and interview behavior to off-board participants if their mental health deteriorated. We also took into account their behavior and responses during interviews. According to our protocol, if participants' data indicate extreme mental distress

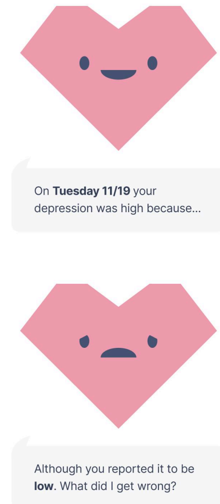
[3]One of the researchers has a background in clinical psychology and was involved in preparing the initial seeds
[4]If excluded, we did not provide participants a precise reason, and instead shared several mental health resources



(a) Visualization shows customized summaries and allows users to reflect deeper high precision on specific times in a day.



(b) Configuration allows users to tune features as positive or negative contributors to their mental-health outcome.



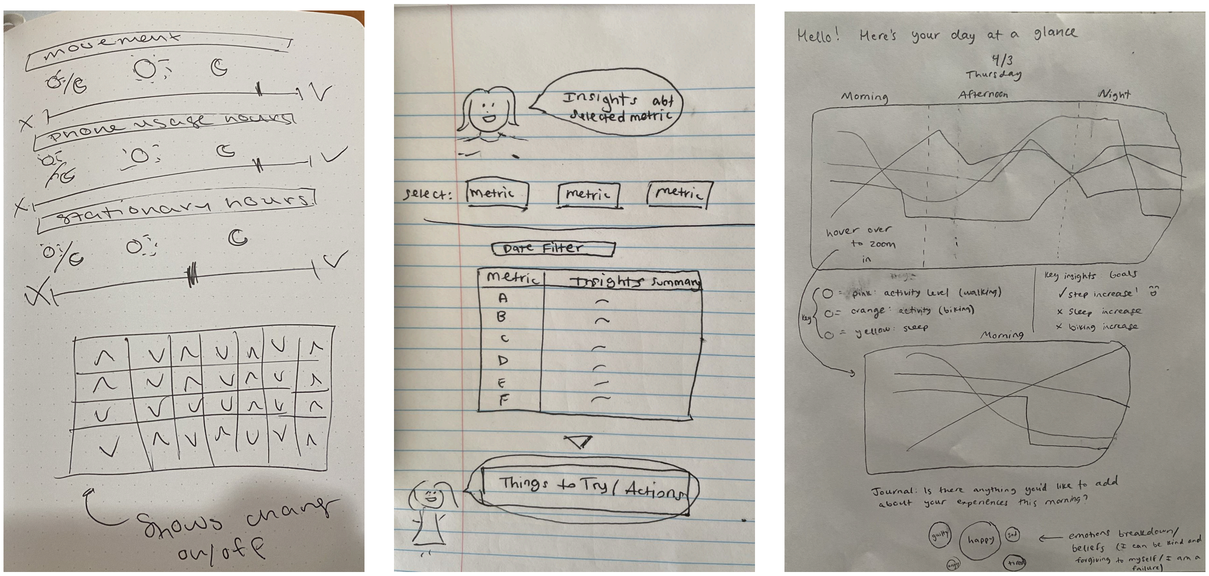
(c) Insights are described through an anthropomorphic agent that accepts natural language queries..

Fig. 5. Three of ten lo-fi seed prototypes we used to initiate the co-design activity with participants. Other seed prototypes can be found in our supplemental files.

(e.g., expressions of self-harm), we would first contact them and confirm their ability to participate further. If severe, we would direct them to local facilities or emergency services.

3.5 Thematic Analysis

The participants successfully completed a total of 58 check-in sessions. One recording was corrupted, leading to 57 transcript sessions available to inspect. We followed Braun and Clarke to perform inductive coding and dissect user experiences over the course of the study [8]. Three authors independently and iteratively coded the same 10 transcripts through open-coding. Each transcript was sourced from a different participant, and the transcripts were evenly distributed across the three check-ins. During this initial batch, the coders regularly synchronized to converge their codes and agree on new codes. Using this initial codebook, the first author coded a majority of transcripts (51%)—which were verified by other coders, and the second and third authors coded the remaining 13 transcripts—which were verified by the first author. Through this method, the first author studied every transcript, while every user's experience was vetted by at least two authors. The coding process involved careful inspection of transcripts, notes logged in DYMOND, and co-design artifacts (Fig. 6). While the transcripts reflected participant perspectives, experience, and opinion, we contextualized every account with empirical data. Therefore, the interview responses were anchored around the actual behaviors sensed and the depression estimates of DYMOND, thus providing naturalistic grounding to subjective evaluations. Moreover, the journal entries provided insight into participants' impression of DPMH in-the-moment and independent of



(a) User008 Configuration Design (b) User009 Insights Design (c) User012 Visualization Design

Fig. 6. Design exercise results showing user-created sketches for different interface components: configuration settings, insights presentation, and data visualization. Additional examples have been included in the supplementary materials.

the interview context. Additionally, these data were then aligned with the participants' co-design artifacts. The open-coding resulted in 180 codes, which we affinity mapped using Miro. Apart from merging similar codes, we also filtered out codes that fell outside the scope of our research questions. To answer our RQs, we oriented the remaining 147 codes into interrelated high-level themes. First, to address RQI, we organized themes describing friction or variation from expectation across different points in the DPMH flow, e.g., *data seams*. Then, to address RQII, we organized themes related to actual user engagement with the identified seams and attempts to modify DYMOND, e.g., *barriers to evaluation*. Finally, with these clusters in place, we identified our themes (primarily from our co-design sessions) to define *design requirements* for future DPMH systems.

Reflexive Considerations: The values and situated identities of the authors shape this study [5]. The lead and senior authors have substantial experience in advancing DPMH research. One author has the necessary clinical background to construct this study with appropriate safeguards. We acted as "critical insiders" [14, 25] who are in the unique position to holistically assess the technical and human-centered aspects of this study.

4 RQI: IDENTIFYING SEAMS IN DPMH: POINTS OF VARIATION, AMBIGUITY, AND UNCERTAINTY

Seamfulness aims to make seams "explicit resources of interaction", such that users are aware of the unevenness and can, therefore, adjust aspects of the digital context (e.g., personalization of the DPMH tool) and their non-digital context (e.g., interpretation of the DPMH estimates) [12]. The DYMOND probe did this in a limited way (revealing inaccuracy and allowing adjustment of features), but it elicited a richer space of seams across the DPMH pipeline. We organized our findings around three phases of the DPMH pipeline: input phase (data for depression), processing phase (modeling an estimate), and output phase (outcome of estimate).

Table 2. Participant PHQ-9 scores on enrollment, and the data streams they toggled during check-in sessions.
NB/TG: Non-binary/Third Gender, AA: African American, DND: Did not disclose
Epochs → ● Late Night ● Daytime ● Evening. • Superscripts indicate session number where the feature was toggled.

ID	Gender	Age	Race	PHQ-9	Configurations
P01	Female	20	White	20	stationary_duration ●^{1} ●^{1} ●^{1}
P02	Female	51	Asian	23	sleep ^{2}, biking_duration ●^{3} ●^{3} ●^{3}
P03	Female	24	Asian	21	running_duration ●^{1} ●^{1}
P04	Female	21	White, Asian	10	
P05	Female	21	Black or AA	17	biking_duration ●^{1} ●^{1} ●^{1}
P06	Male	25	Asian	12	
P08	Female	23	White	18	stationary_duration ●^{2} ●^{1}, unlock_count ●^{3} ●^{3} ●^{3}
P09	Female	32	White	14	sleep ^{2}, running_duration ●^{1} ●^{1} ●^{1}, stationary_duration ●^{1} ●^{1} ●^{1}
P10	NB/TG	23	White	11	
P11	NB/TG	48	White	16	
P12	Female	20	White	12	
P13	Female	27	Asian	10	
P14	Female	22	Asian	12	unlock_count ●^{2} ●^{2} ●^{2}, unlock_duration ●^{2}
P15	Female	31	White	23	distance_traveled ●^{1} ●^{1}
P16	Female	26	White	11	
P18	Female	43	White	11	
P19	Female	26	Black or AA	18	running_duration ●^{1} ●^{1}
P20	Female	25	White	11	stationary_duration ●^{1}
P21	Female	24	DND	19	vehicle_duration ●^{1}, running_duration ●^{2}, biking_duration ●^{1} ●^{1} ●^{1}, unlock_duration ●^{1}, stationary_duration ●^{2}
P22	Female	24	White	22	
P24	Female	25	White	10	
P26	NB/TG	31	White	11	

4.1 Data for Depression

First, we describe seams in the input phase.

> "*[DYMOND] can say that I'm not depressed because I'm moving around a lot in the afternoon. But that's because I went to the grocery store and the laundromat, and I wasn't having pleasure or interest in doing those things. It's just I have to do these things to, you know, function in my life.*" — P08

Much like P08, P03 mentioned that her running may not be positive. Or, P22, who used her phone a lot to make work calls, not because she was depressed. This section elaborates on such imperfect data representations of users, which we refer to as *data seams*.

4.1.1 Missing Personal Context. The first check-in usually revealed that some anticipated behaviors simply did not occur for a user. *"I don't know how to bike"*, said P14 when trying to discern why DYMOND considered it important. She was unclear if no biking would make it appear she was not exercising enough. P22 and P24 also expressed this issue with biking. Strictly speaking, DYMOND's data would show that these users biked zero hours a day. In most models, exercising reduces the likelihood of depression [105, 107], so lack thereof could lead to false positives. Meanwhile, others found vehicle duration to be irrelevant to them as they used public transit (P05, P16, and P18). Another cause of missingness was problems in the sensing instrumentation. Consider P13, who could not use the Garmin because the strap caused her skin to react. In turn, DYMOND read her sleep as zero. Similar instrumentation issues restricted P11 and P12 from sharing wearable data. Unlike behavioral irrelevance, instrumentation failures produced real blindspots in capturing user activity. Taken together, our seam highlights that it is non-trivial to distinguish a zero reading from a *null* value.

4.1.2 Misaligned Personal Context. On viewing their data, many raised an issue regarding the fidelity of DYMOND's contextual data streams. Participants described the data as *"basic"* (P09), *"superficial"* (P02), and *"not hundred percent realistic"* (P24). For some, this was a resolution issue. P15 wanted DYMOND to qualify her sleep, *"How I feel is so closely tied to whether or not I've got a good night's sleep."* P02 wanted DYMOND to distinguish restful sleep from *"nightmares"*. This lack of resolution extended to other data. P14 was not sure if DYMOND captured all physical activities, and P15 knew that DYMOND was not considering *"deep breathing"* as an exercise. This seam shows that feature data on the surface blurs the nuanced constituents of a behavior.

> "*You could hypothetically be running for a long time and still feeling really bad that day, or kind of the opposite, you get a really good day where you're watching movies or doing something pretty stationary... So maybe that's just hard to capture some of that nuance based on what is the nature of the experience*" — P12

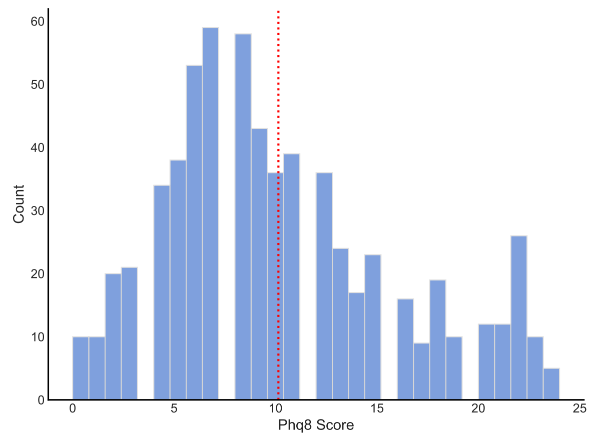


Fig. 7. Distribution of daily PHQ-8 scores self-reported by participants across the six-week study (N=22). Each bar represents the count of daily reports falling within that score range, showing that participants' depression levels spanned the full spectrum rather than clustering at a single severity level.

A key concern of people like P12 was that the data did not embody the role of the activity. Users actually considered certain activities healthy, but DYMOND related these to depression. Most studies indicate that phone usage late at night indicates a higher likelihood of depression [105, 107]. However, P01 relied on her phone to fall asleep on time and, in turn, feel better. Alternatively, P09's lack of phone use was mischaracterized as her feeling fine. The burden of being a primary caregiver did not afford time to leisurely use a phone. *"It's telling me, 'you're not in your phone that's great!', but I'm doing something that's kind of worse."* Another mischaracterized behavior was stationary time. P10 and P11 experienced some chronic pain that made physical movement detrimental to their mental state. Instead, P11 described staying in place as *"restorative"*. Therefore, this seam further reveals the lack of considering mitigating behaviors. As such, this seam captures two compounding failures: lack of nuance within a behavior, and normative assumptions that do not hold across users' lived contexts.

4.1.3 Constrained Personal Context. Across users, we found that activity does not imply agency and, therefore, may not be meaningful to determine mental states. *"There are other people in my life making me do things [...] forcing me to not act on things that I feel"* said P13. We learned that constraints arose from various sources. This seam illuminates the constraints on one's behavior by noting additional observable or reportable phenomena.

> "*Speaking as someone who works rotating day and night shifts, maybe sometimes I'm working night shifts. But then I'm awake during the day on my days off.*" — P04

A piece of information that DYMOND lacked was a user's routine. P04 felt her unique work schedule made her activity rhythms an outlier. P01 noted that on days with schoolwork, she might appear depressed. And P19

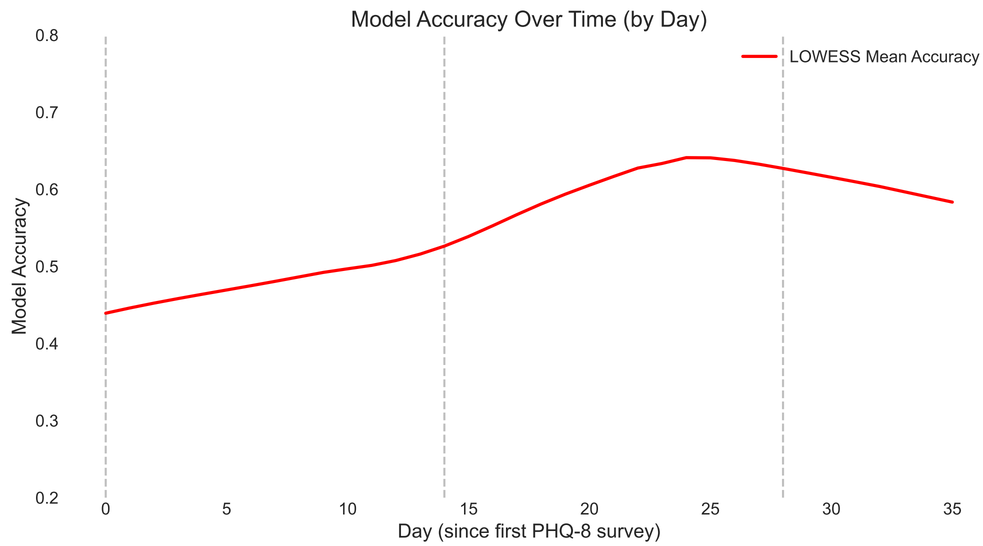


Fig. 8. DYMOND's mean prediction accuracy across all participants increased over the study. Grey vertical lines mark check-ins; the model was personalized after day 14 of data collection. Accuracy reflects the proportion of days where DYMOND's binary depression estimate matched the participant's self-reported PHQ-8 outcome.

had low phone use on days she worked multiple jobs. In fact, work- and school-related influences on behavior were echoed across multiple participants (P11, P12, P15, P16, and P22). This seam underscores real ecological dynamics that influence the sensing data more directly than a user's mental state.

> "*Homeland security showed up on campus, like the watch is not going to be able to track that. But that's something that was a very stressful experience that affected my mental health that day*" — P24

Unlike the routine factors, unforeseen external events may also influence user behavior. P24's experience on a university campus was an extreme case of this. Other externalities included a bad test score (P14) and breaking a mobile (P12). On deeper inspection, we learned that a core unforeseen factor mediating mental health was social events. This could be as innocuous as *"someone saying something in a certain way"* (P01), as indirect as *"my partner having a bad day"* (P16), or as bleak as close relations passing away (P09, P15). Thus, this seam also explains short-term changes in user behavior due to anomalous events in a user's life. Therefore, constraints could emerge from persistent routine factors that systematically shape behavior, as well as unpredictable anomalous external events that can only be reported retroactively.

4.2 Unpacking the Depression Model

Digital phenotyping models tend to be black-boxes for end-users. Sometimes, this might be desirable to achieve seamless systems that "just work". We purposefully designed DYMOND to challenge this assumption by presenting users with more transparency into the model (§3.3).

> " *I've not felt good this entire weekend. And that is what the data is showing. So I think over time it has gone more accurate.*" — P03

Echoing P03, Fig. 8 shows that accuracy improved after the first check-in. This was around the time that DYMOND's model would personalize to the user's data, and the first set of configurations would be set in. Yet, DYMOND also often made mistakes. We learned that users scrutinized the algorithmic estimation, regardless of accuracy, and found ways to use it despite its precision. These hurdles are described as *modeling seams*.

4.2.1 Accuracy Perceived as Trivial. Even the correctness of DYMOND involved friction as users expressed reservations on reliability (P03, P05, P10, P12, P14, P16, P18, P20, P21, P22, and P26). Naturally, some felt *"validated"* (P16) or *"reaffirmed"* (P20) when the estimates were correct. Others, however, remained critical about the accuracy. *"I was feeling kind of fine overall, so the model just happened to be correct"*, said P13, who had the notion that DYMOND was only accurate when she was not depressed. P06 felt that on the days DYMOND correctly estimated high-depression, *"the symptoms are evident enough that even without looking at the model I would have known"*. This seam highlights that despite correctness, users might remain unconvinced of the utility of a DPMH.

4.2.2 Inaccuracy Perceived as Inevitable. A number of participants mentioned DYMOND's inaccuracy in at least one check-in (P01, P04, P06, P08, P09, P10, P11, P13, P15, P18, P22, and P24). Consequently, they wondered if a nebulous concept like depression can ever be objectively estimated.

> "*I'm super depressed. I never become hostile, or I never move slowly, so people can't notice. Some of those questions just don't apply to how my depression comes out.*" — P24

Users like P24 realized that the ambiguity in the model emerged from methodological challenges in estimating subjective constructs like depression. P18 thought that the ground-truth data (self-reported PHQ-8) led to a limited portrayal of her mental state. She contrasted DYMOND as a tool that observes the full day, but the ground-truth is attenuated by the time of day she completed the questionnaire. Users' preconceived notions also mediated their trust in the model. P22 was concerned about any model that could inform health decisions. And P09 got suspicious of DYMOND's eventual accuracy because it was initially inaccurate. In fact, users like P09, P04, and P06 could not fathom how their interaction, without the guidance from a clinician, could lead to improvements. This seam highlights that users' attitudes towards depression modeling can make them hesitant to engage with better DPMH models as well.

4.2.3 Ambiguity Perceived as Useful. Interestingly, users found utility in the model beyond its accuracy.

> "*I would say (DYMOND) has helped me to remember what could have happened that day and what could have contributed to me feeling lower, and with that I'm able to kind of explain better to a counselor or somebody I trust if I'm able to reach out to them. I can better explain in detail why I'm feeling this way.*" — P01

P01's quote indicates that engaging with the model can equip users with important vocabulary to describe their depression. DYMOND reminded P12 that her mental state was linked to her physical behaviors, giving her hope that she can change in the future. DYMOND's model was *"a piece of the story and not the entire story"* (P24) to understand their depression. Any estimate provided by the model, accurate or not, was only viewed as partial information. For P14 and P16, on days DYMOND classified their depression risk to be high, when they reported otherwise, they could still justify how the classification explained some aspects of their day. The seam demonstrates that uncertainty can serve a reflective purpose, which is not accounted for when DPMH systems are over-corrected for diagnostic verdict.

4.3 Outcomes of the Depression Classification

For users to continue providing data to DPMH, they need to perceive value in the system. However, typical systems only provide a classification, which can be reductive. The *outcome seams* present a multitude of ways in which users were in tension with the value of their mental health classification.

4.3.1 *Lack of Plurality.*

Much like prior DPMH studies [108], DYMOND leveraged the *Patient Health Questionnaire* (PHQ) as ground truth to make a binary classifier that indicates depression risk as high or low. Our participants were familiar with the PHQ and what a high score would mean (≥ 10). Yet, they found it hard to calibrate the severity of DYMOND's estimate in relation to their feelings. P03 wanted to know if she was doing better than the previous week, whereas P19 wanted to see how a *"healthy individual"* behaves. They needed a personalized baseline to discern how to act on the classification result.

> "*I don't think—I don't really feel—like I've been depressed. I've been dealing with a lot, but not necessarily depressed... Maybe some of my feelings were interpreted wrongly like that.*" — P18

Users speculated that DYMOND's classification might indicate other mental health symptoms such as stress (P02) and anxiety (P06, P13, and P18). According to P10, the output was optimized to a flawed representation of their depression (via PHQ-8), *"I wouldn't necessarily say that the self-report itself is accurate because of the questions, about being tired, having low energy, and I have a few chronic illnesses; while sometimes that's related to my depression level, it's not always."* Even if DYMOND was not estimating depression, it needed to be clear on what it was representing. During the study, P12 was in a psychology course, and thought DYMOND should be more open in describing negative mental states, *"like affect and intensity."* This seam shows that a single binary output obscures both the severity of depression and the plurality of mental states users actually experience.

4.3.2 *Lack of Actionability.*

As a probe, the explanations were designed to support human-in-the-loop inspection of the algorithm, but users wanted insight into their health.

> "*Powerless would be a good word. I'm looking at my review for yesterday and all of its phone usage. Realistically, that's not gonna go down. It would be great if I could. But I don't think that's gonna happen right now. So to just be consistently told that my big issue is phone usage phone usage phone usage isn't really helpful.*" — P15

On one hand, P15 received a consistent explanation, which would help her systematically evaluate the problem with her phone use data. On the other hand, the explanation provided limited ways to improve her own behaviors. P06 articulated this as a dichotomy in expectations, *"[DYMOND] has an entirely different mental model [that is] trying to surface symptoms instead of trying to focus change in behavior"*. Participants wanted the explanations to not only help them inspect the model, but also make use of the classification. They wanted to learn behaviors they can *"adapt"* (P10) and *"control"* (P11, P14). Examples of such nudges for them were, *"try putting your phone in a different room when you're doing something"* (P13). Moreover, explanations were isolated to interpreting a single day. By contrast, P08 and P16 pointed out that they cannot evaluate an estimation for a day in isolation and instead wanted to see a summary of patterns across days. This seam illustrates that DPMH outputs fall short when users cannot associate the estimates with their symptoms or with methods to mitigate them.

4.3.3 *Lack of Protection.*

Arguably, inaccurate outputs were expected to inconvenience the user. Findings showed these inconveniences manifested as harms that disrupted user mental state. Incorrect labels caused a *"dissonance"* (P21) in the minds of users. At worst, it could even *"gaslight"* the user, as P13 put it, *"it's saying that I wasn't depressed, but I was, and so it might make me feel like I was not actually depressed... which I think is actively bad"*. Alternate forms of harm included DYMOND causing users to obsess over their data. P06 stated this risk both in personal use and in the presence of a clinical stakeholder and did not want this data to be the *"centerpiece"* of his treatment session. This seam highlights that without safeguards, DPMH outputs can actively harm users — reinforcing self-doubt or becoming an unhealthy focal point in their care.

5 RQII: NAVIGATING SEAMS IN DPMH: INTENTIONAL APPROPRIATION

Encountering seams presents an opportunity to adjust usage of a tool, because now the scope of the system is more apparent. To understand how users wanted to adjust DYMOND, we describe our findings via Yang and Newman's framing [109]: users can appropriate technology through *exception flagging*—the user identifies the

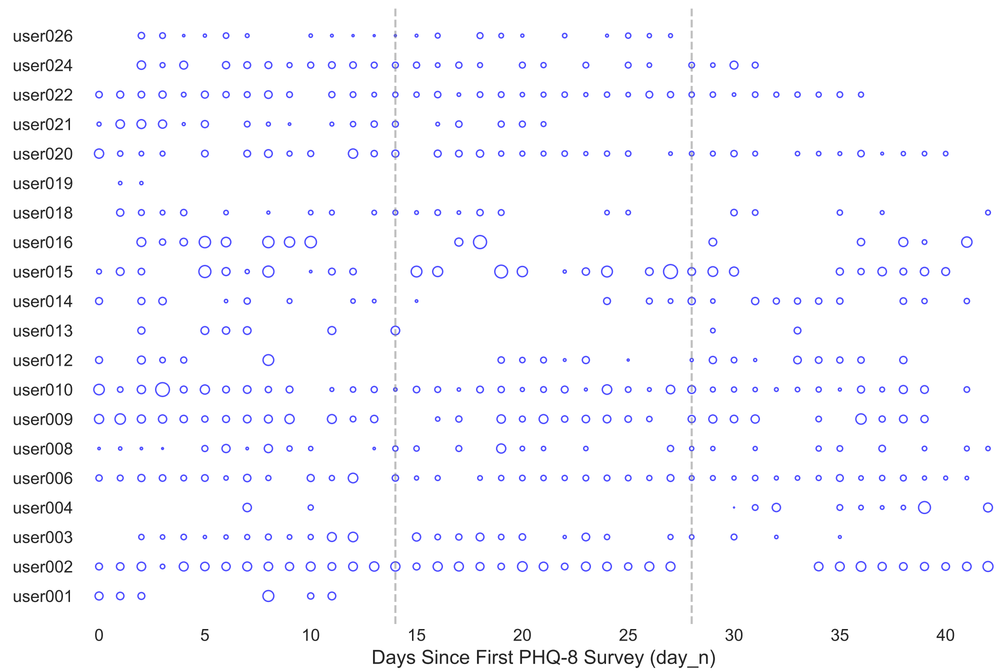


Fig. 9. Timeline of journal entries logged by participants in DYMOND over the six-week study. Each row indicates a participant and each point represents a single entry; point size reflects entry length. The distribution shows that participants engaged in reflection throughout the study, with entries concentrated around check-in periods (weeks 2, 4, and 6).

problems—and through *constrained engagement*—the user occasionally modifies the system. Our probe provided some affordances to navigate these seams. We learned why users wanted to appropriate DPMH tools and how the navigation itself was a meaningful act. However, the probe also revealed the need to reduce excessive data-work from the user. This section characterizes these motivations and challenges. Then we draw from the co-design exercise to address these issues through *design requirements* for seamful DPMH.

5.1 Explanations: Encouraging Users to Inspect DYMOND

To facilitate this *exception flagging*, DYMOND visualized the user's behaviors for each day and also stated a rationale behind its depression classification. Users logged their assessment on the system every day (Fig. 9) and then reflected on the system during the check-ins. Here, we expand on how users found value in inspecting the system, its challenges, and designs to improve the evaluation process.

5.1.1 Evaluating can be Reflective. Many users found the act of evaluating DYMOND as a positive exercise. Users drew comparisons with the act of *"journaling"* (P03, P04), which was familiar to them. The primary purpose of regularly reviewing was to help users make sense of their data (P02, P16). We also found some secondary uses, such as augmenting their memory of behaviors (P01, P10). Another such use was evaluating the model to glean how others might see them. For P24 it was to stay *"aware"*, whereas for P21 it helped them *"admit"* how they felt.

> "*I think reading and breaking it down into little chunks that explain how things are related makes me realize that it's not as ambiguous and giant.*" — P21

P21's experience reminded us that users were not domain experts, and understanding their mental state was overwhelming without the right tools. DYMOND's interface enabled them to understand the model better. By the third check-in, P06 gathered that the system could *"definitely predict depression that manifests in a very particular form"*. Consequently, inspecting the model made users feel like active participants. For P09 and P15, it was a way to stay assured about how their data is represented in models. Meanwhile, others like P05 found an opportunity to *"experiment"*. In the same vein, P14 and P18 found themselves being more mindful and planning future behaviors. Importantly, users felt positive when DYMOND's explanations supported their hypotheses of mental health. P04 noted, *"If the data were to agree with the same conclusions I was coming to, it would make me really feel validated in my emotions."* This was a catalyst for users to better articulate their behaviors in the context of depression (P01) and to help *"zoom in"* on specific aspects the system did not surface (P02). P22 actually discussed her learnings in one of her therapy sessions during the study. While DYMOND's interface was designed to help users scrutinize how algorithms represent their data, we found their primary motivation to assess the model was derived from their need to be more mindful of their mental state.

5.1.2 Evaluating can be Burdensome. Inspecting DYMOND was a challenging task since none of our users had experience with DPMH or data modeling. Given the elevated risk of depression of our users (§3.4.3), the emotional demands of their mental state often impaired their ability to critically think.

> "*Some people are more visual versus reading. It's just very overwhelming. I have to look at really things very carefully... Everything is too much exposure, especially if you have this diagnosis. That even gets you more frustrated, and [then you] don't want to do this.*" — P02

P09 echoed P02's frustration and felt debilitated to argue with DYMOND's explanation. P06 remarked that correlating behaviors to depression took considerable cognitive effort. And P14 found the information to be too simplistic and demanding of interpretation. Several others (P02, P03, P15, and P16) felt that the utility of the explanations was redundant over time. Lastly, even if users were able to digest the information on the dashboard, it was difficult to describe and log issues. P13 stated she needed *"activation energy"* to record where DYMOND was going wrong. And P11 admitted that she *"did not know what to fill out"*. Therefore, drawing critical insight (to suggest changes) is not merely about dashboard design, but compounded by users' mental states.

5.1.3 Design Requirements. Based on the challenges described above, we arrived at the following requirements. These surface the underlying assumptions in DPMH in a manner that users do not passively accept the system, but instead actively inspect and reflect on it.

DR1: Communicate with Compassion: *"Depression and monitoring data sound very scary"*, said P03. The interface needs to provide a warm and inviting space for users to reflect. Users suggested improvements in form, function, and content. After redesigning DYMOND's visual aesthetic, P01 remarked, *"I feel like it's more attentive to my own life, I feel more cared for..."* In terms of function, users imagined DYMOND would appear more caring if embodied by a character or an interactive *"helper"* (P21). *"I liked the idea of interacting with [...] an animal that you get to name, or a plant that you get to name, [and] you form a connection"*, said P24. Utility-wise, it would provide a conversational interface to inspect data (P13, P15, P16). P05 felt such an interface could narratively highlight her highs and lows, *"most active moments of the week or the best days of the year in a story format [which] is more social and [lets you] bring your information to your clinician or your provider"*. Intimidating interfaces present implicit barriers for user evaluation and can deter well-meaning users from purposeful engagement. These design changes stand to mitigate the affective harms of reflecting on one's own behaviors.

DR2: Enable Drilling Down: The interface needs to allow deeper and flexible data exploration. P24 wanted to look deeper than the minimalist visualization, *"I think that it is helpful to break down the data in more than* *just those 3 data points"*. Users also wanted to configure the visualization without interfering with the model. P10 elaborated this in their design, *"you're only looking at a few of the types of data but they are still all being recorded, so you're not actually telling the model to only look at those ones."* Besides selecting data, participants also expressed the need to cluster and synthesize data. P20 wanted to set up multi-modal groups of data, such as combining time of day, sedentary behavior, and phone usage together to distinguish checking her phone from mindlessly scrolling. P26 wanted to cluster data across time to find their own *"fluctuations and patterns"*. Allowing these types of personalization could be key to reducing the cognitive load of interpreting the data.

DR3: Implement a Clinician-Facing Counterpart: To ensure trust and shared meaning, participants suggested companion interfaces for clinical partners to provide expert insight. P13 thought this was essential to maintain transparency between actions they take between check-ins. Alternatively, P09 said a collaborative interface would give her objectivity. She wanted *"the ability to see what data my doctor's looking at [because] maybe she doesn't care about me running, so I know that that's not something I have to stress about."* At the same time, P09 imagined objectivity as bidirectional, *"I think her seeing my comments is great because I would want her to know why [the data] is that way, from my point of view."* Moreover, P16, P21, and P22 found the linking would make them more responsible to test the DPMH themselves between visits. P21 said, *"if this wasn't available, and I had to wait until an appointment [to share data], it would be slowing the progress and not be an opportunity to take on initiative and change [the system]"*. Thus, a parallel interface can provide the psychological safety to interpret data with protection from bias, while keeping users accountable to evaluate and configure the DPMH tool.

5.2 Configurations: Assisting Users to Modify DYMOND

The configurations in DYMOND were key to learning how users experience *constrained engagement* with DPMH. P02 said, *"having a control over something, so understanding it more [...] and [thus] taking charge of it."* The configurations provided users some sense of control through which they could *"advocate"* (P15) independent of how the model or other stakeholders saw them. Here, we expand on what drove users to configure the system, its challenges, and designs to improve this interaction.

5.2.1 Configurations can Promote Care. A majority of participants found the configurations to be useful for DYMOND's tracking. Interestingly, none of the users toggled data due to privacy reasons. P06 said he was already foregoing privacy broadly while enrolling in a system like DYMOND; P20 said she wanted to provide whatever helps their clinician. The key motivation was to improve the insights they were receiving on depression.

> "*I've noticed that since the [check-in] activity we did—turning off the epoch at particular durations of the day—it has been quite accurate. The model has been picking up things more accurately than before.*" — P03

P03, among others, described DYMOND as having improved after setting the configurations. Some other examples: P13 turned off sleep tracking because she was not using the wearable, P24 turned off phone-screen tracking because she was unclear if navigation apps would be counted against her, P16 turned off vehicle duration because she did not drive, and P01 turned off distance traveled because movement data was sufficient. Participants like P08, who turned off the biking data, found the estimates to be *"more correct"*. For P22, turning off the phone-usage data worked because *"that might be more reflective of how much free time I have, and more like ADHD... as opposed to feeling more depressed."* After configuring, users also described DYMOND to be more personalized (P03, P18, P19, P20). Toggling data streams was certainly a limited instrument for control, but users found both functional and perceptual benefit.

5.2.2 Configurations can be Burdensome. While the study enforced configurations under supervision, our probe revealed challenges in doing these by oneself and in collaboration. Table 2 shows that several users did not change the configurations. Generally, users do not expect to *fix* the tool they are using. *"I'm too used to the model of most tracking apps [that] are trying to invoke a change in your activity... and when I switch to this one, I have* *to remind [myself] that I'm here to observe [the tracking app]"*, said P06. Configuring DYMOND was something participants were not familiar with, nor did they have precedent for.

> "*I think just the risk of not understanding what it is that you're changing. I changed this thing on this day, and now it is feeling like it's not as accurate, and maybe you would change it back. And and it wouldn't really matter whether you understand. But that also does depend on you remembering or having a log of when you changed it, and and keeping track of that diligently*" — P10

We can see from P10's example that configurations were vulnerable to derailing the entire DPMH pipeline. Consequently, users asked who should be accountable. P06 argued it is not a user's job to improve monitoring, and P13 pointed out that clinical stakeholders have more expertise. Participants also admitted that learning to configure was cognitively taxing (P02, P09, P14). The wrong configurations had ramifications. For instance, P21 was concerned she might become less mindful of behaviors she stops tracking. And P03 could imagine a possibility where users purposely *"please"* DYMOND to track behaviors for accuracy, even though it does not reflect their lived experience. Users like P24 faced challenges deciding which features to toggle because they had an irregular schedule. As such, configurations risk burdening users with data-work that DPMH was designed to eliminate, which is why affordances for configurations need to be flexible and guided.

5.2.3 Design Requirements. The requirements below address the burden that emerges when a user is expected to configure DPMH systems. These aim to give users flexible scaffolds to act on what the seams reveal without demanding disproportionate effort.

DR4: Allow Augmenting of Context: The interface needs to provide a larger canvas than just toggles for new forms of data to be included or excluded. Users wanted to self-disclose key contextual events. P09's design provided *"the ability to add like any illness at that point: did you get the flu this day? Or if you're having a bad pain day or an injury, and that's why you didn't move"*. She imagined this interface would require *"a full calendar at the bottom, but with the ability to color code the days"*. Similarly, P13 wanted to self-track certain physical activities such as yoga, because *"when I skip my yoga class, I feel bad."* P03 wanted to use a similar design to distinguish behaviors during her regular school schedule, work shifts, and vacations. Context augmentation does not have to be retroactive. P16 envisioned hypothesizing and then correcting. She stated, *"add some of your journal before you see what's happening and then further explain, because of those days where I didn't realize that I was gonna be at a high level of depression"*. Enabling additional forms of context annotation helps situate the passive data in one's ecology and thus better account for it. This approach ensures that users are not limited by data captured by the system and provides additional decision latitude in expressing what DPMH needs to change.

DR5: Provide Granular Configurations: The interface needs to facilitate a variety of operations on data beyond binary toggles. For instance, P14 did not want DYMOND factoring in *"anything [and everything] that makes her heart rate rise."* During the co-design, many participants gravitated towards adjusting the weights of the data. For instance, P13 wanted to adjust the weight of stationary time, such that at night, longer periods would imply she is more likely to be depressed. In her own words, *"if in the evening I'm stationary, maybe that's bad, because I'm supposed to be working out in the evening."* Another operation that users wanted was to define rules. In P08's case, she wanted a different set of configurations for weekdays and weekends to align with her work shift. *"[With] DYMOND, [if] they're spending a lot of time on their phone, because I'm working, so on the weekdays you would put less weight on [screen usage]"*, said P08. More sophisticated configurations allow users to dynamically adjust the system to their lives. As a result, users can perform light-weight alterations at specific blind spots without the risk of completely overhauling the model.

DR6: Guide User Configurations: The explanations should not just expose the model, but explicitly guide users towards future configurations. A method suggested by P24 was to show trends across days for specific behaviors to help justify their inclusion in the model. For instance, P10's pain restricts their movement, but instead of turning off physical activity, the trends could help them teach the model. *"I can be more active up to a certain* *point, then if I go past that point I get really exhausted and in lots of pain, and then I feel worse"* (P10). Importantly, users needed ways to judge if their configurations had an impact. Some wanted to do this retrospectively. P08 proposed, *"first start with the model gathering everything and then after a month evaluate what to turn on or off so you can get a whole picture [...] after the model gets this baseline."* This approach would extend to help users with sensemaking to disentangle the role of different features. Other participants wanted guardrails to configure in a more informed way. P06 combined aspects of multiple designs to propose, *"I can slide around and see the graph change in order to observe which parameters are the most important."* He wanted to see hypotheticals from his own tracking history to learn more about the features without risking experimentation for the future. Thus, supporting users to configure can give them more confidence to make changes. Providing users with insight to generate their own hypotheses enables a systematic approach to tinker with different versions of DPMH.

6 DISCUSSION

User perspectives of DPMH, often, call for bringing humans-in-the-loop (HitL) and giving them greater control [1, 25, 113]. However, few efforts have been made to do this. *Seamful* design is a promising approach to address these needs. Having said that, it remains unclear how seamfulness can be implemented for DPMH. We designed DYMOND as a probe to gather formative insight for making such tools seamful. §4 articulates these *seams* to identify aspects of the DPMH loop where humans should engage and exercise control. §5 describes methods to provide user control while mitigating user burden.

6.1 Theoretical Implications: Reconsidering Passivity of DPMH

We now reflect on what is "passive" in DPMH. Typically, it refers to the nature of sensing, processing, and delivering insights. However, by situating our findings in previous reflexive critiques [1, 25, 114], we explore how "passive" also describes other entities, processes, and expectations in the DPMH pipeline. These descriptions drive our recommendations to think of DPMH differently.

The underlying notion of passive sensing for mental wellbeing assumes users have complete agency over their behaviors—they are active while surrounding data remains passive. Many studies thus focus on mutable behaviors [26]. However, our study reveals that human behavior is often constrained by ecological changes (*data seams: constrained context*—§4.1.3). In attempting objectivity, passive data remains agnostic to what it cannot capture, reinforcing the fallacy of humans in the void [114]. We found that passiveness rarely disambiguates whether users refrained from acting or found action practically impossible (*data seams: missing context*, §4.1.1). Ignoring personal ecology creates incongruence between user experience and model inference—what Das Swain et al. term a semantic gap [23]. For instance, behavioral data may objectively track user situations (e.g., prolonged sedentary behavior) but cannot discern user appraisal (e.g., an enjoyable challenge versus a forced demanding task). This corresponds to *data seams: misaligned context* (§4.1.2). These seams collectively explain why passive sensing appears unreliable. Rectifying these gaps requires annotating data with user intent (**DR4** in §5.2.3). Users need to examine and label their data, to enrich DPMH.

Training DPMH with passive sensing assumes that better estimation with minimal interference implies satisfaction, ironically treating humans as passive data generators without near-term benefit. As Fitzpatrick et al. state, "system approaches that rely solely on passive monitoring or being surrogate data collectors for others' [...] rarely give the [users] the chance to interact with and adjust the system" [31]. Our findings present a way forward. Through the *modeling seams* (§4.2) we see how humans can actively audit the conclusions drawn by DPMH systems. Research shows fitness tracker users constantly struggle to define mental models for accuracy [111]. The *modeling seams: accuracy* (§4.2.1) and *modeling seams: inaccuracy* (§4.2.2) extend this challenge to mental state estimation. Mental health outcomes are complex, and when models overlook this complexity, users disengage. While these *modeling seams* create barriers for users seeking active roles, we discovered users appropriated other ways of engaging (*modeling seams: ambiguity*—§ 4.2.3). Users ultimately want to manage their health outcomes. Self-experimentation with self-tracking has been studied in the past [45, 52]. Similarly, **DR2** (§ 5.1.3) encourages passive sensing to support active hypothesizing and evaluation. Over-correcting for accuracy alone will not change user attitudes. Instead, we must recalibrate DPMH goals toward care delivery, enabling users to take stock of their DPMH as an important pathway.

The original inspiration behind DPMH was the quantified-self idea of "know thyself" [53]. Our investigation of *outcome seams* (§ 4.3) shows that users do not passively accept information. Bogdanova et al. recently assessed DPMH's failure to consider "the relational concepts of the lived experience of mental health" [7]. Users' depression exists in complex interplay with other mental wellbeing symptoms (*outcome seams: plurality*—§ 4.3.1). Our evidence suggests that users could not easily apply the algorithmic estimates to what they consider depression and constantly wondered if it reflects other aspects of their state (e.g., anxiety). The *outcome seams: actionability* showed us that the estimated levels of their depression neither nudged nor elucidated users on how to act, giving a sense that DPMH does not care about actual lives. Our co-design session produced a concrete path to that goal through **DR1** (§ 5.1.3)—embedding compassion through aesthetics and anthropomorphizing; and **DR5** (§ 5.2.3)—enabling users to granularly edit nuance into their data. Hence, DPMH systems must take stock of the users' lives beyond the inference it produces and support their care needs.

6.2 Technical Design Implications: Charting Interactive DPMH

Blending Seamless and Seamful Approaches. To counteract the seamless passive nature of DPMH, DYMOND was purposefully designed with intense user-engagement to surface the seams. A truly seamful implementation of DPMH cannot expect the same degree of data-work. Remember, users are neither data scientists nor mental health practitioners. Moreover, their severe mental state can make nuanced decisions challenging (§ 5.1.2). Cox et al. had proposed *mindful* interactions to improve the wellness outcomes of *mindless* interactions [22]. Their work demonstrates the role of *design friction* as a means to achieve mindfulness during use. However, they emphasize that this purposeful discomfort, although worthwhile, must not be cognitively burdening. In fact, studies have noted the risk of increased invisible labor at the seams [43, 92] to ensure systems remain configurable. Therefore, advancements in the passive aspects of DPMH continue to remain important, so that in most cases mental health monitoring can just work. Note, however, we learned in § 4.2 that model performance alone is not enough to convince user engagement and they can still find meaningful use with suboptimal models. For instance, DYMOND's affordances helped participants become more mindful of their mental state, feel validated in what they already suspected, and better articulate their experience, while also giving them a stake in how DPMH represented them. While it is unclear from our study whether seamfulness improves outcomes for users' mental health or the accuracy of tracking, the design promotes increased agency and control in self-tracking. Critically, seamful and seamless design are not competing paradigms. Inman and Ribes discussed the need for "beautiful seams" that have "coherent design emphasized by seamlessness" and the "appropriability of seamful design" [43]. From the perspective of digital tools, we can draw insight from the design of mixed-initiative interfaces [39] that dynamically adapt control between users and system. Alternatively, perhaps, we can draw from mental health therapy as a metaphor. Patients do not request a new dosage or modification to their therapy every day. The reflection on one's health is planned, occasional, and deliberate. Future DPMH systems could adopt a similar principle to reveal their seams through routine follow-ups that get sparser as the model stabilizes.

Practical Methods for DPMH Appropriation. End-user data evaluation demands that HitL interfaces balance usability with deep utility (§ 5.1.2 and § 5.2.2). This very concern has inspired research in *Interactive Machine Learning* to explore interfaces for human feedback to ML. The standard tools, however, only support interaction with structured data in confined settings (e.g., refining medical splices for computer vision [9]). Yet, as we eventually learned, users wanted to enrich DPMH with multimodal data, some of which is subjective

(DR4—§5.2.3). This gap could be addressed with LLMs. DYMOND leveraged GPT-4o for communicating its model heuristics (§3.3), but this was a one-way interaction. Recent work on sensemaking of personal health informatics presents LLMs as versatile means for open-ended querying of data and interpretation [18, 19]. Extending LLMs to support configurations is a natural step, given that user-defined rules are important for goal-directed health tracking [51, 86, 87]. By including an agentic layer, the system could nudge and contextualize information for users to annotate [65]. Moreover, generative AI could be leveraged to use visual metaphors to communicate mental health [81]. A natural language interface would assist non-experts in creating rules on how their data should be consumed by the DPMH model. It could also illuminate auxiliary context to a user's mental health outcomes (performing retrieval-augmented-generation [110] on a user's work schedule or the weather in their city). LLMs could mediate user engagement without the burden of complicated configuration consoles.

Expert Guardrails to Reduce Bias. A one-on-one interaction between users and DPMH, even if seamful, carries risks. Mental healthcare exists in collaborative settings, and stakeholders, like clinicians, should also be included in HitL setups. Acting alone, users can undermine their self-judgment and accept insights from a DPMH tool [38], which could lead to a distorted sense of self [36]. Our users described this dissonance to be stressful (§4.3.3) and might also have confirmation bias (§5.1) when the system validates them. Although this can "satisfy" a user, it leads to a worse approximation of their health tracking. We suggested mitigating these mismatched mental models through shared interfacing with experts (**DR3**—§5.1.3) and guided experimentation (**DR6**—§5.2.3). Seamfulness can offer a means for users to hypothesize and articulate their mental model to experts and triangulate better feedback. Health tracking research shows value in eliciting users to document unobservable issues as objects of shared meaning with experts [77, 85]. We envision that seamful DPMH can be collaboratively reviewed in the same cadence as natural interactions between stakeholders, such as scheduled clinical appointments. Mental health experts can help users troubleshoot, and future tools might require clinical review and approval before configurations are implemented.

6.3 Limitations and Future Work

Need for Experimental Evaluation: We designed DYMOND as a probe to uncover seamful opportunities for HitL versions of DPMH. Our study protocol (§3.4.1) was intentionally designed to expose friction and condition users to think of ways to navigate these breakpoints. Prior literature suggests that such design approaches are needed for richer modeling of mental health outcomes [37, 100, 101] and that additional agency improves compliance [31, 77] in health monitoring. Having said that, empirically discerning whether and the extent to which user engagement with future systems might vary was beyond the scope of formative inquiry through a technology probe [42]—as we focused intensely on a small sample over a relatively short period. Our study gathered some data on model performance (Fig. 8) and engagement (Fig. 9), but these outcomes and behaviors cannot be disentangled as causal. In the same vein, we found participants reported value in having the ability to interact with seams (§5.1.1 and §5.2.1), but additional research is needed to study sustained motivation and utility in behavior change [50]. We acknowledge this limitation in our method, but find our results contribute to such efforts in the future. Iterating designs of DPMH after deployment can be practically expensive and ethically concerning [24], and our study lays foundational groundwork to design with seamfulness, deploy it, and study user experiences.

Need for Continued Model Improvement: The motivation to develop DPMH was to relieve users from the burdens of excessive data-work of manual logging. DYMOND used a heuristic model to estimate depression. Selecting such a model offered pragmatic advantages (§3.2) for a formative study over contemporary models [108]. More advanced models could potentially reduce the incidence of errors that users encounter with DPMH. For instance, the missing data aspects can be mitigated with better imputation for behaviors [17]. Similarly, recent DPMH studies are beginning to address heterogeneity and inter-person variability of mental health with sophisticated machine-learning [69, 106]. However, real-world systems deployed in natural settings must always account for a degree of uncertainty and ambiguity that emerges from the unstructured nature of lived experiences [98, 114]. Therefore, designing for seamfulness can add to existing seamless approaches as contingency or fail-safes that empower users through occasional recalibration and reappropriation of systems.

Need to Expand Sample and Conditions: Our sample was predominantly female (§3.4.3). This was consistent with known gender differences in depression studies [75], and often stems from differences in mental health help-seeking attitudes [88]. Thus, further research is needed to account for the lived experience of male and gender-diverse populations who might use DPMH. Another aspect worth considering is illness duration and formal treatment status. Our sample was subclinical and recruited on the basis of their PHQ-9 scores, but future studies could capture participant histories for better understanding of their perspective. Note that our study focused on depression because it is known to have heterogeneous manifestations [32, 75], which lead to a disconnect between observed objective states and subjective perceptions and self-reports. Additionally, depression has been extensively studied by the Ubicomp community [105, 108]. However, heterogeneity is not unique to depression and applies to several other mental health outcomes like anxiety, stress, and chronic pain [69, 70, 74, 103]. Our findings with depression can inform seamful design for these conditions, but each outcome is uniquely expressed, differentially prevalent across populations, and necessitates dedicated studies in its own contextual settings.

Need to Include Clinical Stakeholders: Lastly, our study was limited to understanding the seams of a single stakeholder, the data-subject of DPMH. Our own team involved a clinical psychologist who conducts DPMH studies regularly (§3.5), and we iterated DYMOND's design with external researchers with experience developing DPMH (§ 3.3). Yet, other stakeholders might have their own perspective on seams when they face these in practice. For instance, clinicians and caregivers might have their own burdens in collaboratively configuring with end-users [91]. It is essential to weigh their perspective as the seamful DPMH loop hopes to involve competent and collaborative stakeholders. Future work should consider investigating how multiple stakeholders adjust across the seams.

7 CONCLUSION

Better mental health support needs DPMH to give a full picture of users through unobtrusive data monitoring. Yet this passive nature can disengage users from systems they cannot see or manipulate. Bringing humans into the loop offers a path forward — but until now, we lacked the empirical grounding to know where users encounter friction and what they need to navigate it. We addressed this gap by investigating the feasibility of *seamful design* in DPMH. We deployed a technology probe, DYMOND, with 22 participants over six weeks. Our inquiry revealed that users encounter logical points of friction (seams) across the data, modeling, and outcome phases of the DPMH pipeline — points of variation and ambiguity that passive systems leave invisible. Rather than treating these as failures, seamful design principles consider these seams as resources for user agency. Our study showed how participants tried to exercise this agency and configure DPMH for themselves, but also faced new forms of user burdens in interacting with the model. To counteract these challenges, we derived six design requirements to guide how future DPMH systems can expose, interpret, and negotiate seams collaboratively with users. As mental health data becomes more integrated into care, systems that provide users an opportunity to modify them will be essential to building trust, clinical relevance, and sustained engagement.

Acknowledgments

Our research is supported by funds from NIH National Institute of Drug Abuse under award number NIH/NIDA P30DA029926 (through the Center for Technology and Behavioral Health at Dartmouth College), National Science Foundation award number IIS-2442593, and the CORE program at Northeastern University. Das Swain was supported in part by the Khoury Distinguished Postdoctoral Fellowship at Northeastern University. We are grateful to Catherine Stanger from Dartmouth for her guidance in designing the protocol, Qiuyue (Joy) Zhong for assistance in drafting the IRB protocol, and the members of the UbiWell research group at Northeastern University for their feedback and support throughout the study. Lastly, we thank the participants of our study and their willingness to describe their lived experience with depression for research purposes.

References

[1] Daniel A. Adler, Emily Tseng, Khatiya 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. doi:10.1145/3555531

[2] Hege K. Andreassen, Marianne Trondsen, Per Egil Kummervold, Deede Gammon, and Per Hjortdahl. 2006. Patients Who Use E-Mediated Communication With Their Doctor: New Constructions of Trust in the Patient-Doctor Relationship. *Qualitative Health Research* 16, 2 (Feb. 2006), 238–248. doi:10.1177/1049732305284667

[3] Amid Ayobi, Paul Marshall, and Anna L. Cox. 2020. Trackly: A Customisable and Pictorial Self-Tracking App to Support Agency in Multiple Sclerosis Self-Care. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376809

[4] Samuel A. Ball, Kathleen M. Carroll, Monica Canning-Ball, and Bruce J. Rounsaville. 2006. Reasons for dropout from drug abuse treatment: symptoms, personality, and motivation. *Addictive Behaviors* 31, 2 (Feb. 2006), 320–330. doi:10.1016/j.addbeh.2005.05.013

[5] Roni Berger. 2015. Now I see it, now I don't: researcher's position and reflexivity in qualitative research. *Qualitative Research* 15, 2 (April 2015), 219–234. doi:10.1177/1468794112468475

[6] Ananya Bhattacharjee, Joseph Jay Williams, Miranda Beltzer, Jonah Meyerhoff, Harsh Kumar, Haochen Song, David C. Mohr, Alex Mariakakis, and Rachel Kornfield. 2025. Investigating the Role of Situational Disruptors in Engagement with Digital Mental Health Tools. arXiv:2502.09776 [cs]. doi:10.48550/arXiv.2502.09776

[7] Karin Bogdanova, Nazli Cila, Olya Kudina, and Alessandro Bozzon. 2025. Digital Phenotyping as Felt Informatics: Designing AI-Based Mental Health Diagnostic Tools Through Aesthetics. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–16. doi:10.1145/3706598.3714399

[8] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. doi:10.1191/1478088706qp063oa

[9] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300234

[10] Cynthia I. Campbell, Ching-Hua Chen, Sara R. Adams, Asma Asyyed, Ninad R. Athale, Monique B. Does, Saeed Hassanpour, Emily Hichborn, Melanie Jackson-Morris, Nicholas C. Jacobson, Heather K. Jones, David Kotz, Chantal A. Lambert-Harris, Zhiguo Li, Bethany McLeman, Varun Mishra, Catherine Stanger, Geetha Subramaniam, Weiyi Wu, Christopher Zegers, and Lisa A. Marsch. 2023. Patient Engagement in a Multimodal Digital Phenotyping Study of Opioid Use Disorder. *Journal of Medical Internet Research* 25, 1 (June 2023), e45556. doi:10.2196/45556

[11] Scott Carter and Jennifer Mankoff. 2004. Challenges for ubicomp evaluation. CSD-04-1331 (2004).

[12] Matthew Chalmers. 2003. Seamful Design and Ubicomp Infrastructure. In *Proceedings of Ubicomp 2003 workshop at the crossroads: The interaction of HCI and systems issues in Ubicomp*. 577–584. doi:10.1049/ic:20030140

[13] Larry Chan, Vedant Das Swain, Christina Kelley, Kaya de Barbaro, Gregory D. Abowd, and Lauren Wilcox. 2018. Students' Experiences with Ecological Momentary Assessment Tools to Report on Emotional Well-being. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–20. doi:10.1145/3191735

[14] Stevie Chancellor, Eric P.S. Baumer, and Munmun De Choudhury. 2019. Who is the "human" in human-centered machine learning: The case of predicting mental health from social media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). doi:10.1145/3359249

[15] Eun Kyoung Choe, Bongshin Lee, Haining Zhu, Nathalie Henry Riche, and Dominikus Baur. 2017. Understanding self-reflection: how people reflect on personal data through visual data exploration. In *Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth '17)*. Association for Computing Machinery, New York, NY, USA, 173–182. doi:10.1145/3154862.3154881

[16] Shaan Chopra, Katherine Juarez, James Fogarty, and Sean A. Munson. 2025. Engagements with Generative AI and Personal Health Informatics: Opportunities for Planning, Tracking, Reflecting, and Acting around Personal Health Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3 (Sept. 2025), 75:1–75:33. doi:10.1145/3749503

[17] Akshat Choube, Sohini Bhattacharya, Rahul Majethia, Jiachen Li, Vedant Das Swain, and Varun Mishra. 2025. Imputation Matters: A Deeper Look into an Overlooked Step in Longitudinal Health and Behavior Sensing Research. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4 (Dec. 2025), 167:1–167:30. doi:10.1145/3770647

[18] Akshat Choube, Ha Le, Jiachen Li, Kaixin Ji, Vedant Das Swain, and Varun Mishra. 2025. GLOSS: Group of LLMs for Open-ended Sensemaking of Passive Sensing Data for Health and Wellbeing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 3 (Sept. 2025), 1–32. doi:10.1145/3749474

[19] Akshat Choube, Shreeti Shrestha, Ha Le, Jiachen Li, Vedant Das Swain, and Varun Mishra. 2026. DAIMON: Designing AI-Augmented Research Dashboards to Enable Novel Human-AI Collaborative Workflows in Longitudinal Sensing Studies. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 10, 3 (Sept. 2026). doi:10.1145/3832033

[20] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces (IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. doi:10.1145/3397481.3450644

[21] Victor P. Cornet and Richard J. Holden. 2018. Systematic review of smartphone-based passive sensing for health and wellbeing. *Journal of Biomedical Informatics* 77 (Jan. 2018), 120–132. doi:10.1016/j.jbi.2017.12.008

[22] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, San Jose California USA, 1389–1397. doi:10.1145/2851581.2892410

[23] Vedant Das Swain, Victor Chen, Shrija Mishra, Stephen M. Mattingly, Gregory D. Abowd, and Munmun De Choudhury. 2022. Semantic Gap in Predicting Mental Wellbeing through Passive Sensing. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3491102.3502037

[24] Vedant Das Swain, Lan Gao, Abhirup Mondal, Gregory D. Abowd, and Munmun De Choudhury. 2024. Sensible and Sensitive AI for Worker Wellbeing: Factors that Inform Adoption and Resistance for Information Workers. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–30. doi:10.1145/3613904.3642716

[25] Vedant Das Swain, Lan Gao, William A Wood, Srikruthi C Matli, Gregory D. Abowd, and Munmun De Choudhury. 2023. Algorithmic Power or Punishment: Information Worker Perspectives on Passive Sensing Enabled AI Phenotyping of Performance and Wellbeing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3544548.3581376

[26] Vedant Das Swain, Koustuv Saha, Hemang Rajvanshy, Anusha Sirigiri, Julie M. Gregg, Suwen Lin, Gonzalo J. Martinez, Stephen M. Mattingly, Shayan Mirjafari, Raghu Mulukutla, Subigya Nepal, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Andrew T. Campbell, Nitesh V. Chawla, Sidney D'Mello, Anind K. Dey, Kaifeng Jiang, Qiang Liu, Gloria Mark, Edward Moskal, Aaron Striegel, Louis Tay, Gregory D. Abowd, and Munmun De Choudhury. 2020. A Multisensor Person-Centered Approach to Understand the Role of Daily Activities in Job Performance with Organizational Personas. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4 (Sept. 2020), 130:1–130:27. doi:10.1145/3369828

[27] Vedant Das Swain, Shane Williams, Adam Fourney, and Shamsi T. Iqbal. 2022. Two Birds with One Phone: The Role of Mobile Use in the Daily Practices of Remote Information Work. In *Proceedings of the 1st Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '22)*. Association for Computing Machinery, New York, NY, USA, 1–8. doi:10.1145/3533406.3533416

[28] Graham Dove, Marina Roos Guthmann, Leigh Charvet, Oded Nov, and Giuseppina Pilloni. 2024. "Data Is One Thing, But I Want To Know The Story Behind": Designing For Self-Tracking and Remote Patient Monitoring In The Context Of Multiple Sclerosis Care. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference (DIS '24)*. Association for Computing Machinery, New York, NY, USA, 597–618. doi:10.1145/3643834.3661499

[29] Jerry Alan Fails and Dan R. Olsen. 2003. Interactive machine learning. In *Proceedings of the 8th international conference on Intelligent user interfaces (IUI '03)*. Association for Computing Machinery, New York, NY, USA, 39–45. doi:10.1145/604045.604056

[30] Daniel Fitton and Keith Cheverst. 2010. Context sharing in a 'real world' ubicomp deployment. *Pervasive and Mobile Computing* 6, 5 (Oct. 2010), 527–536. doi:10.1016/j.pmcj.2010.04.004

[31] Geraldine Fitzpatrick, Alina Huldtgren, Lone Malmborg, Dave Harley, and Wijnand Ijsselsteijn. 2015. Design for Agency, Adaptivity and Reciprocity: Reimagining AAL and Telecare Agendas. In *Designing Socially Embedded Technologies in the Real-World*, Volker Wulf, Kjeld Schmidt, and David Randall (Eds.). Springer London, London, 305–338. Series Title: Computer Supported Cooperative Work. doi:10.1007/978-1-4471-6720-4_13

[32] Eiko I Fried. 2015. Problematic assumptions have slowed down depression research : why symptoms , not syndromes are the way forward. 6, March (2015), 1–11. doi:10.3389/fpsyg.2015.00309

[33] Elliot M Goldner, Anna Lusted, Michael Roerecke, Jürgen Rehm, and Benedikt Fischer. 2014. Prevalence of axis-1 psychiatric (with focus on depression and anxiety) disorder and symptomatology among non-medical prescription opioid users in substance use treatment: systematic review and meta-analyses. *Addictive behaviors* 39, 3 (2014), 520–531. doi:10.1016/j.addbeh.2013.11.022

[34] Laura E. Gressler, Nabil M. Natafgi, Bruce R. DeForge, Barbarajean Shaneman-Robinson, Christopher Welsh, and Fadia T. Shaya. 2019. What motivates people with substance use disorders to pursue treatment? A patient-centered approach to understanding patient experiences and patient-provider interactions. *Journal of substance use* 24, 6 (2019), 587–599. doi:10.1080/14659891.2019.1620891

[35] Erik Grönvall and Nervo Verdezoto. 2013. Beyond self-monitoring: understanding non-functional aspects of home-based healthcare technology. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 587–596. doi:10.1145/2493432.2493495

[36] Kaely Hall, Max Ohsawa, Vedant Das Swain, and Jennifer G Kim. 2026. LL.me: Supporting Identity Work through Human-AI Alignment. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–17. doi:10.1145/3772318.3790890

[37] Gabriella M. Harari, Sandrine R. Müller, Varun Mishra, Rui Wang, Andrew T. Campbell, Peter J. Rentfrow, and Samuel D. Gosling. 2017. An Evaluation of Students' Interest in and Compliance With Self-Tracking Methods: Recommendations for Incentives Based on Three Smartphone Sensing Studies. *Social Psychological and Personality Science* 8, 5 (July 2017), 479–492. doi:10.1177/1948550617712033

[38] Victoria Hollis, Alon Pekurovsky, Eunika Wu, and Steve Whittaker. 2018. On Being Told How We Feel: How Algorithmic Sensor Feedback Influences Emotion Perception. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3 (Sept. 2018), 114:1–114:31. doi:10.1145/3264924

[39] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. *Conference on Human Factors in Computing Systems - Proceedings* May (1999), 159–166. ISBN: 0201485591. doi:10.1145/302979.303030

[40] Esther Howe, Jina Suh, Mehrab Bin Morshed, Daniel McDuff, Kael Rowan, Javier Hernandez, Marah Ihab Abdin, Gonzalo Ramos, Tracy Tran, and Mary P Czerwinski. 2022. Design of Digital Workplace Stress-Reduction Intervention Systems: Effects of Intervention Type and Timing. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–16. doi:10.1145/3491102.3502027

[41] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[42] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, and Others. 2003. Technology probes: inspiring design for and with families. Pages: 17-24 Publication Title: Proceedings of the SIGCHI conference on Human factors in computing systems. https://dl.acm.org/doi/abs/10.1145/642611.642616

[43] Sarah Inman and David Ribes. 2019. "Beautiful Seams": Strategic Revelations and Concealments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300508

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

[45] Sabrina Zaman Ishita, Sidharth Kaliappan, Mashrur Rashik, Daniel A. Epstein, and Ravi Karkar. 2026. CASEbot: A Conversational Agent for Structuring and Personalizing the Design of Self-Experiments in Personal Health. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–26. doi:10.1145/3772318.3791551

[46] Nicholas C. Jacobson and Yeon Joo Chung. 2020. Passive Sensing of Prediction of Moment-To-Moment Depressed Mood among Undergraduates with Clinical Levels of Depression Sample Using Smartphones. *Sensors* 20, 12 (Jan. 2020), 3572. Number: 12. doi:10.3390/s20123572

[47] Nicholas C. Jacobson and Brandon Feng. 2022. Digital phenotyping of generalized anxiety disorder: using artificial intelligence to accurately predict symptom severity using wearable sensors in daily life. *Translational Psychiatry* 12, 1 (Aug. 2022), 1–7. doi:10.1038/s41398-022-02038-1

[48] Robert Jakob, Samira Harperink, Aaron Maria Rudolf, Elgar Fleisch, Severin Haug, Jacqueline Louise Mair, Alicia Salamanca-Sanabria, and Tobias Kowatsch. 2022. Factors Influencing Adherence to mHealth Apps for Prevention or Management of Noncommunicable Diseases: Systematic Review. *J Med Internet Res 2022;24(5):e35371 https://www.jmir.org/2022/5/e35371* (May 2022). doi:10.2196/35371

[49] Maria Jerónimo, Tiago Guerreiro, and Ruben Gouveia. 2026. Between Control and Uncertainty: Understanding Self-Tracking Practices in Enigmatic Disease Management. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3772318.3791251

[50] Matthew Jörke, Defne Genç, Valentin Teutschbein, Shardul Sapkota, Sarah Chung, Paul Schmiedmayer, Maria Ines Campero, Abby C. King, Emma Brunskill, and James A. Landay. 2025. Bloom: Designing for LLM-Augmented Behavior Change Interactions. *arXiv* (2025). arXiv:2510.05449. doi:10.48550/arXiv.2510.05449

[51] Matthew Jörke, Yasaman S. Sefidgar, Talie Massachi, Jina Suh, and Gonzalo Ramos. 2023. Pearl: A Technology Probe for Machine-Assisted Reflection on Personal Data. In *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*. Association for Computing Machinery, New York, NY, USA, 902–918. doi:10.1145/3581641.3584054

[52] Ravi Karkar, Jessica Schroeder, Daniel A. Epstein, Laura R. Pina, Jeffrey Scofield, James Fogarty, Julie A. Kientz, Sean A. Munson, Roger Vilardaga, and Jasmine Zia. 2017. TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6850–6863. doi:10.1145/3025453.3025480
[53] Kevin Kelly and Gary Wolf. 2007. What is the quantified self. *The Quantified Self* 5 (2007), 2007.
[54] Vera Khovanskaya, Eric P.S. Baumer, Dan Cosley, Stephen Voida, and Geri Gay. 2013. "Everybody knows what you're doing": a critical design approach to personal informatics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 3403–3412. doi:10.1145/2470654.2466467
[55] Hankyung Kim and Youn-Kyung Lim. 2023. Investigating How Users Design Everyday Intelligent Systems in Use. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)*. Association for Computing Machinery, New York, NY, USA, 702–711. doi:10.1145/3563657.3596039
[56] Taewan Kim, Haesoo Kim, Ha Yeon Lee, Hwarang Goh, Shakhboz Abdigapporov, Mingon Jeong, Hyunsung Cho, Kyungsik Han, Youngtae Noh, Sung-Ju Lee, and Hwajung Hong. 2022. Prediction for Retrospection: Integrating Algorithmic Stress Prediction into Personal Informatics Systems for College Students' Mental Health. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–20. doi:10.1145/3491102.3517701
[57] Reuben Kirkham, Tom Bartindale, Michael Evans, and Patrick Olivier. 2015. Build me a Ubicomp: bespoke ubiquitous accessibility in live television production environments. *Personal and Ubiquitous Computing* 19, 5 (Aug. 2015), 853–880. doi:10.1007/s00779-015-0859-7
[58] Rachel Kornfield, Sarah Alexandra Popowski, Emily Tack, Jack Svoboda, Miguel Herrera, Theresa Nguyen, Ashley Arehart Knapp, David Curtis Mohr, and Jonah Meyerhoff. 2025. Designing Digital Mental Health Interventions to Meet the Needs of Older Adolescents: Qualitative Interview and Group Discussion Study. *JMIR Formative Research* 9, 1 (Aug. 2025), e68950. doi:10.2196/68950
[59] Jan-Niklas Kramer, Florian Künzler, Varun Mishra, Shawna N Smith, David Kotz, Urte Scholz, Elgar Fleisch, and Tobias Kowatsch. 2020. Which Components of a Smartphone Walking App Help Users to Reach Personalized Step Goals? Results From an Optimization Trial. *Annals of Behavioral Medicine* 54, 7 (June 2020), 518–528. doi:10.1093/abm/kaaa002
[60] Richard L. Kravitz, Debora A. Paterniti, Ronald M. Epstein, Aaron B. Rochlen, Robert A. Bell, Camille Cipri, Erik Fernandez y Garcia, Mitchell D. Feldman, and Paul Duberstein. 2011. Relational Barriers to Depression Help-Seeking in Primary Care. *Patient education and counseling* 82, 2 (Feb. 2011), 207–213. doi:10.1016/j.pec.2010.05.007
[61] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of general internal medicine* 16, 9 (2001), 606–613. doi:10.1046/j.1525-1497.2001.016009606.x
[62] Morten Kyng. 1994. Scandinavian Design: Users in Product Development. *Human Factors* (1994). doi:10.1145/191666.191673
[63] Florian Künzler, Varun Mishra, Jan-Niklas Kramer, David Kotz, Elgar Fleisch, and Tobias Kowatsch. 2020. Exploring the State-of-Receptivity for mHealth Interventions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4 (Sept. 2020), 140:1–140:27. doi:10.1145/3369805
[64] Kirsten J Langdon, Kathrine Dove, and Susan Ramsey. 2019. Comorbidity of opioid-related and anxiety-related symptoms and disorders. *Current opinion in psychology* 30 (2019), 17–23. doi:10.1016/j.copsyc.2018.12.020
[65] Ha Le, Akshat Choube, Vedant Das Swain, Varun Mishra, and Stephen Intille. 2025. A Multi-Agent LLM Network for Suggesting and Correcting Human Activity and Posture Annotations. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 877–884.
[66] Imogen E. Leaning, Nessa Ikani, Hannah S. Savage, Alex Leow, Christian Beckmann, Henricus G. Ruhé, and Andre F. Marquand. 2024. From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression. *Neuroscience & Biobehavioral Reviews* 158 (March 2024), 105541. doi:10.1016/j.neubiorev.2024.105541
[67] Jiachen Li, Xiwen Li, Justin Steinberg, Akshat Choube, Bingsheng Yao, Xuhai Xu, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2025. Vital Insight: Assisting Experts' Context-Driven Sensemaking of Multi-modal Personal Tracking Data Using Visualization and Human-in-the-Loop LLM. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3 (Sept. 2025), 101:1–101:37. doi:10.1145/3749508
[68] Zachary C. Lipton and Search about this author. 2018. The Mythos of Model Interpretability. *Queue* 16, 3 (June 2018), 31–57. doi:10.1145/3236386.3241340
[69] Lakmal Meegahapola, Marios Constantinides, Zoran Radivojevic, Hongwei Li, Michael Eggleston, and Daniele Quercia. 2026. Stress Mindset Matters: Rethinking Mental Stress Detection with Multimodal Wearable Sensors. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26)*. Association for Computing Machinery, New York, NY, USA, 1–26. doi:10.1145/3772318.3791340
[70] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4 (Jan. 2023), 176:1–176:32. doi:10.1145/3569483

[71] Varun Mishra, Sarah Hong, and David Kotz. 2024. Exploring the Relationship Between Intrinsic Motivation and Receptivity to mHealth Interventions. In *Companion of the 2024 on ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, Melbourne VIC Australia, 437–443. doi:10.1145/3675094.3678498

[72] Varun Mishra, Florian Künzler, Jan-Niklas Kramer, Elgar Fleisch, Tobias Kowatsch, and David Kotz. 2021. Detecting receptivity for mhealth interventions in the natural environment. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 5, 2 (2021), 1–24. doi:10.1145/3463492

[73] David C. Mohr, Ken R. Weingardt, Madhu Reddy, and Stephen M. Schueller. 2017. Three Problems With Current Digital Mental Health Research . . . and Three Things We Can Do About Them. *Psychiatric Services* 68, 5 (May 2017), 427–429. doi:10.1176/appi.ps.201600541

[74] 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), 237. doi:10.1038/s41746-023-00975-9

[75] Matthew D Nemesure, Amanda C Collins, George D Price, Tess Z Griffin, Arvind Pillai, Subigya Nepal, Michael V Heinz, Damien Lekkas, Andrew T Campbell, and Nicholas C Jacobson. 2024. Depressive symptoms as a heterogeneous and constantly evolving dynamical system: Idiographic depressive symptom networks of rapid symptom changes among persons with major depressive disorder. *Journal of psychopathology and clinical science* 133, 2 (2024), 155. doi:10.1037/abn0000884

[76] Subigya Nepal, Wenjun Liu, Arvind Pillai, Weichen Wang, Vlado Vojdanovski, Jeremy F. Huckins, Courtney Rogers, Meghan L. Meyer, and Andrew T. Campbell. 2024. Capturing the College Experience: A Four-Year Mobile Sensing Study of Mental Health, Resilience and Behavior of College Students during the Pandemic. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 1 (March 2024), 38:1–38:37. doi:10.1145/3643501

[77] Francisco Nunes, Tariq Andersen, and Geraldine Fitzpatrick. 2019. The agency of patients and carers in medical care and self-care technologies for interacting with doctors. *Health Informatics Journal* 25, 2 (June 2019), 330–349. doi:10.1177/1460458217712054

[78] Jukka-Pekka Onnela. 2021. Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology* 46, 1 (Jan. 2021), 45–54. doi:10.1038/s41386-020-0771-3

[79] Joonyoung Park and Uichin Lee. 2023. Understanding Disengagement in Just-in-Time Mobile Health Interventions. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 2 (June 2023), 72:1–72:27. doi:10.1145/3596240

[80] Kelly Posner, Gregory K Brown, Barbara Stanley, David A Brent, Kseniya V Yershova, Maria A Oquendo, Glenn W Currier, Glenn A Melvin, Laurence Greenhill, Sa Shen, and others. 2011. The columbia–suicide severity rating scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. *American journal of psychiatry* 168, 12 (2011), 1266–1277. doi:10.1176/appi.ajp.2011.10111704

[81] Roya Rad and Graham Dove. 2025. Making/Unmaking with Metaphors: Generative Co-design with People Living with Chronic Pain. In *Companion Publication of the 2025 ACM Designing Interactive Systems Conference*. 345–350.

[82] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* (Nov. 2020). doi:10.1080/07370024.2020.1734931

[83] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (May 2019), 206–215. doi:10.1038/s42256-019-0048-x

[84] 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. doi:10.1145/3448117

[85] Isabel Schwaninger, Florian Güldenpfennig, Astrid Weiss, and Geraldine Fitzpatrick. 2021. What Do You Mean by Trust? Establishing Shared Meaning in Interdisciplinary Design for Assistive Technology. *International Journal of Social Robotics* 13, 8 (June 2021), 1879–1897. doi:10.1007/s12369-020-00742-w

[86] Yasaman S. Sefidgar, Carla L. Castillo, Shaan Chopra, Liwei Jiang, Tae Jones, Anant Mittal, Hyeyoung Ryu, Jessica Schroeder, Allison Cole, Natalia Murinova, Sean A. Munson, and James Fogarty. 2024. MigraineTracker: Examining Patient Experiences with Goal-Directed Self-Tracking for a Chronic Health Condition. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–19. doi:10.1145/3613904.3642075

[87] Yasaman S. Sefidgar, Matthew Jörke, Jina Suh, Koustuv Saha, Shamsi Iqbal, Gonzalo Ramos, and Mary Czerwinski. 2024. Improving Work-Nonwork Balance with Data-Driven Implementation Intention and Mental Contrasting. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 74:1–74:29. doi:10.1145/3637351

[88] Zac E. Seidler, Alexei J. Dawes, Simon M. Rice, John L. Oliffe, and Haryana M. Dhillon. 2016. The role of masculinity in men's help-seeking for depression: A systematic review. *Clinical Psychology Review* 49 (Nov. 2016), 106–118. doi:10.1016/j.cpr.2016.09.002

[89] Emily Seltzer, Jesse Goldshear, Sharath Chandra Guntuku, Dave Grande, David A. Asch, Elissa V. Klinger, and Raina M. Merchant. 2019. Patients' willingness to share digital health and non-health data for research: a cross-sectional study. *BMC Medical Informatics and Decision Making* 19, 1 (Aug. 2019), 157. doi:10.1186/s12911-019-0886-9

[90] D. V. Sheehan, Y. Lecrubier, K. H. Sheehan, P. Amorim, J. Janavs, E. Weiller, T. Hergueta, R. Baker, and G. C. Dunbar. 1998. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry* 59 Suppl 20 (1998), 22–33;quiz 34–57.

[91] Jiayue Melissa Shi, Dong Whi Yoo, Keran Wang, Violeta J. Rodriguez, Ravi Karkar, and Koustuv Saha. 2026. Mapping Caregiver Needs to AI Chatbot Design: Strengths and Gaps in Mental Health Support for Alzheimer's and Dementia Caregivers. *ACM Transactions on Computing for Healthcare* (March 2026). Just Accepted. doi:10.1145/3803549

[92] Ranjit Singh and Steven J. Jackson. 2017. From Margins to Seams: Imbrication, Inclusion, and Torque in the Aadhaar Identification Project. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 4776–4824. doi:10.1145/3025453.3025910

[93] Claire F. Snyder, Albert W. Wu, Robert S. Miller, Roxanne E. Jensen, Elissa T. Bantug, and Antonio C. Wolff. 2011. The Role of Informatics in Promoting Patient-Centered Care. *Cancer journal (Sudbury, Mass.)* 17, 4 (2011), 211–218. doi:10.1097/PPO.0b013e318225ff89

[94] Victoria Stanhope, Chuck Ingoglia, Bill Schmelter, and Steven C. Marcus. 2013. Impact of person-centered planning and collaborative documentation on treatment adherence. *Psychiatric Services (Washington, D.C.)* 64, 1 (Jan. 2013), 76–79. doi:10.1176/appi.ps.201100489

[95] Mirjam Stieger, Marcia Nißen, Dominik Rüegger, Tobias Kowatsch, Christoph Flückiger, and Mathias Allemand. 2018. PEACH, a smartphone- and conversational agent-based coaching intervention for intentional personality change: study protocol of a randomized, wait-list controlled trial. *BMC Psychology* 6, 1 (Sept. 2018), 43. doi:10.1186/s40359-018-0257-9

[96] Poorna Talkad Sukumar, Gonzalo J Martinez, Ted Grover, Gloria Mark, Sidney K D'Mello, Nitesh V Chawla, Stephen M Mattingly, and Aaron D Striegel. 2020. Characterizing exploratory behaviors on a personal visualization interface using interaction logs. *EuroVis 2020-Short Papers* (2020). doi:10.31219/osf.io/axqwc

[97] Dorothy Szinay, Andy Jones, Tim Chadborn, Jamie Brown, and Felix Naughton. 2020. Influences on the Uptake of and Engagement With Health and Well-Being Smartphone Apps: Systematic Review. *J Med Internet Res 2020;22(5):e17572 https://www.jmir.org/2020/5/e17572* (May 2020). doi:10.2196/17572

[98] Anja Thieme, Danielle Belgrave, and Gavin Doherty. 2020. Machine Learning in Mental Health : A Systematic Review of the HCI Literature to Support the Development of Effective and Implementable ML Systems. 27, 5 (2020).

[99] Edmund R Thompson. 2007. Development and validation of an internationally reliable short-form of the positive and negative affect schedule (panas). *Journal of cross-cultural psychology* 38, 2 (2007), 227–242. doi:10.1177/0022022106297301

[100] John Torous, Jessica Lipschitz, Michelle Ng, and Joseph Firth. 2020. Dropout rates in clinical trials of smartphone apps for depressive symptoms: A systematic review and meta-analysis. *Journal of Affective Disorders* 263 (Feb. 2020), 413–419. doi:10.1016/j.jad.2019.11.167

[101] John Torous, Jennifer Nicholas, Mark E. Larsen, Joseph Firth, and Helen Christensen. 2018. Clinical review of user engagement with mental health smartphone apps: evidence, theory and improvements. *BMJ Ment Health* 21, 3 (Aug. 2018), 116–119. doi:10.1136/eb-2018-102891

[102] Vincent W. S. Tseng, Michael Merrill, Franziska Wittleder, Saeed Abdullah, Min Hane Aung, and Tanzeem Choudhury. 2016. Assessing mental health issues on college campuses: preliminary findings from a pilot study. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct (UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1200–1208. doi:10.1145/2968219.2968308

[103] Renato Vellucci. 2012. Heterogeneity of Chronic Pain. *Clinical Drug Investigation* 32, 1 (Feb. 2012), 3–10. doi:10.2165/11630030-000000000-00000

[104] Viswanath Venkatesh and Fred D. Davis. 2000. Theoretical extension of the Technology Acceptance Model: Four longitudinal field studies. *Management Science* 46, 2 (2000), 186–204. doi:10.1287/mnsc.46.2.186.11926

[105] 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. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1 (March 2018), 43:1–43:26. doi:10.1145/3191775

[106] Yi Xiao, Harshit Sharma, Sawinder Kaur, Dessa Bergen-Cico, and Asif Salekin. 2025. Human Heterogeneity Invariant Stress Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 3 (Sept. 2025), 142:1–142:42. doi:10.1145/3749465

[107] Xuhai Xu, Prerna Chikersal, Janine M Dutcher, Yasaman S Sefidgar, Woosuk Seo, Michael J Tumminia, Daniella K Villalba, Sheldon Cohen, Kasey G Creswell, J David Creswell, Afsaneh Doryab, Paula S Nurius, Eve Riskin, Anind K Dey, and Jennifer Mankoff. 2021. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1 (2021), 1–27. doi:10.1145/3448107

[108] 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. 2022. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (Dec. 2022), 1–34. doi:10.1145/3569485

[109] Rayoung Yang and Mark W. Newman. 2013. Learning from a learning thermostat: lessons for intelligent systems for the home. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing (UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 93–102. doi:10.1145/2493432.2493489

[110] Rui Yang, Yilin Ning, Emilia Keppo, Mingxuan Liu, Chuan Hong, Danielle S. Bitterman, Jasmine Chiat Ling Ong, Daniel Shu Wei Ting, and Nan Liu. 2025. Retrieval-augmented generation for generative artificial intelligence in health care. *npj Health Systems* 2, 1 (Jan. 2025), 2. doi:10.1038/s44401-024-00004-1

[111] Rayoung Yang, Eunice Shin, Mark W. Newman, and Mark S. Ackerman. 2015. When fitness trackers don't 'fit': end-user difficulties in the assessment of personal tracking device accuracy. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. Association for Computing Machinery, New York, NY, USA, 623–634. doi:10.1145/2750858.2804269

[112] Dong Whi Yoo, Jiayue Melissa Shi, Violeta J. Rodriguez, and Koustuv Saha. 2026. AI Chatbots for Mental Health Self-Management: Lived Experience–Centered Qualitative Study. *JMIR Mental Health* 13, 1 (April 2026), e78288. doi:10.2196/78288

[113] Dong Whi Yoo, Hayoung Woo, Viet Cuong Nguyen, Michael L. Birnbaum, Kaylee Payne Kruzan, Jennifer G Kim, Gregory D. Abowd, and Munmun De Choudhury. 2024. Patient Perspectives on AI-Driven Predictions of Schizophrenia Relapses: Understanding Concerns and Opportunities for Self-Care and Treatment. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–20. doi:10.1145/3613904.3642369

[114] Jichen Zhu, Pedro Sanches, Vasiliki Tsaknaki, Willem van der Maden, and Irene Kaklopoulou. 2025. The Centers and Margins of Modeling Humans in Well-being Technologies: A Decentering Approach. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. arXiv:2503.19132 [cs]. doi:10.1145/3706598.3713940