UNIVERSITY OF CALIFORNIA

Los Angeles

Machine Learning-Based Detection of Depression Symptoms with Smartphones and Consumer Wearable Devices

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Medical Informatics

by

Samir Akre

2024

© Copyright by

Samir Akre

2024

ABSTRACT OF THE DISSERTATION

Machine Learning-Based Detection of Depression Symptoms with Smartphones and Consumer Wearable Devices

by

Samir Akre

Doctor of Philosophy in Medical Informatics

University of California, Los Angeles, 2024

Professor Alex Anh-Tuan Bui, Chair

Consumer wearable devices and smartphones are ubiquitous and generate valuable health-related data that remain underexplored. These data have the potential to enhance our understanding of depression by bridging gaps left by traditional methods that often rely on retrospective self-reports. By leveraging machine learning, we identified relationships between specific passively measured behaviors and retrospective self-reports related to depression severity, reward functioning, and sleep quality. Focusing on sleep quality, our findings indicate that self-reported and physiologically measured sleep quality assess different constructs and offer distinct insights into depression. Anomaly detection (AD) methods were examined and aimed at identifying correlations between deviations from typical behavior, as recorded by mobile health (mHealth) devices, and changes in depression severity and symptoms. Although

no significant relationship was found, the AD methods effectively detected multivariate anomalies, indicating potential applications beyond depression. Additionally, real-time data from wearable devices proved effective in detecting momentary reward functioning and affect, with models performing above random chance and performance varying across demographic and clinical groups. This dissertation highlights the importance of nuanced approaches in using consumer device-generated data to passively detect depression symptomology.

The dissertation of Samir Akre is approved.

Arash Naeim

Loes Marlein Olde Loohuis

Nelson B. Freimer

Alex Anh-Tuan Bui, Committee Chair

University of California, Los Angeles

2024

To my parents, Sunil and Anupama Akre

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I am deeply grateful to everyone who has supported me throughout this Ph.D. journey. Your guidance, encouragement, and companionship have been invaluable.

First and foremost, I would like to express my heartfelt gratitude to my advisor, Prof. Alex Bui. You have been an absolute marvel of an advisor, striking a perfect balance between offering poignant advice when I was lost and allowing me to carve my own path when it was appropriate. Throughout this process, I have questioned if I was a real scientist or if I could do the research, but I never doubted that I was well-supported. Alex, your knowledge and capabilities are astounding, and I hope to one day give back to another what you have given to me.

Prof. Nelson Freimer, thank you for providing me with opportunities to gain experience that have made me essential in this field. Your support has been instrumental in my development. Prof. Loes Olde Loohuis, I am grateful for your guidance through roadblocks in research and helping me work forward from null results. Your insights have been crucial in overcoming challenges. Prof. Arash Naeim, your monthly check-ins, confidence in me as a professional, and looking out for what's best for me have been greatly appreciated. Thank you for your unwavering support.

I would also like to extend my gratitude to other researchers who have significantly impacted my journey. Dr. Eliza Congdon, you were one of my first introductions to working with the DGC. You balance being an exemplar of competence while still being cool. Amelia Wellborn, thank you for answering every question I've had about the OPTIMA study and being an incredible force in getting the data that has become the foundation of my research. Prof. Zachary D. Cohen, your gentle guidance

as I navigated my way into research in psychology has been invaluable. You have connected me to incredible people and become a great friend and colleague. Prof. Diana H. Taft, thank you for being one of my first mentors in computational bioinformatics. Your trust in me when I had done little to prove I deserved it has been a cornerstone of my journey.

To the students who have shared this journey with me, Kerneau Seok, it has been fascinating working together on our project, and I look forward to seeing your amazing future work. Henry Zheng, being there through the inception of the Medical Informatics program and joining the same lab has been an enriching experience. Your unique perspective and knowledge have been invaluable.

My family and community are the backbone of all that I do. To my wife, Sharvari Bhide, your unwavering support and confidence in me have been the pillar I leaned on when things got tough. I don't think it’s common for PhD students to have such fulfilling personal lives, but you’ve made these last five years filled with love and the best I've had.

Mummy and Baba, your support and love have been true constants in my life. Mummy, your care and advice throughout our phone calls and visits have been a pillar of strength. Baba, I’m following in your footsteps getting this degree. You have been an inspiration and role model. Sagar, you are one of my favorite humans, it’s so convenient that you are my brother. Your insights have helped me consider my research through a design lens I would never have otherwise.

Sonal, Parag, and Shaunak Bhide, having a whole new family has been an amazing comfort. Thank you for your warmth, care, and support.

Sunny Karnan, your presence in my life has been one of those unexpected blessings that have fundamentally changed how I operate and how these last five years have gone. I'll always look forward to trying a new bean, discussing a new chapter, and getting some rounds in.

Thank you to those that have supported me over these last five-plus years. I am so grateful for the group of amazingly fun, intelligent, and kind people I get to spend my time with. Thank you, Ana Silverstein-Png, Shaun Silverstein-Png, Meredith Buganski, Jonatan L. Hervoso, L. Sebastian Ojeda, Krishna Basude, Shiv Bhandari, and Manohar Boppana. And thank you, Noodle Akre-Bhide. You will never read this, nor will you ever know how critical your presence has been to me completing this journey. I will give you greenies and pets.

Thank you all.

VITA

2021-2022	Biodesign Fellowship – University of California, Los Angeles
2019-2021	T32 National Institutes of Health Training Grant – University of California, Los Angeles
2019	Research programmer – USC Keck School of Medicine
2018-2019	Software & bioinformatic pipeline development – Stanford University School of Medicine
2017	Algorithm Development Intern – Lumo Bodytech Inc.
2014-2018	B.S. in Biomedical Engineering – University of California, Davis

PUBLICATIONS

Akre, S., Cohen, Z., Welborn, A., Zbozinek, T., Balliu, B., Craske, M. & Bui, A. Comparison of self-reported and physiological sleep quality from consumer devices to depression and neurocognitive performance. Under review at *NPJ Dig. Med.* (2024). doi:10.21203/rs.3.rs-4769246/v1

Langener, A. M., Siepe, B. S., Elsherif, M., Niemeijer, K., Andresen, P. K., **Akre, S.,** Bringmann, L. F., Cohen, Z. D., Choukas, N. R., Drexl, K., Fassi, L., Green, J., Hoffmann, T., Jagesar, R. R., Kas, M. J. H., Kurten, S., Schoedel, R., Stulp, G., Turner, G. & Jacobson, N. C. A template and tutorial for preregistering studies using passive smartphone measures. Behav. Res. Methods 1–19 (2024). doi:10.3758/s13428-024-02474-5

Holstein, V. L., **Akre, S.,** Leenings, R., Chung, Y., Hahn, T. & Baker, J. T. Predicting dimensions of depression from smartphone data. medRxiv 2024.01.08.23300679 (2024). doi:10.1101/2024.01.08.23300679

Akre, S., Balliu, B., Cohen, Z. D., Flint, J., Welborn, A., Bui, A. A. T., Zbozinek, T. D. & Craske, M. G. Detection of Symptoms of Depression Using Data From the iPhone and Apple Watch. 2023 IEEE Int. Conf. Bioinform. Biomed. (BIBM) 1818–1823 (2023). doi:10.1109/bibm58861.2023.10385797

Wong, M. S., Wells, M., Zamanzadeh, D., **Akre, S.,** Pevnick, J. M., Bui, A. A. T. & Gregory, K. D. Applying Automated Machine Learning to Predict Mode of Delivery Using Ongoing Intrapartum Data in Laboring Patients. Am J Perinat (2022). doi:10.1055/a-1885-1697

Akre, S., Liu, P. Y., Friedman, J. R. & Bui, A. A. T. International COVID-19 mortality forecast visualization: covidcompare.io. JAMIA Open 4, ooab113 (2021).

Friedman, J. & **Akre, S**. COVID-19 and the Drug Overdose Crisis: Uncovering the Deadliest Months in the United States, January–July 2020. Am J Public Health e1–e8 (2021). doi:10.2105/ajph.2021.306256

Chapter 1: Introduction

This dissertation centers on using smartphones and consumer wearable devices to detect symptoms of depression. I leveraged machine learning techniques to train models with passive sensor data streams to detect self-reported measures of depression. I additionally developed novel methods to detect abnormal or anomalous behavior that can utilize all the new and changing data streams we obtain from wearable devices. While there are significant limitations inherent in the use of consumer products for research, I find that consumer wearable devices can detect self-reported depression symptomology and can be used to advance the boundaries of how depression is measured and defined.

1.1 Defining and measuring depression

Depression is a condition that has been studied under different names since at least the Greek philosophers and is a leading cause of disability, suffering, and death globally[2]. However, it was only in the mid-20th century that the study of depression formally began in psychology. In fact, the reliability of diagnosing major depressive disorder (MDD) is incredibly low (kappa = 0.25)[3] for such a highly prevalent condition (lifetime prevalence 18.5% in the United States)[4].

To improve clinical outcomes in depression we need a more granular understanding of depression beyond classification, starting with a look at symptomology[5]. The specific symptoms one experiences during a depressive episode are critical to the response to antidepressant treatment[6]. This finding is crucial to

consider, as medication-based treatments have side effects that are themselves symptoms of depression. Attempts are being made to better define depression as a multidimensional measure rather than a simple classification to treat the condition more precisely and effectively. There are two key attempts to redefine how we view mental health in ways that are driven both by empirical data and theory. The Hierarchical Taxonomy of Psychopathology (HiTOP) represents a top-down approach based on empirical data regarding symptoms, diagnoses, and maladaptive behaviors[7]. The taxonomy is based on currently measured symptoms and leverages existing diagnoses, making the framework relatively immediately actionable, and the hierarchical structure has been validated on clinical and biological data[8–10]. Within HiTOP, the most common depressive symptomology falls under the distress subfactor within the internalizing spectra of psychopathology alongside commonly comorbid conditions such as generalized anxiety disorder (GAD). In contrast to HiTOP, the Research Domain Criteria (RDoC) framework from the National Institute of Mental Health (NIMH) represents a more theoretically driven, bottom-up approach to understanding mental health and cognitive functioning[11]. The RDoC framework focuses on domains of basic human neurobehavioral functioning (e.g., negative valance, social processes, reward functioning, etc.) and specific ways in which they can be measured in the context of the environment and developmental stage. It does not immediately have clinically actionable insights but sets a stage for research to build a mechanistic understanding of the brain. By combining current findings from HiTOP that are linked to RDoC, Michelini et al. summarized areas to validate the mechanism from RDoC with empirically

observed phenotypes detailed in HiTOP[12], providing targets for empirical exploration and validation.

1.2 Consumer devices for mental health symptom monitoring

Linking HiTOP phenotypes with domains in RDoC provides clearer targets for exploration with novel forms of data collection, such as mobile health (mHealth), especially where domains traditionally relied on, such as neuroscience[13] and genetics[14], have not shed as much light. With respect to depression, Michelini et al. reported that Negative Valence and Arousal & Regulatory domains in RDoC are associated with the Internalizing HiTOP domain, which contains the bulk of MDD-related symptomology. We may thus expect to validate relationship components in domains such as circadian rhythm, sleep-wake with key depressive symptoms via wearable devices, or smartphone data. By detecting specific symptomology with these devices, we can improve and expand our understanding of how depression is currently defined and link it to new and emerging frameworks for understanding mental health.

This dissertation focuses on how consumer wearable technology and smartphones enable more precise and high temporal monitoring of depression symptoms in the context of such frameworks to investigate mental health. Consumer trends in recent years have shown the growing prevalence of mHealth devices, with most adults owning smartphones and increasing adoption of fitness trackers or smartwatches. In 2019, 21% of American adults used fitness trackers or smartwatches[15], and that number increased in 2023, with 26% of internet users in the United States (US) using a smartwatch (e.g., the Apple watch) and 30% of internet users 16--64 using a smartwatch or fitness tracker (e.g., Fitbit)[16].

Data from wearable devices and smartphones show promise in bridging the time gap between the sparse clinical touchpoints used in traditional assessments of mental health. Mental health trajectories, or changes in symptom severity over time, are highly heterogeneous across populations, and different patterns emerge across age, sex, and socioeconomic status[17]. Given this variation in depression symptomology as well as trajectories, large sample sizes or very clinically homogenous studies are needed to account for the variance. In realizing larger cohorts, consumer wearable devices are particularly useful given their large prevalence. Data from mHealth devices can provide an understanding of the key behaviors that may lead to changes in symptoms, particularly as they occur in real-world settings and daily activities. Such insights will be a critical step for developing targeted and personalized interventions that address diverse needs. The use of research devices can be limited by cost, research staff burden in manually retrieving the data, and lack of clean user interfaces that naturally incentivize the use of the product. And in contrast to research devices, people *want* to use consumer-facing smart watches.

1.3 Contributions

While consumer wearable devices hold significant promise for improving our understanding of mental health, the data they generate are messy, large, and currently not clinically useful. My work centers on finding what signals may be present in these consumer device data in relation to depression in several ways:

- **Aim 1**: *Investigate which key symptoms of depression and reward functioning are detectable with passive sensor data.* Machine learning models were tested to determine if there was any relationship across participants (nomothetic modeling)

between their sensor data and depression symptomology. We find that there is a signal between the device data and depression symptoms (appetite, reward functioning, and sleep quality measures) and find notable discrepancies between self-reported symptoms and objectively measured behavior. We highlight that self-reported sleep disturbances in those with depression are not necessarily a reflection of changes in physiological sleep measures but are still associated with worsening neurocognitive performance.

- **Aim 2:** *Develop anomaly detection techniques to investigate whether deviations from normal behavior, as measured by mHealth devices, are associated with changes in mental health*. Anomaly detection algorithms were developed for use on large multimodal datasets, such as those with consumer wearable and smartphone data, and tested in simulations and with two different datasets with depression outcome measures. The approach to AD used extends the previous methodology, making fewer assumptions on the input data and applying the methods specifically to depression. While the detectors do find behavioral anomalies, the number of these daily anomalies is not found to be significantly associated with changes in depression severity, suggesting that for depression, we need to investigate specific behavioral changes rather than general instability in behavior.
- **Aim 3**: A*ssessment of momentary reward functioning and affect with passive sensor data.* Using passive sensing data, we detect responses to ecological momentary assessments (EMAs) using between 15 minutes and 3 hours of passive sensor data prior to the EMA response. We find that 11 of 15 EMA items related to affect and reward functioning can be detected, but models perform significantly differently

across participants on the basis of demographic characteristics and self-reported mental health. Features not typically investigated or measured in the study of depression, such as environmental audio exposure and basal energy expenditure, are highly important to models detecting momentary reward and affect.

1.4 Dissertation Organization

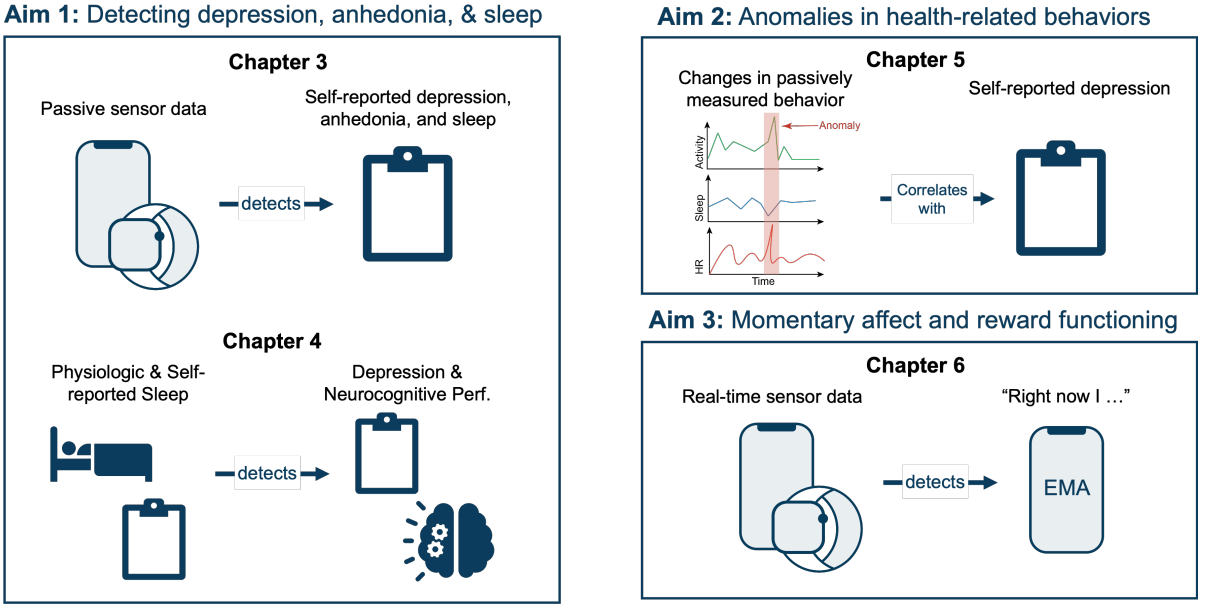

Figure 1.1 Overview of dissertation organization

This dissertation is organized as follows:

- Chapter 2 presents a description of prior investigations of smartphones and wearable devices in relation to depression outcomes. I also describe the analytic methods typically utilized and introduce the primary dataset investigated in this dissertation.

- Chapter 3 (Aim 1) reports on the use of passive sensor data to detect individual question responses to retrospective self-reports related to depression, anhedonia, and sleep quality. This work is a preliminary investigation of which specific depression-related symptoms can be linked to passive sensor data.
- Chapter 4 (Aim 1) builds off chapter 3 by specifically investigating how sleep quality measured via self-reports or by wearable devices and phones is related to depression symptomology. This work is intended to clarify whether physiologically measured sleep from the phone and watch measures the same construct as self-reported sleep quality in relation to depression and neurocognitive performance.
- Chapter 5 (Aim 2) describes the creation of a new anomaly detection method for detecting anomalies in passive sensor data and characterizes whether those shifts from normal behavior are related to changes in self-reported depression symptomology.
- Chapter 6 (Aim 3) looks at a finer temporal resolution of self-reported factors relevant to depression by examining ecological momentary assessment (EMA) responses related to reward functioning and affect. A machine learning modeling approach is taken to determine whether momentary affect and reward functioning can be detected with real-time sensor data and on which population subgroups the models work best.

The concluding section, Chapter 7, highlights the key contributions made to the field of digital sensing in mental health, the limitations of working in the space, and future directions to move the work presented forward.

Chapter 2: Related Work

The field of digital sensing in mental health is relatively new and rapidly evolving. Initial studies using consumer devices for depression monitoring focused on data collectable through smartphones, including location, calls, texts, voice, usage of device apps, websites, and keyboard interactions. Research-grade wearable devices such as the GeneActiv and Actiwatch wrist-worn actigraphy devices have been used for the passive, continuous assessment of sleep and activity patterns since the 1970s[18]. However, outside of sleep and activity, other sensors common to more recent consumer wearable devices, such as heart rate, heart rate variability, and respiratory rate, are not commonly utilized in studies monitoring depression, although there is a growing body of literature investigating heart rate variability and anxiety[19]. **Fig. 2.1** summarizes the key features from phone- and watch-based studies in relation to depression, where all items from wearable devices are used in this dissertation.

Notably, prior work in the field of digital sensing in mental health has tended to involve small sample sizes of fewer than 100 participants and short durations of up to two weeks[20]. In recent years, many smartphone-based studies have leveraged the widespread adoption of smartphones to collect larger sample sizes and longer durations of data than has been done earlier in the field, with many studies involving several hundred participants and lasting months to years[21]. A recent study by Zheng et al. using the All of Us dataset even looked at Fitbit data in 6,785 participants over a median of 4.5 years[22]. Some studies in the field of digital sensing for mental health incorporated wearable devices but often did not focus on metrics derived from them in analysis.

Depression prediction is a task often attempted by studies using passive sensing technologies. In these cases, depression is commonly determined via self-reports, such as a score above a threshold for the Patient Health Questionnaire-8 (PHQ-8)[23]. The performance varies by study, but the average classification area under the receiver operator curve (AUROC) values range from 0.6-0.7. For example, Melcher et al. reported that in their sample of 415 participants with phone data, a logistic regression model had an AUROC of 0.656[24], and Daniel et al., in a study of 112 adults over 2 weeks, were able to predict MDD with an AUROC of 0.72[25]. More often, studies have attempted to identify whether a specific metric or feature from smartphones and wearables is related to self-reported or clinician-assessed depression severity. The findings concerning important features for depression monitoring are summarized below.

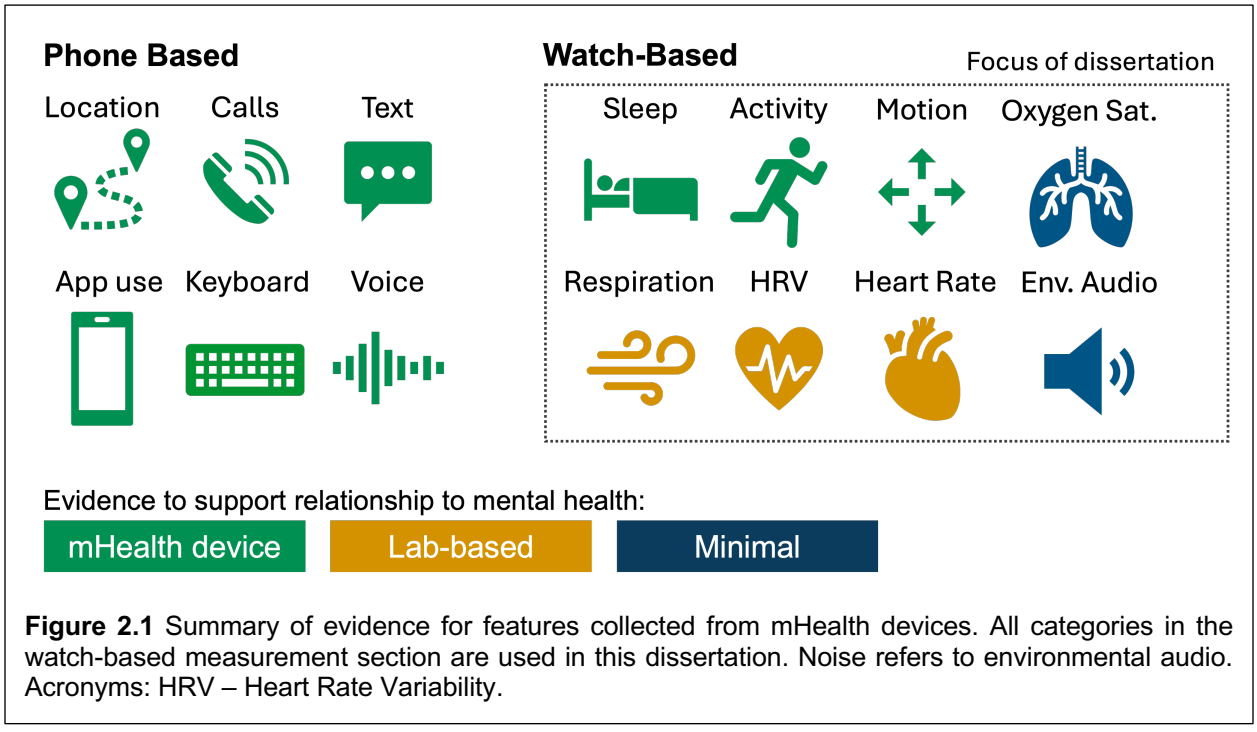

Figure 2.1 Summary of evidence for features collected from mHealth devices. All categories in the watch-based measurement section are used in this dissertation. Noise refers to environmental audio. Acronyms: HRV – Heart Rate Variability.

2.1 Smartphone-based features

2.1.1 Location

Location-based features from smartphones have been among the earliest indicators to show robust associations with depression severity. Their utility was established in 2015 by Saeb et al., who first defined the metrics of location entropy and location variance[26]. Studies tend to compare these metrics to PHQ-8 or PHQ-9 self-report surveys by investigating linear correlations between passively sensed features aggregated over 1–2 weeks and the total score:

- The exact findings from Saeb et al.'s original sample of 28 people[27], such as circadian rhythm-related GPS-derived features, have not been replicated in larger samples[28].
- For the metrics of location entropy and location variance, several larger studies with sample sizes of 1,013[29], 290 participants[30], 182[31], and 164[32] participants confirmed an association with self-reported depression severity.
- A review by Leaning et al. revealed that across studies, higher depression symptom scores are associated with less mobility, as measured via the GPS[21].
- Additionally, Xia et al. reported that the distinctiveness (how unique a pattern is relative to a full sample) of mobility patterns was associated with unstable affect[33].

2.1.2 Call and Text Data

Call and text data are often aggregated as the number of calls and texts an individual sends or receives and the number of unique contacts communicated with:

- Cao et al. reported that higher depression scores were correlated with fewer social interactions on the phone[34].
- Currey et al. reported that when assessing the correlation of mHealth features with depression severity, features from calls were among the most important[35].
- In a multilevel modeling approach with a large sample size of 1013 participants, Stamatis et al. did not find call or text data to be significant predictors of depression severity[29].

Call and text data may represent a feature whose utility may be in specific symptom detection rather than overall depression severity, as individuals largely differ in how they use these phone-based communication tools, especially in different age groups.

2.1.3 Smartphone Usage

Metrics related to smartphone usage, such as screen time, unlock duration, and app category usage, have had mixed efficacy in predicting or correlating with depression severity:

- Sverdlov et al. reported that communication app usage metrics such as “lower count of use” and lower entropy of usage time are correlated with depression severity[36].
- Sun et al. reported that unlock duration of smartphones was correlated with depression. However, Zou et al. reported that very few device use features related to depression severity[37], and Currey et al. reported that screen time was not predictive of depression severity[35].

2.1.4 Voice Data

Voice-derived features have been a promising area for depression biomarker derivation for decades, with early work looking at markers showing relationships with depression severity[38,39]:

- A recent review by Low et al. revealed that prosodic features such as decreases in perceived pitch and vocal range and features related to vocal vibrations such as jitter and shimmer increase with depression severity and psychomotor retardation[40].
- Wasserzug et al. reported that automated vocal depression scores could predict depression in 40 depressed, 104 nondepressed, and 14 participants in remission[41].

With new developments in machine learning and large language models, emotion and vocal features are likely to become more important and useful over time. Initiatives such as Bridge2AI-Voice, which are developing datasets to enable these models, will only further accelerate progress[42].

2.1.5 Keyboard Metrics

Keyboard-derived metrics have promising early data suggesting that they may work as digital biomarkers for cognitive functioning[43]. Several companies, including Mindstrong, which went out of business in 2023, started on this promise and linked these biomarkers to mental healthcare. Since these early results, more recent studies have demonstrated additional promising and more nuanced ideas of what metrics from keyboard use can tell us about mental health and depression, particularly when coupled with phone accelerometer data. Ning et al. reported that the combination of accelerometer and keyboard data was useful for understanding diurnal patterns in cognitive function, a finding that may translate to symptom tracking in mental health disorders[44]. Additionally,

Knol et al. reported that less phone movement during typing was related to increased levels of anhedonia[45].

2.2 Wearable Devices in Mental Health Monitoring

Most studies conducting digital sensing for mental health outcomes rely on smartphones. Those that include fitness trackers or smartwatches are often relatively small sample sizes given the frequent need to supply participants with wearable devices. Studies have focused on either consumer-facing devices or traditional research devices, with most earlier studies focusing on research actigraphy devices comprising a wrist worn accelerometer, gyroscope, and sometimes ambient light sensor. Data from different studies reveal that metrics from these wearable devices can augment our ability to detect depression symptoms beyond smartphone data alone[46].

2.2.1 Wrist Worn Actigraphy

Wrist-worn actigraphy devices have been commonly used in research to assess physical activity, sleep, and circadian rhythm and are considered the gold standard for assessing physical activity:

- In a meta-analysis, Wüthrich et al. reported that actigraphy-measured rest–activity rhythms were associated with depression severity[47], validating the idea that psychomotor slowing can be measured objectively with these devices.
- In a 359-person sample, Difrancesco et al. reported that objective, not subjective, sleep, activity, and circadian rhythm parameters changed in depressed or anxious individuals compared with healthy controls[48].

When investigating sleep specifically, polysomnography (PSG) is the gold standard used to compare sleep parameters such as efficiency, latency, duration, and sleep stages:

- Razjouyan et al. reported that wrist-based actigraphy may not be enough to capture sleep parameters, as it may miss core components of motion in comparison with chest-based devices[49]. However, in a recent white paper, Apple described the ability to capture both sleep timings and staging using accelerometer data from the watch alone[50]. The sensitivity and specificity of these algorithms are better than those of state-of-the-art algorithms that use other wrist-based devices[51].
- In a sample of 2,317 people, Glaus et al. reported significant associations with current MDD and sleep parameters measured by wrist actigraphy[52].
- In sample of 6,785 participants over a median of 4.5 years, Zheng et al. found a 1.75 odds ratio for MDD per hour increase in standard deviation of sleep duration[22]
- Sleep has also been determined with acceptable accuracy via metrics from smartphones alone, with a high Pearson correlation between predicted and self-reported sleep (r=0.83)[53].
- A comparison of self-reported sleep via daily sleep diaries or longer duration retrospective instruments revealed correlations between 0.40 and 0.68[54].

Common devices used in research settings for actigraphy include wrist worn devices from companies such as Empatica, ActiGraph, and GeneActiv. Devices from research wearable device companies commonly include accelerometers and gyroscopes. Proprietary software is used to calculate “activity counts,” a measure of activity, which is then further analyzed by researchers. The software used to calculate

these “activity counts” was recently made public[55] by ActiGraph. Other features common to research-grade devices include ambient light, and some newer research wearables also include measures of electrodermal activity, photoplethysmography (for measures such as heart rate and heart rate variability), and skin temperature readings[56,57].

2.2.2 Vital Signs

Elevated heart rate has been associated with depression severity[58] and may be a byproduct of stress and physical inactivity. Generally, cardiorespiratory fitness can be measured by combining features related to respiratory rate, heart rate, and physical activity from a wearable device.

Heart rate variability (HRV) is typically measured in a laboratory setting and is derived from electrocardiogram (ECG) readings. The intervals between specific peaks in the ECG waveform are analyzed and can provide insight into autonomic nervous system function. Early studies initially reported no difference in R–R interval variability between healthy controls and depressed individuals[59]. However, a body of evidence demonstrating that lower heart rate variability is observed in patients with severe MDD than in healthy controls is emerging[60]. However, this decrease can be modulated or obscured in the presence of psychopharmaceutical treatment[61].

The context in which HR and HRV are measured can influence the utility of the metrics derived. A systematic review by Schiweck et al. revealed that a reduced heart rate and HRV response to stress in validated tasks were found in those with MDD[62]. Wearable devices represent one way to find context-specific heart rates and HRVs. However, few studies have investigated wearable devices that measure heart rate

variability and depression. One example study modeling heart rate variability from an apple watch revealed an association between circadian HRV features and emotional support and resilience[63].

Respiratory rate is known to be related to mental stress and anxiety[64]. However, it is not often measured or reported in studies investigating mental health with mHealth devices. A lab-based study by Kral et al. revealed that a lower respiratory rate after mindfulness training was associated with better well-being[65], suggesting that respiratory rate from mHealth devices may be a useful biomarker in studies of mental health.

2.3 Analytic Methodologies Employed

The analysis of mHealth data for depression-related outcomes generally follows a framework in which passive health data are aggregated prior to self-reported or clinician-assessed measures of depression severity.

2.3.1 Feature Generation and Imputation

Typically, aggregation of features is performed daily (e.g., average daily steps) and then further aggregated into 2nd-order statistics over a period relevant to a particular outcome (e.g., standard deviation of average daily steps during the two weeks prior to a depression screening survey). Alternatively, features are aggregated only via 1st-order statistics over the full relevant timespan for a survey (e.g., average heart rate over the month prior to a retrospective sleep quality survey). This approach can lead to fewer features being generated, improving the interpretability and power of each feature. Imputation can be performed on missing passive sensor data in multivariate, univariate, or not at all. Many studies looking at passive mHealth data for depression detection

either do not disclose how missing data are handled or use simple median imputation[20]. When imputation is typically performed, there are criterion set per study for a minimum amount of information required before imputation will be considered[21]. Research investigating the effects of various imputation methods and their effects on downstream predictive tasks has revealed that simpler imputation methods may be less accurate in reconstructing the true data but are better for predictive modeling[66].

2.3.2 Modeling Approaches

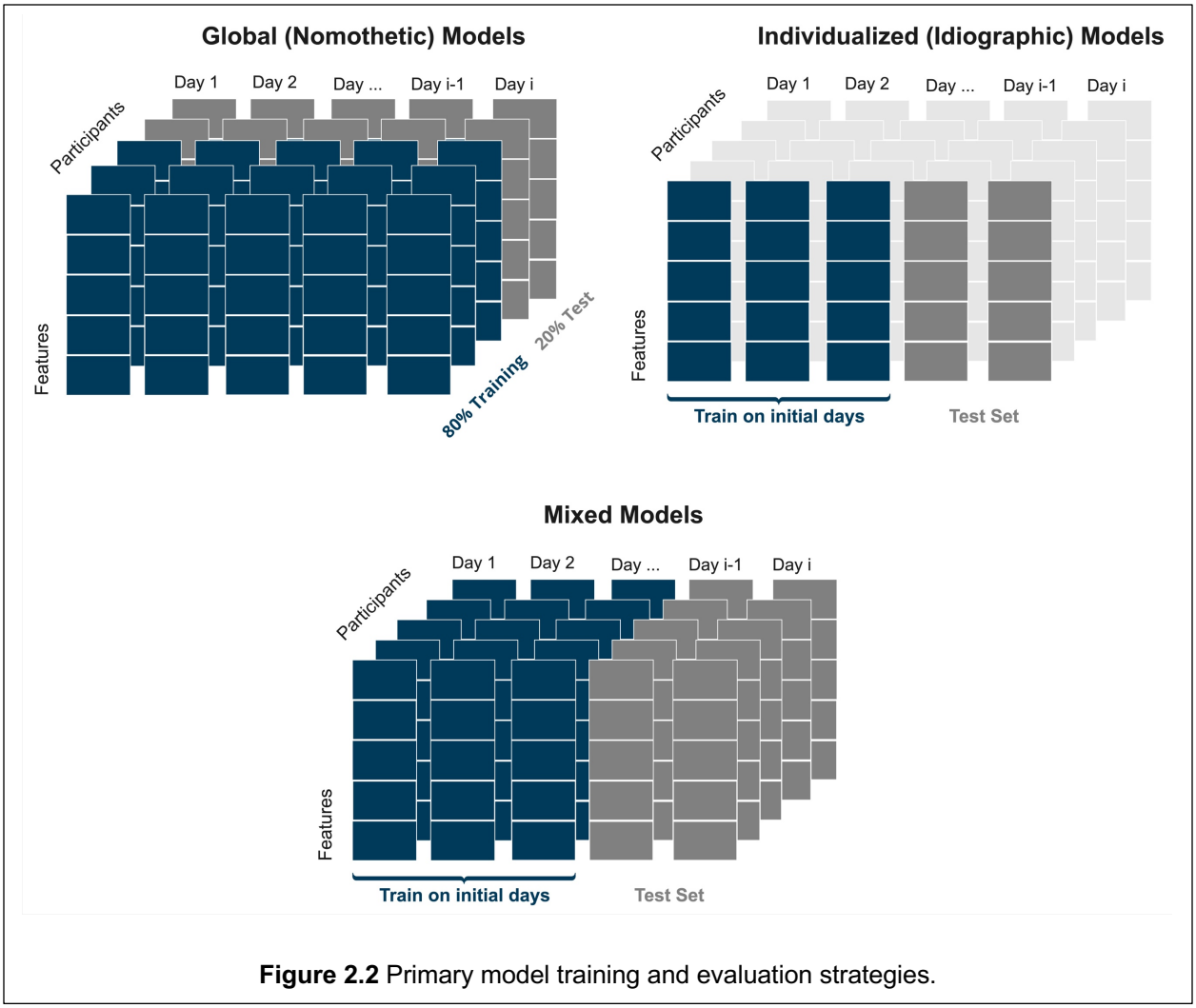

Figure 2.2 Primary model training and evaluation strategies.

A key promise of smartphones and wearable technology is that by leveraging the dense data streams they generate; we can quickly personalize models to detect symptomology and tailor treatments. In this spirit, several studies have shown that personalized modeling approaches, or idiographic modeling, improve the predictive performance for depression and mood detection, as outlined in **Fig. 2.2**[67–70].

Personalized modeling trains a model on a variable amount of data from a participant and uses that same participant's later data to evaluate performance. In contrast, global (or nomothetic) modeling separates individuals and trains models on one set of individuals; evaluating performance on a set the model has never seen. A third approach is mixed effect modeling, where an adjustment is made per individual and a separate model is used to find changes across individuals.

The higher performance of personalized as opposed to nomothetic models may be driven by the model's ability to "remember" the past responses of a participant. When predictive personalized models and mixed effects models are compared to simply predicting a participant's prior response, several studies have found past response to be the best predictor[71,72]. This observation may be driven by the highly identifiable nature of mHealth data[73], allowing models to quickly identify the individuals from whom a set of data comes and predict their prior response set. A more recent 2024 study by Balliu et al., however, revealed that individualized modeling improved predictive performance, even in comparison to participants' prior response[74].

Personalized modeling is aimed at addressing the incredible heterogeneity in the relationship between behaviors and changes in depression symptomology[75]. However, by leveraging nomothetic modeling, and identifying key participant characteristics that

are associated with differences in model performance to detect depression symptoms, we may be able to identify parameters that begin to explain that heterogeneity and improve our understanding and definition of depression.

2.3.3 Anomaly Detection

The data from smartphones and smartwatches are incredibly complex, being multimodal, being sampled dynamically and at different timepoints, and suffering from complex missingness. One way to leverage these data while minimizing assumptions is to use them to detect shifts in behavior. Anomaly detection (AD) represents a framework to look at multimodal data streams and determine if a given day or region is *different* than temporally proximate data points. This approach has not been used to successfully detect changes in the symptomology of depression. However, there has been success from Tourous' lab in linking anomalies to relapse of schizophrenic episodes in a 17-patient sample over 3 months[76], 126 participants over 3-12 months[77], and 132 participants from 3 international sites by Cohen et al.[78]. The study by Cohen et al. did find a relationship between passive sensor changepoints and measures of depression, anxiety, and sleep quality, suggesting that instead of online anomaly detection, the offline changepoint detection used may be of interest in unipolar depression.

Amor et al. investigated the use of AD on medical data streams and reported that principal component analysis (PCA)-based methods are effective and interpretable[79]. In addition, nonnegative matrix factorization (NMF) represents a potential improvement in the interpretability of PCA, especially for use in biology and other physiologic and clinical phenomena[80]. The interpretability of NMF over PCA is due to all components of

an NMF decomposition being unable to cancel each other out, as they can contain no negative values. This is demonstrated in **Fig. 2.3** Lee and Seung, where NMF and PCA were used to analyze facial images[1]. The NMF matrix decomposition of faces shows the isolation of individual facial features, whereas the PCA decomposition is more difficult to interpret.

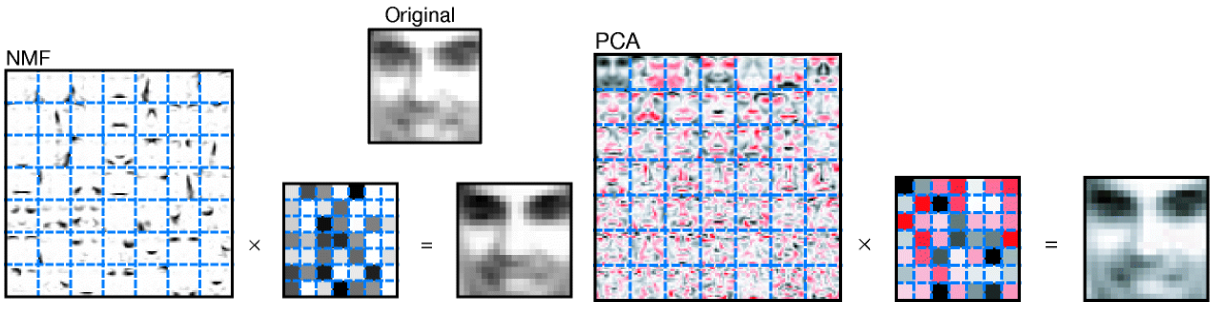

Figure 2.3 Comparison of non-negative matrix factorization (NMF) and principal component analysis (PCA) matrix interpretability adapted from figure 1 of Lee and Seung 1999[1].

2.4 The **O**perationalizing Digital **P**heno**T**yping **I**n the **M**easurement of **A**nhedonia (OPTIMA) Study

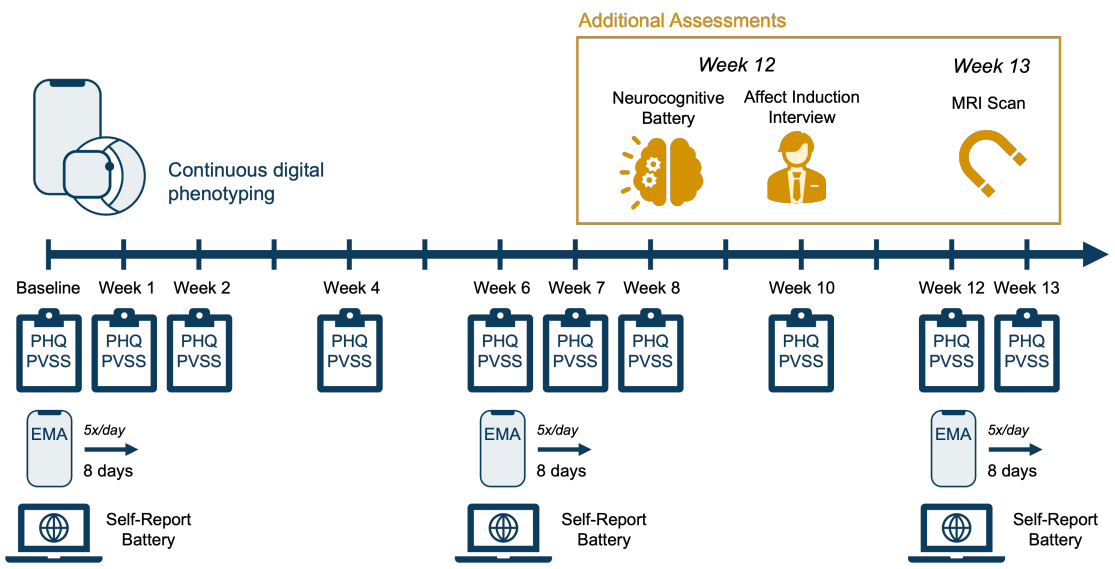

Figure 2.4 Overview of the Operationalizing Digital PhenoTyping In the Measurement of Anhedonia (OPTIMA) study data collection protocol.

The UCLA Depression Grand Challenge is conducting some of the largest studies leveraging mHealth devices to investigate mental health alongside traditional and investigational clinical assessments. The Operationalizing Digital PhenoTyping In the Measurement of Anhedonia (OPTIMA) study recruits participants with high or medium depression severity and high, medium, and low levels of anhedonia. As part of the study, extensive digital phenotyping data were collected from participants via their own iPhone and a study-provided Apple Watch series 7 or higher over the course of 13 weeks; data were collected between October 2022 and April 2024. Notably, the OPTIMA study specifically enriches for those with higher levels of anhedonia, a

symptom of depression that common serotonergic therapeutics often fail to improve[81] and is associated with worse mental health outcomes[82]. The OPTIMA study data collection design is shown in **Fig. 2.4.**

The UCLA Depression Grand Challenge Study App (DGC Study App) built by Avicenna Research is installed on participant iPhones and used to collect digital health data. The OPTIMA study collected data from 343 participants whose demographic data are detailed in **Table 2.1.**

Table 2.1 OPTIMA demographic and treatment history breakdown. One participant is missing baseline demographic data.

Demographics and Treatment History	**Yes**	**No**
Sex – Female	224 (65.5%)	118 (34.5%)
Family Income <100k	191 (55.8%)	151 (44.2%)
Non-Hispanic White	166 (48.5%)	176 (51.5%)
History of Psychotherapy	273 (80.1%)	68 (19.9%)
Depression diagnosis	231 (67.7%)	110 (32.3%)
Psychotherapy (in past 4 weeks)	151 (44.3%)	190 (55.7%)
Currently using medication for mental health	147 (43.1%)	194 (56.9%)
	Mean	**Std**
Age	33.48	12.11

Chapter 3: Detecting Symptoms of Depression with the iPhone and Apple Watch

This chapter is adapted from the paper: "Detection of Symptoms of Depression Using Data From the iPhone and Apple Watch," published in the 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)[83].

Digital health data from consumer wearable devices and smartphones have the potential to improve our understanding of mental illness. However, in conditions such as depression, there is not yet a consistent uniform measurement tool whose results can be reliably used as a gold standard measure of depression severity. This work seeks to specify what symptoms and dimensions of depression can be detected via vitals, activity, and sleep monitored by consumer wearable devices. Machine learning models are fit to digital health data and used to detect responses to individual questions from surveys (self-reports) as well as summary scores from these self-reports. For high-performing models, feature importance is investigated. The analysis was conducted on preliminary data from 99 participants in an ongoing study with data from the Apple Watch and iPhone along with validated self-reports relevant to depression severity, anhedonic severity, and sleep quality. The receiver operating characteristic area under the curve (ROC AUC) and average precision are used to assess model performance. The digital health sensor data investigated were found to significantly detect five of 74 measures, including overall depression severity and specific symptoms such as poor appetite, aspects of anhedonia, and sleep timings (ROC AUC between 0.63 and 0.72).

The features these models use in detection vary per detection task and suggest further areas for investigation to specify the right features to look at per symptom.

3.1 Introduction

Depression is a complex and heterogeneous condition in terms of symptom presentation, with 52 symptoms assessed across seven commonly used depression screening tools[84]. A core symptom of depression is anhedonia—the inability or reduced ability to feel pleasure. This symptom is present in some but not all people with depression. Digital health data from wearable devices represent another domain by which we may improve our characterization of depression, its subtypes, and symptoms through the continuous measurement of human behavior and physiology. Nonetheless, associations between single digital features and overall depression severity vary in direction and magnitude across studies, suggesting that overall depression severity may be too phenotypically heterogeneous to predict accurately. The goal of the current study is to move beyond overall depression severity scores to identify which features derived from wearable devices are related to which symptoms of depression and anhedonia[85].

By using machine learning techniques and interpretability methods, this analysis attempts to identify how simple features derived from wearable devices and smartphones may be associated with self-reported symptoms or dimensions of major depressive disorder. A machine learning model was trained to classify high- or low-item-level responses into participant self-reports, as well as subscale scores and total scores of standardized questionnaires that measure depression and related constructs. The importance of features from the best-performing models was then investigated to

understand which features from digital health sensors contribute to model performance and how they do so.

3.2 Methods

An overview of the analysis pipeline is shown in **Fig. 3.1.**

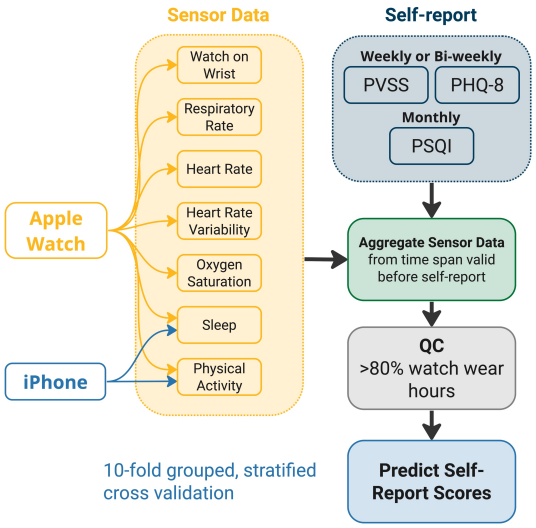

Figure 3.1 Flowchart overviewing methods used for to detect self-reported itemr esponse with passive sensor data

3.2.1 Dataset Overview

From the OPTIMA study, all collected data until July 7th 2023 are used in this analysis; as such, individual participants are contributing between 2 and 13 weeks of data. Participant characteristics of those enrolled in the study as of July 7th 2023 are shown in **Fig. 3.2,** note that this is a subset of participants prior to study completion.

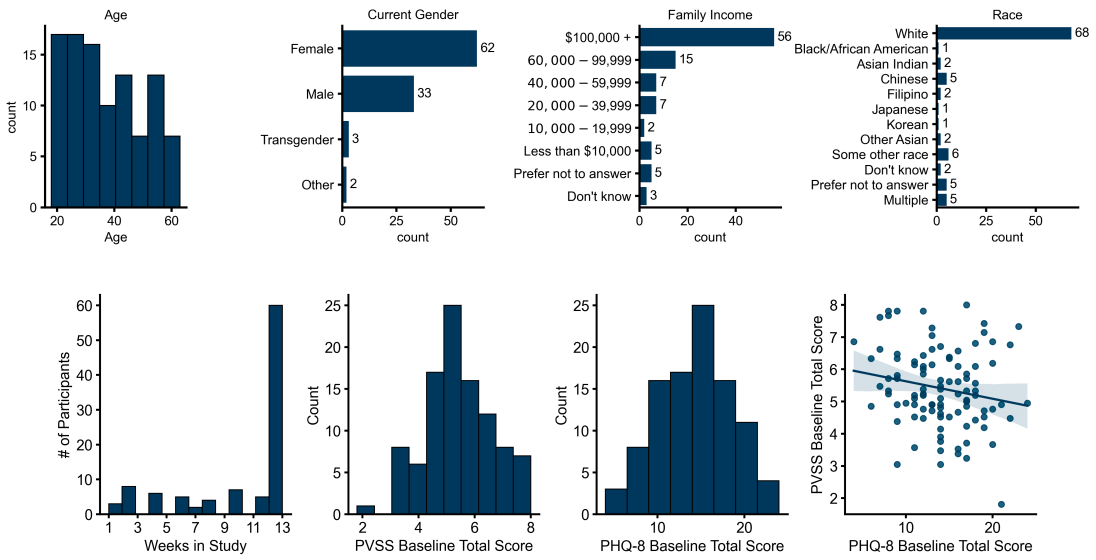

Figure 3.2 Top row: Distribution of demographic characteristics (age, current gender, family income, and race). Bottom row: time in study and self-reported anhedonia (PVSS score) and depression (PHQ-8 score) at the start of the study.

Self-report Measures

For this analysis, responses from three self-report questionnaires are investigated to compare digital health sensor data to depression, and anhedonia, and sleep quality. For all self-reports, baseline (Week 0) administration is excluded from analysis except to stratify model performance on participants with high or low levels of depression and anhedonia at the beginning of the study.

1. A modified version of the Patient Health Questionnaire depression scale 9 (PHQ-9) is used that has 14 total items. The PHQ-9 has high internal reliability (Cronbach's alpha = 0.89)[86]. Modifications include splitting compound symptoms (i.e., appetite decrease vs. overeating, sleep increase vs. decrease, psychomotor agitation vs. retardation, feeling down or depressed vs. feeling hopeless), adding two items to assess irritability and libido (i.e., "little interest in sex"), and removing the suicidality item. A recent IPD-MA demonstrated the equivalence of the PHQ-8 and PHQ-9 for

screening/diagnosis[87]. For this study, a total score representing the PHQ-8 was created by taking the max score of each pair of separated compound symptoms and excluding the two added items. The PHQ-8 is administered nine total times at Weeks 0, 1, 2, 4, 6, 7, 8, 10, and 12.

2. The Positive Valence Systems Scale (PVSS) short form is a 21-item measure of anhedonia that assesses pursuit of and response to a wide range of rewards (with seven reward-specific subscales: Food, Physical Touch, Outdoors, Positive Feedback, Hobbies, Social Interactions, Goals), and across six positive valence system domains (domain-specific subscales: Reward Valuation, Reward Expectancy, Effort Valuation, Reward Anticipation, Initial Responsiveness, Reward Satiation), as well as total score (mean response across all items)[88]. The PVSS has a high internal reliability with Cronbach's alpha between 0.91-0.94[88]. The 14 item PVSS is administered nine total times at Weeks 0, 1, 2, 4, 6, 7, 8, 10, and 12 (same as the PHQ-8).
3. The Pittsburgh Sleep Quality Index (PSQI) asks participants to rate their prior 1-month of sleep and assesses sleep quality and disturbances[89]. The PSQI asks 19 questions which are used to calculate seven subscales: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. The subscales are added to form a global score. The three PSQI domains (sleep duration, use of medication, and sleep quality), which are a discretization of just a single question, are excluded from analysis as they are redundant with the item response. This leaves 19 questions, four subscales, and one total score used as detection targets. The PSQI also has a

high internal reliability (Cronbach's alpha = 0.8)[90]. The PSQI is completed at Weeks 0, 6, and 12.

For the PHQ-8 and PVSS, there are two versions used across the study, one which asks participants to rate their past single week (7/9 administrations), and another that asks about the past two weeks (2/9 administrations). For all three questionnaires, item level responses, subscales, and total scores are converted to binary outcomes based on whether they are greater than the median response across collected data. The threshold for binary classification is modified per item to ensure that there is as close to an even class distribution as possible. On average across items, the true class (values greater than threshold) has 42.7% of responses (minimum 24.1%, maximum 66.2%).

The conversion to a binary classification target lowers our ability to interpret if models would be capable of detecting changes within a participant, especially if participants do not switch between the low or high class of the target variable. To understand how many participants transition from high to low class or vice versa during the study we look at each item per survey and calculate how many participants have responses in both the low and high class over the course of the study. For the PHQ-8 items on average 49% of participants have both a high or low response to items (min 31%, max 60%). For the PVSS items it is an average of 47% of participants (min 38%, max 56%). And for PSQI items, on average only 15% of participants transition class during the study (min 5%, max 26%).

Passive Sensor Features

Digital health sensor features were generated by aggregating sensor data prior to self-report administration. For the PSQI, participants are asked about their last month, so 28

days of digital health sensor data prior to the timestamp of administration are collected per participant response. For the PHQ-8 and PVSS, participants are asked about 7 or 14 days prior to the timestamp of administration based on whether participants were asked about the prior week or prior two weeks. All sensor data is collected utilizing Apple's HealthKit application programming interface (API). A total of 57 features across three categories are generated from sensor data expected to have high availability given expected watch and phone usage according to the study protocol.

1. Vitals (23 features): Respiratory rate, oxygen saturation, heart rate, and heart rate variability (HRV), as standard deviation of the N-to-N interval (SDNN), are filtered to be within plausible ranges (heart rate between 30 and 200bpm, HRV between 0 and 300ms, oxygen between 0 and 100%, respiratory rate between 0 and 40bpm). Vitals are then resampled to median value per hour to account for the dynamic sampling rate of metrics by the Apple Watch, resulting in periods with high or low rates of sample collection. Aggregation is done by taking the mean, median, standard deviation, minimum, and maximum of the given vital during the 1 week, 2 week or 28-day period prior to self-report. For heart rate skew, kurtosis, and the count of samples collected during the timespan are also calculated as it is the most frequently sampled of the vitals.
2. Sleep (26 features): Annotations of bedtime and sleep times from Apple Health annotations are used to calculate bedrest duration, sleep duration, sleep efficiency, and sleep onset latency every night. In combination with heart rate data, sleeping average heart rate and average heart rate variability are also calculated. The median,

minimum, maximum, and standard deviation of these day level features are aggregated across the relevant timespan as well as count of sleep and bedrest logs.

3. Activity (8 features): Apple HealthKit-reported exercise time, active energy burned, and step count are aggregated by collecting the total duration, sum of value, and count of logs. Sum of value is not calculated for exercise time as it is redundant with duration.

Watch wear hours are determined by the number of hours that participants have at least one heart rate log. Percentage of watch wear hours is calculated for the relevant span of data prior to a given self-report. Watch wear is not included as a predictive feature, but rather used for quality control. Self-reports with less than 80% of watch wear hours during the relevant timespan prior to assessment are removed from analysis.

3.2.2 Model Training and Evaluation

Classification models are trained to use the passive sensor features from Section 2C to detect high or low values of the self-report responses in Section 2B. All sensor data collected prior to the timestamp of a self-report administration is used to classify self-report response. Model training and testing is done so that individuals in the training data are not in the testing set.

Data Availability

At the time of analysis, the study has collected 788 responses to the PHQ-8 from 133 participants, 592 of those responses have digital health data that meet quality control measures described in Section 2C. After quality control, any survey response with missing or low-quality data is not considered in analysis; no imputation is done. For the PVSS, there are 786 responses from 133 participants, where 590 responses (99 participants) have sufficient digital health data. The PSQI has 152 responses from 91

participants with 120 responses (72 participants) having sufficient digital health data for analysis.

Model Training

A gradient boosting classifier (XGBoost) using the implementation from scikit-learn 1.2.2[91] with default parameters was utilized as the machine learning model. XGBoost is used because it is a high-performing model capable of learning nonlinear relations on sparse data in a variety of domains[92]. For each of the 74 questions, subscale-scores, and total scores from the PHQ-8, PVSS, and PSQI, a 10-fold grouped, stratified cross-validation was performed. This cross-validation ensures that participants in the training data are not in the testing data and vice versa while also approximately balancing levels of each class (response to item high or low) across the training and testing splits.

Model Evaluation

The metrics of area under the precision recall curve (known as AUPR or Average Precision) and receiver operator characteristic area under the curve (ROC AUC) are calculated for each fold of a cross-validation. ROC AUC is used as the primary performance metric, with Average Precision used to distinguish model performance where there may be heavy class imbalance.

To confirm if models are performing greater than by random chance, a 1-sided t-test was conducted on the ROC AUC value across the 10-folds to determine if average ROC AUC was greater than 0.5. As we are investigating performance of 74 models (one per item response, subscale score, or total score), Benjamini-Hochberg multiple testing correction is applied to control for the false discovery rate (FDR; α=0.05), and

models with a p-value < 0.05 after correction are examined further for feature importance and population subgroup performance.

The performance for items and subscales where models had statistically significant aggregate performance are examined for discrepancies in performance across subgroups. Subgroup evaluation is done across baseline PHQ-8 and PVSS scores as well as current gender identity (at study intake), family income, age, and ethnicity. No correction is done for multiple testing for investigating difference in performance across subgroups. An independent t-test comparing difference in ROC AUC scores is used to assess if performance differences across groups are statistically significant. Statistical testing is done using the Pingouin 0.5.3 python package[93].

Feature Importance

For models whose ROC AUC is significantly greater than 0.5 (FDR<0.05), feature importance is examined to investigate which features appear important to model performance and how they relate to model decision making. Feature importance was analyzed using SHapley Additive exPlanation (SHAP) scores[94]. SHAP feature importance scores help explain each individual prediction from a model, allowing researchers to understand how different feature values impact model decisions in the testing set.

3.3 Results

3.3.1 Overall Performance

Five models were shown to have ROC AUCs significantly greater than 0.5 after multiple testing correction. From the PHQ-8 models, models detected total score (ROC AUC =

0.64, p=0.029) and poor appetite (ROC AUC=0.67, p=0.016). From the PVSS models detected endorsement of, "*A fun activity during the weekend sustained my good mood*" (ROC AUC=0.63, p=0.026). From the PSQI, models detected bedtime (ROC AUC=0.72, p=0.009) and trouble staying awake (ROC AUC=0.71, p=0.029). Results outlined in **Table 3.1**.

Table 3.1 Performance of significantly performing models and class balance within the data

Survey	Item	Average Precision	ROC AUC	Adjusted p-value	% True Class
PHQ-8	PHQ-8 Total Score	0.6	0.64	0.029	0.45
	Poor appetite	0.59	0.67	0.016	0.49
PVSS	A fun activity during the weekend sustained my good mood	0.58	0.63	0.026	0.43
PSQI	Bedtime	0.8	0.72	0.009	0.56
_	Trouble staying awake	0.67	0.71	0.029	0.38

Performance Across Symptom Severity

For most items model performance did not significantly differ when evaluated across those with high vs. low PHQ-8 or PVSS total scores. For the PVSS item, "*A fun activity during the weekend sustained my good mood,*" there was a significant difference in those with a total PHQ score >14 having greater detection performance (p=0.006, ROC AUC 0.74 with high PHQ-8 score vs. 0.56 with low PHQ-8 score).

Performance across demographic characteristics

There were no significant differences in performance across ethnicity (Hispanic vs. non-Hispanic), current gender (male vs. female), or family income (>$100k vs. ≤ $100k annual income). Due to limited sample size, evaluation was not done across race or endorsed genders outside of male and female.

3.3.2 Feature Importance

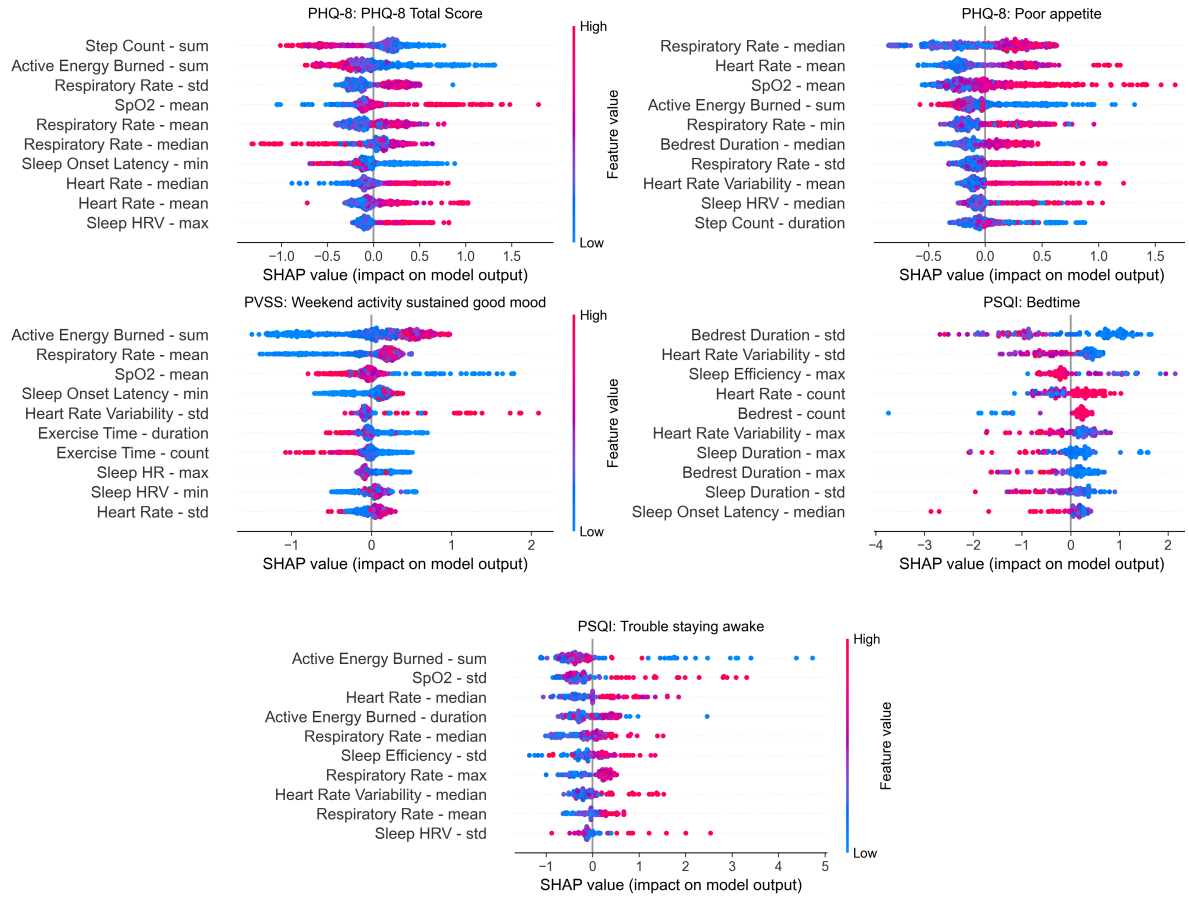

Figure 3.3 SHAP feature importance for top ten features per classification task. Each dot represents a a prediction within the test set of a cross-validation fold. Color indicates relative feature value (e.g, higher mean heart rate will be red, lower mean heart rates would be blue). SHAP value indicates how influential a feature was to a single prediction. The sign (positive or negative) of the SHAP value denote whether importance was for classifying as negative or positive class (low vs high class). Magnitude of the SHAP value indicates its importance to the model for a prediction.

The top ten features from all folds of cross validation were assessed for the five models with high performance. Features and SHAP feature importance scores are shown in **Fig. 3.3**. Physical activity related measures such as step count, and energy expenditure were the most important features across models. The exception being the model predicting the PSQI measure of bedtime which prioritized more sleep related features such as those related to bedrest duration.

3.4 Discussion

3.4.1 Findings

The PHQ-8 total score, a measure of depression severity, was detectable, with an average ROC AUC of 0.642. Features important to this model centered around physical activity, with higher levels of activity linked to lower total PHQ-8 scores. The model detecting the PHQ-8 item for poor appetite had slightly better performance than the total score. This model revealed that poor appetite was related to increased respiratory rate, heart rate, and oxygen saturation and decreased physical activity. The directional association of these features with poor appetite suggests that in the presence of low physical activity, increases in those vital signs may be indicative of poor appetite. However, the XGBoost models were not able to detect overeating, an item normally not separated from poor eating, highlighting the utility in separating out these two directions when screening for symptoms of depression.

The primary feature associated with detecting the PVSS item, "*A fun activity during the weekend sustained my good mood*" was active energy expenditure, with higher values corresponding to the positive class. Interestingly, this item's model ranked active energy expenditure highly but did not rank step count as an important feature, in contrast to the model predicting depression severity (PHQ-8 total score). Step count may act as a proxy for location entropy (a metric derived from GPS data), one of the most consistently related digital health features linked to depression severity[20]. An area for future work would be to determine whether location entropy is related to anhedonia or if it is specific to depression severity.

Notably, only one item from the PVSS was detectable via the digital health data streams. Given the limited set of features centered on sleep, vitals, and activity, incorporating features such as patterns of app usage, screen time, and social connectedness from the phone may improve our ability to predict domains represented by the PVSS.

Compared with the PHQ-8 and the PVSS, the models trained to detect items from the PSQI performed better. For bedtime detection, metrics related to sleep duration were most important. In contrast, energy expenditure and vitals related features were prioritized when detecting trouble staying awake.

The models were unable to detect the PSQI item asking about overall sleep quality (mean ROC AUC 0.509), suggesting that the metrics used in this analysis can detect objective measures but are insufficient for subjective aspects of sleep known to be related to depression[95].

3.4.2 Limitations

In this initial analysis, for model training and evaluation, there is no accounting for time or repeated measures. Similarly, the act of converting ordinal and numeric values into binary categories reduces the model's ability to detect smaller changes in item response. While binary classification is done to simplify the model detection task and gain a preliminary understanding of how digital health features may detect self-reported responses, it obfuscates more nuanced shifts in response items and interpretability of results. With a larger sample size, multilevel modeling or regression will be utilized.

The study population is relatively homogenous in demographic characteristics across racial distributions and income and is restricted to participants from the Los Angeles area.

Furthermore, data is filtered for those with high compliance to study protocol and device wear (>80% of hours). Consequently, findings from this analysis may not generalize to other populations.

The analysis conducted is with data from an ongoing study. As such the recruited sample to date is not representative of the full final study cohort. Due to limited sample size no holdout test set was designated and only cross-validation was used to assess model performance leading to potential for over-fitting.

The sleep features used in analysis are derived from sleep phase annotations reported by Apple HealthKit and not derived from raw data. While relying on these annotations is convenient, it limits the generalizability of findings both across different devices and sleep annotation algorithms.

3.4.3 Conclusion

Personal sensing data from consumer phones and wearable devices show promise for improving our understanding of mental and physical health. However, adequately characterizing how data from these devices fit into the larger landscape of research on depression requires us to move beyond the detection of summary scores for depression severity.

This work demonstrates that there are specific features from personal sensing data that relate to dimensions of depression that are different than overall depression detection. Moreover, investigations of feature importance in machine learning models can help uncover potential nonlinear relationships between passive sensor data that are worthy of further exploration.

Features related to vital, activity, and sleep appear insufficient to detect most aspects of anhedonia measured by the PVSS, suggesting potential value in increasing the types of features derived from watch and phone data. Measures of physical activity and vital signs are important for models predicting depression severity, whereas features more directly tied to sleep performance are important for the detection of sleep timing. Vitals during sleep (e.g., average heart rate during sleep) were used by models detecting poor appetite and sustained mood due to weekend activity.

With increased participant populations and more domains of passively sensed features, approaches leveraging machine learning techniques show promise in showing how real-world participant-generated data can improve our understanding of depression.

Chapter 4: Comparison of self-reported and physiological sleep quality from consumer devices to depression and neurocognitive performance

This chapter is adapted from the manuscript: "Comparison of self-reported and physiological sleep quality from consumer devices to depression and neurocognitive performance" under review at NPJ Digital Medicine (submitted July 19, 2024).

This study examines the relationship between self-reported and physiologically measured sleep quality in individuals with depression and its impact on neurocognitive performance. Using data from 249 participants with medium to high depression monitored over 13 weeks, sleep quality was assessed via retrospective self-reports and physiological measures from consumer smartphones and smartwatches. The correlations between self-reported and physiological sleep measures were generally weak. Machine learning models revealed that self-reported sleep quality could detect all depression symptoms measured via the Patient Health Questionnaire-14, whereas physiological measures detected only "sleeping too much" and low libido. Notably, only self-reported sleep disturbances correlated significantly with neurocognitive performance. Physiological sleep was able to detect changes in the self-reported sleep quality domains of sleep medication use and sleep latency. These findings emphasize that self-reported and physiological sleep quality do not measure the same construct and that both are important for monitoring sleep quality in relation to depression.

4.1 Introduction

Sleep disturbance is a core symptom of depressive episodes, and sleep disorders commonly co-occur with major depressive disorder (MDD)[96]. While polysomnography is the gold standard for assessing sleep quality, it is challenging to use in naturalistic settings or over the extended periods of time typical of depressive episodes. Actigraphy from wrist-worn research-grade devices, which has been compared to polysomnography[51], represents one alternative that is deployable in naturalistic settings. However, these research actigraphy devices are not as easy to use or as prevalent as their consumer wearable device counterparts[15,16]. Researchers thus often rely on self-reported sleep quality, using tools such as daily sleep diaries or retrospective questionnaires such as the Pittsburgh Sleep Quality Index (PSQI), and find that changes in self-reported sleep quality are associated with changes in depression severity[97].

However, there is low concordance between self-reported and physiological sleep measurements[98,99], especially in groups with greater depression severity[54]. Differences between self-reported and physiological sleep measures (i.e., misappraised poor sleep) are associated with worse neurocognitive functioning[100], and decreased neurocognitive functioning is associated with depression severity[101]. This evidence suggests the need to better understand how physiological and self-reported subjective sleep quality measures relate to each other and to both self-reported depression and neurocognitive measures.

Prior work investigating physiological and self-reported sleep quality has typically utilized a single night of physiologically measured sleep, although some studies have

measured up to nine nights of sleep[54,102]. With growing evidence for the comparability of sleep-related metrics from the gold standard polysomnography and actigraphy with consumer wearable devices[50,103,104], we can now leverage large sample sizes over longer durations for studying physiological sleep quality in relation to depression.

In the present report, we studied the relationship between physiological and self-reported measures of sleep quality in populations with depression. The dataset used comes from 342 participants who were monitored via passive sensing and self-reported measures over 13 weeks as part of a larger trial investigating features of anhedonic depression (Wellcome Leap MPsych). A subset of 249 participants were used in this analysis on the basis of the high availability of sleep annotation data from the iPhone and Apple watch. From the passively sensed sleep annotation data, sleep quality features are extracted per night (sleep duration, bedrest duration, onset, efficiency, latency, etc.). Self-reported depression symptoms are taken from individual items in the Patient Health Questionnaire-14 (PHQ-14), and subjective sleep quality is assessed via the Pittsburgh Sleep Quality Index (PSQI).

First, we investigated correlations between self-reported sleep quality measures that have a direct correspondence with physiological sleep quality measurements (sleep duration, bedrest onset, sleep offset, sleep latency, sleep efficiency, and nightly awakenings) and found weak correlations ($|r| < 0.5$) for all measures except wakeup time ($r=0.78$). Second, building from our prior findings[83] and extending prior work associating depression with self-reported sleep quality, we use either physiological measurements of sleep quality from smartphones and wearable devices or self-reported sleep quality from the Pittsburgh Sleep Quality Index (PSQI) to detect self-reported

symptoms of depression. All 14 measured depression symptoms were detectable via self-reported sleep quality, whereas only “sleeping too much” and “little interest in sex” were detectable with respect to physiological sleep quality. Third, as there is a tautological connection between self-reported sleep quality and self-reported depression severity (i.e., both forms of data are acquired via the same method with overlap in item topic, which could artifactually correlate these measures to a greater degree compared to correlating methodologically mismatched measures, such as self-report questionnaires and physiological measurements), we investigate how both subjective and objective measures of sleep quality correlate with neurocognitive performance as a proxy for an area impacted by depression (i.e., neurocognitive performance is measured methodologically differently from both self-report measures and physiological measures). We found that only self-reported sleep disturbances were significantly correlated with any measured aspect of neurocognitive performance. Finally, to determine how physiological measurements may enable the detection of changes in subjective sleep quality, physiological sleep quality is used to detect future self-reported sleep quality (controlling for self-reported sleep quality at the time of physiological sleep quality measurement). We find that physiological sleep can enable models to detect changes in self-reported sleep medication use and daytime dysfunction due to sleepiness.

These findings underscore that self-reported and physiologically measured sleep quality do not measure the same construct. Self-reported sleep quality cannot be readily substituted with physiological measurements, as it remains significantly related to key constructs in depression research, such as neurocognitive performance. Furthermore,

the utilization of passive physiological sleep monitoring provides a valuable tool for bridging gaps between self-reported assessments, facilitating the detection of notable changes in self-reported sleep quality. This integrative approach holds promise for advancing our understanding of sleep-related processes in depression and enhancing the precision of future studies.

4.2 Results

The data used in this analysis are from the Operationalizing Digital PhenoTyping in the Measurement of Anhedonia (OPTIMA) study, which collected data between October 2022 and April 2024. OPTIMA aims to measure behaviors related to anhedonia in the context of depression, relating observations to neural markers of anhedonia. This analysis uses data from 249 participants comparing their physiologically measured sleep quality to self-reported measures of sleep quality, depression, and TestMyBrain-based[105] neurocognitive performance. Participant demographics are outlined in **Table 4.1.**

Table 4.1 Participant demographics and treatment history. There were 249 participants used in the analysis; however, one participant was missing baseline assessments, including demographic data.

Demographics and Treatment History	Yes	No
Sex - Female	160 (64.5%)	88 (35.5%)
Family Income <100k	125 (50.4%)	123 (49.6%)
Non-Hispanic White	120 (48.4%)	128 (51.6%)
History of Psychotherapy	201 (81.0%)	47 (19.0%)
Depression diagnosis	172 (69.4%)	76 (30.6%)
Psychotherapy (in past 4 weeks)	112 (45.2%)	136 (54.8%)
Currently using medication for mental health	116 (46.8%)	132 (53.2%)
	Mean	**Std**
Age	33.78	11.86

In the study population, 83.5% (207) of 249 participants had moderate to severe depression at baseline, per PHQ-14 total score greater than or equal to 10. Note that the PHQ-14 is a self-report questionnaire adapted from the Patient Health Questionnaire (PHQ)[86] by Cohen, Cohen, & Fried (see Depression Symptom Response Project OSF site: https://osf.io/j6r3q/) that disentangles confounded items from the PHQ-9 (e.g., “Trouble falling or staying asleep, or sleeping too much” is split into two items: “Trouble falling or staying asleep” and “sleeping too much”) and adds two symptoms (libido and irritability).

4.2.1 Physiological Sleep Quality Correlation with Self-reported Sleep Quality

Physiological sleep parameters measured over 28 days were correlated with six self-reported items or domains from the PSQI that had direct correspondences from 224 participants with 369 total responses. Differences in the number of self-reported

responses used are due to insufficient sleep annotation data prior to self-reports for some assessments. The mean of physiological and self-reported sleep quality was taken for those with multiple responses. Among the six self-reported sleep quality items, three were significantly correlated with physiological sleep parameters after Bonferroni correction, with family-wise error rate of 0.05 (**Fig. 4.1. A-C**): sleep duration (Spearman's r=0.37, p=1.21e-07), bedtime (Spearman's r=-0.35, p=4.23e-07), and wakeup time (Spearman's r=0.78, p=1.36e-45). Habitual sleep efficiency, time to fall asleep, and nightly awakenings were not significantly correlated with their physiological sleep counterparts (**Fig. 4.1. D-F).** The correlation between self-reported and physiological sleep duration of 0.37 is lower than that reported by Matthews et al. in their full population (r=0.40) but greater than that reported in the highest quartile of depression symptoms (r=0.23)[54].

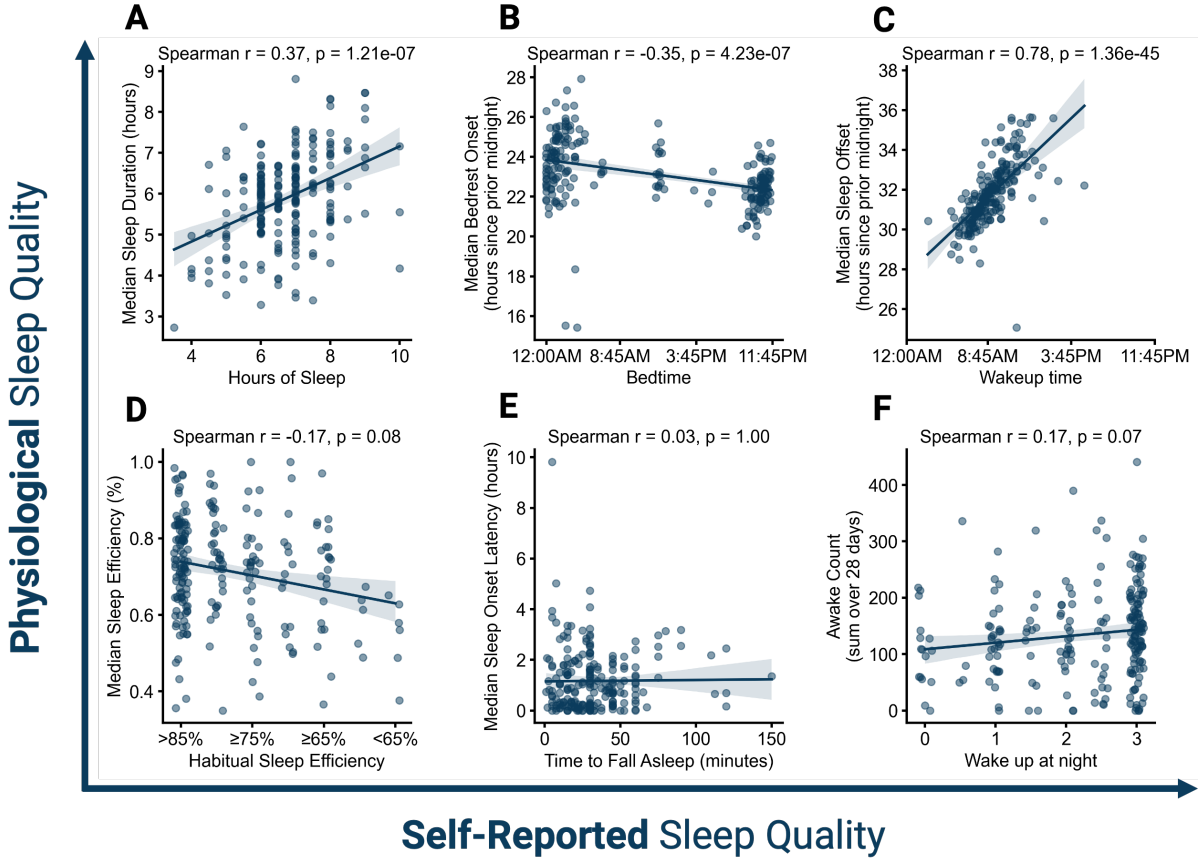

Figure 4.1 Correlation between physiological sleep parameters (Y-axes) and self-reported sleep quality items or domains (X-axes). Comparisons are of **A**) of median sleep duration to self-reported hours of sleep, **B)** median bedrest onset to self-reported bed-time, **C)** median sleep onset to self-reported wakeup time, **D)** median sleep efficiency to self-reported habitual sleep efficiency, **E)** median sleep onset latency to self-reported time to fall asleep, and **F)** number of recorded night time awakenings to self-reported nightly awakenings where response is defined as: 0, Not during the past month; 1, Less than once a week; 2, Once or twice a week; 3, Three or more times a week.

4.2.2 Detecting Depression Symptoms with Self-reported or Physiological Sleep Quality

Machine learning models were trained to detect symptoms of depression from the PHQ-14 (**Fig. 4.2.A**) using either self-reported sleep quality (all individual items, domains, and total score) from the PSQI taken on the same day as the PHQ-14 (performance in **Fig.**

4.2.B) or using physiological sleep parameters taken from the 8 days prior to PHQ-14 administration (performance in **Fig. 4.2.C**). The models were trained to detect the presence (or absence) of each PHQ-14-measured symptom (item response >0). Tenfold cross-validation was used to generate a distribution of model performance via the area under the receiver operator curve (AUROC). In cross-validation, no data from individuals in the test set are present in the training set. The performance of the models in the prediction of PHQ-14 symptoms was tested to determine which symptoms could be detected more accurately than random chance (AUROC>0.5). Correction for multiple comparisons was performed via the Benjamini–Hochberg method, with a false discovery rate (FDR) of 0.05. The design and validation of these models are described further in the **Methods**.

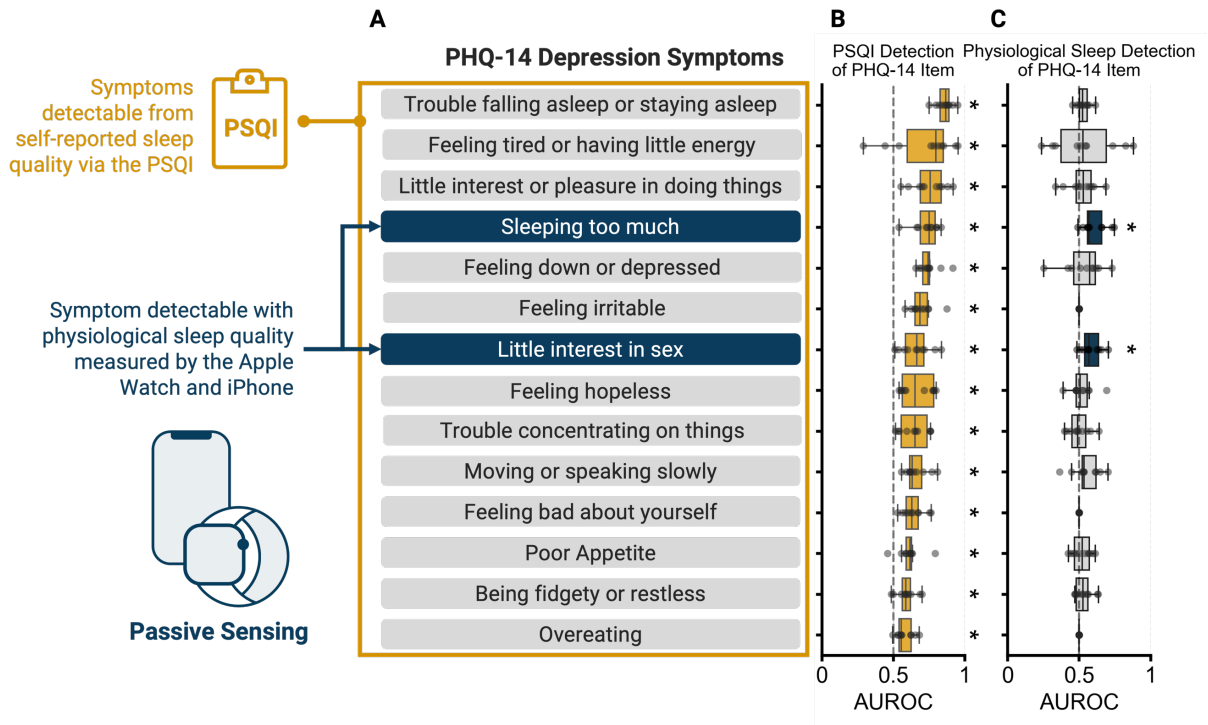

Figure 4.2 A) Symptoms of depression listed by order of relevance to self-reported sleep quality. Symptoms within yellow box are detectable with self-reported sleep quality. Dark blue colored symptoms are detectable with physiological sleep data beyond random chance. **B)** Performance of machine learning models to classify presence or absence of PHQ-14 symptoms using item-level PSQI responses. **C)** Performance of machine learning models to classify presence of PHQ-14 symptoms using physiological sleep features. If AUROC is greater than 0.50 with a p-value < 0.05 after correcting for FDR < 0.05 a star (*) is annotated and the associated box is colored dark gold or dark blue.

Of the 14 self-reported depression symptoms assessed on the PHQ-14, all were detectable above random chance via item-level responses to the PSQI. In contrast, only the symptoms “sleeping too much” (median AUROC=0.568, Wilcoxon signed rank test FDR-adjusted p=0.029) and “little interest in sex” (median AUROC=0.568, Wilcoxon signed rank test FDR-adjusted p=0.041) were detectable above random chance using physiological sleep quality. Model performance is further described in **Table 4.2.**

Table 4.2 Model performance for predicting PHQ-14 item responses with either self-reported sleep quality or physiological sleep quality. LR = logistic regression, RF = random forest, GB = gradient boosting, D = dummy, underline = FDR adjusted p-value < 0.05 for model AUROC > 0.5.

	PSQI (n=249 users, 705 responses)			Physiological Sleep (n=247 users, 1565 responses)		
PHQ-14 Item	**Model**	**Median AUROC**	**Adjusted P-value**	**Model**	**Median AUROC**	**Adjusted P-value**
Trouble falling asleep or staying asleep	RF	0.867	0.015	GB	0.523	0.686
Feeling tired or having little energy	GB	0.798	0.015	LR	0.526	1.000
Little interest or pleasure in doing things	RF	0.758	0.015	RF	0.530	1.000
Sleeping too much	LR	0.751	0.015	RF	0.568	0.029
Feeling down, depressed	LR	0.743	0.015	RF	0.574	1.000
Feeling irritable	LR	0.685	0.015	D	0.500	1.000
Little interest in sex	LR	0.665	0.015	LR	0.568	0.041
Feeling hopeless	LR	0.652	0.015	GB	0.506	1.000
Trouble concentrating on things	RF	0.651	0.015	LR	0.490	1.000
Moving or speaking slowly	LR	0.631	0.015	GB	0.534	0.961
Feeling bad about yourself	LR	0.629	0.015	D	0.500	1.000
Poor appetite	LR	0.616	0.015	LR	0.521	1.000
Being fidgety or restless	LR	0.588	0.015	GB	0.525	1.000
Overeating	LR	0.556	0.015	D	0.500	1.000

For models that use physiological sleep, whose AUROC is significantly greater than 0.5 (FDR <0.05), feature importance is examined to investigate which features are important for model performance and how they are related to model decision making. Feature importance was analyzed via SHapley Additive exPlanation (SHAP) scores[94]. SHAP feature importance scores help explain each individual prediction from a model, allowing researchers to understand how different feature values impact model decisions in the testing set. We find that higher sleep offset and bedrest offset times with longer bedrest durations and quieter bedtime environments are used by the model to detect "sleeping too much." While maximum sleep duration is also a highly ranked feature by the model, its association with self-reported sleeping too much does not appear linear

(i.e., higher max sleep duration is not consistently used by models to detect sleeping too much; **Fig. 4.3.A**). The model for "little interest in sex" revealed that low sleep efficiency, less variable sleep onset, and longer duration of bedrest were associated with the depression symptom (**Fig. 4.3.B**).

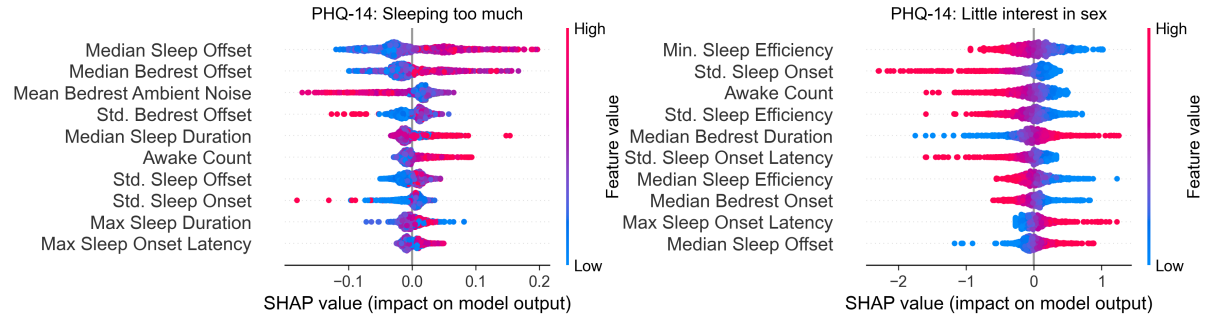

Figure 4.3 Feature importance via SHAP value for models predicting PHQ-14 items **A)** "Sleeping too much" and **B)** "Little interest in sex". Each dot represents a prediction within the test set of a cross-validation fold. Color indicates relative feature value (e.g, higher mean heart rate is red, lower mean heart rates is blue). SHAP value indicates how influential a feature was to a single prediction. The sign (positive or negative) of the SHAP value denote whether importance was for classifying as negative or positive class (low vs high class). Magnitude of the SHAP value indicates its importance to the model for a prediction.

For models using physiological sleep to detect PHQ-14 item responses that had performance greater than random chance, performance was compared across baseline depression severity (PHQ-14 total score ≥10), baseline anhedonia (PVSS total score <5), family income (≥100k USD), race (non-Hispanic white vs. all), and sex at birth. No significant differences in AUROC were found in performance between groups.

4.2.3 Physiological and Self-reported Sleep Correlation to Neurocognitive Performance

Neurocognitive performance was measured via TestMyBrain: a series of computer-based tasks or games. Measures from TestMyBrain were correlated with self-reported

sleep quality domains from the PSQI taken at the same day and physiological sleep parameters aggregated over the prior 8 days (**Fig. 4.4)**. Only the sleep disturbances domain was found to correlate with a neurocognitive performance measure of processing speed, the Digit Symbol Coding (DSC) test rate correct score (Spearman's r=-0.29, FDR adjusted p-value=0.009).

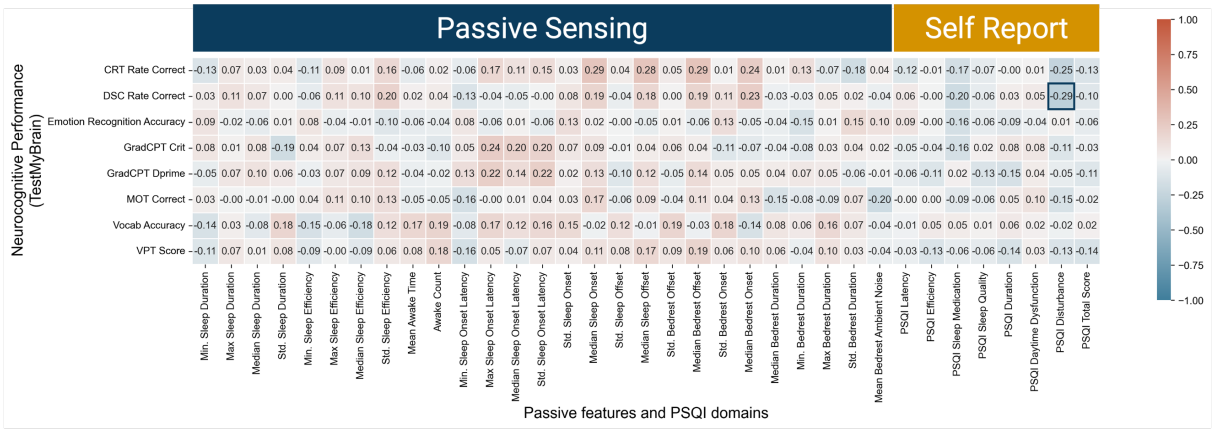

Figure 4.4 Spearman correlation of watch-derived median features over a 28-day period and PSQI domains and total score with TestMyBrain neurocognitive performance metrics. Correlations with a p-value < 0.05 after multiple testing correction (FDR < 0.05) have a box drawn around them.

In contrast to Gualtieri et al.[101], we did not find a significant correlation between depression severity (PHQ-14 total score) and the tested measures of neurocognitive performance (FDR-adjusted p-values all > 0.05). This finding may be driven in part by the small number of participants with low depression severity in the OPTIMA study.

4.2.4 Detecting Changes in Self-reported Sleep Quality with Physiological Sleep Data

Several domains of self-reported sleep quality can be detected with physiological sleep alone. These domains are daytime dysfunction due to sleepiness (median

AUROC=0.602, FDR-adjusted p=0.008) and sleep duration (median AUROC=0.632, FDR-adjusted p=0.021).

To assess whether physiological sleep quality can enable the detection of changes in self-reported sleep quality, models were trained using only physiological sleep quality, physiological and 6-week prior self-reported answers, and only prior self-reported answers. For two domains, i.e., daytime dysfunction due to sleepiness and sleep medication, performance improved when passive data were added (**Fig. 4.5**), with median AUPRC differences with and without passive data = 0.046 (FDR-adjusted p=0.008) and 0.064 (FDR-adjusted p=0.039), respectively, indicating that physiological data can help models better detect the positive class without sacrificing overall performance. No significant difference was observed in performance, as measured via the AUROC.

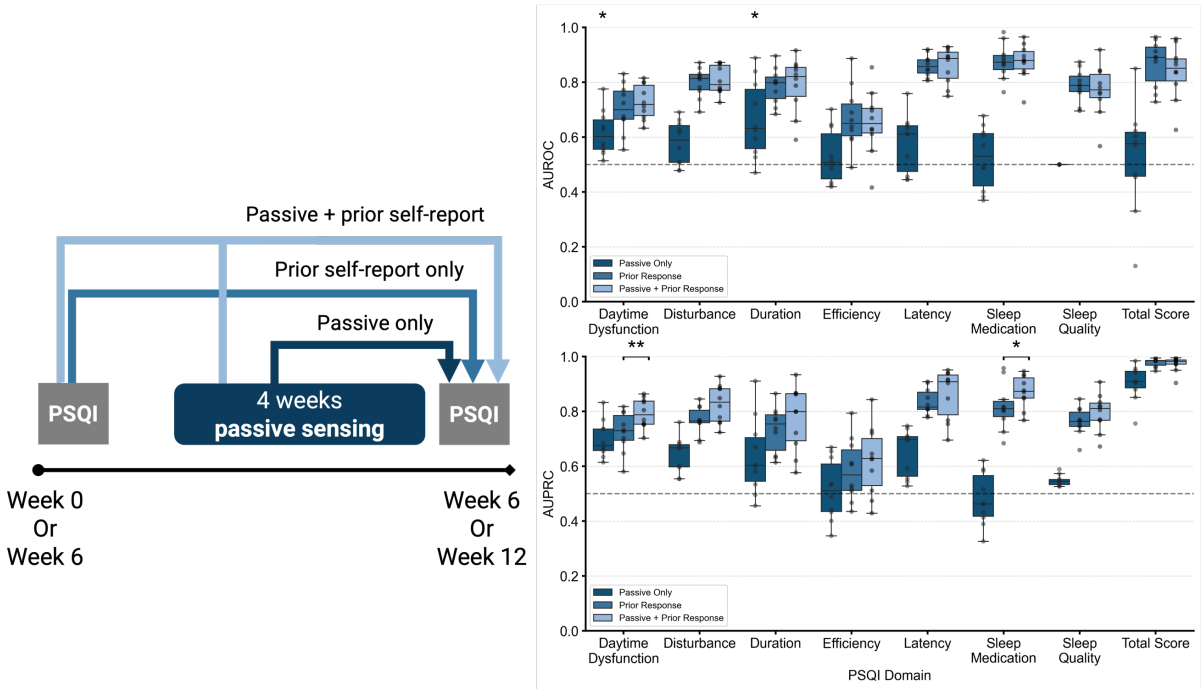

Figure 4.5 Performance of predictive models at detecting domains and total scores of the Pittsburgh Sleep Quality Index (PSQI) with different input feature sets. Top plot shows AUROC score value; if p-value for passive-only detection greater than random chance is < 0.05 after controlling FDR < 0.05, a star is annotated. Bottom plot shows average precision performance; if passive data with prior response improves performance over just prior response, with a p <0.05, it is annotated with a star (*) or two stars if p <0.01 (**).

4.3 Discussion

The finding that all 14 symptoms of depression from the PHQ-14 are detectable via self-reported sleep quality highlights the interconnections among different self-reported measures related to depression. While sleep quality-related items are part of the PHQ-14 (e.g., sleeping too much), many other items, such as “little interest or pleasure in doing things,” are not directly related to sleep and yet are detectable from self-reported sleep quality with good performance (AUROC values above 0.75). In comparison, physiological sleep was only able to detect the PHQ items of “sleeping too much” and “little interest in sex.” This observation suggests that physiological sleep is not a sufficient detector for most depressive symptoms that may require other social behavior

features measurable from phones and watches, as illustrated in prior work[20,21,26,74,106–108].

The model detecting “little interest in sex” is associated with low sleep efficiency but high bedrest duration with the endorsement of symptoms. Prior studies have revealed associations between self-reported sleep duration and libido[109,110]. However, the modeling work performed here suggests that in cohorts with depression, the relationship between libido and sleep quality may be driven more by sleep efficiency and bedrest duration than by sleep duration. The detection of “sleeping too much” as opposed to “trouble falling or staying asleep” also shows the benefit of separating traditionally combined symptoms on self-report scales. “Trouble falling or staying asleep” may be associated with rumination or bouts of wakefulness during sleep, whereas “sleeping too much” is more likely to be reflected in later sleep and bedrest offset times that can be objectively measured, a finding reflected in the top features used by the model to detect “sleeping too much.”

As a validation of consumer wearable-based sleep assessment, we also confirmed the expected ability to detect self-reported sleep duration as measured by the PSQI. Prior work has established this correlation[54], which is also observed in the OPTIMA dataset.

4.3.1 Relationship with Neurocognitive Test Performance (TestMyBrain)

Zavecz et al. reported no relationship between overall subjective sleep quality and working memory[111]. We did not find any significant correlations between either overall subjective sleep quality (PSQI total score) or physiological measures of sleep duration, efficiency or latency and neurocognitive performance. However, we did find that higher

levels of self-reported sleep disturbances were correlated with decreased processing speed.

4.3.2 Combining Self-Reported and Physiological Data to Detect Sleep Quality Domains

In traditional mental healthcare, patient interactions with clinical professionals or the completion of self-reports occur infrequently. By employing continuous passive monitoring, we can detect changes in self-reported sleep medication use and sleep latency, allowing for the identification of significant variations in these areas. These findings illustrate how integrating self-reported and objectively measured sleep quality parameters can enhance patient care.

In the analysis presented in **Fig. 4.5**, increases in model performance as measured by AUPRC occur when adding passive sensing data to self-reported sleep quality, although the same does not occur for the AUROC. An improvement in AUPRC without a significant change in AUROC suggests that the model improvement is centered on better identifying the positive class. For example, it can be important to detect endorsement of sleep medication use, as it serves as an objective index of having experienced sleep problems; incorporating passive sensing data can significantly enhance the detection of sleep medication usage.

4.3.3 Overall Clinical Utility

The data reported here suggest that physiological sleep from consumer devices could augment how self-reported sleep quality measures are interpreted and enable continuous passive sensing of sleep quality-relevant symptomology over time.

Ultimately, passive data could one day be used to recommend that users with specific digital phenotype profiles seek a mental healthcare professional for further evaluation (e.g., to complete self-report questionnaires, obtain diagnostic evaluations, or receive treatment). This approach could greatly increase the user's awareness of their own functioning, increase the number of people who receive treatment, and reduce the delay between symptom onset and treatment initiation. The data could also be used by the user and healthcare professional to monitor treatment progress and mechanisms, as well as inform when treatment termination is advised or if lapses/relapses occur.

4.3.4 Limitations

Sleep stage annotation (rapid-eye movement, deep, etc.) data are available for participants, but throughout the duration of the study, it is uncertain how changes in operating system versions and updates to annotation algorithms from Apple HealthKit may have influenced the comparability of these annotations. For that reason, this work centers on measures such as sleep efficiency, duration, latency and others that rely only on a differentiation of sleep vs. bed rest. Similarly, of the 342 participants in the study, only 249 had any sleep annotation data available from Apple HealthKit, limiting the sample size used in this work. This missing data was caused by a dependency on setting up approximate bedtime and intended wakeup times within iOS, which were not known at the onset of the study but were included within onboarding instructions once discovered.

It is important to contextualize these results in the population represented by the parent study, which recruited participants with medium to severe depression and across the full spectrum of anhedonia severity, resulting in a symptomatically distinct

population with the goal of understanding anhedonic depression. The models trained and tested here perform differently on a sample that is more representative of the American population or even a population of participants with depression. Additionally, there is a heavy skew in the socioeconomic characteristics of this population, with 50.4% of participants having annual family incomes >$100K USD.

As shown in **Table 4.3**, several PHQ-14 items like “feeling tired or having little energy” have class imbalances (97% positive class) when converted to binary outcomes in this study population. This imbalance likely contributes to being unable to predict presence or absence of the symptom greater than random chance. One potential remedy for future analyses would be useful to identify what a meaningful difference in item response is per participant.

Table 4.3 Distribution of self-reported survey responses within OPTIMA study, including baseline assessments.

Survey	Question	Mean	Std	Threshold	% True	# Responses
PHQ-14	Total Score	12.78	4.92	10	73%	2162
	Little interest or pleasure in doing things	1.53	0.84	1	91%	2162
	Trouble concentrating on things	1.66	0.96	1	87%	2162
	Moving or speaking slowly	0.34	0.69	1	24%	2162
	Being fidgety or restless	0.83	0.96	1	52%	2162
	Feeling irritable	1.5	0.9	1	88%	2162
	Little interest in sex	1.6	1.12	1	78%	2080
	Feeling down, depressed	1.57	0.86	1	92%	2162
	Feeling hopeless	1.19	0.95	1	74%	2162
	Trouble falling asleep or staying asleep	1.64	1.06	1	83%	2162
	Sleeping too much	0.96	1.03	1	56%	2162
	Feeling tired or having little energy	2.13	0.86	1	97%	2162
	Poor appetite	0.87	0.95	1	56%	2162
	Overeating	1.01	1.01	1	60%	2162
	Feeling bad about yourself	1.51	1	1	83%	2162
PSQI	Total Score	8.89	3.46	5	90%	705
	Daytime Dysfunction	1.7	0.71	2	61%	705
	Disturbance	1.58	0.63	2	54%	705
	Duration	0.7	0.85	1	50%	705
	Efficiency	0.69	0.97	1	41%	705
	Latency	1.67	1.02	2	57%	705
	Sleep Medication	0.9	1.22	1	41%	705
	Sleep Quality	1.64	0.72	2	57%	705

4.3.5 Conclusion

With improvements to consumer phone and wearable device measurements, the field of digital sensing in mental health is positioned to leverage these devices for larger longitudinal assessments of physiological sleep quality. This work shows that physiological sleep parameters can augment self-reported sleep quality to more thoroughly characterize how sleep quality is related to depression severity.

Furthermore, our work highlights that in populations with higher levels of depression, there is larger discordance between self-reported sleep quality measures and objectively monitored sleep and physiology, emphasizing the need to investigate both in research studies aiming to use sleep to characterize depression.

4.4 Methods

4.4.1 Dataset and Study Description

The parent study recruited participants with high or medium depression severity and high, medium, and low levels of anhedonia. As part of the study, there is extensive digital phenotyping data collected from participants using their own iPhone and a study-provided Apple Watch series 7 or higher over the course of 13 weeks. All collected data until February 6th, 2024, are used in this analysis; as such, individual participants are contributing between 2-13 weeks of data. Participant characteristics of those enrolled in the study are shown in **Table 4.1**. The UCLA Depression Grand Challenge Study App (DGC Study App) built by Avicenna Research is installed on participant iPhones and used to collect digital health data. The DGC Study App uses HealthKit and SensorKit APIs for passive measures and deploys ecological momentary assessments (EMAs).

4.4.2 Self-report measures

For this analysis, responses from two self-report questionnaires are investigated to compare digital health sensor data to depression and sleep quality.

1. A modified version of the Patient Health Questionnaire Depression Scale 9 (PHQ-9) is used that has 14 total items, referred to as the PHQ-14. The PHQ-9 has high internal reliability (Cronbach's alpha = 0.89)[86]. Modifications include splitting

compound symptoms (i.e., appetite decrease vs. overeating, sleep increase vs. decrease, psychomotor agitation vs. retardation, feeling down or depressed vs. feeling hopeless), adding two items to assess irritability and libido (i.e., “little interest in sex”), and removing the suicidality item. A recent individual participant data meta-analysis (IPD-MA) demonstrated the equivalence of the PHQ-8 and PHQ-9 for screening/diagnosis[87]. For this study, a total score representing the PHQ-8 was created by taking the max score of each pair of separated compound symptoms and excluding the two added items, where higher scores indicate greater depression. The PHQ-14 is administered 9 total times at Weeks 0, 1, 2, 4, 6, 7, 8, 10, and 12.

2. The Pittsburgh Sleep Quality Index (PSQI) asks participants to rate their prior 1-month of sleep and assesses sleep quality and disturbances[89]. The PSQI asks 19 questions which are used to calculate 7 subscales: subjective sleep quality, sleep latency, sleep duration, habitual sleep efficiency, sleep disturbances, use of sleeping medication, and daytime dysfunction. The subscales are added to form a global score, where higher scores indicate worse sleep. The 7 subscales and total score are used as detection targets. The PSQI has a high internal reliability (Cronbach’s alpha = 0.8)[90]. The PSQI is completed at Weeks 0, 6, and 12.

For the PHQ-14 there are two versions used across the study, one which asks participants to rate their past single week (7/9 administrations), and another that asks about the past two weeks (2/9 administrations). For all questionnaires, item level responses, subscales (also referred to as domains), and total scores are converted to binary outcomes enabling binary classification machine learning models to be trained. For the PHQ-14, questions are binarized to endorsing or not endorsing a symptom, and

total score is converted to binary based on a score ≥10 to match depression screening guidelines[23]. The PSQI domains and total scores are converted to binary as follows:

- Daytime Dysfunction**:** Requires a cause of daytime dysfunction weekly
- Disturbances: Requires at least 3 causes of sleep disturbance >=3 times a week
- Duration**:** Requires <7 hours of sleep on average. Picked because recommendations for adults is 7-9 hours
- Efficiency (HSE): Requires sleep efficiency <85%. Picked because recommendations for adults is >80%
- Latency**:** Requires >=30 minutes to fall asleep less than once a week AND average time to sleep >=15 minutes
- Sleep Medication: Any vs no sleep medication
- Sleep Quality**:** Good vs bad distinction (scores of 3 and 2 correspond to very bad and fairly bad)
- PSQI total score ≥ 5 is used for sleep disorder screening[89].

Class balance for PHQ-14 and PSQI items is shown in **Table 4.3**.

4.4.3 TestMyBrain Measures

TestMyBrain includes a set of standardized computer-based tasks used to assess neurocognitive performance in several domains[105,112]. OPTIMA's protocol utilizes a subset of these assessments, the digit symbol coding (DSC) test, choice reaction time (CRT) test, multiple object tracking (MOT) test, emotional recognition test (ERT), vocabulary accuracy test (VAT), gradual onset continuous performance test (GradCPT), and verbal paired association (VPT) test. From each of these tests the output metrics suggested for use by TestMyBrain are calculated:

1. Accuracy (for VAT and ERT) is the proportion of correct responses per test in a trial
2. Rate correct score (for DSC and CRT) is a combined metric of speed and accuracy. It is calculated as $1000 * \frac{Accuracy}{median\ Reaction\ Time}$
3. D-prime (d'; for GradCPT) is a measure of user's sensitivity and is reported from TestMyBrain
4. Crit (for GradCPT) is a measure user bias reported from TestMyBrain
5. Correct (for MOT) is the proportion of correctly identified targets during the task

4.4.4 Physiological Sleep Feature Generation

Digital health sensor features were generated by aggregating sensor data prior to self-report administration. For the PSQI, participants are asked about their last month, so 28 days of digital health sensor data prior to the timestamp of administration are collected per participant response. For the PHQ-14 participants are asked about 7 or 14 days prior to the timestamp of administration, however, to make input features comparable 8 days of sensor data are aggregated prior to assessment both when participants are asked about the last week or 2 weeks. A timespan of 8-days has been shown by Sun et al., to be a useful minimum span when using smartphone data to predict the PHQ-8[113]. All sensor data is collected utilizing Apple's HealthKit application programming interface (API). A total of 27 physiological sleep features are generated. Distributions of features after aggregation prior to PHQ-14 and PSQI assessments in the supplementary materials.

Annotations of bedtime and sleep times from Apple Health annotations are used to calculate bedrest duration (time in bed), sleep duration, sleep efficiency, sleep onset latency, and night awake time each day between 3pm the day prior to 3pm the day of

metric reporting. These values are aggregated over the 28- or 8-day period prior to PSQI or PHQ-14 administration. Sleep duration, bedrest duration, sleep efficiency, and sleep onset latency are aggregated by taking the minimum, maximum, median, and standard deviation per day. Nightly awakenings are aggregated as the mean hours and count of awakenings. Sleep onset, sleep offset, bedrest onset, and bedrest offset are aggregated by taking the median and standard deviation. Noise during sleep is aggregated as the mean noise during bedrest periods over the aggregation time window.

Watch wear hours are determined by the number of hours that participants have at least one heart rate log. Percentage of watch wear hours is calculated for the relevant span of data prior to a given self-report. Watch wear is not included as a predictive feature, but rather used for quality control. Self-reports with less than 80% of watch wear hours during the relevant timespan prior to assessment are removed from analysis. If participants did not set their sleep schedules via the phone operating system, automatic sleep detection would not occur and there would be no sleep annotations for a participant even if the watch was worn during sleep. For the purposes of this analysis which centers on sleep quality measurements, records without sleep annotation data are removed. Missing data is described in **Table 4.4.**

Table 4.4 Availability of physiological sleep parameters aggregated prior to self-report administration of the PSQI and PHQ-14.

	PSQI (28-day)			PHQ-14 (8-day)		
Passive Sensing Feature	**Count**	**Missing**	**% Missing**	**Count**	**Missing**	**% Missing**
Min. Sleep Duration	364	0	0%	1565	0	0%
Max Sleep Duration	364	0	0%	1565	0	0%
Median Sleep Duration	364	0	0%	1565	0	0%
Std. Sleep Duration	355	9	3%	1503	62	4%
Min. Sleep Efficiency	364	0	0%	1565	0	0%
Max Sleep Efficiency	364	0	0%	1565	0	0%
Median Sleep Efficiency	364	0	0%	1565	0	0%
Std. Sleep Efficiency	355	9	3%	1503	62	4%
Mean Awake Time	364	0	0%	1565	0	0%
Awake Count	364	0	0%	1565	0	0%
Min. Sleep Onset Latency	364	0	0%	1565	0	0%
Max Sleep Onset Latency	364	0	0%	1565	0	0%
Median Sleep Onset Latency	364	0	0%	1565	0	0%
Std. Sleep Onset Latency	355	9	3%	1503	62	4%
Std. Sleep Onset	355	9	3%	1503	62	4%
Median Sleep Onset	364	0	0%	1565	0	0%
Std. Sleep Offset	355	9	3%	1503	62	4%
Median Sleep Offset	364	0	0%	1565	0	0%
Std. Bedrest Offset	362	2	1%	1551	14	1%
Median Bedrest Offset	364	0	0%	1565	0	0%
Std. Bedrest Onset	362	2	1%	1551	14	1%
Median Bedrest Onset	364	0	0%	1565	0	0%
Median Bedrest Duration	364	0	0%	1565	0	0%
Min. Bedrest Duration	364	0	0%	1565	0	0%
Max Bedrest Duration	364	0	0%	1565	0	0%
Std. Bedrest Duration	362	2	1%	1551	14	1%
Mean Bedrest Ambient Noise	301	63	21%	1298	267	21%

4.4.5 Correlation Analysis

To compare watch-derived sleep and self-reported sleep with neurocognitive performance, watch features were aggregated from 28-days prior starting from the

timestamp of the PSQI administration taken the day participants performed the TestMyBrain assessment. In this way features from the watch are representing the same timespan that the self-report is intended to measure. Each watch feature and PSQI domain is correlated with each of the TestMyBrain performance measures using a Spearman correlation and p-values are adjusted using the Benjamini–Hochberg method, controlling the false discovery rate correction (FDR; at α=0.05).

4.4.6 Machine Learning Pipeline

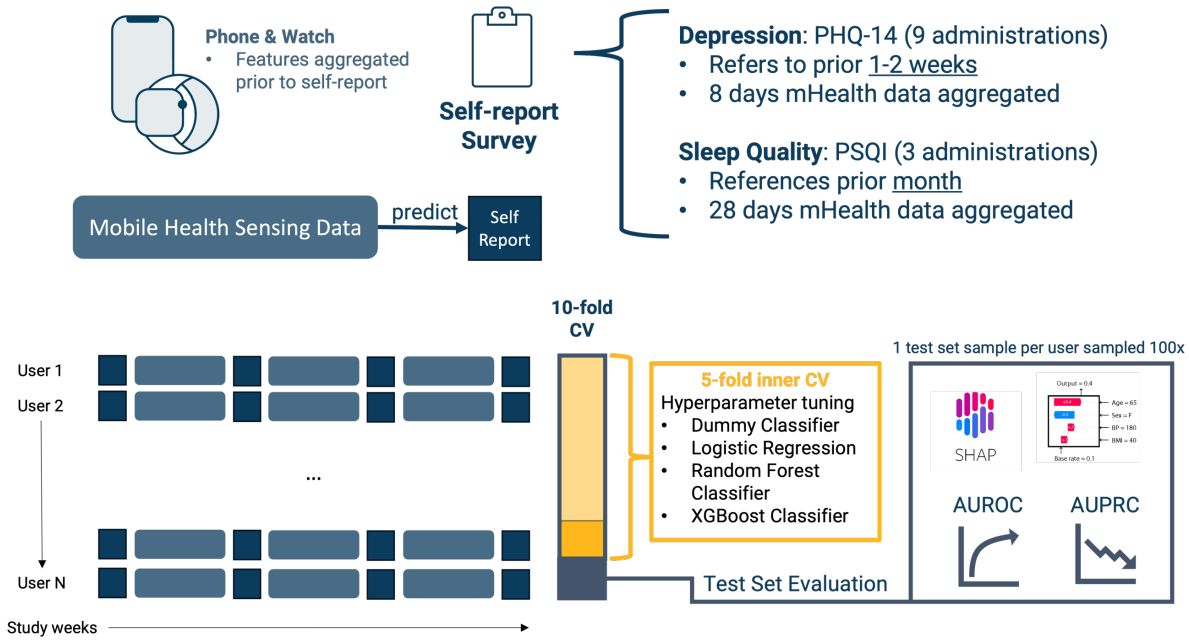

Figure 4.6 Machine learning pipeline for using physiological sleep data to detect self-report items and summary scores.

Classification models are trained to use the passive sensor features, prior survey response, or both to classify survey item and total score responses; an overview of the machine learning methods is shown in **Fig. 4.6**. All sensor data collected eight days (for PHQ-14) or 28 days (for PSQI) prior to the timestamp of a self-report administration is used to classify self-report response. Although the PHQ-14 asks participants about

either their prior week or two weeks, eight days of aggregation is chosen based on findings from Sun et al. predicting PHQ-8[113]. Models are trained in a way that splits participants used in training from those used to evaluate a model. Given how uniquely identifiable individuals are from their vitals taken by mobile health devices[73], if the participant level split is not done, models may simply learn to predict a participant training set score.

Of the 342 OPTIMA participants who completed the study, 249 have sleep annotation data from the Apple watch and are included in this analysis. All 249 participants with any sleep annotation data have PHQ responses totaling 1850 assessments (excluding baseline assessment). After quality control for watch wear covering at least 80% of hours for 8 days prior, 1565 PHQ-14 responses from 247 participants are used in analysis. There are 242 participants with PSQI responses totaling 457 PSQI assessments. After quality control (80% of hours for 28 days prior) there are 365 PSQI responses from 221 participants are used in analysis. The above numbers do not include baseline or intake assessments as they were at the beginning of the study and do not have associated physiological sleep data. Participants have up to 8 responses to the PHQ-14 with associated sensor data, and up to 2 responses to the PSQI. When using PSQI item level responses to predict PHQ-14 symptoms, baseline assessments are included allowing for 705 responses from 249 participants used in the machine learning modeling.

The 249-participant subset was selected for presence of sleep annotation data, other missing data elements in the physiological sleep data are filled in using median

imputation within cross-validation folds as described below. No imputation is done of self-reported measures.

For each classification task k-fold nested cross-validation was done to select the best model and hyperparameter combination using a random search cross-validation approach. Internal cross-validation folds were created using a standard 5-fold cross-validation. The outer fold was a stratified group 10-fold cross-validation keeping participants split across train and test sets. The pipeline for model training comprised median imputation, variance thresholding (features must have >0 variance in train set), feature selection, and robust scaling (5 to 95th percentile) before features reached the classification model. All implementation comes from standard functions in scikit-learn 1.2.2[91].

Models used were the gradient boosting classifier (XGBoost), random forest (RF) classifier, logistic regression, and a dummy classifier (predicts mean of train data responses) to act as a baseline. For RF and XGBoost classifiers, hyperparameters were the number of estimators (100, 200, or 500), max depth (5, 10, or 20), minimum samples per leaf (2, 5, 10), max features (none, square root of total, log base 2 of total). All models other than the dummy classifier also had parameters for select-K-best feature selection. These were number of feature (10 or all), and feature selection scoring function (mutual information based or F-score based).

4.4.7 Statistical Analysis

The metrics of area under the precision recall curve (known as AUPRC or average precision) and receiver operator characteristic area under the curve (AUROC) are calculated for each fold of a cross-validation. AUROC is used as the primary

performance metric, with AUPRC used to distinguish model performance where there may be heavy class imbalance. To account for differing number of repeated measurements in the test set per individual, one test set sample per individual is drawn 100 times per fold. For each test set fold AUROC and AUPRC are calculated.

To confirm if models using passive data only are performing greater than by random chance, a 1-sided Wilcoxon signed rank test was conducted on the AUROC value across the 10-folds to determine if median AUROC was greater than 0.5. As we are investigating performance of 22 models (one per item response, domain score, or total score), Benjamini–Hochberg method is applied per survey to control for the false discovery rate (FDR; α=0.05), and models with a p-value less than 0.05 after correction are examined further. For the PSQI to determine if passive data can enable detection of changed response with data on a participant's prior response, a paired t-test comparing if models with the feature "prior response only" perform worse than "prior response with passive data" was performed for each PSQI domain and total score on both AUPRC and AUROC. Multiple testing correction is applied to results with a FDR of 0.05. Performance of models with the addition of passive data was considered significantly better than prior response only if the adjusted p-value was less than 0.05.

The performance for items and subscales where models had statistically significant aggregate performance are examined for discrepancies in performance across subgroups. Subgroup evaluation is done across baseline depression severity (PHQ-14 total score ≥ 10) and anhedonia (PVSS total score < 5), family income (≥ 100k USD), race (non-Hispanic white vs all), and sex at birth. No correction is done for multiple testing for investigating difference in performance across subgroups. To assess

performance differences across groups, each test set fold is separated based on participants belonging to one of the groups. To account for repeated samples per participant, 1 sample is taken per participant 100 time before calculating AUROC per fold. Mann-Whitney U-test is performed to determine if median AUROC is different across groups. All statistical testing is done using the Pingouin package version 0.5.3[93] in Python version 3.11.6.

Code availability

The code that supports the findings of this study is available online at https://github.com/akre96/OPTIMA_sleep_quality

Chapter 5: Reconstruction error-based anomaly detection to detect changes in depressive symptoms

Major depressive disorder (MDD) is not fully understood using existing clinical tools and standards. Mobile health (mHealth) data, such as those from smartphones and wearable devices, may help fill the gap in our understanding of MDD. However, these devices generate large, dense amounts of data with high variance across users and with myriad reasons for periods of missingness. Anomaly detection (AD) algorithms can help filter and highlight data points worth investigating further to understand how mHealth data can generate insights into mental health. Here, we characterize how AD algorithms work on a simulated dataset and apply them to the GLOBEM and OPTIMA datasets. We find that in some settings, detected anomalies are slightly correlated ($\rho < 0.4$) with measures of interest related to mental health trajectories. However, we do not find a relationship between detected anomalies and symptoms of depression in either dataset.

5.1 Introduction

Changes in health-related behaviors are known to correspond to changes in mental health and depression[114]. Wearable devices and smartphones can track health-related behaviors such as physical activity and sleep but generate many health-related metrics, and when consumer devices are used, events such as software version updates or device differences can limit our ability to interpret a given metric across individuals and over time[115].

Anomaly detection (AD) algorithms show promise for finding valuable and personalized insights from an individual's mHealth data by discovering deviations from expected behaviors without making strong assumptions on which metrics are input and what they truly represent. For example, such algorithms have been demonstrated to detect anomalies more frequently prior to schizophrenia relapse[76]. However, it is unclear how to interpret the results of AD algorithms for mHealth datasets, what types of anomalous behavior they detect, and how they perform relative to one another. Additional work is needed in understanding how these algorithms perform on data with varying underlying generation patterns to confirm that these methods work in surfacing data points worthy of further investigation.

Building on these earlier efforts, this work hypothesizes that the number of anomalies detected in a period is correlated with changes in one's mental health; the hypothesis is based on the premise that deviation from one's normal behavioral patterns may indicate changes in mental health. To verify this premise, we investigate how anomaly detection algorithms perform under a variety of situations with simulated data given different data generating processes (number of features, relationship between features, frequency of anomalous days, and autocorrelation of features). The AD methods are then applied to two different observational datasets with mHealth data and depression measures. The GLOBEM mHealth dataset is used to compare the number of anomalies detected to measurements of depression and anxiety (PHQ-4, BDI-II), and the OPTIMA dataset is used in comparison to measures of depression and anhedonia (PHQ-14 and PVSS, respectively). To validate that anomaly detectors appear to emphasize features expected to be relevant, we find that the wake time after sleep onset (WASO) is

prioritized in the detection of sleep disturbance from the Pittsburgh Sleep Quality Index (PSQI). We find that in the simulation, the anomaly detectors tested work best when the input features are uncorrelated, and the anomaly is represented simultaneously in a higher percentage of the features. The simulation findings suggest curating health-relevant metrics to those related to behaviors of interest. Notably, on the GLOBEM dataset, social support-related measures correlated most strongly with the detected anomalies. However, in the OPTIMA, we find no significant relationship between anomalies and self-reported depression or anhedonia severity and symptoms. These findings are in line with recent evidence that behavioral anomalies from the passive sensing of phone and watch data are related to schizophrenic relapse but not depression or anxiety[78].

5.2 Methods

5.2.1 Data Simulation

Mobile health data can be gathered at varying frequencies but are often aggregated into daily features (e.g., average resting heart rate for a day), which is how consumer wearables such as the Fitbit often report metrics. To understand the impact of different temporal representations of data, we created several simulated datasets with varying numbers of daily features (5, 10, 24, 100, 200), autocorrelation timespan (features all autocorrelated for 28 days, or between 0 and 28 days), and feature correlation. Correlations between features are simulated to be either independent, linearly correlated with one another, or nonlinear functions of one another. When simulating data with feature correlations, for each independent feature, there are two correlated

features. Linear correlation is modeled by setting one feature equal to another and adding Gaussian noise scaled to 10% of that feature's mean value. Nonlinear dependencies are modeled by making one feature equal to another while that feature is greater than its mean value. Below the mean value, the nonlinearly related variable is set to 0. Given these degrees of variation, seven different datasets were created, each with 100 "subjects" and each with 120 days of continuous data, to sufficiently power the analyses. For one subject and one feature, a given day was simulated as a pull from a normal distribution in the following manner:

```
function generateDailyFeature(history: ArrayLike, std: float, max: float,
min: float, initial_value: float) -> float:
    if history is not empty:
        center = mean(history)
    else:
        center = initial_value

    feature_value = random.normal(loc=center, scale=std)

    if feature_value > max:
        return max
    elif feature_value < min:
        return min
    else:
        return feature_value
```

In this simulation, a feature has parameters for its standard deviation (*s*) and length of autocorrelation, (length of history variable), and if a determined value is above or below the feature's range, it is capped to the feature specified range. During data generation, an anomaly is simulated at a specified periodicity per feature (e.g., every seven days). Data for an anomalous day are simulated as a pull from a uniform distribution (uncorrelated to previous days of data).

5.2.2 Anomaly Detection

We applied anomaly detection methods to each simulated individual's data independently. Anomaly detection was performed via models based on nonnegative matrix factorization (NMF), principal component analysis (PCA), one-class support vector machines (SVMs), and isolation forests (IFs). Notably, PCA and NMF are methods for dimensionality reduction but can also be used for AD by assessing the representations' reduced fidelity in reconstructing an unobserved data point. A baseline model (RollingMean) that reconstructed data as the mean value in the training data was assessed for comparison. In contrast to other methods, the baseline model does not consider multivariate relations when labeling anomalous days. Days are labeled anomalous if the error in reconstruction is two standard deviations above the mean reconstruction error in the training data. For the SVM-based models, the radial basis function (RBF), 5^{th}-degree polynomial, and sigmoid kernels were tested. The isolation forest algorithms were run with 100 estimators (trees). In simulation settings with five features, three components were set for PCA and NMF; in all other simulation conditions, five components were used.

Anomaly detection methods were implemented via a rolling window. Each subject, an anomaly detector was trained on *n* days of data, and an anomalous day was labeled if the *n+1* day was labeled anomalous by the detector, as illustrated in **Fig. 5.1**. We used this approach so that multiple anomaly detection models could be run with varying window sizes. Notably, different sizes of rolling windows should be able to detect anomalies relevant at different temporal scales. For example, a 7-day rolling

window detector may find that weekends are anomalous days, whereas a 28-day rolling window detector may find a sudden sick day in a month as anomalous.

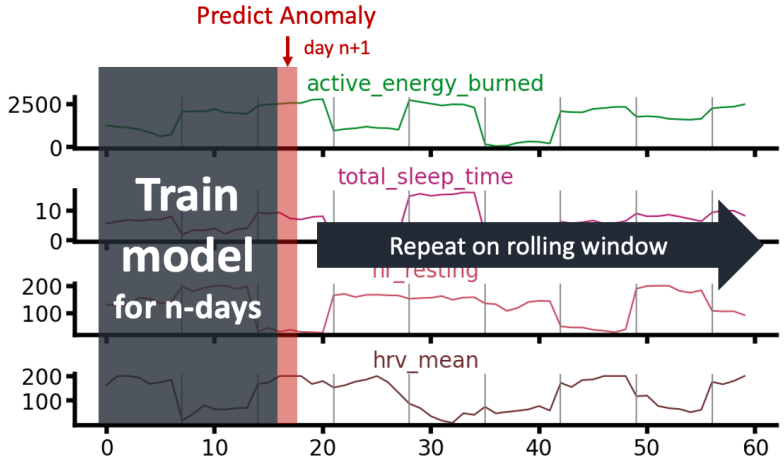

Figure 5.1 Training schema for anomaly detection models, showing four example features simulated over 60 days with anomalous days every 7 days highlighted by gray vertical bars.

5.2.3 Simulation Performance

In mHealth datasets there are often hundreds of derived daily features that are highly correlated, including potential nonlinear interactions. In our simulation, the number of features, their autocorrelation, and relationship to one another were varied to isolate and investigate the effect of common data relations in mHealth studies. Data simulation conditions are summarized in **Table 5.1**. Each day of data per subject is classified as anomalous by an AD algorithm and compared to its true label (i.e., whether an anomaly was simulated on that day). Due to the class imbalance (more normal days than anomalous days) average precision (a method to calculate area under the precision-recall curve) was used as the primary performance metric. F1 score, sensitivity, and specificity are also reported to further characterize predictions.

Table 5.1 Data simulation conditions. Anomaly period refers to how often an anomaly is simulated (every n-days). Window size is the length of the rolling window used for training anomaly detection algorithms.

# Features	Anomaly period	Window size	Feature Autocorrelation	Feature Relation
5	2, 7, 14, 28 days	7, 14, 28 days	28 days all features, different/feature (0-28 days)	Independent
5, 10, 25, 50, 100	28 days	14 days	28 days all features	Independent
24	28 days	14 days	28 days all features	Independent, linear, nonlinear
5, 10, 25, 50, 100	28 days (anomalies only in first 5 features)	14 days	28 days all features	Independent

5.2.4 Performance on Real-world Data

GLOBEM Dataset

To understand how these anomaly detection methods may relate to trajectories in mental health, the same AD algorithms were run on the GLOBEM dataset[116]. The GLOBEM dataset contains four years of data with three months of data collected per year from 497 unique participants. Smartphones were used to collect self-reported data (e.g., PHQ-4, BDI-II, etc.) as well as location, call logs, screen state, and Bluetooth connections. Participants were given Fitbit devices to assess steps and sleep. The Patient Health Questionnaire-4 (PHQ-4) is a validated four-item questionnaire used to quickly screen for depression and anxiety with a total score ranging from 0 to 12[117] and was administered weekly in the GLOBEM study. The Beck Depression Inventory-II (BDI-II) is a well-established survey to detect depressive symptoms[118] and is administered at the beginning and end of a 10-week period in the GLOBEM study. Features were filtered to those that are present at the daily level (357 features) as well as missingness indicators for each of the categories of features (GPS, calls, sleep, activity), thus resulting in a total of 361 features. AD was also assessed using a

representative subset of 12 features across feature types, establishing 16 features after including missingness indicators (**Table 5.2**). Participants were filtered to those with greater than 4 days containing call, location, sleep, and step data before imputation. For the Year 2 dataset, this resulted in 192 of the 218 participants being analyzed (26 removed); and for Year 3, this resulted in using 128 of 137 participants (9 removed). Data was imputed using the *scikit-learn* iterative imputer[119] in a manner that avoids temporal data leakage. To do this, the imputer was trained on all prior data for a participant before a given day. The first seven days of data per participant are not imputed.

Table 5.2 Variables that are part of a reduced 16-feature set in analysis of anomaly detector performance.

Type	Variable Name
Location	f_loc:phone_locations_doryab_locationentropy:allday
	f_loc:phone_locations_barnett_circdnrtn:allday
Activity	f_steps:fitbit_steps_intraday_rapids_sumsteps:allday
	f_steps:fitbit_steps_intraday_rapids_sumdurationactivebout:allday
Sleep	f_slp:fitbit_sleep_intraday_rapids_sumdurationasleepunifiedmain:allday
	f_slp:fitbit_sleep_intraday_rapids_countepisodeasleepunifiedmain:allday
	f_slp:fitbit_sleep_summary_rapids_firstbedtimemain:allday
	f_slp:fitbit_sleep_summary_rapids_avgefficiencymain:allday
Calls	f_call:phone_calls_rapids_missed_count:allday
	f_call:phone_calls_rapids_incoming_count:allday
	f_call:phone_calls_rapids_outgoing_count:allday
	f_call:phone_calls_rapids_outgoing_sumduration:allday
Missingness	sleep_missing
	steps_missing
	location_missing
	call_missing

The AD methods are evaluated first on Year 2, and the results are confirmed on Year 3 of the GLOBEM dataset. The anomaly detection algorithms were run with varying rolling window sizes, model hyperparameters, and input feature sets, as summarized in **Fig. 5.2**. Fine tuning of these parameter choices was performed via the year 2 dataset, with model performance validated on Years 3 and 4. The number of

detected anomalies is compared with the PHQ-4 score (at the end of the period) over a 1-week period via Spearman correlation. Performance was also calculated by assessing the correlation of anomalies detected to change in self-reported assessment scores per participant over the course of the full 10 weeks.

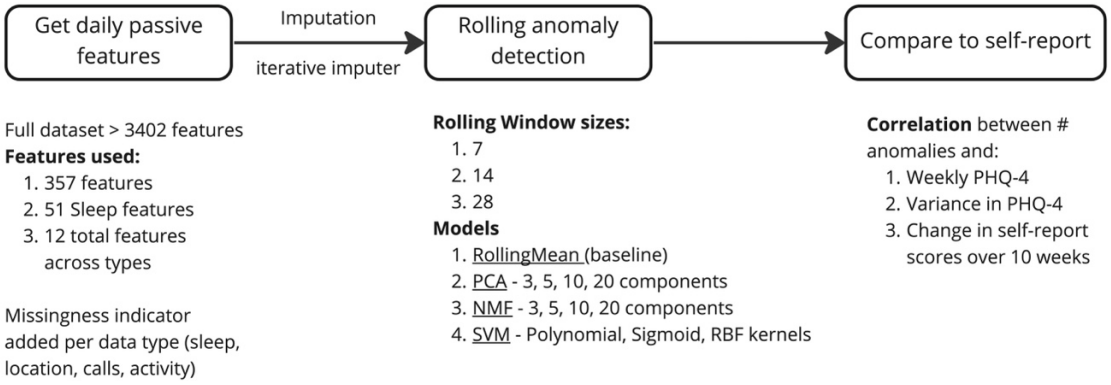

Figure 5.2 Analysis procedure for the GLOBEM dataset using Years 2 and 3 of data.

OPTIMA Dataset

The OPTIMA dataset used for anomaly detection included 343 participants monitored for up to 13 weeks. The dataset is not preprocessed like the GLOBEM study to have day-level features, so data from different frequencies of measurement are aggregated down to the daily level. A total of 59 features are generated and summarized in **Table 5.3**.

Table 5.3 Summary of daily features generated from the OPTIMA study

Feature Type	Context	Aggregations	# Features
Heart Rate	daily, sleep	mean, median, std, min, max	10
Heart Rate Variability (SDNN)	daily, sleep	mean, median, std, min, max	10
Oxygen Saturation	daily, sleep	mean, median, std, min, max	10
Respiratory Rate	daily, sleep	mean, median, std, min, max	10
Ambient Noise	daily, sleep	mean, duration	4
Active Energy Expenditure	daily	mean, duration	2
Sleep Stage (light, REM, deep, awake)		duration	4
Sleep Quality		Sleep onset, sleep offset, bedrest onset, bedrest offset, sleep efficiency, sleep latency, wake after sleep onset, sleep duration, bedrest duration	9

To validate that the models highlighted the expected important features, reconstruction error for the feature wake time after sleep onset (WASO) was correlated with self-reported sleep disruption from the Pittsburgh Sleep Quality Index (PSQI) subscale for sleep disruption. A one-sided Wilcoxon rank sum test was applied to compare the correlation of the feature importance (reconstruction error) between WASO and PSQI sleep disruption with the correlation between all other features and PSQI sleep disruption. The same procedure was used for the PSQI total score to confirm that WASO is specifically important for sleep disruption as opposed to general sleep. Data are available for 62 participants, comprising 95 PSQI responses that have 28 days of digital sensing data prior to assessment.

5.3 Results

5.3.1 Simulation Performance

Under most conditions (**Figs. 5.3-5.4**), the PCA-based anomaly detector outperforms the other models and is closely followed by NMF. Conversely, SVM models with

polynomial and sigmoid kernels overpredict anomalous days, as indicated by their higher sensitivity and lower specificity. The polynomial kernel appears to perform better than the other kernels and is closer to the baseline RollingMean model and PCA-based methods. The isolation forest performed similarly to the RBF kernel SVM model but was not analyzed in subsequent tests (**Fig. 5.4**) because of its longer computational time.

All the models performed best when the features were independent (**Fig. 5.4a**), with comparable performance between linear and nonlinear feature correlation. When anomalies are present in all the features, more features improve the predictive performance of all the models except the SVM with an RBF kernel **(Fig. 5.4b**). When anomalies are present in only five features, increasing the number of features without anomalies decreases the performance of all models (**Fig. 5.4c**).

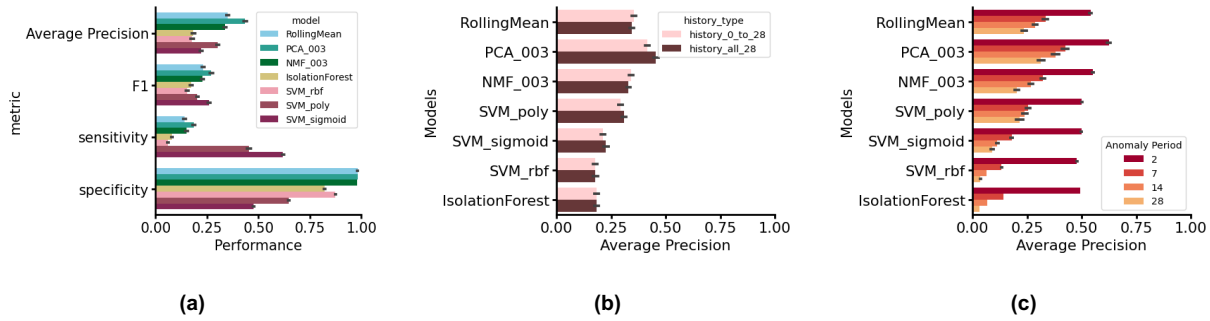

Figure 5.3 Performance of anomaly detectors across simulation conditions. Models are trained on a rolling window of 7, 14, and 28 days, with anomalies at a periodicity of every 2, 7, 14, or 28 days: **(a)** shows a bar plot of each detector's overall performance across all four calculated metrics; **(b)** shows the average precision of models on data simulated to have 5 features all with 28 days of autocorrelation (history_all_28) or between 0 and 28 days of autocorrelation (history_0_to_28); **(c)** depicts difference in model performance as period of anomalies increases from every 2 days to every 28 days.

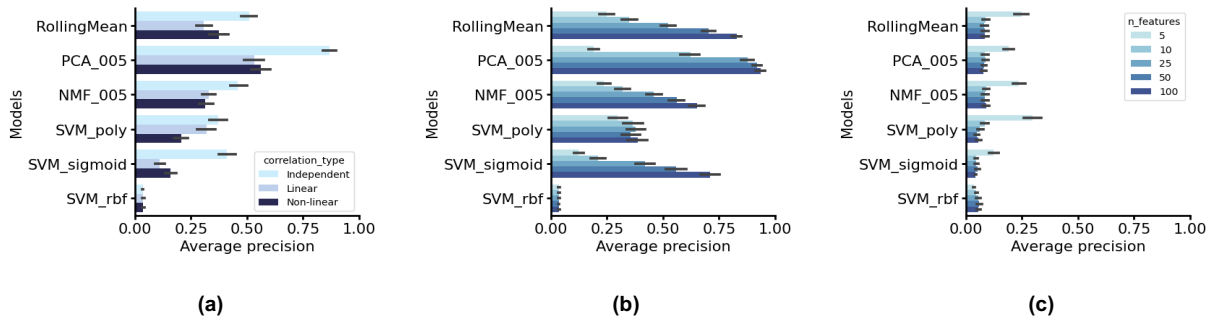

Figure 5.4 Performance of anomaly detectors across simulation conditions. Anomalies are at a 28-day period and a 14-day rolling window is used for training: **(a)** shows model performance when features are independent, linearly correlated, or nonlinearly correlated; **(b)** and **(c)** shows performance of anomaly detectors as more features are added. In **(c)** anomalies are only present in the first five features, showing the effect of signal dilution on detector performance.

5.3.2 Feature Importance Validation on OPTIMA

Feature importance for wake time after sleep onset was significantly more correlated with sleep disruption than importance of other features ($p=2.37x10^{-4}$). WASO feature importance was not significantly more correlated to PSQI total score ($p=0.57$)) as seen in **Fig. 5.5.**

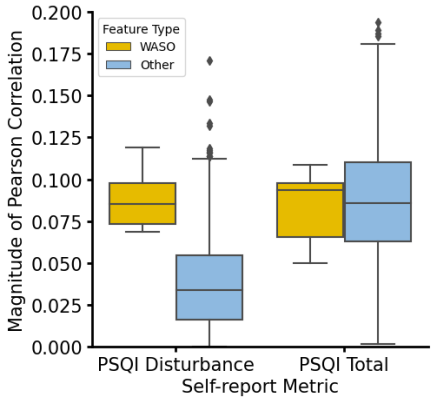

Figure 5.5 Magnitude of the correlation between feature importance to anomaly detectors to self-reported sleep disturbance (PSQI Disturbance) and overall sleep quality (PSQI Total). Wake after sleep onset (WASO) feature importance correlation to each metric is seperated from all other feature importance correlations.

5.3.3 GLOBEM Data Performance

RollingMean, PCA-, NMF-, and SVM-based anomaly detectors were run on the GLOBEM Year 2 and Year 3 datasets with varying hyperparameters such as the number of components and the kernel type. Varying the models and hyperparameters influenced the overall performance of the AD methods measured by correlating the number of detected anomalies with weekly PHQ-4 measures, as shown in **Fig. 5.6**. The 16-feature dataset was used in the following analyses to limit the influence of noise on the anomaly detection performance, as it appears that the methods were sensitive to such noise in the simulation study. These 16 features comprise 12 selected features around location, sleep, physical activity, and calls with one missing data indicator per feature type. PCA with three components performed best, with the highest median squared correlations of number anomalies to the end PHQ-4 score in a 1-week period (R^2 = 0.114), and the rolling mean baseline model performed second best, with a median R^2 of 0.111.

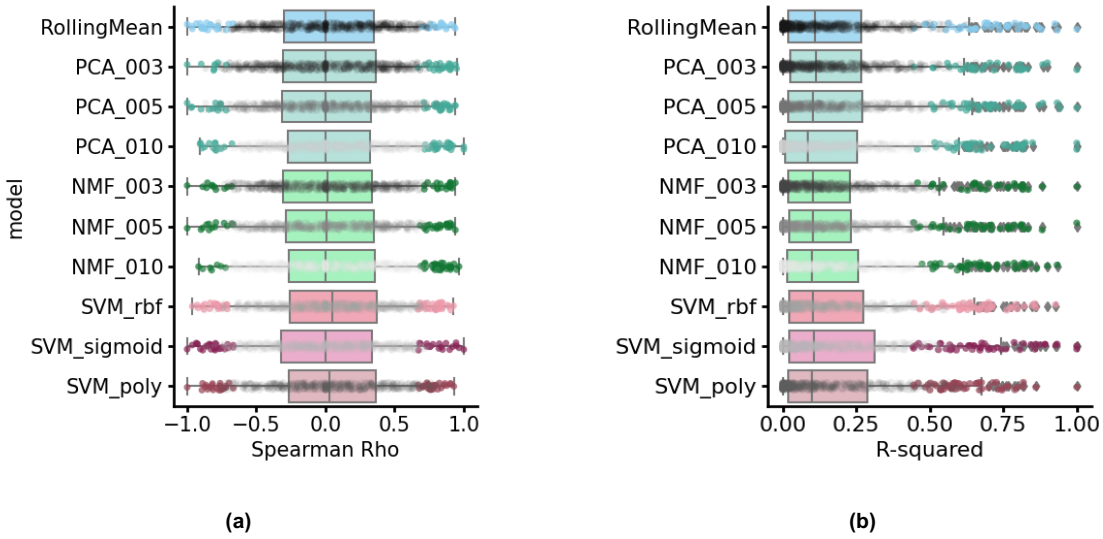

Figure 5.6 Box plot overlaid with a strip plot of Spearman rho values correlating the number of detected anomalies in a week to the PHQ-4 score at the end of the week, with one correlation calculated per participant. **A** shows the raw correlation, and **B** shows the R squared value. The gray dots on the strip plot indicate insignificant correlations ($p > 0.05$), and the box plot indicates the distribution of all the correlations for a given anomaly detector.

At the full 10-week study period level as seen in **Fig. 5.7**, the NMF anomaly detector with 10 components had the largest magnitude of correlation with the "2-way social support scale giving instrumental support section" at Year 2 ($\rho = -0.39$; $-\log_{10}$ p-value = 3.4). In Year 3 of the GLOBEM study, the NMF 3-component model had significant positive correlation with the 2-way social support receiving emotional support, but no model had a significant correlation with other social support elements.

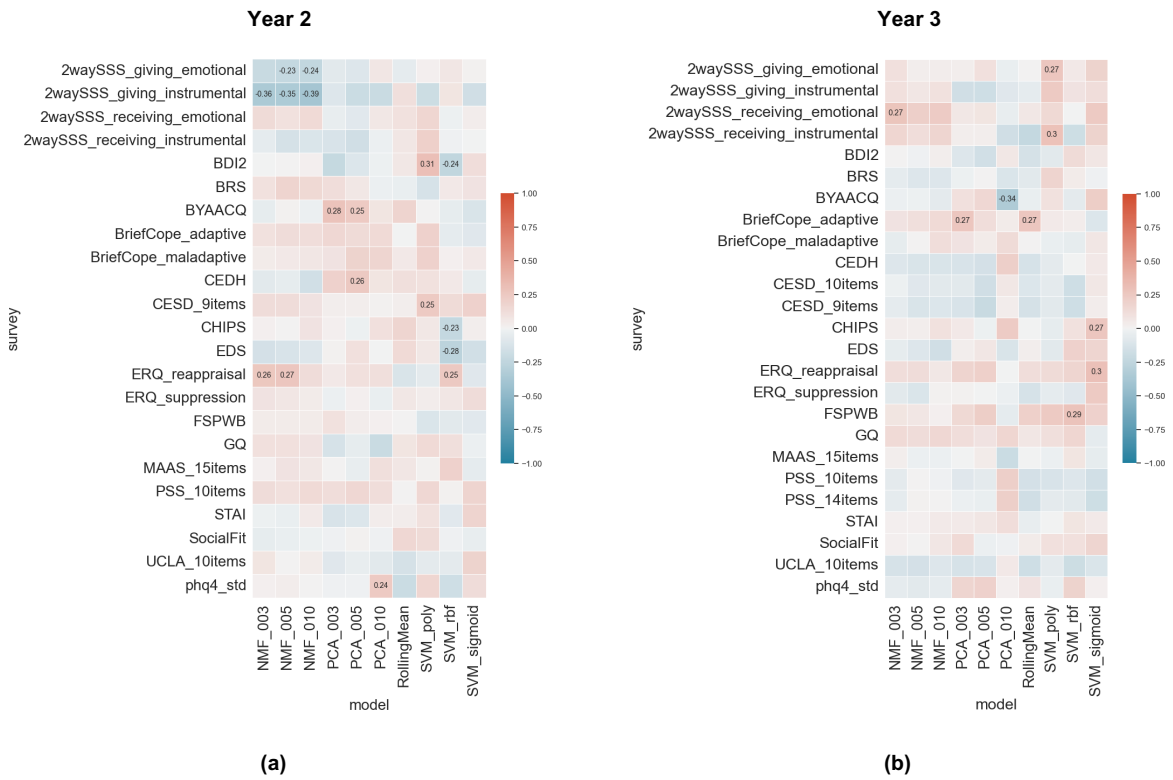

Figure 5.7 Study length correlation of counted anomalies to changes in survey scores. Surveys from the GLOBEM study on the vertical axis correlated with the number of anomalies from the detectors listed on the horizontal axis. Squares annotated with correlations have a p-value < 0.05. **(a)** shows the correlation for year 2 of the GLOBEM study, and **(b)** shows the same analysis performed for year 3. PHQ4_std represents the standard deviation of participants' PHQ-4 replies over the 10-week study.

To investigate how anomaly detection works on participants, **Fig. 5.8** shows a subset of participant data with high positive, zero, and negative correlations between detected anomalies and PHQ-4 score. These are participants from Year 2 of the GLOBEM study.

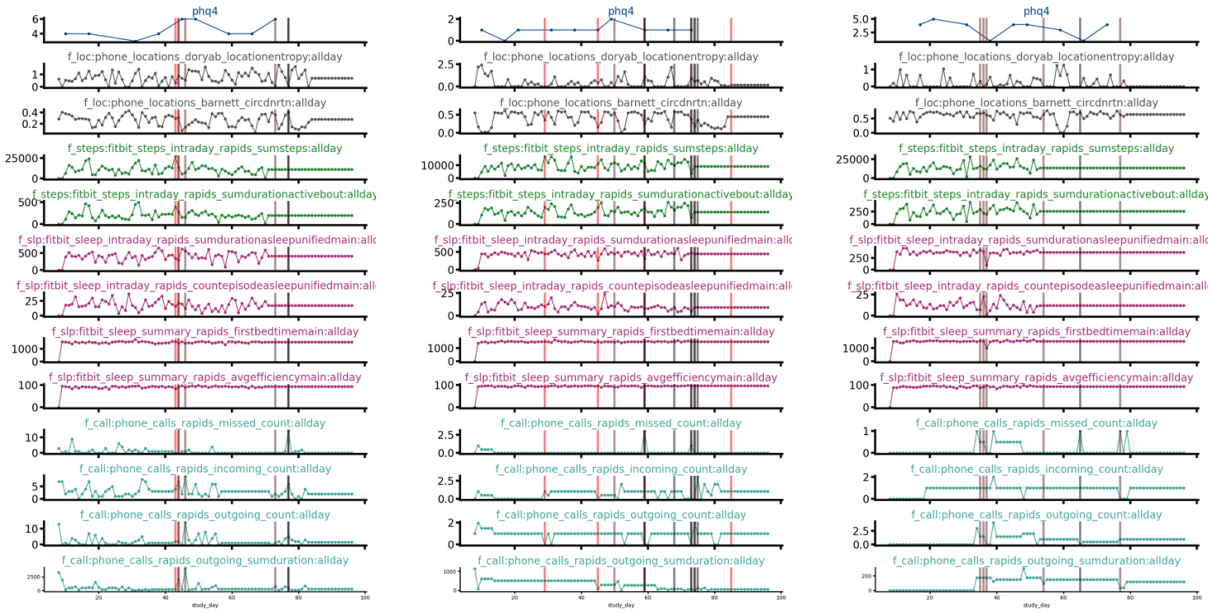

Figure 5.8 (a) An example participant with a ρ=0.90 between the detected anomalies and the end-of-week PHQ-4 score with anomalous days labeled via a 3-component PCA. **(b)** shows a participant with ρ=0 using a rolling mean-based anomaly detector. **(c)** shows a participant with ρ=-0.91 using a 10-component PCA-based anomaly detector. The light red bars correspond to a 7-day rolling window, the maroon bars correspond to a 14-day window, and the dark gray bars correspond to a 28-day rolling window.

5.3.4 OPTIMA Dataset Performance

There was no significant correlation between the number of detected anomalies in the OPTIMA dataset relative to PVSS total score, PVSS subscales, PVSS individual items, or PHQ-14 total score and individual item responses. These results are shown in **Fig. 5.9-5.10.**

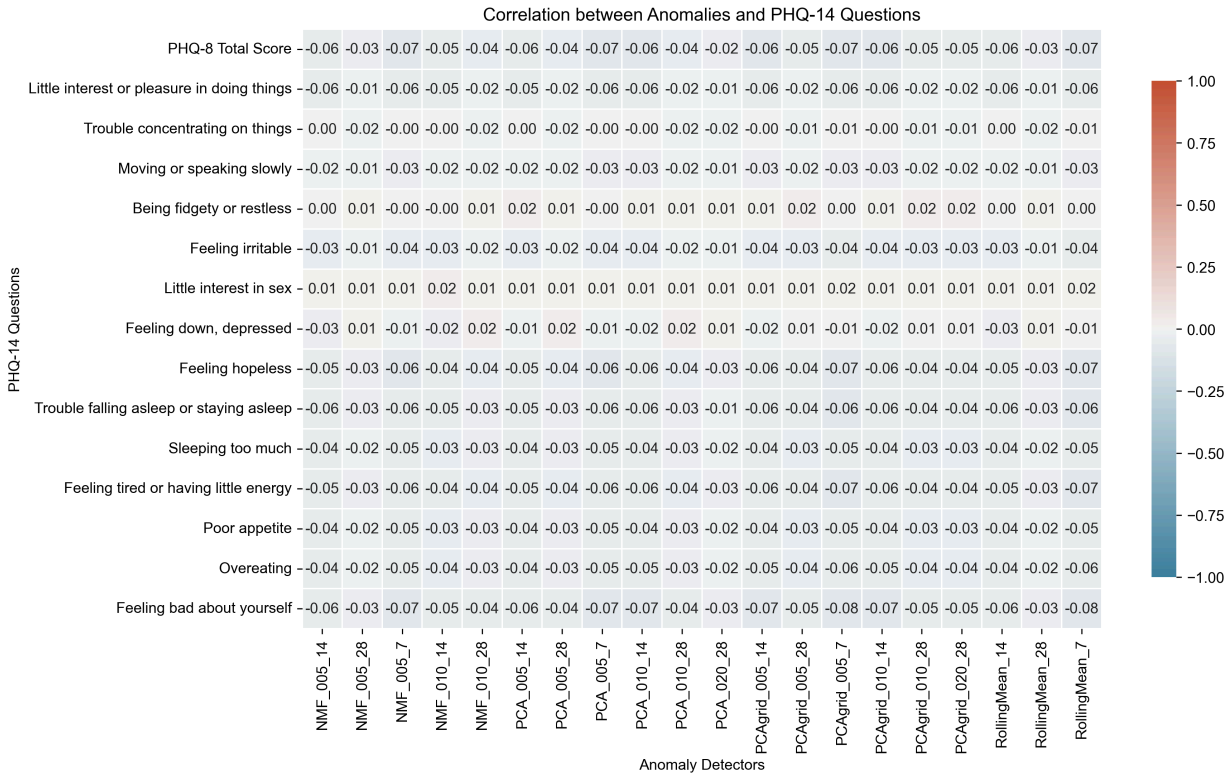

Figure 5.9 Spearman correlation of counted anomalies for the week prior to survey scores. Items and total scores from the PHQ-14 in the OPTIMA study on the vertical axis correlated with the number of anomalies from the detectors listed on the horizontal axis.

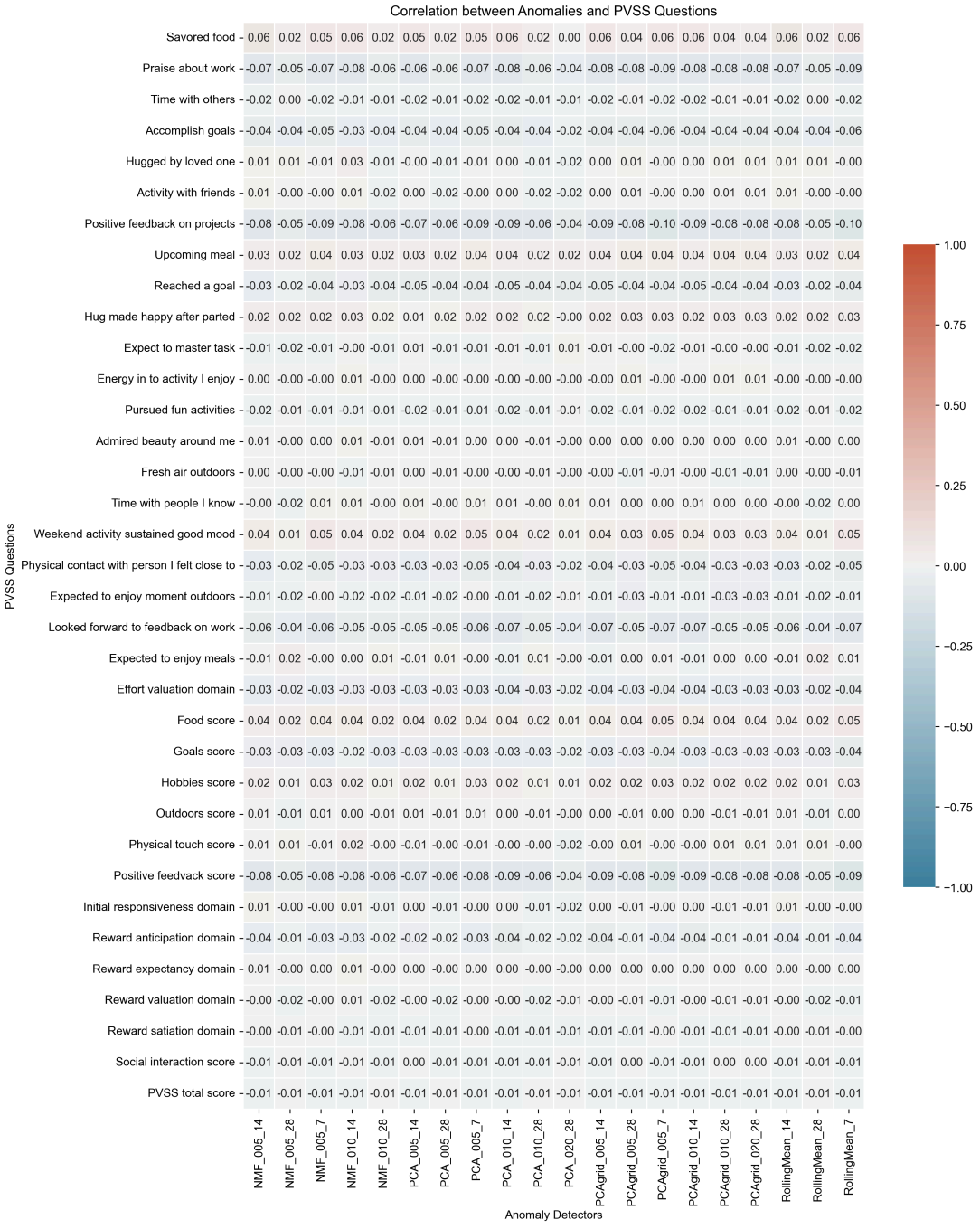

Figure 5.10 Spearman correlation of counted anomalies for the week prior to survey scores. Items and total scores from the PVSS in the OPTIMA study on the vertical axis correlated with the number of anomalies from the detectors listed on the horizontal axis.

5.4 Discussion

This work characterizes anomaly detectors in a variety of simulated conditions before applying them to a real mHealth dataset. Our goal in using anomaly detection methods with mHealth data is to filter out and highlight specific days and behaviors that may be important for further exploration, including providing potential areas for understanding one's general and mental health. This analysis is especially relevant as the digital health research community continues to assess the key mHealth features that serve as proxies for mental health and related behaviors and as the underlying technologies (sensors and devices) continue to evolve. Through this simulation work, we find that anomaly detectors vary dramatically in how they perform both relative to one another and relative to which conditions underlie data generation. Assessing these algorithms first on a simulated (i.e., controlled) dataset is a prerequisite to understand if methods are robust to specific data generation patterns and assumptions that we may be making of mHealth data. A key observation from the simulation experiments is that all algorithms tested performed much better when anomalies were present concurrently across most of the features. This finding suggests the need for testing more AD methods that work well in a low signal-to-noise environment as well as conducting feature selection when they are applied to real mHealth datasets, which can have thousands of features. However, we find that anomaly detectors prioritize the expected features in relation to sleep disturbances, confirming that anomaly detectors appear to work as expected in real datasets. Furthermore, different AD-based techniques may be required depending on the underlying nature of the data stream and behavior—there is no "one-size-fits-all" strategy. Feature selection was initially not performed before applying AD methods to

the GLOBEM dataset, but improved performance of AD methods was found by restricting analysis to a subset of 16 out of 361 features.

In the GLOBEM data, we do not find large correlations between detected anomalies and self-reported scores; however, modest signals are detected, with correlations between 0.2 and 0.3. Notably, the highest correlations found in the paper describing the dataset are on the order of 0.1 when comparing a passively sensed feature to a change in self-reported score[116]. Similarly, in the OPTIMA dataset, we do not observe a significant correlation between detected anomalies and individual items or total scores from the PHQ-14 and PVSS.

Little to no correlation between anomalies and self-reported depression severity may be expected. Recent work by Cohen et al. in a study with 132 participants (76 with schizophrenia, 50 controls) revealed that anomalies in mHealth data were predictive of relapse. However, they found that using a separate technique, offline changepoint detection did significantly correlate with depression and anxiety measures. Given this finding with changepoint detection, it may be of use to investigate how the rate of change of reconstruction error rather than the count anomalies detected relates to depression. The work by Cohen et al. and confirmation in this dataset suggest that unlike schizophrenia, depression may not be a condition in which the count of behavioral anomalies are related to a consistent change in symptomology. The anomalies detected only indicate changes, not whether those changes are beneficial or detrimental to an individual. Additionally, all studies investigated only have at most 13 weeks of data, it is possible that with longer datasets, anomalies detected with training

windows of over 6-months or more may be able to detect anomalous behavior of relevance to depression.

However, although the number of anomalies does not correlate with a change in depression severity, the days highlighted provide a starting point to discuss and evaluate key moments in their trajectory. For example, the first anomaly detected in **Fig. 5.8b** appears to coincide with a large change in phone call behavior and a commensurate increase in physical activity. Nevertheless, this work does not suggest that detected anomalies can be used as a clinical diagnostic. We believe that AD can instead be useful in identifying initial points where wearable device and smartphone data can be looked at in relation to different behavioral trajectories, identifying changes, and that significant deviation from past behaviors may prove a useful means to contextualize a change in mood. By way of illustration, some of the most significant self-report correlations with anomalies in **Fig. 5.7a** were related to social support; similarly, many of the anomalies detected via manual inspection appeared to fluctuate around phone call behavior, which might be linked to changes in social support. Notably, the PHQ-4[117] and even the DSM-V diagnostic criteria[120] for MDD do not have items asking specifically about social support. This represents one potential area of future investigation around feature curation and evaluation to use mHealth data to improve how we understand and define MDD.

In addition to filtering data for human use, AD methods may also help in the development of additional machine learning tools for the prediction of mental health trajectories. Training models on individual participant data has been shown to be more effective at tasks such as mood prediction[121]; however, this approach also reduces the

amount of data available for a model to be trained on. Improving the effectiveness of these remaining data for model training may help filter out anomalous days. By removing these anomalous days from training data, we may be better able to determine the relationships between metrics from mHealth devices and mental health trajectories. Similarly, the label of an anomalous day may itself be a useful predictive feature for models.

5.4.1 Limitations

Anomaly detection methods need further development to be truly useful at filtering user data to find the days that are worth further investigation for mental health trajectories: there are many improvements to be made in future work. The performance of AD algorithms on even the simulated data is suboptimal, leaving room for improvement without necessarily overfitting to the simulated conditions. Potential venues to improve the tested AD algorithms include removing anomalous days from the training of detectors on future days, as done in Ren et al.[122], and the use of more complex models, such as autoencoder-based methods[123].

As the signal-to-noise ratio appears to be an important indicator of performance, feature selection processes should also be explored. In this study, imputation was performed via one type of iterative imputer; however, the imputation technique may dramatically change the performance of an anomaly detector. Investigating how imputation and missingness patterns affect anomaly detectors is of critical interest, as mHealth data suffer from high degrees of missingness.

5.4.2 Conclusion

This work characterizes reconstruction error-based anomaly detection techniques on simulated data and then uses them to detect changes in depression severity and symptoms. The anomaly detectors appear to highlight days that deviate from a participant's usual behavior; however, those deviations do not consistently correlate with an increase or decrease in depressive symptoms. As such, while the detectors investigated in this work cannot detect depression, anomaly detectors may present a useful first pass for filtering and investigating data points of interest when presented with large digital health datasets.

Chapter 6: Detecting momentary reward and affect with real-time passive sensor data

This study explores the capability of mobile health (mHealth) data from smartphones and smartwatches to predict self-reported momentary affect and anhedonia/reward system functioning through ecological momentary assessment (EMA). Utilizing data from 245 participants generating 23,812 EMA sessions, we evaluated whether behaviors and physiological factors measured by passive mHealth data can detect subjective states. Despite generally low performance (AUROC <0.6), the models exceeded random chance, suggesting detectable signals between passive measures and subjective states. The optimal aggregation periods for sensor data ranged from 15 minutes to 3 hours, with no single window consistently outperforming others across all affect and reward items. Subgroup analyses revealed variations in model performance on the basis of demographics, depression severity, and anhedonia severity, with notable performance improvements when sleep data were incorporated for some EMA items. The findings highlight the influence of individual characteristics on model performance and the potential for passive digital sensing to monitor mental health on a large scale, although further refinement and personalized modeling approaches are necessary for improved accuracy.

6.1 Introduction

Digital sensing is frequently compared to self-report surveys in the context of monitoring mental health. However, digital sensing provides the ability to examine much more

granular time series data than does the broader, retrospective scope of weeks and months typically covered by self-reports. Experience sampling or ecological momentary assessment (EMA) methods enable the collection of ecologically valid, moment-to-moment self-reported mental state data from participants at a much finer temporal resolution[124,125]. The use of the EMA as an outcome measure to evaluate digital sensing aligns more closely with the inherent time scale of these technologies.

In this study, we investigate whether digital sensing data can inform us about self-reports of momentary affect and anhedonia/reward system functioning from EMAs. Here, we specifically focus on this task in the context of a population with depression and a spectrum of anhedonia severity. Paired with EMA, mobile health (mHealth) data from smartphones and smartwatches allow us to observe whether an individual's behavior and physiology measurably relate to their subjective state. This approach can provide an understanding of the behaviors and physiological factors associated with specific changes in affect and reward functioning, which, when chronic and extreme, become disordered. There are two key considerations when creating features from digital sensor data. First, we investigate performance differences based on aggregation period for sensor data prior to an EMA response, ranging from 15 minutes to 3 hours. Second, we assess whether incorporating information about the previous night's sleep enhances model performance. Our analysis utilizes data from 245 participants who generated 23,812 EMA sessions.

Prior work on using passively sensed data at the time of EMA has looked at phone accelerometers and gyroscopes at the time of the EMA session[126] or GPS and weather data from the hour prior to EMA[127]. A recent preprint (posted July 3rd 2024) by

Siepe et al. investigated associations between aggregation windows as short as 15 minutes with EMA prompts related to stress, tiredness, and sleep[128]. The analysis from Siepe et al. relies on heavily processed metrics such as the “body battery” or “stress sensor” reported by the Garmin smart watch, and they do not find associations between self-reported tiredness or stress and the “body battery” or “stress sensor”, respectively. An informal review of the literature failed to identify published prior work reporting on the use of less than one hour of data to predict response to EMA data from consumer wearable devices using features such as vital signs (heart rate, heart rate variability (HRV), respiratory rate) and ambient noise.

Our findings indicate that nomothetic (as opposed to personalized or idiographic) models using passive mHealth data can predict responses to both affect- and reward-related EMA items in a cohort with depression. While performance is generally low (often AUROC <0.6), it exceeds random chance, demonstrating that the models can detect signals between passive measures and subjective momentary states. For analytic purposes, it would significantly reduce analytic complexity of investigating the relationship of EMA to mHealth data if a single time window prior to EMA response consistently predicted EMA item response. However, in alignment with previous findings for depression self-report prediction[20,29], we do not find a single time window prior to an EMA response that consistently outperforms any other to predict all affect and reward items, suggesting that analyses utilizing passive sensor data to predict EMA items will need to explore time aggregation windows on which sensor features are used and what the predicted EMA outcome is.

To understand how the performance of models varies as a function of cohort characteristics, performance is stratified by subgroups of demographics, depression severity, and anhedonia severity. Subgroup analyses revealed that for certain populations and specific EMA-related outcomes, performance can be as high as 0.68 (e.g., the inactive enjoy EMA response in males using two hours of sensor data and the past night's sleep as features). The significant performance differences across demographic, anhedonia, and depression severity subgroups suggest that our ability to passively detect EMAs may benefit from incorporating additional data sources to account for population variance.

6.2 Methods

The data for this analysis were obtained from 342 participants who were monitored via passive sensing and self-reported measures over 13 weeks as part of the Operationalizing Digital PhenoTyping in the Measurement of Anhedonia (OPTIMA) study investigating features of anhedonic depression, which collected data between October 2022 and April 2024. Self-reported depression severity is taken from responses to the Patient Health Questionnaire-14 (PHQ-14) (see Depression Symptom Response Project OSF site: https://osf.io/j6r3q/), and anhedonia is assessed via the Positive Valence Systems Scale (PVSS)[88].

The OPTIMA study is part of the Wellcome Leap Multi-Channel Psych Program, a consortium of studies focused on anhedonic depression. A subset of 245 participants was used in this analysis. This subset was selected on the basis of having greater than 30% of expected days with EMAs containing at least 2 of 5 responses, more than 30% of the first 90 days in the study with at least 1 log of heart rate, and greater than 0

variability in the individual's EMA responses (at least one EMA item has greater than 0 variability per participant). The participants were screened to all be right-handed, but no instructions were given as to whether watches would be worn on the nondominant hand or not. The participants were given the Apple watch series 7 or higher and used their own iPhone. This analysis is exploratory and was not preregistered, but its reporting incorporates relevant expected information described in the "Template for studies using passive smartphone measures" by Langener et al[129].

6.2.1 Ecological Momentary Assessment Processing

EMAs were administered to participants via the UCLA Depression Grand Challenge Study App (DGC Study App) for three separate 8-day bursts (5 times per day) during the study (at baseline, week 6, and week 12). The 15 EMA items used in this analysis correspond to either affect or reward functioning. The EMA begins by instructing participants: "The following questions will ask you to describe your feelings and experiences **right now**. "Right now" means right before you began this survey.

Affect-related EMA items are of the form "How ___ do you feel right now?" where the blank is one of 9 items:

1. Sad
2. Stressed
3. Anxious
4. Annoyed/Irritated
5. Energetic
6. Happy
7. Motivated

8. Engaged
9. Lonely

The participants responded on a 5-point Likert scale with the following mapping:

1. Not at all
2. Slightly
3. Moderately
4. Very
5. Extremely

All affect-related items are converted to binary responses if their answer is greater than 1 (not at all). This results in affect-related EMA items being converted to the classification of any endorsement or none. Reward-related items are binary (true or false) responses. The questions are in the format “Right now, I…” and end with (underlines represent the term used in figures and tables to describe the EMA item):

1. am looking forward to an upcoming activity (anticipatory)
2. am feeling good after doing something (consummatory)
3. am putting effort into planning something
4. could be doing something positive but am not because I don’t think I’d enjoy it (inactive enjoy)
5. could be doing something positive but am not because it feels like too much effort (inactive effort)
6. am feeling a sense of meaning and purpose.

The presentation order was randomized within the affect and reward item blocks. An EMA session is an event associated with administering the full set of EMA questions.

The participants were instructed that they had 30 minutes from the time the EMA was sent to submit their responses. EMA sessions that lasted shorter than 10 seconds or longer than 5 minutes were not used in the analysis. This filtering results in the removal of 461 (1.9%) sessions, with 23,812 remaining.

6.2.2 Contextualizing Population

The OPTIMA study recruits a population enriched with individuals with depression and high or low levels of anhedonia. To understand the psychometrics of the PVSS in the general population (especially the distribution of anhedonia across the spectrum depression, age, and sex-at-birth) and help establish the study's inclusion criteria and recruitment strategy, a dataset comprising self-reported PHQ, PVSS, and EMA responses was collected in the planning phase of OPTIMA via Amazon's Mechanical Turk (MTurk) service (Cohen, Forbes, Khazanov, & Fried, in prep; see https://osf.io/6xnv2/). Target recruitment was N=500, stratified to be equally distributed across five age buckets (18 to 25, 26 to 35, 36 to 45, 46 to 55, 56 to 65) and sex-at-birth (50% Male, 50% Female). Of 520 response, 512 have PHQ-8 and PHQ-14 responses, 520 have PVSS responses, and 510 having EMA responses. The distributions of age, sex at birth, PVSS score, PHQ-8 total score (comparable to the PHQ-8 score calculated from the PHQ-14), EMA affect, and reward item responses were compared between this more general sample and the OPTIMA cohort. As the MTurk collected dataset has only 1 EMA session response, when comparing distributions, the median EMA response is taken per participant per EMA item from the OPTIMA dataset, and the PVSS and PHQ-14 scores are taken from the end of the study.

A Wilcoxon rank-sum test was used to compare distributions for affect-related EMA items, age, PVSS total score, and PHQ-8 total score. Chi-square independence tests are used to compare distributions for the binary responses to reward-related EMA items and sex at birth. P-values are corrected via Bonferroni adjustment with a familywise error rate (FWER) < 0.05.

6.2.3 Dataset Split

To ensure an even distribution of important participant-level characteristics between the training data sample and the test set, the participants were split to ensure approximately equal distributions across 10 variables, with an approximately 80–20 split of participants (195 train, 50 test). This split leads to 19,016 EMA responses in the training data and 4796 in the testing data. The full list of variables used to split the data into training and test sets were:

1. Age: distribution of integer years
2. Sex at birth: male or female
3. race and ethnicity: operationalized as identifying as non-Hispanic white or not
4. Study participation: Number of study timepoints with self-reported data available
5. PHQ-14 total score at the end of the study
6. PVSS total score at the end of the study
7. Availability of EMA responses: percentage of expected days (24) within the first 90 days of the study with at least two EMA responses
8. Availability of watch data: percentage of days with at least one heart rate entry within the first 90 days of the study

9. Availability of sleep annotation data: percentage of days with sleep duration calculatable within the first 90 days of the study
10. Variance of EMA responses: average standard deviation of EMA responses across all 15 items used as prediction targets

Assignment to training or test splits was performed via the automated randomization of multiple traits for study design (ARTS) package[130]. The distributions of the variables between the training and test splits are shown in **Figure 6.1.**

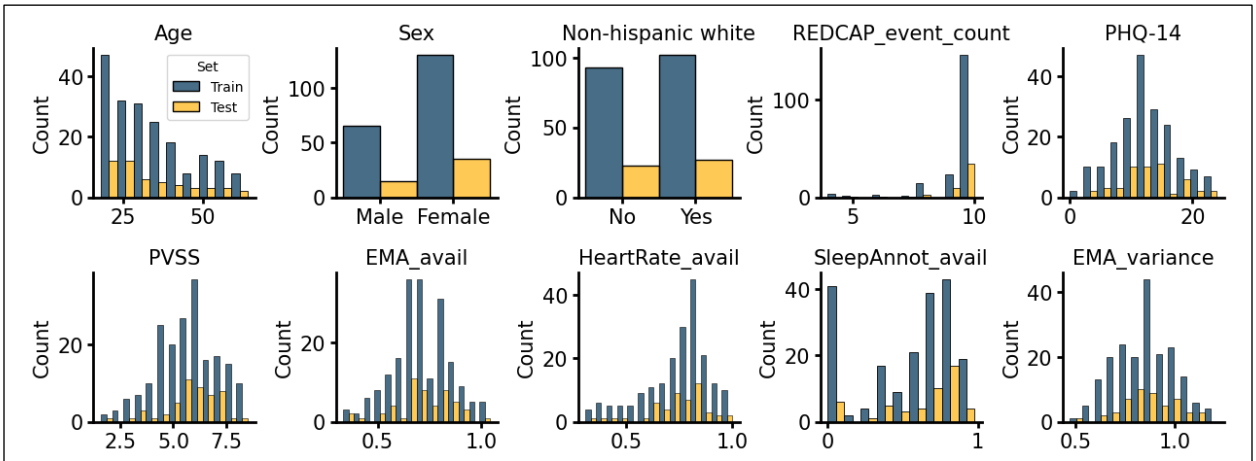

Figure 6.1 Histograms of train and test set user characteristics.

6.2.4 Passive Sensor Features

Passive sensor features are created relative to each timestamp when each EMA session starts. A total of 29 momentary features from vital signs, activity, and ambient noise are generated during the time window prior to the start of the EMA session, and an additional 11 features related to sleep quality from the previous night are generated. Details of heart rate monitoring from consumer devices are reported where possible in accordance with guidelines from Nelson et al.[131]. These features are as follows:

1. Vital signs (13 features): Prior to aggregation, vital signs are resampled as the median at 5-minute intervals to account for dynamic sampling. Heart rate is the most frequently sampled vital sign (the median sampling time is every 2 minutes), so more aggregations are run for it than others. The mean, standard deviation, minimum, maximum, number of entries, slope, and intercept (from linear model fit) are used as features. For respiratory rate, heart rate variability (measured as the standard deviation of the N-to-N interval), and oxygen saturation, mean and number of entries are captured.
2. Activity (12 features): Active energy expenditure, basal energy expenditure, exercise time, and step count, are aggregated by taking the sum value, count of entries, and sum duration of entries.
3. Ambient noise (4 features): Audio exposure events from Apple HealthKit are aggregated by taking the total number of entries, the number of 5-minute intervals with entries, the mean decibel level, and the sum duration of entries. Ambient noise is sampled at a median rate of once every 30 minutes.
4. Sleep (11 features): Sleep features are calculated on the basis of HealthKit sleep annotations from 3pm the prior day until 3pm of the day of EMA assessment. We generate sleep duration, bedrest duration, sleep efficiency, sleep onset latency, sleep onset, sleep offset, bedrest onset, bedrest offset, number of nighttime awakenings, duration of awakenings, and average noise during bedrest.

The median time difference between samples for vitals, activity, and ambient noise is shown in **Figure 6.2** to provide context for what data may be missing on shorter time aggregation windows.

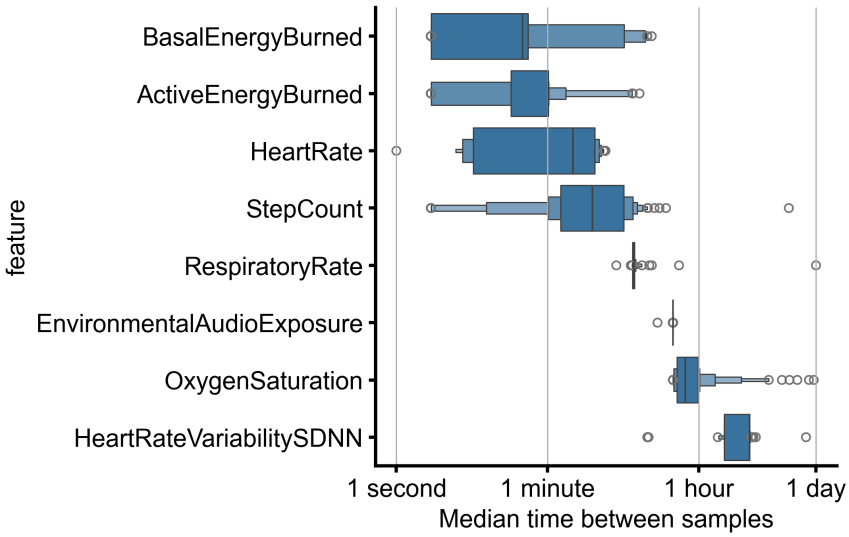

Figure 6.2 Median time between samples of a feature per user. Boxen plot shows distribution across participants. X-axis is logarithmically scaled.

Entries where calculation of mean heart rate was not possible (no heart rate entry was found) were excluded from the analysis. Reporting of a heart rate measurement from the watch is used as a proxy for watch wear. Other missing features are imputed as the median value per feature in the training set. The percentage of missing features per aggregation period is shown in **Figure 6.3** after the removal of EMA sessions, as described in the **Ecological Momentary Assessment Processing** section.

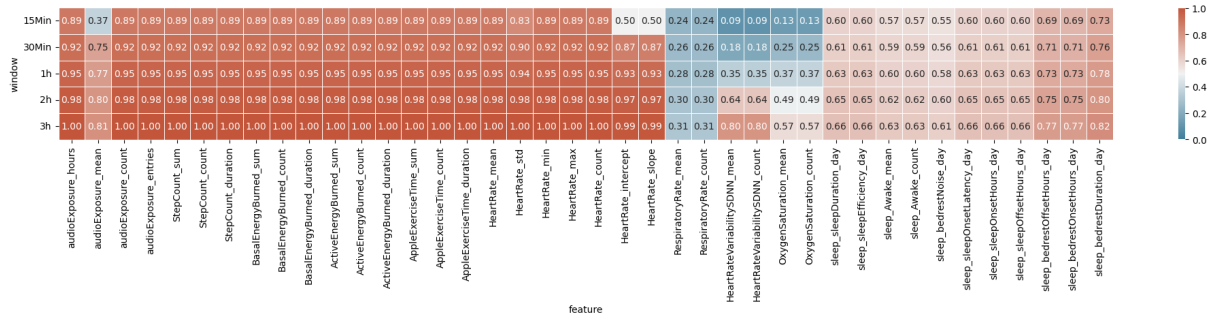

Figure 6.3 Missing passive sensor features based on availability of data prior to EMA session start. Note Sleep features refer to availability of the prior night's sleep data.

6.2.5 Machine Learning Modeling

Machine learning was applied to characterize whether passive sensor-derived features can detect responses to EMA items. Automated machine learning (AutoML) was used on the training data to select the best model for the prediction of EMA binary response via 5-fold stratified grouped cross validation. Prior to model training, the data were scaled via a standard scaler and imputed via simple median imputation. Grouping was performed to ensure that the participants used to train the models were not in the validation set for hyperparameter tuning and model selection. The models evaluated were the gradient boosting classifier (XGBoost), random forest (RF), light gradient boosting classifier (LGBM), logistic regression with L1 loss (LRL1), also known as LASSO, and logistic regression with L2 loss (LRL2), also known as Ridge. Automated machine learning was performed via the fast and lightweight AutoML library (FLAML) 2.1.2[132], with a time budget set to 200 seconds and early stopping and scikit-learn 1.2.2[91].

6.2.6 Model Evaluation

Models were evaluated for performance on the test dataset of 50 participants and 4,796 EMA responses via the metric of area under the receiver operating characteristic curve (AUROC). To account for the unequal number of EMA responses per participant and to generate confidence intervals, a bootstrapped sampling approach was taken. Each bootstrapped sample consisted of 50 EMA responses per individual in the test set sampled randomly with replacement, and for each bootstrapped sample, the AUROC was calculated. Bonferroni adjustment is used when calculating confidence intervals with a false discovery rate set to 0.05 and 150 comparisons (15 outcomes, 5 aggregation window sizes, 2 feature sets).

The best performing model per outcome with respect to aggregation duration (15 minutes, 30 minutes, 1 hour, 2 hours, or 3 hours) and feature set (momentary features, or momentary and sleep features) was selected per detected outcome. For these models, performance was evaluated stratified by the following user characteristics: age, race and ethnicity, sex at birth, anhedonia, and depression. Each characteristic was converted to a binary variable. Age, anhedonia severity, and depression severity were split on the basis of whether the participants were above or below the median value in the test set. Anhedonia was assessed as participants with end-of-study PVSS scores less than the median (5.95), and depression was assessed as end-of-study PHQ-14 scores greater than the median (13.5). Racial-ethnic background is converted into a binary by determining whether an individual self-identifies as non-Hispanic white. A comparison of the AUROC between each group was performed via the Wilcoxon

signed-rank test. The feature importance for the best performing model is assessed via SHapley Additive exPlanation (SHAP) scores[94].

To assess the sensitivity of the models to missing data, the test set performance was evaluated on the full test set as well as a subset of EMA sessions missing fewer than 10 of 29 features (n=2716, 56.6% of EMA sessions in the test set). Differences between the AUROC on the full test set relative to the high data availability test set are compared per outcome via a Wilcoxon rank-sum test corrected for multiple testing via the Bonferroni adjustment with the FWER set to 0.05.

6.3 Results

6.3.1 Distribution of EMA Responses

Participants in the OPTIMA study were significantly different from the population sampled using MTurk in terms of sex at birth (chi-squared 57.24; Bonferonni adjusted p-value = 1.5e-13), age (mean difference = 3.49 years, Bonferonni adjusted p-value = 1.12e-06), PHQ-8 total score (mean difference 5.68, Bonferonni adjusted p-value = 6.11e-30), and PVSS total score (mean difference = 0.89, Bonferonni adjusted p-value = 2.12e-16). This finding demonstrated that, compared with the population in the MTurk dataset, the OPTIMA study population included participants who were significantly younger, more depressed, more anhedonic, and more female **(Fig. 6.4A**).

Additionally, we see significant differences in the distributions for responses to 7 of 9 affect EMA items and 4 of 6 reward EMA items. These differences are shown in **Fig. 6.4.B-C** and **Table 6.1.**

Table 6.1 EMA response means in the OPTIMA study versus the population sample from Amazon MTurk. P-values from Wilcoxon rank-sum test between OPTIMA and general means adjusted with Bonferroni method FWER < 0.05.

EMA item	OPTIMA mean	General mean	Mean difference	Adjusted p-value
Reward (1-5)				
sad	1.784	1.837	-0.054	1.00E+00
stressed	2.427	2.078	0.348	2.01E-08
anxious	2.335	1.951	0.384	6.09E-10
annoyed	1.827	1.731	0.095	4.42E-03
energetic	2.114	2.839	-0.725	3.57E-18
happy	2.127	3.010	-0.883	1.00E-23
motivated	2.198	3.310	-1.112	7.28E-40
engaged	2.382	3.575	-1.193	2.39E-44
lonely	1.776	2.010	-0.234	1.00E+00
Affect (0-1)				
anticipatory	0.480	0.735	-0.255	4.11E-11
consummatory	0.427	0.715	-0.289	2.07E-14
effort	0.302	0.639	-0.336	8.55E-17
inactive enjoy	0.169	0.265	-0.096	6.89E-02
inactive effort	0.253	0.346	-0.093	1.55E-01
meaning	0.359	0.715	-0.356	2.25E-19

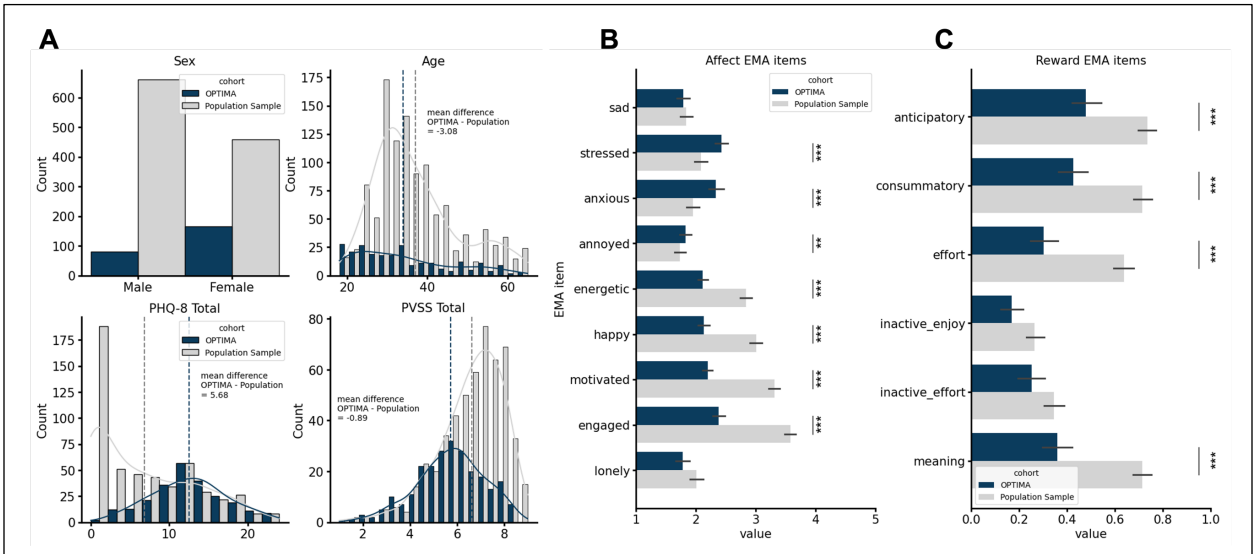

Figure 6.4 Distribution of demographics, depression severity, anhedonia, and EMA responses between OPTIMA and a population sample collected on Amazon MTurk. **A**) Sex at birth, age, PHQ-8 score (depression severity, directly comparable to PHQ-14 total score), PVSS (anhedonia; lower means more anhedonia). All measures significantly difference at $p < 0.001$ between groups. **B)** Bar plot comparing affect EMA item response and **C**) reward EMA responses. Significant differences between items in **B** and **C** annotated if $p < 0.05$ after Bonferroni adjustment with FWER < 0.05. Annotation legend: * < 0.05, ** < 0.01, *** < 0.001.

6.3.2 Model Performance

We find that models can detect EMA responses greater than random chance (AUROC > 0.5 via the Mann–Whitney U test; Bonferroni correction at FDR = 0.05) for 11 out of 15 EMA item outcomes (sad, stressed, anxious, annoyed, energetic, motivated, anticipatory, consummatory, inactive enjoy, inactive effort, and meaning). The performance details are outlined in **Figure 6.5.** The performance stratified by either the window of aggregation or feature set is depicted in **Figure 6.6.**

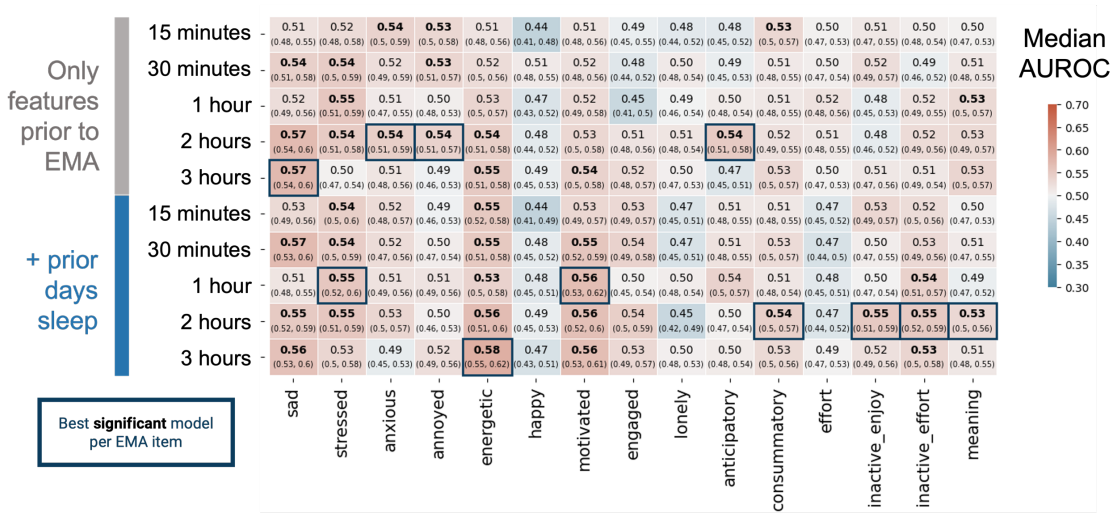

Figure 6.5 Model performance (median AUROC ± 95% Bonferroni-adjusted bootstrapped confidence interval) on the held-out test set of 50 users for each feature set (rows) and each outcome (columns). Bold values indicate model AUROC performance for an outcome where p<0.05 for testing that AUROC > 0.5 after Bonferroni adjustment of Mann Whitney U-test. Blue outline indicates best significantly performing model for a given outcome column.

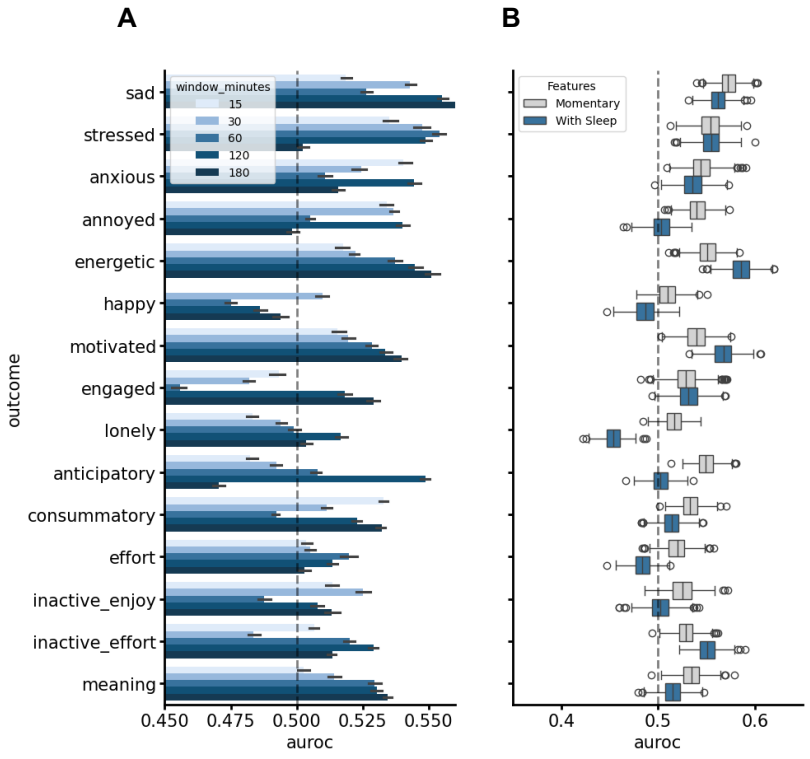

Figure 6.6 Performance of EMA detection models across **A)** passive data aggregation window with momentary features only and **B)** inclusion or exclusion of sleep data from the prior night based on peak aggregation window per outcome shown in **A**.

For the best performing model per outcome, performance is assessed stratified by age, race and ethnicity, sex at birth, depression severity, and reward functioning. Model performance split by these characteristics is shown in **Figure 6.7.** Performance is notably higher in males for 7 of 9 affect related EMA items (sad, anxious, annoyed, energetic, happy, motivated, and engaged) and 2 of 6 reward related EMA items (inactive enjoy and inactive effort) while only higher in females for the meaning EMA item. Stressed, anxious, engaged, and inactive effort EMA items are better detected in older participants. The highest performing item is inactive enjoy for males with an AUROC of 0.684.

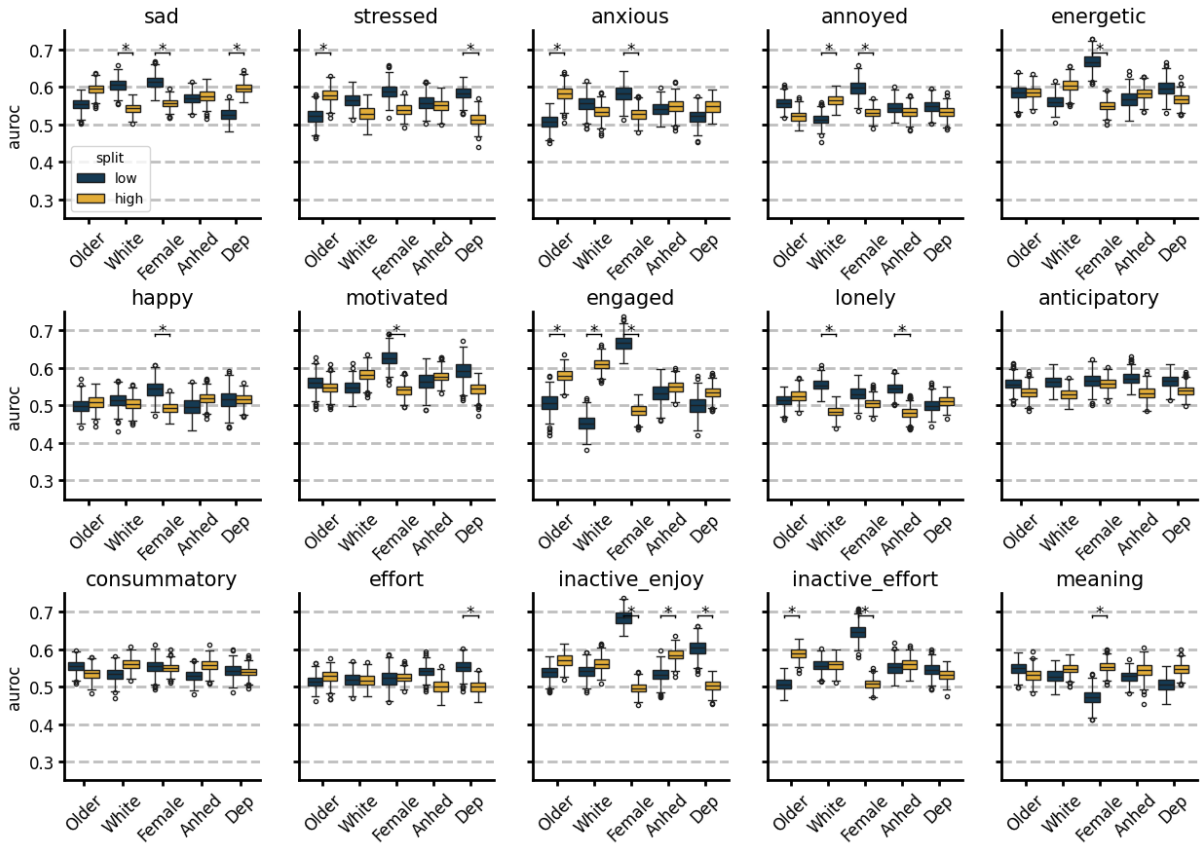

Figure 6.7 Performance of EMA detection models split by clinical and demographic parameters. * Indicates median difference between low (False) and high (True) group are >0.05 AUROC and Wilcoxon rank sum test Bonferroni adjusted p<0.05.

To determine how models work SHAP analysis was performed on the 11 outcomes whose performance was greater than an AUROC of 0.5. Magnitudes of SHAP values are shown in **Figure 6.8** across models to determine which features appear to be consistently important to different models**.** Notably, basal energy expenditure, step count, and environmental audio features rank near the top features used by most models.

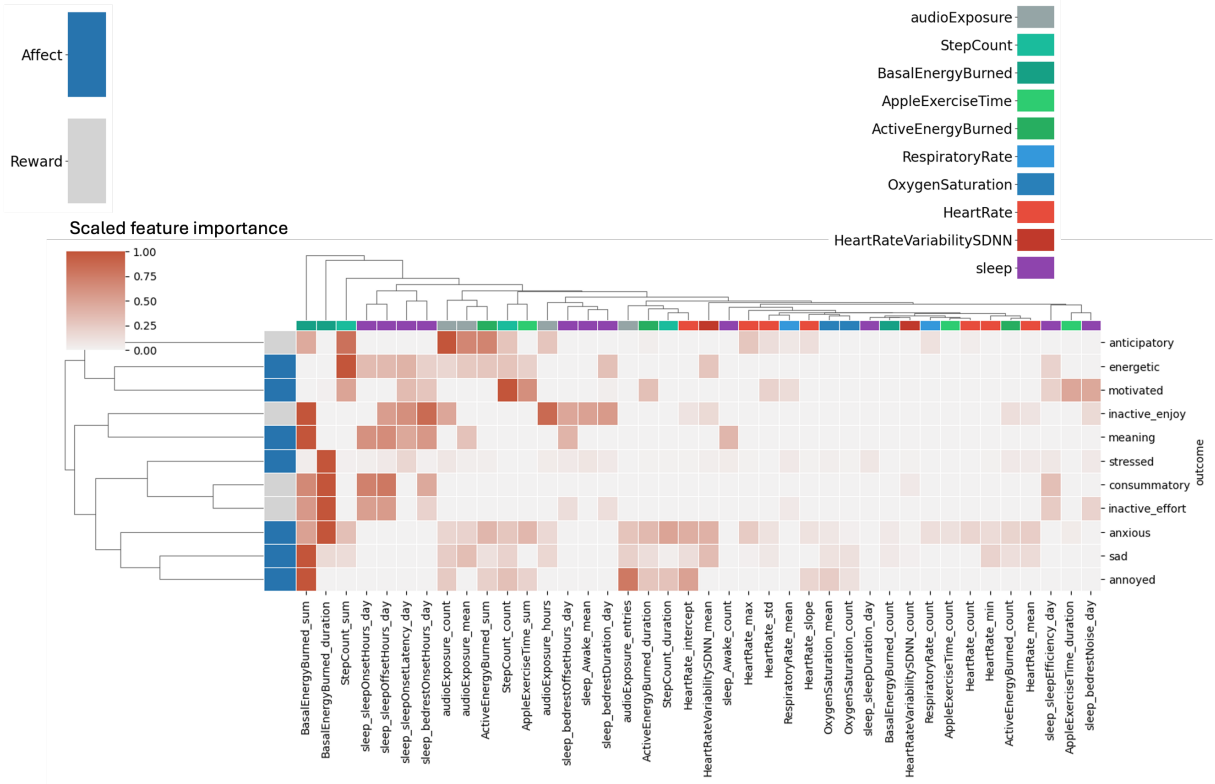

Figure 6.8 Clustered heatmap of the average magnitude of SHAP feature importance for models using passive sensor features (x-axis) to detect EMA outcomes (y-axis). Higher values indicate more importance of a given feature to a model. SHAP values scaled to maximum value per outcome. Row colors indicate if item is related to affect or reward functioning. Column colors indicate type of sensor feature.

To further investigate what basal energy expenditure may represent, **Figure 6.9** illustrates the median value of basal energy expenditure across a 24-hour period split by users' sex at birth.

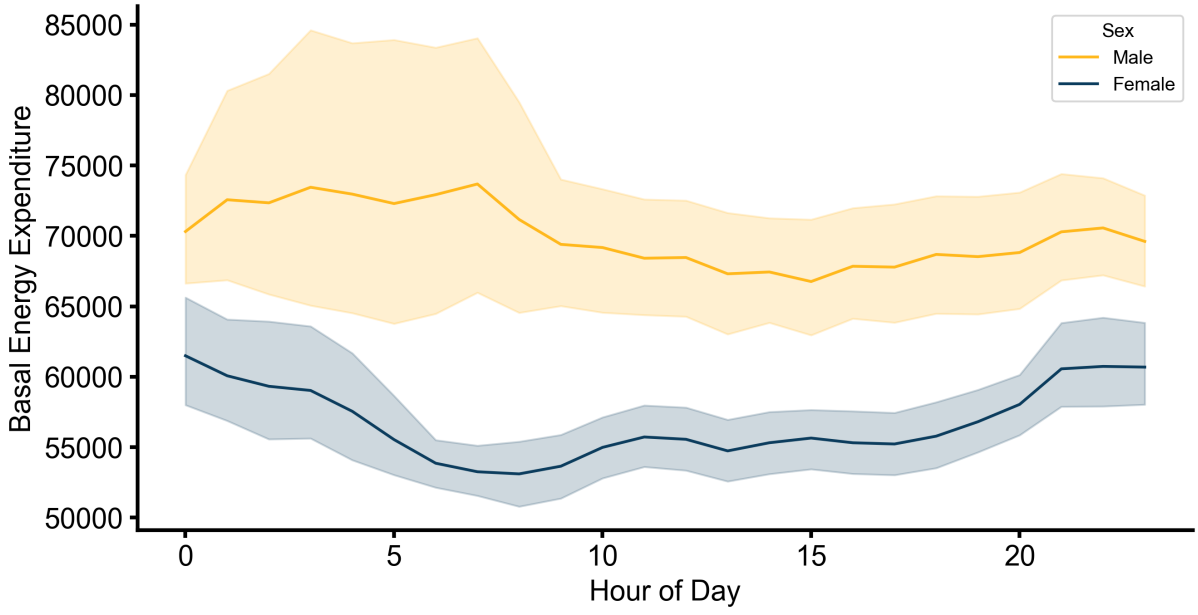

Figure 6.9 Median value per user of basal energy expenditure per hour of the day split by sex at birth. Error bounds represent 95% confidence intervals.

6.3.3 Sensitivity to Missing Data

There are significant differences in performance on the basis of data availability. For two of the EMA items with performance greater than random chance (anxious and energetic), performance is significantly greater in the full sample than in the sample missing fewer data points, suggesting that the highly missing features are not useful to the models when present. In contrast, the performances of the models that detect EMA items for consummatory, sad, inactive enjoy, anticipatory, and meaning are all higher in the sample with less missing data. The full results of the sensitivity analysis are summarized in **Table 6.2**. This may be explained in part as models whose performance degrades, such as the sad EMA rely on features that are sampled less frequently (see **Figure 6.2**) and are more likely to be missing, such as HRV (see **Figure 6.8**).

Table 6.2 Difference in performance of models detecting EMA item response in the test set on either the full test set, or a subset of responses with less than 10 of 29 features missing. Bold values indicate FWER adjusted p-value < 0.05 for difference and the full sample model performed greater than random chance (AUROC > 0.5 and adjusted p-value < 0.05).

EMA Item	Features	Window	Mean AUROC		AUROC Difference	Adjusted P-value
			Full sample	Missing < 10 features		
anxious	momentary	2h	0.545	0.510	0.035	**1.59E-26**
consummatory	with sleep	2h	0.543	0.559	-0.016	**9.04E-17**
sad	momentary	3h	0.572	0.584	-0.012	**1.18E-11**
inactive enjoy	with sleep	2h	0.551	0.563	-0.012	**5.13E-10**
anticipatory	momentary	2h	0.549	0.557	-0.008	**1.68E-04**
meaning	with sleep	2h	0.532	0.539	-0.007	**2.88E-04**
energetic	with sleep	3h	0.586	0.580	0.006	**1.54E-02**
annoyed	momentary	2h	0.540	0.536	0.005	5.13E-02
motivated	with sleep	3h	0.567	0.562	0.005	1.53E-01
stressed	with sleep	1h	0.554	0.556	-0.002	1.00E+00
inactive effort	with sleep	2h	0.550	0.551	-0.001	1.00E+00

6.4 Discussion

Our findings indicate that there is no single best approach to detect all 15 EMA items related to affect and reward functioning. Instead, each EMA item is best predicted by different feature sets and time windows for feature aggregation, and there does not appear to be a rule-of-thumb aggregation window or feature set. For most models, at least two hours of passive data aggregation are required for the best results; however, this may not apply if the sampling frequency of measures like HRV increase in future device or operating system versions. Many items that we might expect to be influenced by prior night's sleep do indeed show improved performance with sleep quality data from the prior night included. For example, four of the top five features used by the model for detecting "energetic" are related to sleep measures. Similarly, as we may expect elevated heart rate to occur when an individual feels anxious, heart rate is a key

feature utilized by the models predicting the “anxious” EMA item, and specifically, a higher heart rate is associated with the endorsement of feeling anxious.

Unexpectedly, basal energy burned emerged as a prominent feature in most models. The authors could not find documentation from Apple on how this HealthKit data type is calculated. The most common formula to calculate basal energy expenditure is the Harris–Benedict equation, which considers it a constant function of age, sex, body weight, and height[133]; however, in HealthKit, we observe that basal energy is not constant per individual, with circadian fluctuations in this value throughout a day. Models may perhaps be learning sex at birth from the basal energy burned feature as the baseline value is significantly different across males and females in **Figure 6.9**.

The significant differences in model performance based on symptom severity and demographic characteristics have implications for future modeling efforts. Many models perform better for older and male participants. This is not an effect of imbalanced data, as only 32.7% of the participants used in the analysis (both training and test) are male, and the definition of older was based on the median age in the test set. There is additionally no significant difference in the item response distribution between males and females in the test set for items that have significantly different performance except for the inactive effort EMA item. Additionally, recent work from Adler et al. confirms that the relationship between smartphone sensed behavior and depression risk can even invert in different demographic groups[75]. These performance discrepancies suggest that (for these models), the measurable behaviors from digital health devices, as assessed in this study, are most informative for the older male population, potentially reflecting a

bias in device or sensor design. Other sensors and features may enable broader detection of these items.

The items for which performance is better in the low depression severity group may represent models that can better generalize to a more general (less depressed) population, as shown in **Figure 6.4**. This includes items such as stress and effort. Notably, the sad EMA response distribution was not significantly different between the OPTIMA cohort and the MTurk-generated sample, but performance in detecting the item with mHealth data was better in those with depression, suggesting that the behaviors linked to the endorsement of momentary sadness are different in those with more severe depression than in a population sample even if the frequency of endorsement was similar.

This work focuses on using nomothetic models to understand how passive sensor features are related to momentary affect and reward. Models in digital sensing for mental health are often trained in a personalized fashion (e.g., idiographic or n-of-1) because performance without personalization can be low or minimal due to sample size restrictions and the complexity of the detection task[68,69,121]. The use of nomothetic modeling demonstrates that these sensors provide metrics that are useful in detecting momentary states across individuals. By investigating performance differences by demographic characteristics, depression severity, and anhedonia, we find that these parameters significantly affect model performance and begin to explain some of the characteristics that drive the heterogeneity associated with linking behavior to mental health outcomes. This finding serves as a baseline for building future models that can leverage participant-specific information to further enhance model performance.

6.4.1 Limitations

Missing sensor data are prevalent, especially for vital signs such as respiratory rate and heart rate variability during smaller aggregation windows. This is a limitation of the sampling frequency from the consumer wearable device, which occurs dynamically. Future work may look at what the relevant aggregation period is per sensor type. With the sample size available from EMA data, temporally aware deep learning models might be better suited to this task, as they can work with the raw time series data and implicitly select the most salient time span for a sensor relative to the specific EMA response.

6.4.2 Conclusion

We found that it is possible to detect responses to subjective momentary assessments of affect and reward functioning using only passive data and, importantly, without any prior information on a participant. As it precludes the need for the collection of burdensome person-specific training data, this finding represents an early step toward enabling scalable continuous passive mental health monitoring for the large existing user base of consumer wearable devices and smartphones. Another goal of this work was to determine the optimal time window across which to aggregate passive sensor data relative to EMA response items. Perhaps unsurprisingly, the optimal window varied across EMA items (with a general minimum of 2 hours) based on which sensor features were included. This finding highlights the limitations of one-size-fits-all analytic approaches when leveraging passive data from wearables to detect momentary affect and reward functioning. Ultimately, we find that there is signal in passively collected mHealth data to detect momentary affect and reward in those with depression and highlight key distinguishing participant characteristics influencing model performance.

Chapter 7: Conclusion

7.1 Summary of Research

The work of my dissertation begins by attempting to make few assumptions about the underlying passive data used and investigating in broad strokes what passive sensor data may tell us about depression. Chapter 3 leverages machine learning to find relationships between device-measured behaviors and retrospective self-reports on depression severity and related constructs. In Chapter 3, we find that several specific items are detectable with passive data and additionally note that although sleep quality is measured by the watch and phone, few items relevant to self-reported sleep quality are detectable.

In Chapter 4, sleep quality, as measured physiologically by the phone and watch or via self-reports, is assessed in relation to depression severity. Chapter 4 demonstrates that, particularly in the depressed population recruited in the OPTIMA study, physiologically measured sleep and self-reported sleep do not measure the same construct, emphasizing the need to measure both when studying the role of sleep in depression.

In the analyses presented in Chapter 4, there are two key underutilized qualities of mHealth data for the study of depression: 1) the multidimensional nature of mHealth data, which comprises many data streams, and 2) the fine-grained temporality of mHealth data, which allows real-time monitoring. In Chapter 5, the anomaly detection methods explored were aimed at leveraging the multidimensional nature of mHealth data to identify whether the number of deviations an individual has from their normal

behavior (as measured by mHealth devices) is correlated with changes in depression severity and depression symptoms. Although no significant relationship was found, the AD methods appear able to detect multivariate anomalies in a manner that picks up expected signals and may have uses outside of depression. In the case of depression, being unable to label an anomaly as a positive or negative event limits how anomalies may be related to symptom progression.

Finally, in Chapter 6, we investigate leveraging the real-time fine resolution of data that consumer wearable devices deliver to detect reward functioning and affect in the moment. We find that many machine learning models trained to detect these EMA items can do so above random chance, and performance varies drastically across different demographic, depressive, and anhedonic groups. This finding lays a foundation of considerations when building momentary predictive models for depression and key factors associated with heterogeneity in the presentation of depression.

7.2 Key Findings

Across my work in this dissertation, I find that while mHealth data hold potential for detecting symptoms of depression, considerable advancements are necessary to realize certain ambitious use cases, such as passive preventative monitoring. With consumer device data, we can identify meaningful signals between subjective internal states and passive mHealth data without requiring personalization. These signals are evident over both extended periods, such as weeks, and at momentary levels. Additionally, mHealth data assist in clarifying discrepancies between self-reported and actual behaviors, a critical issue in mental health studies. For example, significant differences between perceived and actual sleep quality can be independently assessed

via mHealth data, a principle that can be applied to other areas that often rely on retrospective self-reports due to the challenges of naturalistic measurement. While most data captured by these devices may not be directly related to mental health, their utility increases when they are combined with other data streams over time.

However, several challenges remain. The performance of models across participants, though better than random chance, is not sufficiently reliable for deployment. The data predominantly come from watches, which, despite their growing use, are less widespread than smartphones alone. Additionally, the infrastructure and processing requirements for mHealth data present a complex, interdisciplinary task that is not yet standardized. This complexity introduces multiple layers of noise and inconsistency, limiting the replicability of findings.

7.3 Future Directions

The work of this dissertation is intended to lay a foundation of what depression symptoms we might be able to use modern consumer wearable device data to detect, specifically by leveraging sleep, vitals, physical activity, and ambient noise. There is significant work to be done to translate the findings presented into solutions that enhance patient care, as any application will be specific to the community it is aimed at helping.

7.3.1 Standardization of digital sensing for mental health

Consumer technology is subject to changing versions of software and hardware over time that meaningfully impact the features they generate[115,134,135]. Research in digital sensing for mental health that can meaningfully build off prior studies and replicate

findings will require a degree of standardization. Fortunately, this work is in progress with efforts such as open mHealth[136] and discussions between involved stakeholders such as the “Workshop on Advancing the Utility of Digital Sensing Tools for Mental Health Research” held in March 2023.

7.3.2 Leveraging Deep Learning

The large datasets generated in digital mental health studies present an opportunity to leverage more advanced machine learning methods than those utilized in this dissertation. I focus on relatively simple methodologies such as random forest models as a first pass to understand the potentially high value targets to focus on. A more complex methodology leveraging temporally aware deep learning frameworks or transformer-based modeling approaches may be able to uncover associations between mHealth data and depression symptomology that are not possible with simpler approaches. These more computationally expensive approaches hold more promise as the digital mental health datasets being generated grow larger and explainability tools become more advanced, lowering the tradeoff between model complexity and interpretability.

References

1. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

2. Solomon, A. *The Noonday Demon: An Atlas of Depression.* (Scribner/Simon & Schuster, New York, NY, US).

3. Regier, D. A. *et al.* DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *Am. J. Psychiatry* **170**, 59–70 (2013).

4. Lee, B. *et al.* National, State-Level, and County-Level Prevalence Estimates of Adults Aged ≥18 Years Self-Reporting a Lifetime Diagnosis of Depression — United States, 2020. *Morb. Mortal. Wkly. Rep.* **72**, 644–650 (2023).

5. Fried, E. Moving forward: how depression heterogeneity hinders progress in treatment and research. *Expert Rev. Neurother.* **17**, 423–425 (2017).

6. Chekroud, A. M. *et al.* Reevaluating the Efficacy and Predictability of Antidepressant Treatments: A Symptom Clustering Approach. *JAMA Psychiatry* **74**, 370 (2017).

7. Kotov, R. *et al.* The Hierarchical Taxonomy of Psychopathology (HiTOP): A Quantitative Nosology Based on Consensus of Evidence. *Annu. Rev. Clin. Psychol.* **17**, 1–26 (2021).

8. Kotov, R. *et al.* Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): I. Psychosis superspectrum. *World Psychiatry* **19**, 151–172 (2020).

9. Krueger, R. F. *et al.* Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): II. Externalizing superspectrum. *World Psychiatry* **20**, 171–193 (2021).

10. Watson, D. *et al.* Validity and utility of Hierarchical Taxonomy of Psychopathology (HiTOP): III. Emotional dysfunction superspectrum. *World Psychiatry* **21**, 26–54 (2022).

11. Cuthbert, B. N. & Insel, T. R. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. *BMC Med.* **11**, 126 (2013).

12. Michelini, G., Palumbo, I. M., DeYoung, C. G., Latzman, R. D. & Kotov, R. Linking RDoC and HiTOP: A new interface for advancing psychiatric nosology and neuroscience. *Clin. Psychol. Rev.* **86**, 102025 (2021).

13. Thompson, P. M. *et al.* ENIGMA and global neuroscience: A decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl. Psychiatry* **10**, 100 (2020).

14. Waszczuk, M. A. *et al.* Redefining Phenotypes to Advance Psychiatric Genetics: Implications From Hierarchical Taxonomy of Psychopathology. *J. Abnorm. Psychol.* **129**, 143–161 (2020).

15. Vogels, E. A. 21% of Americans use a smart watch or fitness tracker. *Pew Research* https://www.pewresearch.org/short-reads/2020/01/09/about-one-in-five-americans-use-a-smart-watch-or-fitness-tracker/.

16. Kemp, S. The Rise of Smartwatches. *DataReportal – Global Digital Insights* https://datareportal.com/reports/digital-2023-deep-dive-the-rise-of-wearables.

17. Musliner, K. L., Munk-Olsen, T., Eaton, W. W. & Zandi, P. P. Heterogeneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes. *J. Affect. Disord.* **192**, 199–211 (2016).

18. Ancoli-Israel, S. *et al.* The Role of Actigraphy in the Study of Sleep and Circadian Rhythms. *Sleep* **26**, 342–392 (2003).

19. Hickey, B. A. *et al.* Smart Devices and Wearable Technologies to Detect and Monitor Mental Health Conditions and Stress: A Systematic Review. *Sensors* **21**, 3461 (2021).

20. Angel, V. D. *et al.* Digital health tools for the passive monitoring of depression: a systematic review of methods. *Npj Digital Medicine* **5**, 3 (2022).

21. Leaning, I. E. *et al.* From smartphone data to clinically relevant predictions: A systematic review of digital phenotyping methods in depression. *Neurosci. Biobehav. Rev.* **158**, 105541 (2024).

22. Zheng, N. S. *et al.* Sleep patterns and risk of chronic disease as measured by long-term monitoring with commercial wearable devices in the All of Us Research Program. *Nat. Med.* 1–9 (2024) doi:10.1038/s41591-024-03155-8.

23. Kroenke, K. *et al.* The PHQ-8 as a measure of current depression in the general population. *J. Affect. Disord.* **114**, 163–173 (2009).

24. Melcher, J. *et al.* Toward a Mobile Platform for Real-world Digital Measurement of Depression: User-Centered Design, Data Quality, and Behavioral and Clinical Modeling. *JMIR Ment Heal* **8**, e27589 (2021).

25. Daniel, D. M. *et al.* Automated Screening for Social Anxiety, Generalized Anxiety, and Depression From Objective Smartphone-Collected Data: Cross-sectional Study. *J Med Internet Res* **23**, e28918 (2021).

26. Saeb, S. *et al.* The Relationship between Clinical, Momentary, and Sensor-based Assessment of Depression. in *Proceedings of the 9th International Conference on Pervasive Computing Technologies for Healthcare* 229--232 (ICST, 2015). doi:10.4108/icst.pervasivehealth.2015.259034.

27. Saeb, S. *et al.* Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. *J Med Internet Res* **17**, e175 (2015).

28. Braund, T. A. *et al.* Smartphone Sensor Data for Identifying and Monitoring Symptoms of Mood Disorders: A Longitudinal Observational Study. *JMIR Ment. Heal.* **9**, e35549 (2022).

29. Stamatis, C. A. *et al.* Differential temporal utility of passively sensed smartphone features for depression and anxiety symptom prediction: a longitudinal cohort study. *npj Ment. Heal. Res.* **3**, 1 (2024).

30. Zhang, Y. *et al.* Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study. *Jmir Ment Heal* **9**, e34898 (2022).

31. Ware, S. *et al.* Predicting depressive symptoms using smartphone data. *Smart Heal.* **15**, 100093 (2020).

32. Laiou, P. *et al.* The Association Between Home Stay and Symptom Severity in Major Depressive Disorder: Preliminary Findings From a Multicenter Observational Study Using Geolocation Data From Smartphones. *JMIR mHealth uHealth* **10**, e28095 (2022).

33. Xia, C. H. *et al.* Mobile footprinting: linking individual distinctiveness in mobility patterns to mood, sleep, and brain functional connectivity. *Neuropsychopharmacol* 1–10 (2022) doi:10.1038/s41386-022-01351-z.

34. Cao, J. *et al.* Tracking and Predicting Depressive Symptoms of Adolescents Using Smartphone-Based Self-Reports, Parental Evaluations, and Passive Phone Sensor Data: Development and Usability Study. *JMIR Ment Heal* **7**, e14045 (2020).

35. Currey, D. & Torous, J. Digital phenotyping correlations in larger mental health samples: analysis and replication. *Bjpsych Open* **8**, e106 (2022).

36. Sverdlov, O. *et al.* A Study of Novel Exploratory Tools, Digital Technologies, and Central Nervous System Biomarkers to Characterize Unipolar Depression. *Front. Psychiatry* **12**, 640741 (2021).

37. Zou, B. *et al.* Sequence Modeling of Passive Sensing Data for Treatment Response Prediction in Major Depressive Disorder. *IEEE Trans. Neural Syst. Rehabilitation Eng.* **31**, 1786–1795 (2023).

38. Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K. & Geralts, D. S. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology. *J. Neurolinguistics* **20**, 50–64 (2007).

39. Mundt, J. C., Vogel, A. P., Feltner, D. E. & Lenderking, W. R. Vocal Acoustic Biomarkers of Depression Severity and Treatment Response. *Biol. Psychiatry* **72**, 580–587 (2012).

40. Low, D. M., Bentley, K. H. & Ghosh, S. S. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investig. Otolaryngol.* **5**, 96–116 (2020).

41. Wasserzug, Y. *et al.* Development and validation of a machine learning-based vocal predictive model for major depressive disorder. *J. Affect. Disord.* **325**, 627–632 (2023).

42. Bensoussan, Y., Elemento, O. & Rameau, A. Voice as an AI Biomarker of Health—Introducing Audiomics. *JAMA Otolaryngol.Head Neck Surg.* **150**, (2024).

43. Dagum, P. Digital biomarkers of cognitive function. *Npj Digital Medicine* **1**, 10 (2018).

44. Ning, E. *et al.* Smartphone-derived Virtual Keyboard Dynamics Coupled with Accelerometer Data as a Window into Understanding Brain Health. *Proc. 2023 CHI Conf. Hum. Factors Comput. Syst.* **2023**, 1–15 (2023).

45. Knol, L. *et al.* Smartphone keyboard dynamics predict affect in suicidal ideation. *npj Digit. Med.* **7**, 54 (2024).

46. Moshe, I. *et al.* Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. *Frontiers Psychiatry* **12**, 625247 (2021).

47. Wüthrich, F., Nabb, C. B., Mittal, V. A., Shankman, S. A. & Walther, S. Actigraphically measured psychomotor slowing in depression: systematic review and meta-analysis. *Psychol. Med.* **52**, 1208–1221 (2022).

48. Difrancesco, S. *et al.* Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: A 2-week ambulatory assessment study. *Dépress. Anxiety* **36**, 975–986 (2019).

49. Razjouyan, J. *et al.* Improving Sleep Quality Assessment Using Wearable Sensors by Including Information From Postural/Sleep Position Changes and Body Acceleration: A Comparison of Chest-Worn Sensors, Wrist Actigraphy, and Polysomnography. *J. Clin. Sleep Med.* **13**, 1301–1310 (2017).

50. Apple. Estimating Sleep Stages from Apple Watch. https://www.apple.com/healthcare/docs/site/Estimating_Sleep_Stages_from_Apple_Watch_Sept_2023.pdf (2023).

51. Patterson, M. R. *et al.* 40 years of actigraphy in sleep medicine and current state of the art algorithms. *npj Digit. Med.* **6**, 51 (2023).

52. Glaus, J. *et al.* Objectively assessed sleep and physical activity in depression subtypes and its mediating role in their association with cardiovascular risk factors. *J. Psychiatr. Res.* **163**, 325–336 (2023).

53. Langholm, C., Byun, A. J. S., Mullington, J. & Torous, J. Monitoring sleep using smartphone data in a population of college students. *npj Ment. Heal. Res.* **2**, 3 (2023).

54. Matthews, K. A. *et al.* Similarities and differences in estimates of sleep duration by polysomnography, actigraphy, diary, and self-reported habitual sleep in a community sample. *Sleep Heal.* **4**, 96–103 (2018).

55. Neishabouri, A. *et al.* Quantification of acceleration as activity counts in ActiGraph wearable. *Sci. Rep.* **12**, 11958 (2022).

56. Ronca, V. *et al.* Wearable Technologies for Electrodermal and Cardiac Activity Measurements: A Comparison between Fitbit Sense, Empatica E4 and Shimmer GSR3+. *Sensors* **23**, 5847 (2023).

57. E4 wristband | Real-time physiological signals | Wearable PPG, EDA, Temperature, Motion sensors. https://www.empatica.com/research/e4/.

58. Alvares, G. A., Quintana, D. S., Hickie, I. B. & Guastella, A. J. Autonomic nervous system dysfunction in psychiatric disorders and the impact of psychotropic medications: a systematic review and meta-analysis. *J. Psychiatry Neurosci.* **41**, 89–104 (2016).

59. Yeragani, V. K. *et al.* Heart rate variability in patients with major depression. *Psychiatry Res.* **37**, 35–46 (1991).

60. Agelink, M. W., Boz, C., Ullrich, H. & Andrich, J. Relationship between major depression and heart rate variability. Clinical consequences and implications for antidepressive treatment. *Psychiatry Res.* **113**, 139–149 (2002).

61. Geiss, L., Beck, B., Stemmler, M., Hillemacher, T. & Hösl, K. M. Heart rate variability during inpatient treatment of depression. *J. Mood Anxiety Disord.* **6**, 100059 (2024).

62. Schiweck, C., Piette, D., Berckmans, D., Claes, S. & Vrieze, E. Heart rate and high frequency heart rate variability during stress as biomarker for clinical depression. A systematic review. *Psychol. Med.* **49**, 200–211 (2019).

63. Hirten, R. P. *et al.* Factors Associated With Longitudinal Psychological and Physiological Stress in Health Care Workers During the COVID-19 Pandemic: Observational Study Using Apple Watch Data. *J Med Internet Res* **23**, e31295 (2021).

64. Masaoka, Y. & Homma, I. Anxiety and respiratory patterns: their relationship during mental stress and physical load. *Int. J. Psychophysiol.* **27**, 153–159 (1997).

65. Kral, T. R. A. *et al.* Slower respiration rate is associated with higher self-reported well-being after wellness training. *Sci. Rep.* **13**, 15953 (2023).

66. Zamanzadeh, D. J. Imputation Is a Hyperparameter: Imputation Deep Learning Model Selection and Evaluation on Large Clinical Datasets. (University of California, Los Angeles, 2023).

67. Taylor, S., Jaques, N., Nosakhare, E., Sano, A. & Picard, R. Personalized Multitask Learning for Predicting Tomorrow's Mood, Stress, and Health. *IEEE T Affect Comput* **11**, 200–213 (2017).

68. Sükei, E., Norbury, A., Perez-Rodriguez, M. M., Olmos, P. M. & Artés, A. Predicting Emotional States Using Behavioral Markers Derived From Passively Sensed Data: Data-Driven Machine Learning Approach. *JMIR mHealth and uHealth* **9**, e24465 (2021).

69. Yu, H. & Sano, A. Passive Sensor Data Based Future Mood, Health, and Stress Prediction: User Adaptation Using Deep Learning. *2020 42nd Annu Int Conf IEEE Eng Medicine Biology Soc EMBC* **00**, 5884–5887 (2020).

70. Li, B. & Sano, A. Early versus Late Modality Fusion of Deep Wearable Sensor Features for Personalized Prediction of Tomorrow's Mood, Health, and Stress*. *2020 42nd Annu Int Conf IEEE Eng Medicine Biology Soc EMBC* **00**, 5896–5899 (2020).

71. Lewis, R. A. *et al.* Mixed Effects Random Forests for Personalised Predictions of Clinical Depression Severity. *arXiv* (2023) doi:10.48550/arxiv.2301.09815.

72. Rashid, H. *et al.* Predicting Subjective Measures of Social Anxiety from Sparsely Collected Mobile Sensor Data. *Proc Acm Interact Mob Wearable Ubiquitous Technologies* **4**, 1–24 (2020).

73. Chikwetu, L. *et al.* Does deidentification of data from wearable devices give us a false sense of security? A systematic review. *Lancet Digit. Heal.* **5**, e239–e247 (2023).

74. Balliu, B. *et al.* Personalized mood prediction from patterns of behavior collected with smartphones. *npj Digit. Med.* **7**, 49 (2024).

75. Adler, D. A. *et al.* Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data. *npj Ment. Heal. Res.* **3**, 17 (2024).

76. Barnett, I. *et al.* Relapse prediction in schizophrenia through digital phenotyping: a pilot study. *Neuropsychopharmacol* **43**, 1660–1666 (2018).

77. Henson, P., D'Mello, R., Vaidyam, A., Keshavan, M. & Torous, J. Anomaly detection to predict relapse risk in schizophrenia. *Transl Psychiat* **11**, 28 (2021).

78. Cohen, A. *et al.* Relapse prediction in schizophrenia with smartphone digital phenotyping during COVID-19: a prospective, three-site, two-country, longitudinal study. *Schizophrenia* **9**, 6 (2023).

79. Amor, L. B., Lahyani, I., Jmaiel, M. & Drira, K. Anomaly Detection and Diagnosis Scheme for Mobile Health Applications. *2018 IEEE 32nd Int Conf Adv Information Netw Appl AINA* 777–784 (2018) doi:10.1109/aina.2018.00116.

80. Devarajan, K. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. *Plos Comput Biol* **4**, e1000029 (2008).

81. Nutt, D. *et al.* The other face of depression, reduced positive affect: the role of catecholamines in causation and cure. *J. Psychopharmacol.* **21**, 461–471 (2007).

82. Uher, R. *et al.* Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. *Psychol. Med.* **42**, 967–980 (2012).

83. Akre, S. *et al.* Detection of Symptoms of Depression Using Data From the iPhone and Apple Watch. *2023 IEEE Int. Conf. Bioinform. Biomed. (BIBM)* 1818–1823 (2023) doi:10.1109/bibm58861.2023.10385797.

84. Fried, E. I. The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *J Affect Disorders* **208**, 191–197 (2017).

85. Mohr, D. C., Zhang, M. & Schueller, S. M. Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning. *Annu Rev Clin Psycho* **13**, 23--47 (2017).

86. Kroenke, K., Spitzer, R. L. & Williams, J. B. W. The PHQ-9: validity of a brief depression severity measure. *J. Gen. Intern. Med.* **16**, 606–613 (2001).

87. Wu, Y. *et al.* Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. *Psychol. Med.* **50**, 1368–1380 (2020).

88. Khazanov, G. K., Ruscio, A. M. & Forbes, C. N. The Positive Valence Systems Scale: Development and Validation. *Assessment* **27**, 1045–1069 (2020).

89. Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R. & Kupfer, D. J. The Pittsburgh sleep quality index: A new instrument for psychiatric practice and research. *Psychiatry Res.* **28**, 193–213 (1989).

90. Carpenter, J. S. & Andrykowski, M. A. Psychometric evaluation of the pittsburgh sleep quality index. *J. Psychosom. Res.* **45**, 5–13 (1998).

91. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* **12**, 2825–2830 (2011).

92. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *CoRR* **abs/1603.02754**, (2016).

93. Vallat, R. Pingouin: statistics in Python. *J. Open Source Softw.* **3**, 1026 (2018).

94. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc.).

95. Muzni, K., Groeger, J. A., Dijk, D. & Lazar, A. S. Self-reported sleep quality is more closely associated with mental and physical health than chronotype and sleep duration in young adults: A multi-instrument analysis. *J. Sleep Res.* **30**, e13152 (2021).

96. Staner, L. Comorbidity of insomnia and depression. *Sleep Med. Rev.* **14**, 35–46 (2010).

97. Chang, Q. *et al.* Association Between Pittsburgh Sleep Quality Index and Depressive Symptoms in Chinese Resident Physicians. *Front. Psychiatry* **12**, 564815 (2021).

98. Kaplan, K. A. *et al.* When a gold standard isn't so golden: Lack of prediction of subjective sleep quality from sleep polysomnography. *Biol. Psychol.* **123**, 37–46 (2017).

99. Faerman, A., Kaplan, K. A. & Zeitzer, J. M. Subjective sleep quality is poorly associated with actigraphy and heart rate measures in community-dwelling older men. *Sleep Med.* **73**, 154–161 (2020).

100. Baril, A. *et al.* Misappraisal of sleep quality is associated with lower cognitive functioning. *Alzheimer's Dement.* **17**, (2021).

101. Gualtieri, C. T., Johnson, L. G. & Benedict, K. B. Neurocognition in Depression: Patients on and Off Medication Versus Healthy Comparison Subjects. *J. Neuropsychiatry Clin. Neurosci.* **18**, 217–225 (2006).

102. Lauderdale, D. S., Knutson, K. L., Yan, L. L., Liu, K. & Rathouz, P. J. Self-Reported and Measured Sleep Duration. *Epidemiology* **19**, 838–845 (2008).

103. Cheng, P., Walch, O., Hannay, K., Roth, T. & Drake, C. 0004 Using Apple Watch to predict circadian phase in night shift workers. *SLEEP* **46**, A2–A2 (2023).

104. Zambotti, M. de *et al.* State of the science and recommendations for using wearable technology in sleep and circadian research. *SLEEP* zsad325 (2023) doi:10.1093/sleep/zsad325.

105. Singh, S. *et al.* The TestMyBrain Digital Neuropsychology Toolkit: Development and Psychometric Characteristics. *J. Clin. Exp. Neuropsychol.* **43**, 786–795 (2021).

106. Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P. & Mohr, D. C. The relationship between mobile phone location sensor data and depressive symptom severity. *Peerj* **4**, e2537 (2016).

107. LiKamWa, R., Liu, Y., Lane, N. D. & Zhong, L. MoodScope: building a mood sensor from smartphone usage patterns. *Proceeding 11th Annu Int Conf Mob Syst Appl Serv - Mobisys '13* 389–402 (2013) doi:10.1145/2462456.2464449.

108. Torous, J., Kiang, M. V., Lorme, J. & Onnela, J.-P. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *Jmir Ment Heal* **3**, e16 (2016).

109. Kalmbach, D. A., Arnedt, J. T., Pillai, V. & Ciesla, J. A. The Impact of Sleep on Female Sexual Response and Behavior: A Pilot Study. *J. Sex. Med.* **12**, 1221–1232 (2015).

110. Kohn, T. P., Kohn, J. R., Haney, N. M., Pastuszak, A. W. & Lipshultz, L. I. The effect of sleep on men's health. *Transl. Androl. Urol.* **9**, S178–S185 (2019).

111. Zavecz, Z., Nagy, T., Galkó, A., Nemeth, D. & Janacsek, K. The relationship between subjective sleep quality and cognitive performance in healthy young adults: Evidence from three empirical studies. *Sci. Rep.* **10**, 4855 (2020).

112. Passell, E. *et al.* Digital Cognitive Assessment: Results from the TestMyBrain NIMH Research Domain Criteria (RDoC) Field Test Battery Report. (2019) doi:10.31234/osf.io/dcszr.

113. Sun, S. *et al.* Challenges in Using mHealth Data From Smartphones and Wearable Devices to Predict Depression Symptom Severity: Retrospective Analysis. *J. Méd. Internet Res.* **25**, e45233 (2023).

114. Clayborne, Z. M. & Colman, I. Associations between Depression and Health Behaviour Change: Findings from 8 Cycles of the Canadian Community Health Survey. *Can. J. Psychiatry* **64**, 30–38 (2019).

115. Woolley, S. I., Collins, T., Mitchell, J. & Fredericks, D. Investigation of wearable health tracker version updates. *BMJ Heal. Care Inform.* **26**, e100083 (2019).

116. Xu, X. *et al.* GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. *Arxiv* (2022) doi:10.48550/arxiv.2211.02733.

117. Kroenke, K., Spitzer, R. L., Williams, J. B. W. & Lowe, B. An Ultra-Brief Screening Scale for Anxiety and Depression: The PHQ-4. *Psychosomatics* **50**, 613–621 (2009).

118. Beck, A. T., Steer, R. A., Ball, R. & Ranieri, W. F. Comparison of Beck Depression Inventories-IA and-II in Psychiatric Outpatients. *J Pers Assess* **67**, 588–597 (1996).

119. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

120. Association, A. P. Diagnostic and Statistical Manual of Mental Disorders, DSM-5. (2013) doi:10.1176/appi.books.9780890425596.

121. Shah, R. V. *et al.* Personalized machine learning of depressed mood using wearables. *Transl Psychiat* **11**, 338 (2021).

122. Ren, H., Ye, Z. & Li, Z. Anomaly detection based on a dynamic Markov model. *Inform Sciences* **411**, 52–65 (2017).

123. Harush, S., Meidan, Y. & Shabtai, A. DeepStream: Autoencoder-based stream temporal clustering and anomaly detection. *Comput Secur* **106**, 102276 (2021).

124. Bolger, N., Davis, A. & Rafaeli, E. Diary Methods: Capturing Life as it is Lived. *Psychology* **54**, 579–616 (2003).

125. Fritz, J. *et al.* So you want to do ESM? Ten Essential Topics for Implementing the Experience Sampling Method (ESM). *AMPPS* doi:10.31219/osf.io/fverx.

126. Hart, A., Reis, D., Prestele, E. & Jacobson, N. C. Using Smartphone Sensor Paradata and Personalized Machine Learning Models to Infer Participants' Well-being: Ecological Momentary Assessment. *J Med Internet Res* **24**, e34015 (2022).

127. Jacobson, N. C. & Chung, Y. J. Passive Sensing of Prediction of Moment-To-Moment Depressed Mood among Undergraduates with Clinical Levels of Depression Sample Using Smartphones. *Sensors* **20**, 3572 (2020).

128. Siepe, B. S., Tutunji, R., Rieble, C., Proppert, R. K. K. & Fried, E. I. Associations between ecological momentary assessment and passive sensor data in a large student sample. (2024) doi:10.31234/osf.io/ybns6.

129. Langener, A. M. *et al.* A template and tutorial for preregistering studies using passive smartphone measures. *Behav. Res. Methods* 1–19 (2024) doi:10.3758/s13428-024-02474-5.

130. Maienschein-Cline, M. *et al.* ARTS: automated randomization of multiple traits for study design. *Bioinformatics* **30**, 1637–1639 (2014).

131. Nelson, B. W. *et al.* Guidelines for wrist-worn consumer wearable assessment of heart rate in biobehavioral research. *Npj Digital Medicine* **3**, 90 (2020).

132. Wang, C., Wu, Q., Weimer, M. & Zhu, E. FLAML: A Fast and Lightweight AutoML Library. *arXiv* (2019) doi:10.48550/arxiv.1911.04706.

133. Harris, J. A. & Benedict, F. G. A Biometric Study of Human Basal Metabolism. *Proc. Natl. Acad. Sci.* **4**, 370–373 (1918).

134. Shcherbina, A. *et al.* Accuracy in Wrist-Worn, Sensor-Based Measurements of Heart Rate and Energy Expenditure in a Diverse Cohort. *J Personalized Medicine* **7**, (2017).

135. Kainec, K. A. *et al.* Evaluating Accuracy in Five Commercial Sleep-Tracking Devices Compared to Research-Grade Actigraphy and Polysomnography. *Sensors* **24**, 635 (2024).

136. Zeng, B. *et al.* Standardized Integration of Person-Generated Data Into Routine Clinical Care. *Jmir Mhealth Uhealth* **10**, e31048 (2022).