Ecole polytechnique fédérale de Lausanne

Predicting changes in depression using person-generated health data

A Master thesis by
Mariko Makhmutova

Supervised by
Dr. Ieuan Clay
Prof. Martin Jaggi

Machine Learning and Optimization Laboratory, EPFL
Evidation Health

August 19, 2021

Abstract

Predicting changes in depression using person-generated health data

Depression is a leading cause of disability, impacting the lives of an increasingly large number of individuals worldwide. Despite a range of treatment options, a substantial fraction of individuals experiencing depressive symptoms do not seek or receive treatment. Person-generated health data (PGHD), including self-reported data and consumer-grade wearable technology, can be used to observe individual behaviour and improve outcomes in depression. In this thesis, we propose PSYCHE-D (Prediction of SeveritY CHange - Depression), a two-phase classification model that predicts longitudinal change in depression severity using sparse depression severity labels, self-reported survey responses and consumer wearable data, trained on a large and diverse cohort comprising more than 10,000 samples from more than 4,000 individual participants. In the first phase, the model enhances the density of depression severity labels by generating intermediate monthly labels. The generated monthly labels, combined with the existing labels, self-reported survey responses and consumer-grade wearable data are used as inputs for the second phase, where PSYCHE-D predicts whether an individual experiences an increase in depression severity over a 3-month period, achieving a sensitivity of 55.4% and a specificity of 65.3%. We demonstrate a promising PGHD-based approach that can be used as a foundation for a low-burden consumer-facing system that could minimize barriers to depression diagnosis and treatment, and eventually improve outcomes in depression.

Acknowledgements

The work presented in this thesis could not have been accomplished without some spectacular people.

I would first like to thank Dr Ieuan Clay, my internship supervisor. His support and encouragement throughout the internship have inspired me far beyond the scope of this thesis. A problem can be approached in an infinite number of ways, but to solve it one must find the Middle Way [1].

I would like to thank Prof. Martin Jaggi, my EPFL Master's project supervisor, for his advice and support throughout the internship. This climb would have been much more difficult without his guidance.

I would like to thank everyone at Evidation, particularly the ReAL team and everyone who worked on the DiSCover project for their continuous support and advice throughout the project. I would particularly like to give a special thanks to three magnificent colleagues: Marta Ferreira, for her thoughtful advice and help throughout the internship; Raghu Kainkaryam, for helping me debug my ideas and reading through my code countless times; and Jae Min, for her incredible editing, enthusiasm, and help throughout the internship. You're the ReAL MVPs.

Finally, I would like to thank my friends and family for their unconditional support and encouragement throughout my studies.

Contents

Table of Abbreviations

The following table describes the meaning of various abbreviations used throughout the thesis.

Abbreviation	**Meaning**
AUPRC	Area under the precision-recall curve
AUROC	Area under the receiver operating curve
CI	Confidence interval
DART	Dropouts meet multiple additive regression trees (LightGBM boosting implementation)
GBDT	Gradient boosting decision tree (LightGBM boosting implementation)
HC	Highly correlated
LightGBM	Light gradient boosting machine
LMC	Lifestyle and medication changes
PGHD	Person-generated health data
Phase 1 (2) model	Initial implementation of the first (second) phase model of PSYCHE-D
Phase 1c (2c) model	Final "Combined" implementation of the first (second) phase of PSYCHE-D after eliminating data leakage
PHQ-9	Patient Health Questionnaire-9
PSYCHE-D	Prediction of SeveritY CHange - Depression, the two-phase approach presented in this work
SM0, SM1, SM2, SM3	Sample month 0-3
XGBoost	Extreme gradient boosting

Chapter 1

Introduction

1.1 Background

Depression is a leading cause of disability, impacting the lives of more than 264 million people worldwide, according to the World Health Organization [2, 3]. The COVID-19 pandemic has further increased the number of people suffering from depressive symptoms [4, 5]. The symptomatology of depression is variable both within and across individuals, and long-standing untreated depression, especially at higher severity, can become a serious health condition [2, 3].

Despite the risks associated with depression and its prevalence, depression, among other mental disorders, often remains undiagnosed and untreated. In 2017, an estimated 17.3 million adults in the US experienced at least one major depressive episode, with 35% of them not receiving any treatment [6]. In addition, a study of more than 240,000 adult patients with newly diagnosed depression found that only 35.7% of the patients initiated treatment within the 90 days following their diagnosis [7].

The under-diagnosis and under-treatment of depression has been attributed to many factors, including personal and perceived stigma surrounding mental health, lack of knowledge about depression, limited access to medical care and barriers due to cost [2, 3], [8, 9]. Leaving depression undiagnosed and untreated can have a significant impact on quality of life, negatively affecting factors such as physical health, productivity, and social relationships [10]. Additionally, undiagnosed and untreated depression has significant economic consequences, adding an economic burden of over $200 billion annually in the US alone [11]. It is thus essential to make detecting and monitoring depressive symptoms both easier and more affordable.

An increasingly explored and promising way to accomplish this is through person-generated health data (PGHD), defined as wellness and/or health-related data that has been created, recorded, or gathered by individuals [12]. PGHD can come from a variety of sources, including wearable devices, phones, electronic surveys, and mobile applications. Collecting PGHD through consumer-grade wearable devices is a low-burden approach to generating objective behavioural data, and can be applied at a large scale. PGHD allows researchers to observe longitudinal changes within a person's behaviour, as well as to perform comparative analyses within a population or cohort. In the context of depression, PGHD is being explored as a method to monitor depressive symptoms using consumer-grade wearable devices and smartphones [13, 14, 15]. Monitoring these symptoms over time is especially important as the rate of recurrence of major depressive symptoms is high even after full remission [16]. Additionally, delivering treatment in a timely manner is crucial, as the benefits of early intervention have been established for both older [17] and younger [18] individuals.

Monitoring and tracking changes in depression severity could thus be facilitated using low-burden PGHD approaches employed on a large scale. In this thesis we present a possible implementation of such a low-burden PGHD approach for predicting depression severity changes.

1.2 Related work

The use of PGHD in the form of wearable device data to observe and predict relationships with mental health status indicators is becoming increasingly popular. Multiple studies show that consumer-grade wearable devices can remotely measure activity features that can be used as measures of depression in the real world [19, 20].

One recent study used consumer wearable devices to study associations between sleep quality and depressive symptom severity for 368 participants [19]. Using linear mixed regression models, adjusted for socio-demographic factors, the authors calculated Z scores to evaluate the significance of different sleep features on depression severity. Their findings showed that insomnia, changes in sleep architecture, a high percentage of time spent awake while in bed, mean sleep offset time, and hypersomnia are all significantly associated with the degree of depressive symptom severity, with Z scores up to 6.19 ($P < 0.001$). Sleep efficiency was found to be significantly negatively correlated with depressive symptom severity. The authors noted that as the study was conducted in a real-life setting with a low-burden approach, there was an issue of missing data, as some participants did not consistently wear their devices or did not complete all study questionnaires, which was required for the model. This had to be taken into account and reduced the amount of data available for analysis, and is of high relevance for the work presented here.

In a more controlled context, another recent study used PGHD in the form of smartphone and consumer wearable data from 138 college students to identify students with depressive symptoms, and students whose depressive symptom severity worsened over the course of a semester [21]. The binary classification models achieved 85.7% accuracy in predicting the presence of post-semester depressive symptoms and 85.4% accuracy in predicting a worsening of symptom severity. The study found that absolute behavioural features (e.g. the total number of outgoing calls) are better than relative behavioural features (e.g, the change in number of outgoing calls) in detecting post-semester depression. This study also recognized the limits of a low-burden data collection process regarding missing data, and, where possible, missing data was imputed in a way to represent its missingness.

Another recent longitudinal study used smartphone and wearable device data collected over a 30-day period to predict depression and anxiety symptoms of 60 adult participants [22]. The study found that GPS and wearable device data, including total sleep time and time spent in bed, were good predictors of mental health status for depression. The study did not find physical activity measures, such as step data, to have a significant association with depression or anxiety. The authors acknowledged the sociodemographic bias of the small cohort, which may have caused a bias in the relationships between variables.

Other studies have shown that early indicators of changes in depression can be detected from PGHD in the form of social media use [23, 24] or physical activity patterns [25]. We can see that the use of low-burden PGHD collection to determine mental health status is an increasingly popular and growing field of research that is still being developed, with studies being performed on cohorts limited in size and biased in sociodemographic factors. The current work aims to further contribute to the growth and development of this field.

1.3 Aims and objectives

Depression is a common and serious health condition that affects an increasing number of individuals worldwide. Despite a large range of treatment options, the majority of individuals experiencing depressive symptoms do not seek or receive treatment. PGHD, including self-reported data and consumer-grade wearable technology, can be leveraged to observe individual behaviour, increase engagement and improve outcomes in depression. In this thesis, we propose a low-burden approach that leverages person-generated health data to predict and monitor changes in depression severity.

More specifically, we develop PSYCHE-D (Prediction of SeveritY CHange - Depression), a two-phase classification model that uses PGHD to predict longitudinal change in depression severity. Using sparse depression severity labels, self-reported survey responses and consumer-grade wearable data, the goal of the first phase of the model is to enhance the density of the labels by generating intermediate depression severity labels. These generated labels, combined with the existing labels, self-reported survey responses and consumer-grade wearable data, are inputs for the second phase model, where PSYCHE-D predicts whether an individual experiences an increase in depression severity over a 3-month period.

We aim for the work presented in this thesis to be used as a foundation for a low-burden consumer-facing system that could improve engagement, minimize barriers to depression diagnosis and treatment, and eventually improve outcomes in depression.

1.4 Outline

Before commencing the description of the contributions of this work, we provide the reader with an overview of the structure of the document. We first describe the source dataset, the DiSCover Project developed by Evidation Health (ClinicalTrials.gov Identifier: NCT03421223; [26]), as well as its exploration and processing stages. The final processed dataset is publicly available on Zenodo [27].

In the following sections we describe the development of the two separate phases of PSYCHE-D, as well as the development of the combined pipeline consisting of the two chained phases. Additional details on the first phase model can be found in [28]. In the subsequent sections we present and discuss the results of each of the two phases, as well as the final results of PSYCHE-D as a complete pipeline. The final sections discuss the impact and scientific contributions of this thesis, as well as the further outlook for our presented work.

Chapter 2

Methodology

2.1 Data collection

The data used in this thesis is part of the DiSCover Project developed by Evidation Health (ClinicalTrials.gov Identifier: NCT03421223; [26]). The DiSCover Project is a 1-year long longitudinal study consisting of 10,036 individuals who wore consumer wearable devices throughout the study and completed regular surveys about their mental health and/or lifestyle changes. Detailed design and baseline participant characteristics are described in [29].

More specifically, the data subset used in this work comprises the following (and is illustrated in Figure 2.1):

- *Wearable PGHD*: minute-level step and sleep data from the participants' consumer-grade wearable devices (Fitbit) worn throughout the study
- *Screener survey*: prior to the study, participants were requested to respond to screening and baseline surveys regarding socio-economic and demographic information, as well as comorbidities and life events (e.g. trauma, birth)
- *Lifestyle and medication changes (LMC) survey*: every month, participants were requested to complete a survey reporting changes in their lifestyle and medication over the past month (e.g. changes in eating habits or activity levels, starting new medication)
- *Patient Health Questionnaire-9 (PHQ-9) score*: every 3 months, participants were requested to complete the PHQ-9, a 9-item questionnaire that has proven to be reliable and valid to measure depression severity [30], in order to track changes in their depression levels. The questionnaire is provided in Appendix A

The total age range of participants is 18-85 years, the majority of participants in the cohort are young and middle-aged adults (mean=36.8 years; SD=10.5), 80.5% of the participants are Non-Hispanic White, and 73.7% are female. During the screening stage, participants were asked about pre-existing chronic conditions. There is a high

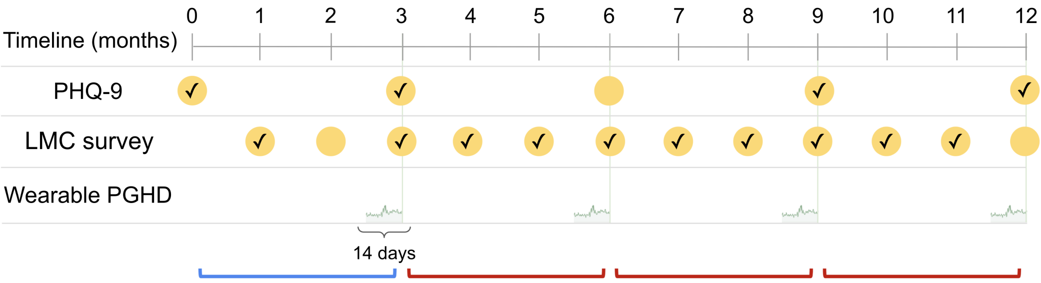


Figure 2.1: *Data collection timeline and survey completion criteria. Yellow circles represent data collection for a given month and data type, and check marks indicate completed surveys for an example participant. Wearable PGHD is collected in the 14 days prior to PHQ-9 completion date. Quarters marked in blue satisfy survey completion criteria, quarters marked in red do not. The first quarter (months 0-3), marked in blue, fulfills the selection criteria, as the baseline and final PHQ-9 were completed at months 0 and 3, and the LMC survey was completed at month 3. The second quarter, marked in red, is not valid as the final PHQ-9 (month 6) was not completed. The third quarter is not valid as the baseline PHQ-9 (month 6) was not completed. The fourth quarter is not valid as the LMC survey for the final month (month 12) was not completed.*

prevalence (58%) of participants reporting chronic pain conditions including migraines, osteoarthritis, fibromyalgia and peripheral nerve pain. In further analysis, we did not differentiate between participants who have been diagnosed with chronic pain conditions and those who haven't, as participants who report not having been diagnosed with chronic pain conditions may still have underlying conditions. Nonetheless, chronic pain comorbidities were included as features derived from the screener survey.

Figure 2.1 describes the data collection timeline. At the beginning of the study, participants completed the screener/baseline survey and the PHQ-9. For months 1 through 12, participants were asked to complete the LMC survey documenting their lifestyle and medication changes over the past month. At months 3, 6, 9 and 12 participants were additionally asked to complete the PHQ-9.

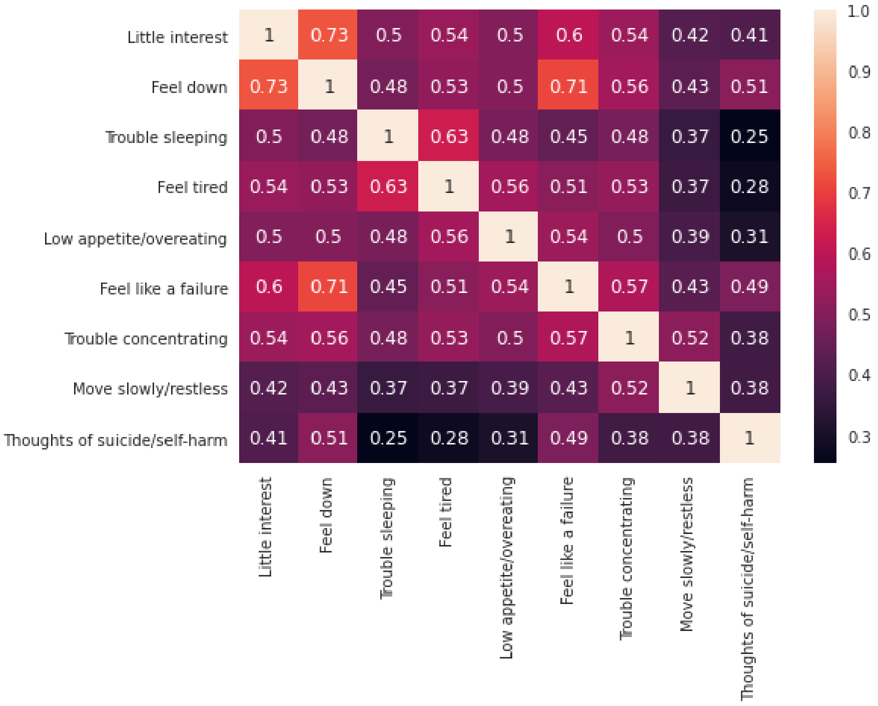
The goal of the first phase of PSYCHE-D is to predict monthly depression severity levels, and the goal of the second phase is to predict increase in depression severity over a 3-month period, using the labels generated by the first phase, as well as additional PGHD. We used the five predefined PHQ-9 score categories [30], representing a scale of depressive symptom severity that ranges from minimal to severe depression. As the PHQ-9 score aims to summarize depression severity over the past two weeks, we only include wearable PGHD over the 14 days prior to the PHQ-9 completion date.

As the first phase generates monthly predictions for intermediate months of the 3-month samples in the second phase, we define two different types of samples: in the first phase, we generate depression severity categories for SM1 and SM2, denoting "sample month 1" and "sample month 2", respectively. In the second phase model, we use 3-month samples to predict a deterioration in depression status. For this we use the "baseline" PHQ-9 category at SM0 (as participants complete the PHQ-9 once every three months), the generated PHQ-9 categories and the probabilities of each output category by the first phase model for SM1 and SM2, screener survey response data, as well as LMC survey responses and wearable PGHD collected at SM3.

2.2 Data exploration

To have a better understanding of the dataset, we performed an initial data exploration to observe trends and correlations between socio-demographic and health factors, and evolution of mental health status.

First, to select a metric for depression severity, we observed the trends in PHQ-9 responses. We considered the options of using PHQ-8 scores [31] (calculated from the first eight questions of the PHQ-9), a single PHQ-9 question response, or the total PHQ-9 score as measures of depression severity. The PHQ-9 questions can be referred to in Appendix A. Figure 2.2 displays the correlations between responses to the PHQ-9 questions for all selected participants. Three correlated questions (with a correlation higher than 0.7) were those related to losing interest, feeling down and feeling like a failure. The



	Little interest	Feel down	Trouble sleeping	Feel tired	Low appetite/overeating	Feel like a failure	Trouble concentrating	Move slowly/restless	Thoughts of suicide/self-harm
Little interest	1	0.73	0.5	0.54	0.5	0.6	0.54	0.42	0.41
Feel down	0.73	1	0.48	0.53	0.5	0.71	0.56	0.43	0.51
Trouble sleeping	0.5	0.48	1	0.63	0.48	0.45	0.48	0.37	0.25
Feel tired	0.54	0.53	0.63	1	0.56	0.51	0.53	0.37	0.28
Low appetite/overeating	0.5	0.5	0.48	0.56	1	0.54	0.5	0.39	0.31
Feel like a failure	0.6	0.71	0.45	0.51	0.54	1	0.57	0.43	0.49
Trouble concentrating	0.54	0.56	0.48	0.53	0.5	0.57	1	0.52	0.38
Move slowly/restless	0.42	0.43	0.37	0.37	0.39	0.43	0.52	1	0.38
Thoughts of suicide/self-harm	0.41	0.51	0.25	0.28	0.31	0.49	0.38	0.38	1

Figure 2.2: *Correlation heatmap for the responses for each of the Patient Health Questionnaire-9 (PHQ-9) questions.*

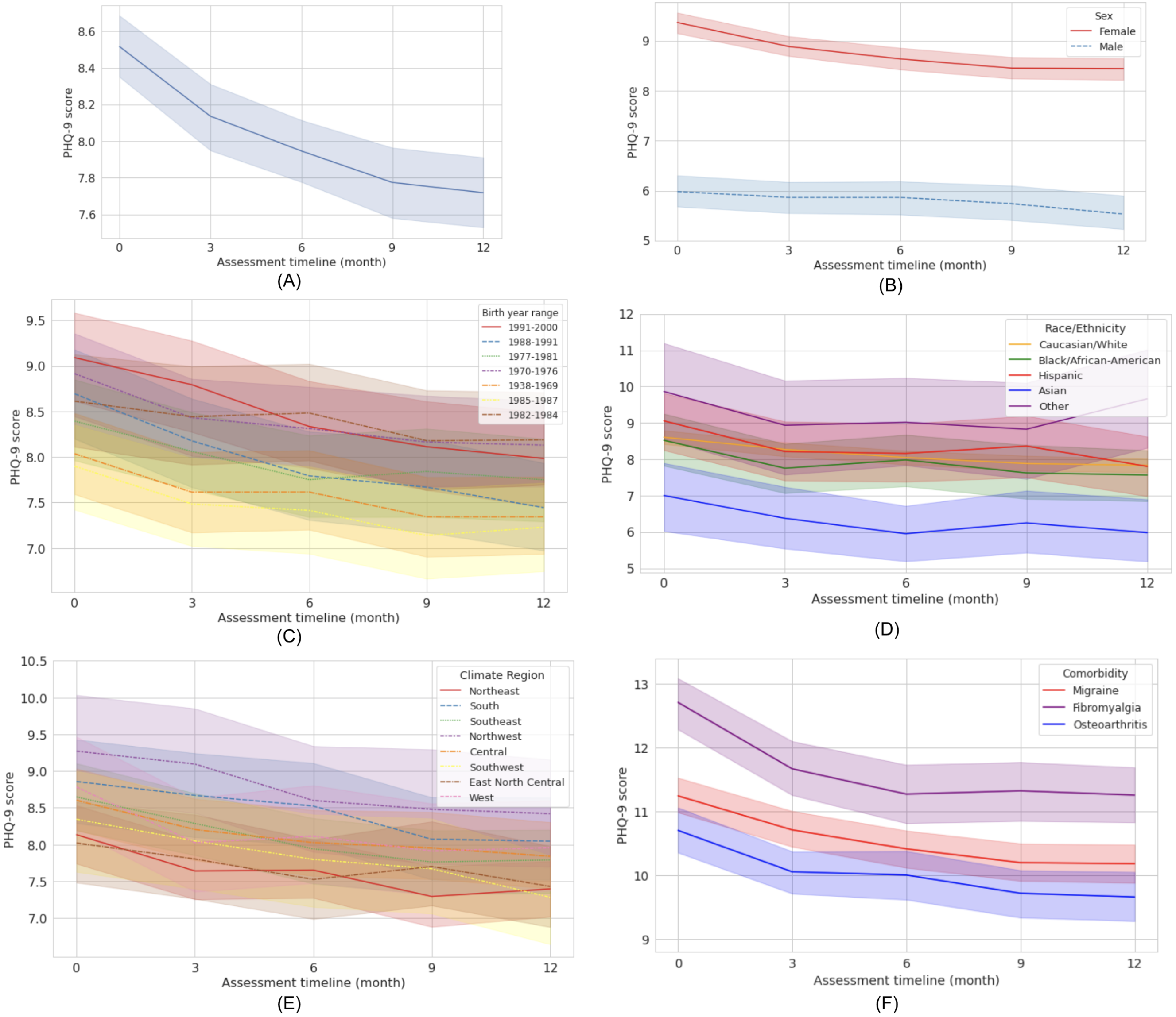
last question, regarding thoughts of suicide and self-harm, was not strongly correlated with any of the other responses. We chose the PHQ-9 score as the representative metric for depression severity as it has been extensively validated as a measure of depression severity [30], and it seemed to provide a more complete picture of a participant's mental health status than any one single question. We chose PHQ-9 over PHQ-8 as we found the ninth question to be an important component of the questionnaire and the responses did not show a strong correlation with any other question.

In order to consider independence of samples, i.e. if it would be correct to assume that participants' quarterly changes in depression severity are independent of each other, we tested the presence of the Hawthorne effect. The Hawthorne effect explains the impact of the awareness of being studied on behaviour [32] and can be observed in longitudinal observational studies similar to the DiSCover project. For instance, participants may set health goals in the start of the study that they may act upon in the start and then remember them again towards the end, and attempt to reach these goals before the end of the study. Such behavioural changes could impact mental health both in the start and the end of the study, implying dependent quarterly changes in mental health status. To test if such an effect was present, we studied the evolution of depression severity across the population throughout the study timeline. In particular, we analyzed the rate at which mental health status changes were taking place, by analyzing participants who completed all five PHQ-9 surveys at baseline, months 3, 6, 9 and 12, and compared the relative PHQ-9 score changes for each 3-month interval.

We performed a Shapiro-Wilk test on each set of PHQ-9 scores, and found that all of the score distributions significantly deviated from normality assumptions ($p < \alpha = 0.01$). To test whether there is a significant difference in distributions between the PHQ-9 scores for consecutive 3-month intervals, we performed a Kolmogorov-Smirnov test. Test results show no evidence of differences in score distributions between consecutively collected PHQ-9 scores ($p > \alpha = 0.01$). We thus assume that there is no significant Hawthorne effect to influence our hypothesis of independent 3-month interval changes in depression severity among participants.

We additionally studied the longitudinal change in PHQ-9 scores, with 95% confidence intervals (CI), for the selected participants, grouped by demographic characteristics, comorbidities and location, shown in Figure 2.3. To obtain more representative confidence intervals, only trends of groups consisting of at least 100 participants were visualized. Figure 2.3A shows a decrease in PHQ-9 score by approximately one point across the entire population over the course of the study. The same trend is observed when participants are grouped by sex, birth year range, climate region and comorbidity[1]. The groups show different absolute levels of PHQ-9 scores, but the overall evolution of self-reported depression status remains consistent across the assessment timeline for the whole population, regardless of the group. Females tend to have higher PHQ-9 scores than males (Figure 2.3B), and there is a clear difference in absolute scores for participants with different comorbidities (Figure 2.3F). We also note that race and ethnicity do not seem to represent changes in PHQ-9 scores very well, with wide confidence intervals.

[1] Only the three most common comorbidities of the cohort are visualized.



(A)

(B)

(C)

(D)

(E)

(F)

Figure 2.3: *Longitudinal change in PHQ-9 scores for participants that have completed all five PHQ-9, at baseline and months 3, 6, 9 and 12. The lines show average trends, with error bars displaying 95% confidence intervals. Panel A shows the overall evolution of PHQ-9 score for participants throughout the year-long assessment period, with the PHQ-9 score reducing on average by one point throughout the year. Panel B shows the longitudinal change in score, split by sex. Panel C shows the longitudinal change in score grouped by birth year range into seven groups, using a quantile split. Panel D shows the longitudinal change in score grouped by race and ethnicity. Panel E shows the longitudinal score change, grouped by climate region of the participants, at baseline. Panel F displays the longitudinal change in PHQ-9 score grouped by diagnosed comorbidity, for the three most prevalent comorbidities in the cohort. Only groupings comprising over 100 participants are visualized.*

The only participants who seem to have an increase in PHQ-9 scores over time are those that identify themselves as “other” for race/ethnicity, shown in Figure 2.3D, although the CI is very wide for this group. One can also observe that Asian participants have significantly lower absolute PHQ-9 scores than participants of other races and ethnicities.

2.3 Data processing

2.3.1 Data filtering process

The participant filtering process is illustrated in Figure 2.4. Of the approximately 25,000 initially screened participants for the DiSCover project, 10,036 were enrolled into the study, and 9,961 passed the survey response quality control.

Then, based on survey completion, only participants who completed the PHQ-9 for at least two contiguous quarters, as well as the LMC survey for the same month as the second PHQ-9 were retained. The selection criteria based on survey completion are illustrated in Figure 2.1. These requirements were set as we aim to study the evolution of depression severity, and recent lifestyle changes, such as changes in medication, are important factors in determining depression status.

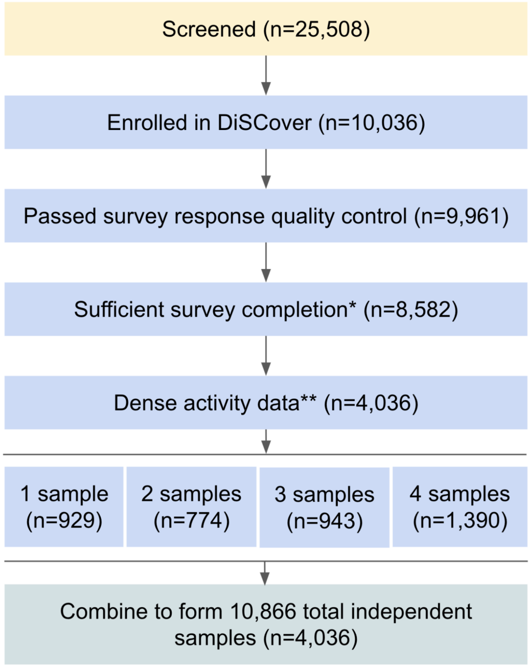

Screened (n=25,508)
Enrolled in DiSCover (n=10,036)
Passed survey response quality control (n=9,961)
Sufficient survey completion* (n=8,582)
Dense activity data** (n=4,036)
1 sample (n=929)
2 samples (n=774)
3 samples (n=943)
4 samples (n=1,390)
Combine to form 10,866 total independent samples (n=4,036)

FIGURE 2.4: *Illustration of the participant filtering process. (*): completion of PHQ-9 for the current quarter, PHQ-9 for the previous quarter, and LMC survey for the current month. (**): ≥ 10 hours daily wear time for ≥ 4 days per week in the 2 week interval.*

The data collection process required little interaction and effort from participants: the participants were prompted to complete monthly surveys on a smartphone app, which also collected the consumer-grade wearable device data in the background. Due to the low-burden nature of this process, we were naturally faced with data scarcity issues, as some participants did not regularly complete questionnaires and did not regularly wear their devices. Nonetheless, we chose not to impute missing data, particularly wearable data, because the nature of the missingness of wearable data is likely not random: the device may have low battery, a person may choose not to wear their device because of a fashion choice or religion, or if they are feeling unwell and do not want to record their data. Non-random missing data features have been shown to be informative in other chronic health condition-related settings [33]. One of the key steps of our project was to filter data in a way that minimizes the amount of missing data, while still keeping the filtering criteria not too strict so as to make the algorithm usable in a real life setting with some missing data, and not to lose too many participants in the dataset.

Based on this reasoning, participants were filtered based on the density of their available activity data in the two weeks matching the PHQ-9 recall period, according to standards proposed in literature [34, 35]. Participants were included if they had at least four valid days of activity data in each of the two weeks. A day is considered as “valid” if the participant wears the device for at least 10 hours that day.

At this stage, we had 10,886 quarterly samples from 4,036 unique participants. As described in the previous section, the overall evolution of depression severity was stable throughout the year when grouping by demographic variables, such as sex, age, race, and geographic location. Based on this observation, we proceeded under the hypothesis that each of the 3-month interval samples from the participants are independent of each other.

Table 2.1 presents the baseline characteristics of the selected participants, in comparison to the enrolled participants. The general trends are consistent, implying that the selected participants are a good representative subset of the enrolled population.

Baseline characteristics	**Selected participants (N=4,036)**	**Enrolled participants (N=10,036)**
Age, mean years (SD)	37.4 (10.0)	37.2 (10.5)
Female, n (%)	3049 (75.5%)	7392 (73.7%)
BMI, mean (SD)	31.3 (8.0)	31.2 (8.3)
College graduate, n (%)	2631 (65.2%)	5849 (58.3%)
Race/Ethnicity:		
Caucasian/White, alone %	84.7%	82.4%
Black/African-American, alone %	4.5%	5.2%
Asian, alone %	2.8%	2.8%
Hispanic, %	5.6%	6.8%
Other, %	2.6%	3.3%

Table 2.1: *Description of the baseline socio-demographic characteristics of the selected participants, in comparison to the enrolled participants.*

Feature extraction and engineering was performed separately for survey responses and wearable PGHD, and is described in the following sections.

2.3.2 Survey responses

We split the survey response data into two categories: static and dynamic features. Features originating from the screener survey were considered as static as these features were not tracked for changes throughout the study. These features thus represent a participant's status at baseline. This includes features such as sex, race, weight, having an insurance and diagnosed comorbidities.

Lifestyle and medication changes (LMC) features were considered as dynamic, as we observed self-reported monthly changes in lifestyle and medication throughout the study. LMC features include changes in medication dosage or reduction in alcohol consumption in the past month. We note that these self-reported lifestyle changes are quite broad: a self-reported change in medication dosage does not specify what the medication is treating, or what the exact change in dosage is.

The full list of considered static and dynamic features is available in Appendix B.

2.3.3 Wearable PGHD

Wearable PGHD was collected from the participants' Fitbit wearable devices worn throughout the study.

Although we had wearable PGHD for participants throughout the year-long study period, we filtered wearable PGHD trends for the two-week intervals prior to PHQ-9 completion. This decision stemmed from the fact that the PHQ-9 questions specifically ask participants to recall their experiences over the past two weeks, so we chose to also analyze PGHD trends in the same two-week recall period.

Wearable PGHD was collected from the participants' Fitbit wearable devices worn throughout the study. The preprocessed DiSCover wearable data consisted of three channels: sleep, step and heart rate. We prioritized using sleep and step data for numerous reasons. Firstly, sleep patterns have been shown to be strongly associated with depressive symptoms. According to one study, approximately 80% of individuals with a current major depressive episode experience insomnia symptoms [36]. Multiple studies have also found strong associations between individuals having short or long sleep duration and depressive symptom severity [19, 37]. Physical activity has shown to reduce depressive symptoms [38, 39], and daily step data can be used as a measure of physical activity, as the frequency of steps can imply light, moderate or vigorous activity levels [40].

Heart rate data was not analyzed in this work for multiple reasons. Heart rate patterns are very noisy and can change based on age, genes, exercise, sleep, weight and medication use [41], among other factors. As the cohort in the DiSCover study is quite varied, controlling heart rate for all of these factors would be difficult with at most likely a low increase in performance. We are also faced with the problem of sparse data, as not all wearable devices were able to accurately collect heart rate data. Based on

Source	**Day-level channel**
Steps	Number of steps taken while awake
	Number of minutes with "not moving" range steps ($<$ 50 steps per minute)
	Number of minutes with light physical activity (LPA) range steps (50-100 steps per minute)
	Number of minutes with moderate-to-vigorous physical activity (MVPA) steps ($>$ 100 steps per minute)
	Maximum rolling number of steps taken in a 6 minute interval
	Number of distinct step streaks ($\geq$ 10 minutes walking)
Sleep	Total number of minutes asleep
	Total number of minutes spent in bed
	Number of minutes spent awake during sleep
	Number of "awake regions" during sleep (i.e. the distinct number of times the user is awake during sleep)
	Start hour of the main sleep
	Sleep efficiency score during the main sleep (calculated using the Fitbit API)
	Number of naps taken

Table 2.2: *Description of aggregated day-level channels, split into sleep and step channels, used for wearable PGHD feature generation.*

initial analyses, after processing for data inconsistencies, we still had 20% more days with missing heart rate data, compared to sleep and step data. We thus focused on trends in sleep and step data.

Sleep and step data, initially provided at a minute-level granularity, were aggregated at day level. The day-level channels selected for feature generation, split into step and sleep channels, are summarized in Table 2.2.

Three different approaches were used to generate feature sets. The first feature set was based on general statistical trends, such as mean, median, interquartile range, and range. The statistical trend features were aggregated over three time windows, relative to PHQ-9 completion date: 4, 7, and 14 days prior to completion date. Examples of such computed features include the range of the start hour of the main sleep, interquartile range of the number of minutes of MVPA range steps, and the ratio of the number of minutes spent asleep over the number of minutes spent in bed on weekends.

The second set of features was designed to observe behavioural changes over the course of the 14 days. For two time windows – 7 days and 14 days – we fit linear regression models for various day-level channels, including the number of minutes spent in bed and the number of steps taken. For each fitted linear regression model, we used the resulting score, intercept and coefficient as features.

The third strategy consisted of defining threshold-based features. Specifically, we defined hypersomnia days (at least 10 hours of sleep), hyposomnia days (less than 5 hours of sleep), active days (at least 10,000 steps walked), and sedentary days (fewer than 5,000 steps walked). We counted the number and percentage of days for each of these categories, aggregated over two time windows: 7 and 14 days.

We have thus obtained a total of 72 sleep-related features and 52 step-related features. A complete list describing processed wearable PGHD features can be referred to in Appendix C.

2.3.4 PHQ-9 depression severity labels

We aggregated PHQ-9 scores based on five predefined categories to measure the severity of depression in participants [30]. Table 2.3 presents the PHQ-9 score categories, representing levels of depression severity, along with the number of samples in each category in our final dataset. The table shows that the category distribution is imbalanced, with the majority of samples representing minimal to mild depression, as expected in a random population sample.

The goal of the second phase is to predict an increase in depression severity, i.e. an increase in PHQ-9 score category over a 3-month period. Of the 10,886 3-month samples consisting of two consecutive PHQ-9 assessments obtained from selected participants, 2,252 (20.7%) samples corresponded to an increase in depression severity level.

PHQ-9 total score	**Depression severity**	**Number of samples (%)**
0-4	Minimal	4202 (38.7%)
5-9	Mild	3220 (29.6%)
10-14	Moderate	1941 (17.9%)
15-19	Moderately severe	981 (9%)
20-27	Severe	522 (4.8%)

Table 2.3: *Description of the PHQ-9 score categories, representing depression severity levels, and the dataset category distribution.*

2.4 Model overview

The design of the PSYCHE-D model was strongly influenced by the data collection method. As we are provided with sparse data, with 3-month intervals between PHQ-9 completion, the model was developed in a way to reduce data sparsity in a first phase, followed by predicting whether there is an increase in depression severity in a second phase.

Figure 2.5 illustrates the data used in phase 1 and phase 2, to make a clear distinction between the goals of the two phases. In phase 1, we developed a model that predicts PHQ-9 score category using collected screener survey and LMC survey responses, as well as wearable PGHD data for a given month. Using this model, we generated intermediate monthly PHQ-9 score categories for SM1 and SM2. In phase 2, we predicted an increase in PHQ-9 category using the person's collected PHQ-9 results at the start of the 3-month period, SM0, generated predictions for intermediate PHQ-9 categories using the phase 1 model, and collected LMC survey responses and wearable PGHD for SM3. In addition

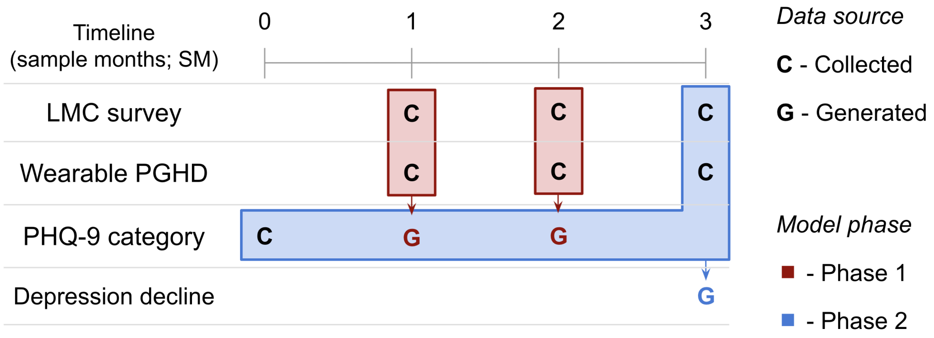


FIGURE 2.5: *A schematic overview of the PSYCHE-D approach. Phase 1 uses screener survey responses (regarding demographics, and chronic comorbidities), self-reported LMC (lifestyle and medication change) surveys from the month in which the PHQ-9 label is generated, as well as data from consumer-grade wearables to categorise each individual's likely PHQ-9 category. In a second phase, this generated information is then combined with the initial PHQ-9 category, screener survey responses, additional LMC self-reports and consumer-grade wearable PGHD to make the final prediction of whether the individual is likely to have experienced increased depression severity over the 3-month period.*

to these data, we also used the screener survey responses as input features for both models, to control for socio-demographic factors.

In the following sections, we explain the modeling process for the phase 1 and 2 models, and the combination of the two phases into one pipeline.

2.5 Modeling: phase 1

The goal of the first phase model was to predict participants' PHQ-9 score categories from socio-demographic information, self-reported lifestyle and medication changes, as well as objective wearable PGHD.

For this model, each 1-month sample from the dataset is defined as one observation of PHQ-9, one set of screener survey responses, one set of LMC survey responses, and wearable PGHD for a minimum of 8 and a maximum of 14 days. Using the selection criteria described in Section 2.3.1, we obtained 10,866 labeled training and testing samples from 4,036 unique participants. We additionally had 18,673 intermediate labels to generate for the second phase samples, using the fitted phase 1 model.

As we aimed to predict depression severity measured by PHQ-9 score categories, the first phase consisted of a multi-class classification problem. In addition to this, as identified in Section 2.3.4, we were faced with a significant class imbalance, as the majority of participants had minimal to mild depression. To mitigate the effects of class imbalance, we performed sample stratification during training, hyperparameter tuning and testing, so as to optimize performance across all categories. In addition to this, we used sample weighting when fitting our model, in order to reduce overfitting to over-represented classes, and improve overall model performance.

We used a rigorous feature selection process in order to optimize model performance through dimensionality reduction. We removed highly correlated features, and used recursive feature elimination and cross-validated selection (RFECV; [42]) in order to eliminate features that had lower contributions to model performance.

A common problem with wearable PGHD is inconsistent and missing data. Participants may choose not to wear their devices for various reasons, so we did not want to impose assumptions on their activity through imputation. We also did not filter participants in a way that only participants wearing their devices every day were chosen, because sparse data is inevitable in the type of real-life setting that we would like to apply the presented model to. We instead added a mask feature to represent “missing days” as a means to use the missingness of PGHD in an informative way, as presented in previous studies [33].

The decision to avoid imputing missing data motivated our classification algorithm selection – the Extreme Gradient Boosting (XGBoost) algorithm [43]. XGBoost efficiently constructs ensembles of decision trees, and is able to handle sparse data. It is also interpretable and allows us to observe feature importance in the decision making process of the model, to understand which features are the most important in determining a participant’s depression severity level. We used a logistic regression model with zero imputation as the baseline to compare to the performance of XGBoost.

2.5.1 Performance metrics

Model performance was primarily measured using quadratic weighted Cohen’s Kappa. Weighted Kappa computes the level of agreement between the predicted and target values, using a distance-based weight penalty [44]. The quadratic weighted Cohen’s Kappa is calculated as follows:

$$\kappa = 1 - \frac{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} O_{ij}}{\sum_{i=1}^{k} \sum_{j=1}^{k} w_{ij} E_{ij}}$$

In our case, $k = 5$ as we have five depression severity categories. O_{ij} and E_{ij} represent elements in the observed and expected matrices, respectively. The observed matrix is the confusion matrix representing the predicted and target classes generated by the model, and the expected matrix is the confusion matrix representing the same confusion matrix generated by chance. The quadratic weight penalty is denoted by w_{ij}, and is calculated using:

$$w_{ij} = 1 - \left(\frac{|i - j|}{k} \right)^2$$

Thus, if the predicted value corresponds to the true value, there is no penalty. If the distance between the predicted and true value increases, the weight penalty gets quadratically larger. The final Kappa score ranges from -1 to 1, with a perfect score of 1 signifying full agreement, and -1 signifying a maximal distance between predicted and target values.

We also used the following secondary performance metrics:

- *Adjacent accuracy*, also referred to as precision at k for $k = 2$, which denotes the fraction of samples predicted at most one off from the target value
- *Balanced accuracy*, which is the average of recall obtained on each class
- *Weighted F1-score*, which calculates the F1-score for each class and computes the average, weighted by support (i.e. the number of true instances of the class), accounting for class imbalance

2.5.2 Hyperparameter tuning

We performed randomized search 5-fold cross-validation over 50 iterations to tune the hyperparameters of our XGBoost model. We performed this procedure on feature subsets of survey response data, step data, sleep data, and combinations of each. The decision to select randomized search cross-validation was made to optimize computational resources, as the performance of a complete grid search would require significantly more resources, at a marginal increase in model performance. We reported the performance metrics of the best tuned models with 95% confidence intervals across 5 training runs (5 outer shuffle splits).

2.6 Modeling: phase 2

Once we generated intermediate monthly depression severity labels using the first phase model, we posed our original aim as a binary classification problem for the second phase model: can we predict increased depression severity? We thus developed a model that predicts whether a person's depression severity level has increased over a 3-month period, using PGHD and the intermediate PHQ-9 category classification generated by the phase 1 model.

We specifically chose to design the second phase model as a binary classification model that predicts increase in depression severity, as opposed to additionally identifying participants who report a decrease in depression severity. Our aim is to use the model as an alert system that identifies participants with worsening mental health, and we believed that simplifying the problem to build a model that focuses solely on this task would result in better performance. Nonetheless, identifying participants whose mental health has improved in a multi-class classification problem could also be an interesting future work.

The output target variable for the second phase model is thus a binary value that represents whether there is an increase in PHQ-9 score category between SM0 and SM3. Samples in the phase 2 model also included the following input features:

- The baseline PHQ-9 category at SM0

- Intermediate generated PHQ-9 categories and the individual probabilities of each category at SM1 and SM2, generated using the phase 1 model
- LMC survey responses collected at SM3
- Wearable PGHD collected over the 14 days prior to final PHQ-9 completion at SM3
- Screener survey responses

The construction of the second phase model was optimized across possible input feature sets and the algorithm used. Using cross-validation, with random assignment of samples across all participants to training and testing sets, we assessed in parallel both a range of possible algorithms and optimal input feature sets. In the following sections, we describe key performance metrics and the model construction process in further detail, as well as our participant-based validation strategy.

2.6.1 Performance metrics

The goal of the second phase model was to effectively identify and monitor participants whose depression level has increased, so naturally the primary metric used to evaluate model performance was the sensitivity score. Sensitivity, or recall, measures the proportion of positives that have been correctly identified and is calculated as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

Where TP designates the number of true positives, and FN designates the number of false negatives. In practice, using sensitivity as the primary performance metric means that we optimized our model in order to maximize the proportion of participants with increasing depression severity to be correctly identified.

We also wanted the model to be good at distinguishing between the two classes of participants, so we considered secondary performance metrics, including specificity and Area Under the Precision-Recall Curve (AUPRC) [45]. We used AUPRC for model comparison, as it measures how well one model is able to distinguish between classes, especially when they are imbalanced. The baseline performance of the AUPRC metric is the fraction of samples with increased depression severity level, which corresponds to 21% in our dataset.

Specificity, which measures the proportion of correctly identified negative samples, is calculated as follows:

$$\text{Specificity} = \frac{TN}{TN + FP}$$

Where TN designates the number of true negatives, and FP designates the number of false positives. Comparing specificity to sensitivity allows us to consider the tradeoff between being able to correctly predict the target minority class (participants with increased depression severity), without ignoring the majority class of participants reporting stable or reduced depression severity. As we primarily targeted higher sensitivity performance we expected to have a lower specificity value, as in an eventual use case of our algorithm it would be better to perform an unnecessary check on people whose mental health is stable, rather than risk not checking in on participants whose mental health is declining.

2.6.2 Model construction and testing

As both the demographic and target class data are highly imbalanced in our dataset, we focused on building a robust model, such that it is not susceptible to data imbalance. Prior to model construction, we performed a stratified shuffle split on the data to ensure that the training and testing data would have a similar distribution of positive and negative target classes. We tested multiple model approaches to find the best performing classification algorithm to suit our needs.

The model construction process took place in two steps: feature selection and model fitting. In initial feature selection experiments, we compared three feature selection methods: RFECV (as in phase 1), a meta-transformer that selects a subset of features based on their relative importance weights, and forward sequential feature selection. Forward sequential feature selection is a greedy algorithm that selects features based on which additional feature brings the most performance to the model [46]. Forward sequential feature selection consistently gave the best results in our preliminary experiments, so we chose to proceed with this feature selection method for the phase 2 model, using 5-fold cross-validation on the training set. Forward sequential feature selection has also previously been used with success in longitudinal studies using digital measures to predict mental health symptoms [47].

In the model selection step, we assessed the following classification models: as reference, we used logistic regression and random forests [48], and we tested two gradient boosting algorithms: Light Gradient Boosting Machine (LightGBM) and Extreme Gradient Boosting (XGBoost) [43]. We tested two implementations of LightGBM: Gradient Boosting Decision Tree (GBDT), a widely used gradient boosting method, shown to be effective across a diverse range of tasks [49] and Dropouts meet Multiple Additive Regression Trees (DART), an ensemble model of boosted regression trees with dropout, which has been shown to deliver high accuracy in comparable classification tasks [50]. We also tested the XGBoost algorithm, as it performs well on large-scale data with a high degree of missingness and sparsity, and was the selected model in the phase 1 generation of intermediate depression severity labels.

As missing data is an important component of our problem, the choice of selected models reflects this. We did not want to make assumptions as to why the data was missing, particularly wearable data, so LightGBM and XGBoost were favourable, as they are able to handle missing data without imputation. For the random forest, we represented missing values with "-1", so that the model could deduce that these are indeed missing values and use this as additional information.

Ensemble methods are also favourable in our problem because they have been shown to generalize well [51, 52] as they combine multiple weak learners. Ensemble methods generally build multiple classification models, and combine them in a way that creates a stronger model. Random forests, for instance, create multiple weak learners in parallel, and then decide on the output value based on a majority decision from the weak learners. In this work, we focus more on gradient boosting machines, a specific type of ensemble model, which consecutively fit new models that learn from the errors of the previous ones. One of the main reasons we use gradient boosting machines is because of their ability to generalize well [53], which corresponds to the end goal of the project: to use the presented approach on previously unseen participants outside of the DiSCover cohort.

Once we obtained the most important features through forward sequential feature selection, we used this feature subset of the training set to fit various models. We repeated the model construction process five times with different randomized training and test set splits, in order to test the models for robustness. An 80:20 train:test split was used. The model construction and testing strategy is outlined in Figure 2.6.

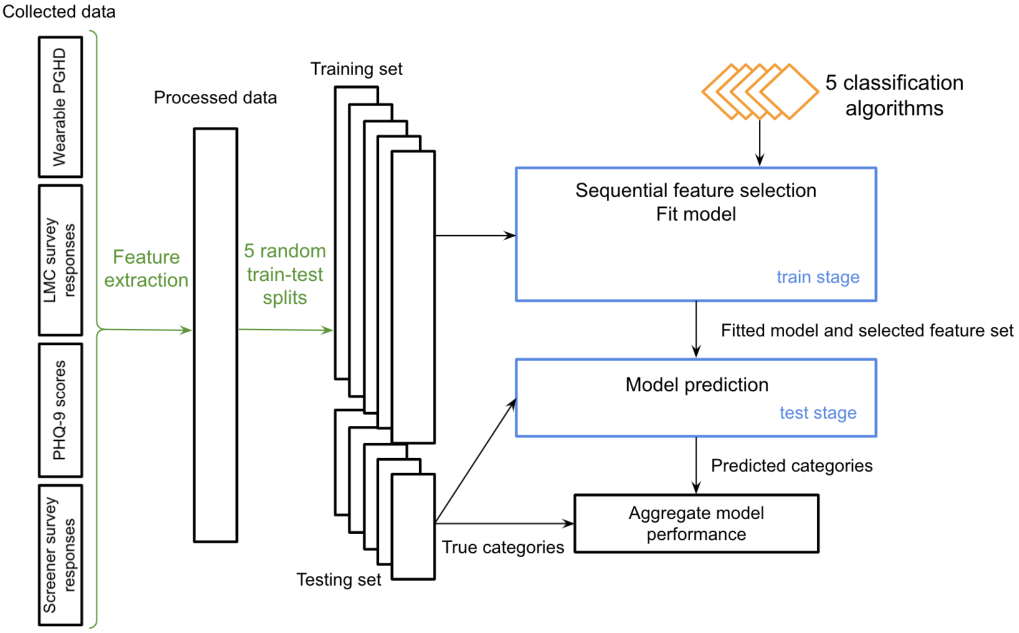


Figure 2.6: *Schematic representation of the phase 2 model construction architecture. Input sources are featurized to create the processed input data, which is assigned to training and test sets in five random splits, with an 80:20 train:test ratio. The training set is used as input for the train stage which includes sequential feature selection for five possible classification algorithms. Model performance is then assessed in the test stage versus the reserved testing set, and aggregate model performance is summarized across the five random splits. Green arrows = transformation; black arrows = input to next block; orange blocks = algorithms; blue blocks = model construction; black blocks = data.*

2.6.3 Feature importance

The overall most important features in determining whether there is an increase in depression severity were identified by a combination of two key metrics: gain importance and split importance [54, 55]. Gain importance, also referred to as Gini importance is the improvement in accuracy brought by a feature to the branches it is on. Split importance, also called permutation importance, summarizes the number of times the feature is used in a model.

As LightGBM is an ensemble tree-based learning model, features correspond to splitting nodes in each tree of the model. Gain importance for a given feature in the input feature set corresponds to the mean decrease in the feature node's Gini impurity, proportional to the number of samples that the node splits (i.e. the number of samples that have reached that node) [56]. This gives an absolute measure of feature importance.

Split importance was calculated as follows: model performance with a validation set (using 5-fold cross-validation) was recorded as baseline performance. Then, for each feature in the input feature set, the rows of the feature column were randomly permuted in the input feature set, and model performance on the input dataset with the permuted column was computed. The difference in baseline performance and the latter performance was thus the split performance of that feature [57]. Consequently, we obtain a relative measure of feature importance.

Additionally, we recorded the total number of times that a given feature was selected during cross-validation.

2.6.4 Participant-based splitting strategy

In order to gain an indication of generalizability, we took the five best performing model/number of input features pair combinations (ranked by sensitivity) and repeated forward sequential feature selection, model fitting and testing (i.e. the full model construction process outlined above), with the exception that rather than randomly assigning samples to the training and test subsets, we assigned all samples from a given participant to only one of the two subsets. This means that the fitted model is naive to all participants in the test subset, and gives us an indication of how the fitted model might perform 'in the wild' where a trained, production algorithm is exposed to data from new individuals.

The participant-based split test sets were created by first randomly selecting 7% of samples, and then assigning all additional samples from participants contained in the random 7% to the test set. We found this method to consistently create a test set containing 20% of all samples, as with the random train/test splits, while ensuring that data from each participant was partitioned into either the training or testing sets.

We present the results of the model performance, as well as feature importance based on both random and participant-based splitting strategies in Chapter 3.

2.7 Modeling: combined pipeline

After having developed and implemented both the phases of the PSYCHE-D model, we wanted to combine the two phases into a single pipeline. When preparing to combine the two phases into a single pipeline, we identified a data leakage between the two phases. Figure 2.7 visualizes the existing data leakage, and Figure 2.8 shows how the PSYCHE-D approach implementation was adapted in order to fix the leakage and test performance. We refer to the adapted versions of the phase 1 and phase 2 models in the "combined" pipeline as phase 1c and phase 2c, respectively.

Figure 2.7 shows that for a given participant, denoted by red circles, samples generated by the participant can be in the training set or in the test set for the phase 2 model. However, the training set for the phase 1 model consists of all the reference labels in the dataset, including the reference labels of the red participant. Thus, samples from the red participant are always part of the training set of the phase 1 model, and can be in the training or test set of the phase 2 model, resulting in data leakage.

To eliminate this data leakage, we split the training and test sets differently, as presented in Figure 2.8. The participant-based train:test split is defined prior to training the phase 1 model. We illustrate two participants by red and yellow circles, representing any given participants that appear in the training and test sets, respectively. The phase 1 model is trained on a subset of the training set and generates intermediate labels for participants both in the training set and in the test set. Phase 2 model samples are built using these intermediate labels, keeping the phase 1 training set participants in the training set and the phase 1 test participants in the test set. The phase 2 model is then trained, and tested using this participant-based data split. We repeat the procedure over five random participant-based splits of the training and test data, to obtain the confidence intervals for the PSYCHE-D final combined pipeline performance.

The combined pipeline implementation of PSYCHE-D, as described in Figure 2.8, uncovered an unexpected issue in the model's performance. Having removed data leakage, using the phase 1 and phase 2 models, we observed a performance of 60.1% (± 2.6%) sensitivity, 57.4% (±0.8%) specificity, 0.282 (± 0.022) AUPRC. Although the sensitivity is still adequate, the AUPRC indicates that this model now predicts an exceedingly high proportion of samples as positive, and that they are often incorrect. Upon further study of the source of the issue, we found that the phase 1 model was overfitting to participants it had seen, so the issue of data leakage had a significantly higher impact on our results than expected. Using participant-based splitting for the phase 1 model, we found that its performance dropped to a Kappa of 0.443 (± 0.029), when presented with new participants that it had not been trained on, with an adjacent accuracy of 75.7% (± 0.8%). This implies that the intermediate generated labels were of low quality for unseen participants, and thus the category probabilities generated by the phase 1 model were not of good quality for the phase 2 model to be able to learn from, for new participants.

To resolve this issue, we designed the combined pipeline using only participant-based splitting, to avoid the issues of overfitting models and testing on previously seen

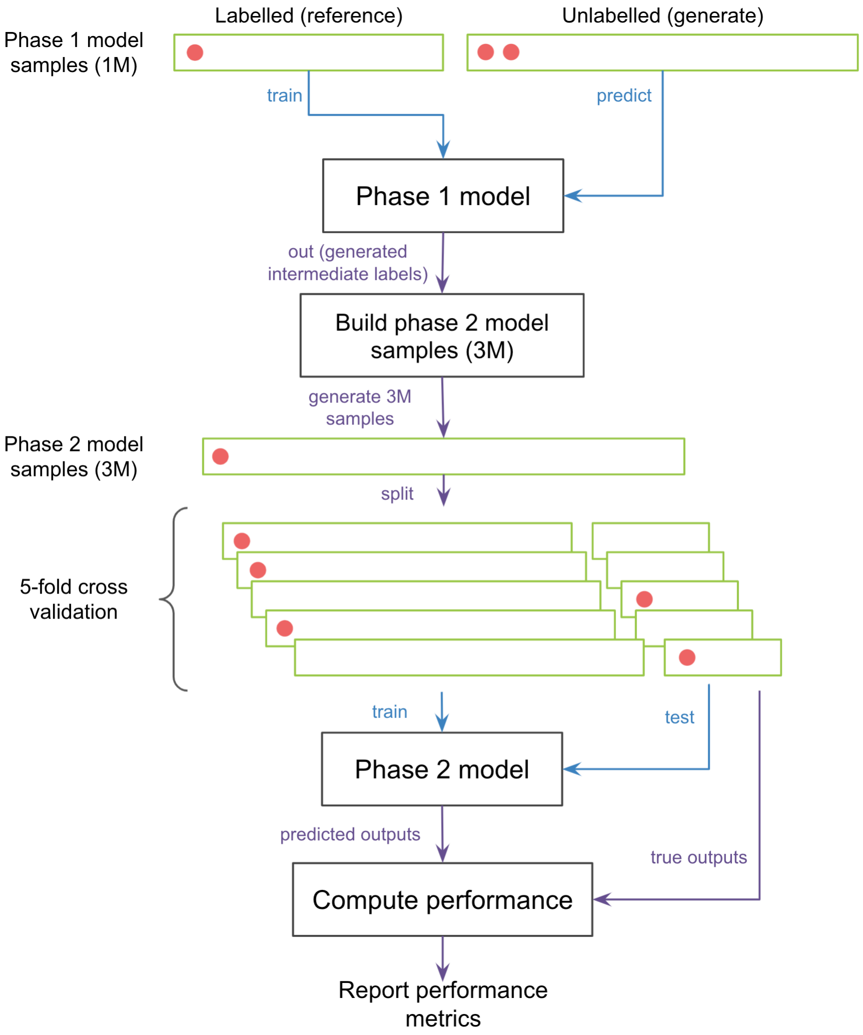


FIGURE 2.7: *Diagram visualizing the initial concatenation of phase 1 and phase 2 models, resulting in data leakage. The phase 1 model was trained on all reference labels (from 1-month) and generated intermediate labels for unlabelled month samples. These were used to build phase 2 model samples, spanning over three months (denoted by "3M"), and the phase 2 model was trained and tested over five CV folds. Red circles represent samples from a given participant. Data leakage is visible, as the red participant is part of the training set for the phase 1 model, and is part of the training or test sets for the cross-validation folds that are used to evaluate model performance. Green blocks represent data, black blocks represent models and data processing stages. Blue arrows represent input to classification models for training or predicting, and purple arrows represent the passage of data for other purposes (e.g. splitting data).*

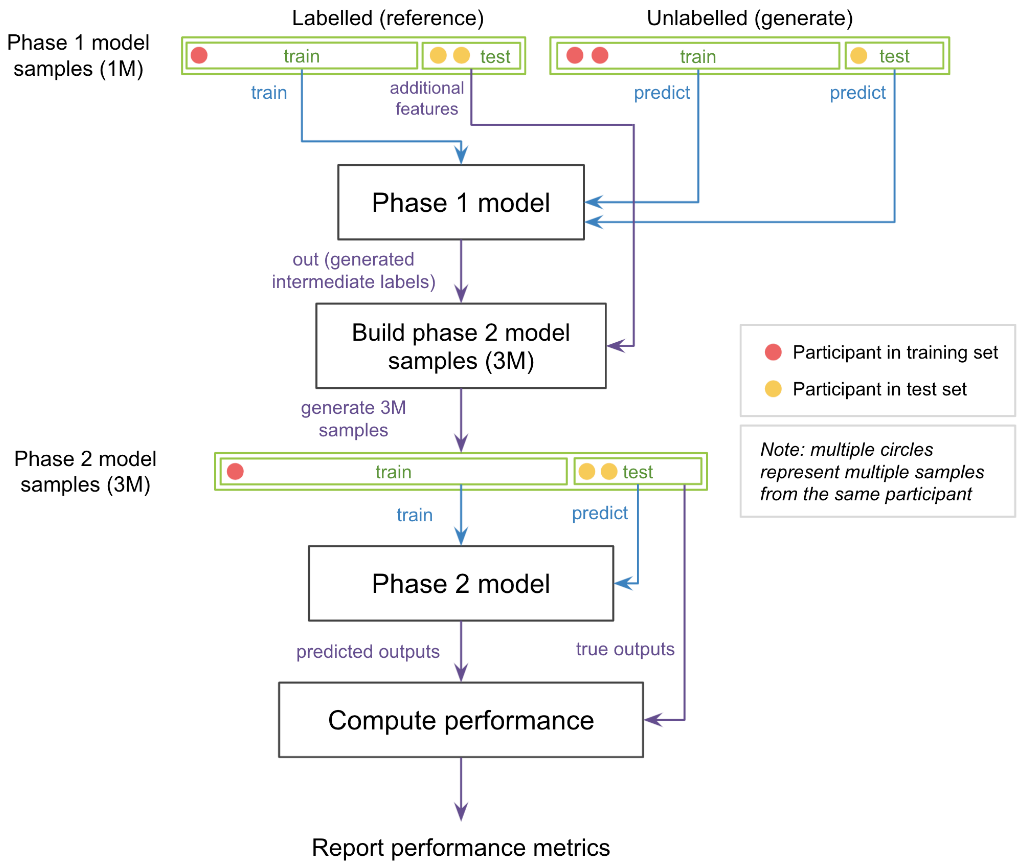

Labelled (reference)
Unlabelled (generate)
Phase 1 model samples (1M)
train
test
train
test
train
additional features
predict
predict
Phase 1 model
out (generated intermediate labels)
Build phase 2 model samples (3M)
generate 3M samples
Phase 2 model samples (3M)
train
test
train
predict
Phase 2 model
predicted outputs
true outputs
Compute performance
Report performance metrics
● Participant in training set
● Participant in test set
Note: multiple circles represent multiple samples from the same participant

FIGURE 2.8: *Diagram visualizing the data leakage fix implementation in the combined pipeline. The phase 1c model is trained on a subset of participants in the training set, and predictions for the training and test set participants are made. The phase 2c model has the same participant split for the training and test set. Red and yellow circles represent samples from two different participants. All samples from the red participant are in the training set and all samples from the yellow participant are in the test set for both phases 1c and 2c. Green blocks represent data, black blocks represent models and data processing stages. Blue arrows represent input to classification models for training or predicting, purple arrows represent data passage for other purposes (e.g. providing true output values for testing). This procedure is repeated over five random participant-based splits of the training and test data, to obtain confidence intervals for the combined pipeline performance.*

participants. We adapted the implementation of the phase 1 model to reduce overfitting and thus improve generalization to new participants, while keeping the performance metrics and overall approach of the model as consistent as possible with the original

PSYCHE-D approach. The modification consisted of implementing a similar two-step procedure for the phase 1 model, as for the phase 2 model: for each of the five train:test split folds that we trained model versions on, we built a pipeline that performs feature selection followed by model training. We performed a randomized grid search on the hyperparameters of the classification algorithm, which we changed from XGBoost to LightGBM DART, as LightGBM DART has an additional dropout parameter that we were able to tune in order to reduce overfitting.

We performed the hyperparameter search with cross-validation, selecting features using recursive feature selection, followed by fitting a LightGBM classifier. In an attempt to reduce overfitting, we varied hyperparameters, such as dropout rate, the number of estimators used, and the maximum tree depth. We thus obtained the combined pipeline version of the phase 1 model — the phase 1c model.

We also made modifications to the phase 2 model in the combined pipeline, producing the phase 2c model. The phase 2c model construction process contained two steps: feature selection and model fitting. In the initial implementation of the phase 2 model with randomized splitting and data leakage, we performed forward sequential selection as a feature selection method on all input features. However, in an attempt to repeat the same procedure, we obtained suboptimal sensitivity (approximately 30%), compared to using the features selected by the initial phase 2 model. This was likely caused by the significantly better results obtained by the phase 1 model using randomized splits, which additionally resulted in a higher feature importance for generated PHQ-9 features, while the PGHD features remained consistently noisy throughout both phase 2 and 2c model constructions.

We chose to further break down the feature selection process into two steps: reducing the initial number of features to choose from, followed by random feature selection. To reduce the number of total input features, we removed highly correlated features, and then used RFECV to identify the most important features for the prediction of increase in depression severity out of the three largest subgroups of PGHD features: sleep data, step data and screener survey response data. This step allowed us to identify a subset of features that were key to the algorithm's performance, and minimise noise added by other features, as well as increase computational efficiency.

Using this subset of features, in combination with LMC features, baseline PHQ-9 category, and generated PHQ-9 features, we proceeded with the same model construction approach as for the initial model. First, we performed forward sequential feature selection, with 5-fold cross-validation on the training set. We then performed model fitting using LightBGM DART, for each of the five participant-based train:test splits. We tested different numbers of features to select during the feature selection method, to ultimately select the best performing phase 2c model of the PSYCHE-D combined pipeline approach.

Chapter 3

Results

The results of the phase 1 and 2 models, as well as the results of the phase 1c and 2c models of the combined pipeline that eliminates data leakage are presented here.

3.1 Phase 1 results

In this section, we present an expanded exploration of the published results for the phase 1 models [28]. We will discuss the performance of baseline phase 1 models compared to tuned XGBoost classification models for various input feature sets, with a 95% confidence interval for each model, averaged over five random train:test splits.

The phase 1 model uses samples that span a 1-month long period. The target value is the PHQ-9 score category at the end of the month. Initial input features include screener survey responses, LMC survey responses for the given month, and wearable PGHD collected over the 14 days prior to PHQ-9 completion.

The experiment structure began using a divide and conquer approach: we initially tested feature selection methods on survey response features, sleep features and step features separately, before combining the best-performing subsets, in order to generate our final input samples.

We found that when applying RFECV to survey response features, the majority of lifestyle changes features were removed, whereas static features and medication changes features were kept. After additional manual backward sequential feature elimination experiments on the remaining survey response features, we obtained a final set of selected survey response features, used in further testing. The final set of selected survey response features is presented in Appendix D.

For wearable PGHD features, with sleep and step features treated separately, we observed similar model performance when using RFECV or when selecting a small subset of features that have a higher correlation with PHQ-9 score categories, based on the training set. To reduce noise from additional metrics, we proceeded using only these more highly correlated (HC) features when fitting models with both survey and wearable data.

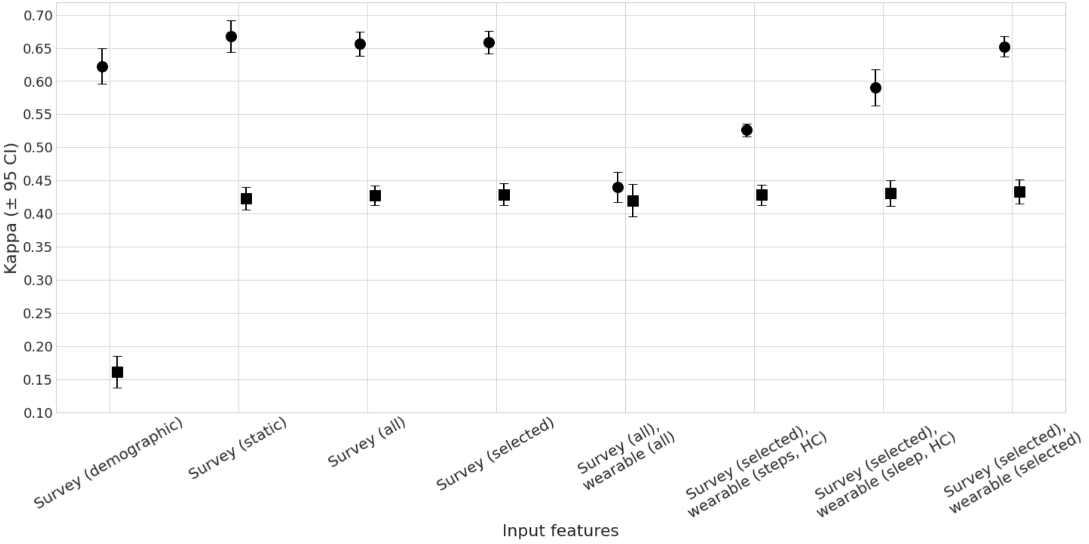


FIGURE 3.1: *Kappa score comparison for models predicting PHQ-9 score category using survey response and wearable data as input features, with 95% confidence intervals. Circular markers represent XGBoost models, and square markers represent logistic regression models.*

The initial set of experiments treating survey and wearable data separately is presented in Appendix E. Based on the experimental results, we proceeded with the best selected subset of survey response features in combination with wearable features to find the optimal model to predict depression severity levels. We present the summarized results in Figure 3.1, with a 95% confidence interval, and the corresponding results table in Appendix F.

We can see that there is a clear improvement from using XGBoost compared to logistic regression, with a mean increase of 0.2 for Kappa scores.

The survey response (demographic) XGBoost model uses only naive demographic features from the screener survey as input features. We were surprised to see that its performance was comparable to more complex models. One reason for this could be that demographic features, such as age or sex, are stronger predictors of absolute depression levels, compared to socio-demographic changes. This is indeed supported by the initial data exploration in Section 2.2, where we found different trends in depression severity among participants grouped by sex and birth year range. However, in retrospect, we find that this was likely due to the use of the random splitting strategy. As the majority of samples in our dataset correspond to no increase in depression severity (21% correspond to an increase, 56% of samples correspond to no change, and 23% correspond to a decrease), it is possible that the model essentially mapped certain static socio-demographic combinations to correspond to certain levels of severity, instead of learning from the probabilities and dynamic features, such as LMC surveys and wearable PGHD.

In general, we see that XGBoost models perform well using only survey responses as input features. Using a selected set of both static and variable survey responses however slightly reduces the CI size. We observe a comparable performance for the model that uses the same survey response features in combination with wearable features, but with a lower CI, implying increased model robustness.

To thoroughly compare the performance of these two best XGBoost models, we performed more extensive hyperparameter tuning, with a higher number of iterations in randomized search cross-validation. We present the results in Table 3.1.

Table 3.1 shows that both models have a very similar performance, with the model using only survey response features obtaining a higher mean Kappa and F1-score, but with a wider CI for both scores, consistent with the previous results. The means of each metric for both models fall into the confidence interval of the other model, signifying an almost identical performance, with wearable features bringing more robustness to the second model, as the model has a narrower CI. The wearable features added to the second model are the number of active days in the past 7 days, and the number of hypersomnia days in the past 7 days.

Input features	**Adjacent accuracy (± 95 CI)**	**Balanced accuracy (± 95 CI)**	**Kappa (± 95 CI)**	**F1-score (± 95 CI)**
Survey responses (selected)	0.889 (± 0.004)	0.472 (± 0.022)	0.661 (± 0.021)	0.543 (± 0.016)
Survey responses (selected), wearable PGHD (threshold-based, selected)	0.889 (± 0.006)	0.464 (± 0.017)	0.655 (± 0.015)	0.542 (± 0.014)

TABLE 3.1: *Performance comparison of the best phase 1 XGBoost models after extended hyperparameter tuning, with 95% confidence intervals.*

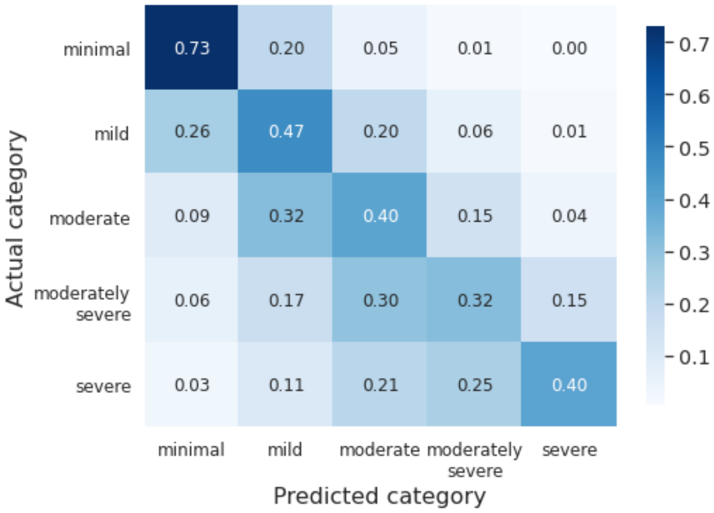


FIGURE 3.2: *Confusion matrix showing the best phase 1 model's PHQ-9 score category accuracy distribution.*

The confusion matrix displaying average category accuracy is also visualized in Figure 3.2. We note that although the model performance is significantly better for the more prevalent categories representing lower depression severity, the adjacent accuracy is consistently high across all classes. This implies that even for less frequently occurring participants with more severe depression, the model tends to categorize them at most one category off, and is not likely to seriously misclassify the depression severity category. We also remark that the misclassification tends to be towards a lower depression severity, e.g. a participant with moderate depression is more likely to be predicted as mild depression severity, rather than moderately severe. This is likely due to the class imbalance, with the majority of participants reporting minimal to mild depression severity.

3.2 Phase 2 results

In this section, we present the results of the phase 2 model that uses participant PGHD, as well as generated intermediate PHQ-9 categories by the best performing phase 1 model, to predict increase in depression severity over a 3-month time period. The phase 2 model uses samples consisting of screener survey responses, the baseline PHQ-9 score category at SM0, LMC survey responses collected at SM3, wearable PGHD collected over the two weeks prior to PHQ-9 completion at SM3, and the intermediate generated PHQ-9 categories at SM1 and SM2, as well as the probabilities of each of the PHQ-9 categories for SM1 and SM2, in order to predict whether there is an increase in PHQ-9 category between SM0 and SM3.

We initially tested the performance of the LightGBM model on only the previous PHQ-9 categories: collected at baseline sample month 0 (SM0), as well as the generated categories and the probabilities of each category at SM1 and SM2. We compared this to the performance of models using only PGHD features, summarized in Table 3.2. Collected and generated PHQ-9 category features performed strongly in terms of sensitivity, so we chose to add PGHD features as additions to these to improve the final performance. We remark that the table shows AUROC (Area Under the Receiver Operating Curve) is used in lieu of AUPRC, as AUPRC was only added as a secondary performance metric later in the experiments, because we found it to be more appropriate than AUROC due to class imbalance [58].

The results using collected and generated PHQ-9 category features looked very promising, so we continued with our experiments, using all PGHD features in combination with PHQ-9 category features as the inputs to the forward sequential feature selection method. We tested the performance of the feature selection method and model fitting using five different classification algorithms. Table 3.3 summarizes the best performance observed for each classification algorithm based on sensitivity; extended performance results are presented in Appendix G.

We used a range of metrics to assess performance, but prioritized sensitivity as the key metric, as our primary goal in this work is to correctly identify the highest proportion of individuals reporting increased depression severity. As the dataset is highly

Input features	**Sensitivity (±95 CI)**	**Specificity (±95 CI)**	**AUROC (±95 CI)**
Baseline PHQ-9 category (SM0)	0.752 (± 0.018)	0.353 (± 0.010)	0.570 (± 0.016)
Baseline PHQ-9 category (SM0), generated PHQ-9 categories (SM1, SM2)	0.743 (± 0.017)	0.876 (± 0.012)	0.863 (± 0.011)
Baseline PHQ-9 category (SM0), generated PHQ-9 categories and probabilities of each category (SM1, SM2)	0.854 (± 0.017)	0.820 (± 0.014)	0.917 (± 0.011)
Sleep (all features)	0.212 (± 0.053)	0.785 (± 0.026)	0.508 (± 0.014)
Step (all features)	0.247 (± 0.016)	0.777 (± 0.013)	0.523 (± 0.019)
LMC (all features)	0.537 (± 0.019)	0.521 (± 0.025)	0.532 (± 0.021)
Screener (all features)	0.399 (± 0.026)	0.586 (± 0.029)	0.499 (± 0.008)

Table 3.2: *Summarized LightGBM model performance for predicting longitudinal change in depression severity, using different input feature sets. For each algorithm we report the best model obtained, including the number of features selected and the following aggregated performance metrics: Sensitivity (Recall), Specificity and Area Under the Receiver Operating Curve (AUROC). 95% confidence intervals are included for each metric. Sensitivity is prioritized as the key performance metric to select the best performing models.*

Model	**Sensitivity (±95% CI)**	**Specificity (±95% CI)**	**AUPRC (±95% CI)**	**Number of features selected**
Logistic Regression	0.799 (± 0.020)	0.764 (± 0.011)	0.609 (± 0.035)	11
Random Forest	0.733 (± 0.012)	0.875 (± 0.016)	0.644 (± 0.020)	5
XGBoost	0.781 (± 0.025)	0.851 (± 0.007)	0.705 (± 0.027)	13
LightGBM (GBDT)	0.840 (± 0.024)	0.830 (± 0.009)	0.723 (± 0.041)	13
LightGBM (DART)	0.882 (± 0.018)	0.808 (± 0.014)	0.739 (± 0.043)	14

Table 3.3: *Summarized model performance for predicting longitudinal change in depression severity. For each algorithm we report the best model obtained, including the number of features selected and the following aggregated performance metrics: Sensitivity (Recall), Specificity and Area Under the Precision-Recall Curve (AUPRC). 95% confidence intervals are included for each metric. Sensitivity is prioritized as the key performance metric to select the best models.*

imbalanced, with 21% of individuals in the dataset reporting increased depression severity, we optimized for performance for both majority and minority classes. We thus took into account specificity and AUPRC as secondary performance metrics, to observe the tradeoff in performance for each class.

Based on this, the best performing model was LightGBM (DART) with 14 selected input features, with sensitivity of 88.2% ± 1.8% (95 %CI), specificity of 80.8% ± 1.4% and AUPRC of 0.739 ± 0.043.

3.2.1 Feature importance

LightGBM (DART) consistently gave the best sensitivity, thus we examined the best performing model to identify the selected features and their relative importance, in order to better understand which components are important in determining longitudinal depression severity changes.

Feature importance for gradient boosting models can be assessed by two key metrics: gain importance and split importance [54, 55]. Gain importance measures the improvement in accuracy that a feature provides, while split importance considers the number of times the feature is used in a model. Taken together, these metrics help us understand which features contribute the most to the 'decisions' that the model makes.

Figure 3.3A summarizes feature importance in the best performing model. Note that all selected features are important, but vary in their relative importance. Less important features included static features, such as the presence of chronic comorbidities or events like pregnancy, as well as dynamic LMC features like starting/changing/stopping medication. PGHD features, such as trends in light and sedentary activity or naps had intermediate importance. The most important features however, were intermediate col-

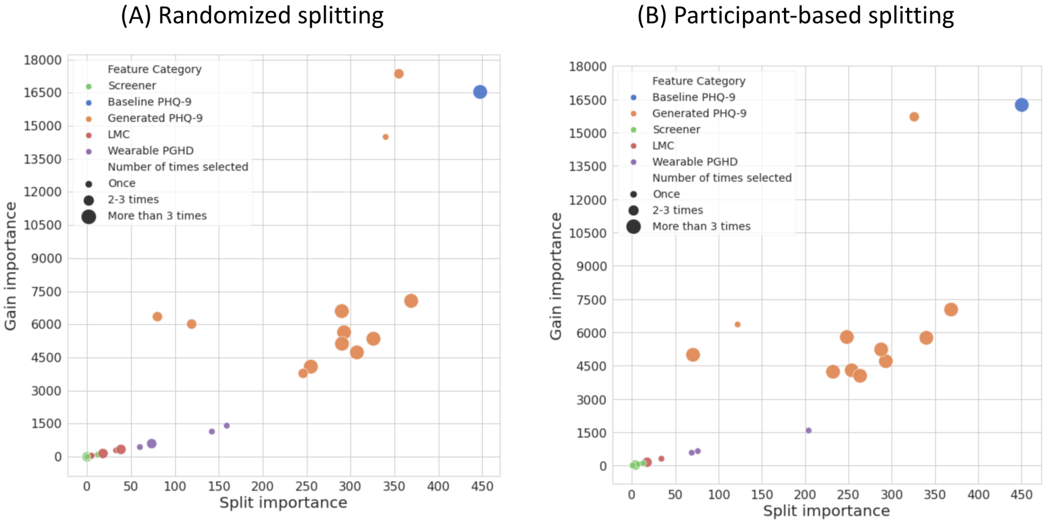


FIGURE 3.3: *Relative importance of selected features in the best performing models for (A) randomized and (B) participant-based splitting strategies. Gain importance is plotted on the vertical axis and split importance on the horizontal axis. Each point represents the mean gain and split importance for a single feature, observed across 5-fold cross-validation. The size of the points reflects the number of times a feature has been selected over the random split (larger = more frequently selected), and color represents the type of feature (screener demographic features = green, LMC features = red, wearable PGHD = purple, baseline PHQ=9 = blue, generated PHQ-9 features = orange).*

lected and generated PHQ-9 category related features. Specifically, the self-reported baseline PHQ-9 category collected at SM0, and the generated PHQ-9 categories and the individual probabilities of being in each category at SM1 and SM2, based on the intermediate labels generated by the phase 1 model. These generated intermediate PHQ-9 category features also incorporate input from across the PGHD sources. We note that the PHQ-9 category features were also selected the most frequently, whereas wearable PGHD features were selected least frequently, but provided higher importance than other PGHD features when they were selected.

3.2.2 Model generalizability to new individuals

Although the observed performance was very encouraging, we wanted to additionally assess performance in a manner representative of how such a model might be deployed. In a real world situation, a trained model (e.g. as part of a smartphone application) would need to make predictions for participants that the model is naive to, i.e. people who have just downloaded the application and perhaps only filled out the baseline assessments, and did not contribute data used in the model construction. We therefore adjusted our cross-validation approach from a randomized splitting strategy to a participant-based strategy, where all samples from a subset of individuals are reserved for testing, rather than a preset fraction of randomly selected samples.

We repeated model construction including feature selection, model fitting and testing of performance as previously described, focusing on LightGBM DART as this classification algorithm consistently performed best in randomized splitting. The best performing model, ranked by sensitivity, had 13 input features, with which we achieved sensitivity of 87.3% ± 1.7%, specificity of 83.1% ± 2.0% and AUPRC of 0.751 ± 0.044. The feature importance for the best performing model is visualized in Figure 3.3B, and the extended performance results are presented in Appendix H.

We examined the most important features in the model and observed that the selected features were very similar to the randomized splitting strategy, with a few small exceptions. Both train:test split strategies considered the most important features to be the baseline PHQ-9 category, as well as the generated PHQ-9 category probabilities based on the intermediate generated labels from phase 1. Of the activity-based features, both strategies also selected the number of sedentary days as important, however the randomized strategy considered the variability and change in sleep habits to be more important, whereas the participant-based strategy selected measures of central tendency (e.g. mean, median) as more important indicators of increased depression severity. The static demographic features and LMC features selected were also similar for both strategies, focusing on the presence of chronic comorbidities, pregnancy/birth and ethnicity.

3.3 Combined pipeline results

This section presents the results of the combined pipeline, consisting of phase 1c and phase 2c models. We first present the results of the phase 1c model, adapted to reduce overfitting, followed by the final results of the phase 2c model, representing the final performance of the PSYCHE-D approach, as well as a feature importance analysis for each of the two phases.

3.3.1 Phase 1c model results

We present the results of a participant-based splitting strategy, with an 80:20 train:test split, where 800 of 4306 unique participants were selected randomly for the test split, resulting in an approximate 20% of participants in the test set, on average. Sample stratification was not performed during this split, as we consistently saw that due to the large dataset volume, the random participant-based split resulted in a distribution of absolute and relative depression severity levels that was similar to the overall trends in our dataset.

Using a pipeline consisting of recursive feature elimination feature selection followed by a classification model, we used randomized grid search cross-validation to optimize the following hyperparameters: the number of features chosen (best values ranging from 22 to 28), dropout rate, maximum tree depth and number of estimators. The classification model chosen was changed from XGBoost to LightGBM DART, as it allowed us to control the dropout rate in order to improve model generalization. We also took care to perform a group K-fold when performing cross-validation for the grid search, in order to have a participant-based split in each fold of the cross-validation process.

We performed the randomized grid search for 50 iterations over 5 folds. We then checked if performance could be improved by manually testing some hyperparameter values close to the selected ones, as it was significantly less computationally expensive than performing a second full grid search on the selected hyperparameter value ranges. Ultimately, we achieved a model with a Kappa score of 0.476 (± 0.017), adjacent accuracy of 77.6% (± 2.0%), balanced accuracy of 35.3% (± 1.0%) and a weighted F1-score of 0.41 (± 0.012).

The corresponding confusion matrix representing average category accuracy achieved by the phase 1c model is visualized in Figure 3.4. When comparing performance to the previous confusion matrix in Figure 3.2, we note that although the overall performance is worse with the participant-based splitting strategy here, it is mainly worse for the participants with mild to moderate depression, and the performance for participants with minimal or severe depression is not as impacted. We also observe that the general trend that predicts participants to be in a lower category than they actually are is still present, and the adjacent accuracy is lower than with randomized splitting based on the phase 1 model.

We plot the relative split vs gain importances of the selected features of the phase 1c model in Figure 3.5. The full list of selected features is available in Appendix I. We observed that the majority of LMC features have lower importance than screener

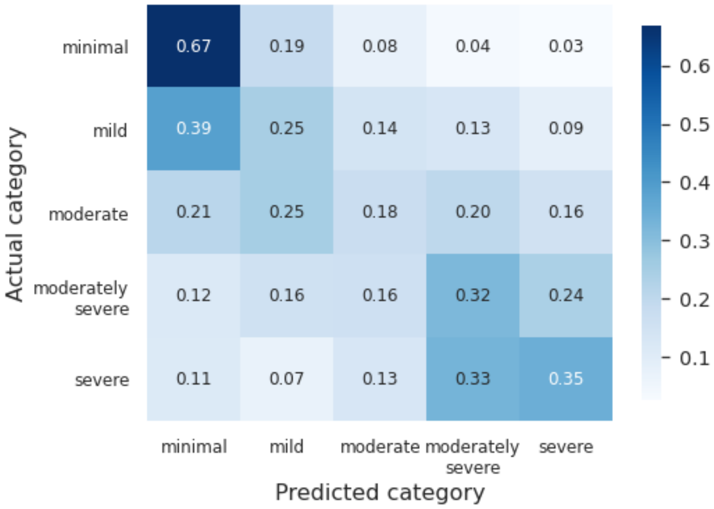


FIGURE 3.4: *Confusion matrix showing the phase 1c model's PHQ-9 score category accuracy distribution. Performance overall is weak, but adjacent accuracy is high, and classification performance of high and low severity samples is relatively high, compared to performances seen for intermediate severity samples.*

features, and that the screener features tend to appear more often among different folds, making them reliable predictors of absolute depression severity.

We do not visualize one wearable PGHD feature in this plot — the mean ratio of the time spent asleep to the time spent in bed over the past four days — because the split importance is a magnitude higher than for the other features.

Three wearable PGHD features were selected to predict participants' absolute depression severity categories, with dates relative to the PHQ-9 completion date. All three features were related to sleep trends: the number of hypersomnia days in the past 7 days, the range of the hour at which the participant went to sleep over the past 14 days, and the mean ratio of the time spent asleep to the time spent in bed over the past 4 days. We remark that these features are all absolute and not relative behavioural features, i.e. they do not represent changes in trends of participants, but instead the absolute statistics that are comparable between participants.

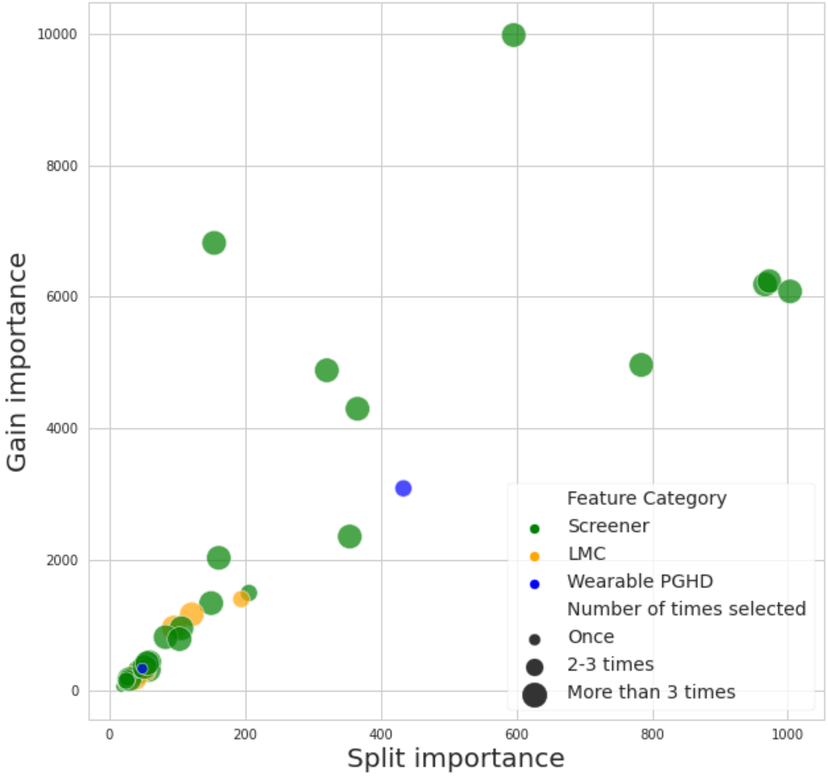


FIGURE 3.5: *Relative importance of selected features in the best performing phase 1c model, using participant-based splitting. Gain importance is plotted on the vertical axis and split importance on the horizontal axis. Each point represents the mean gain and split importance for a single feature, observed across 5-fold cross-validation. The size of the points reflects the number of times a feature has been selected over the random split (larger = more frequently selected), and color represents the type of feature (screener features = green, LMC features = yellow, wearable PGHD features = blue).*

3.3.2 Phase 2c model results

The key component of the combined PSYCHE-D pipeline is the elimination of data leakage between the two phases. Removing data leakage, i.e. employing a participant-based splitting strategy that remains consistent throughout both phases, resulted in a performance of 60.1% (± 2.6%) sensitivity, 57.4% (±0.8%) specificity, 0.282 (± 0.022) AUPRC using the phase 2 model. We note the very low AUPRC score, as an AUPRC score of 0.21 corresponds to baseline performance (as 21% of samples are labeled as increased depression severity, or “positive”). These scores indicate that the model classifies an exceedingly high proportion of samples as positive, and that the classifications are often incorrect. To improve model performance, we modified the phase 2 model to obtain the phase 2c model, starting with rerunning the complete feature selection process.

Number of features selected	**Sensitivity (±95% CI)**	**Specificity (±95% CI)**	**AUPRC (±95% CI)**
12	0.546 (± 0.018)	0.662 (± 0.044)	0.309 (± 0.019)
13	0.554 (± 0.008)	0.653 (± 0.042)	0.310 (± 0.024)
14	0.542 (± 0.030)	0.675 (± 0.035)	0.313 (± 0.024)
15	0.538 (± 0.052)	0.645 (± 0.041)	0.309 (± 0.022)
16	0.533 (± 0.039)	0.653 (± 0.041)	0.309 (± 0.015)

Table 3.4: *Summarized phase 2c LightGBM (DART) model performance for predicting longitudinal change in depression severity for different numbers of features selected in the feature selection process. We report the following aggregated performance metrics: Sensitivity (Recall), Specificity and Area Under the Precision-Recall Curve (AUPRC). 95% confidence intervals are included for each metric. Sensitivity is prioritized as the key performance metric to select the best model.*

This time, in the model selection process, we gave higher importance to AUPRC scores than before, as we found that this was an issue with our previous implementation. We nevertheless took into account sensitivity scores, as our primary goal in this work is to correctly identify individuals reporting increased depression severity.

The phase 2c model selection experiments focused on selecting the optimal number of features for forward sequential feature selection, followed by fitting a LightGBM DART classification model, as this was consistently the best classification algorithm in the phase 2 model development. The feature selection process selects from a subset of all available features, consisting of the collected baseline PHQ-9 category, all LMC features, all generated PHQ-9 features, and the most important subsets of wearable PGHD and screener survey response features as described in Section 2.7.

Table 3.4 presents the results of the phase 2c model selection process with several numbers of selected features. We observe that the performances of the different models are comparable in terms of mean sensitivity. The final best performing model selects 13 input features and achieves a sensitivity of 55.4% (± 0.8%), a specificity of 65.3% (±4.2%) and an AUPRC score of 0.31 (±0.024). We selected this model as it has the highest sensitivity, and a distinctly lower sensitivity CI compared to the other models.

We also analyzed the most frequently selected features and their importances. Figure 3.6 visualizes the mean split feature importances of features selected in at least three of five train:test splits in the second phase model. We note that the majority of the most common features are LMC features. In particular, medication-related features represent 3 of the 11 most frequently selected features, implying that there is likely a notable relationship between participants' medication habits and depressive symptom severity. Participants however do not specify which medications they change, so we cannot reach further conclusions regarding the medication features. We also remark that the two most important features are the baseline PHQ-9 category, and the generated PHQ-9 category at SM1. It is interesting to note that two static socio-demographic features are present: sex and whether the participant has insurance.

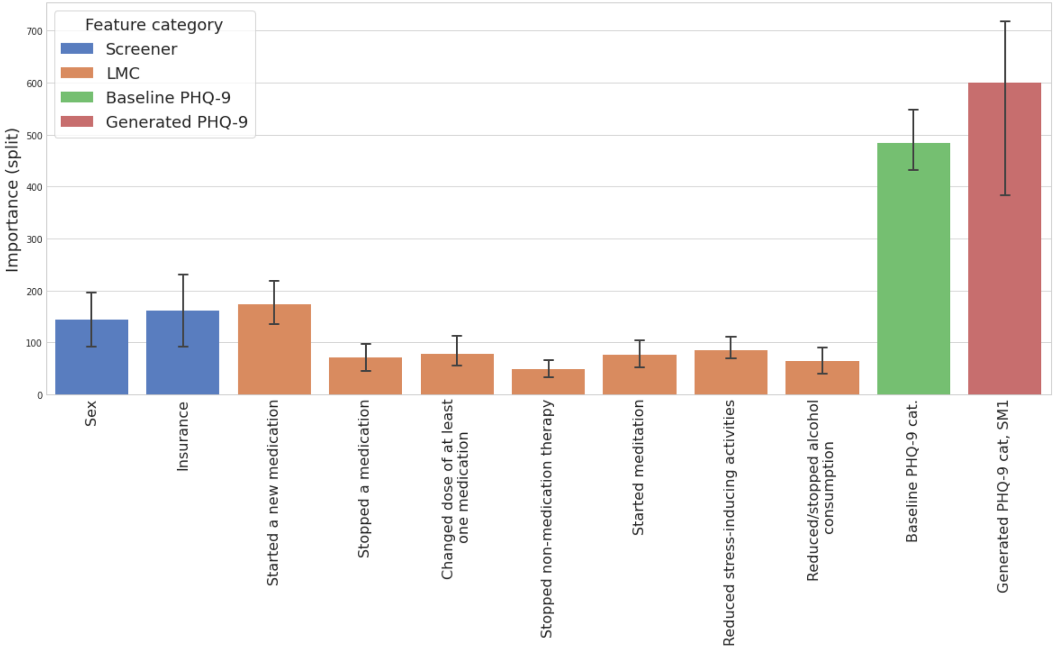


FIGURE 3.6: *Split feature importance of the features selected in at least three of five train:test splits in the best performing phase 2c model of the combined pipeline. Colour represents the type of feature (screener features = blue, LMC features = orange, baseline PHQ-9 = green, generated PHQ-9 features = red). 95% confidence intervals for the split feature importance are also visualized.*

Figure 3.7 shows the relative gain and split importance for all the features selected in the best performing phase 2c model. There is a positive linear correlation between the gain and split importances. The sizes of the points reflect how frequently the feature was selected for a split: we can see that the most consistently frequent features were LMC features. Only one wearable PGHD feature is selected: the slope of a 7 day linear regression fit on the number of minutes a participant spends awake during sleep time. This feature has very high split and gain importance, but is selected in only one split.

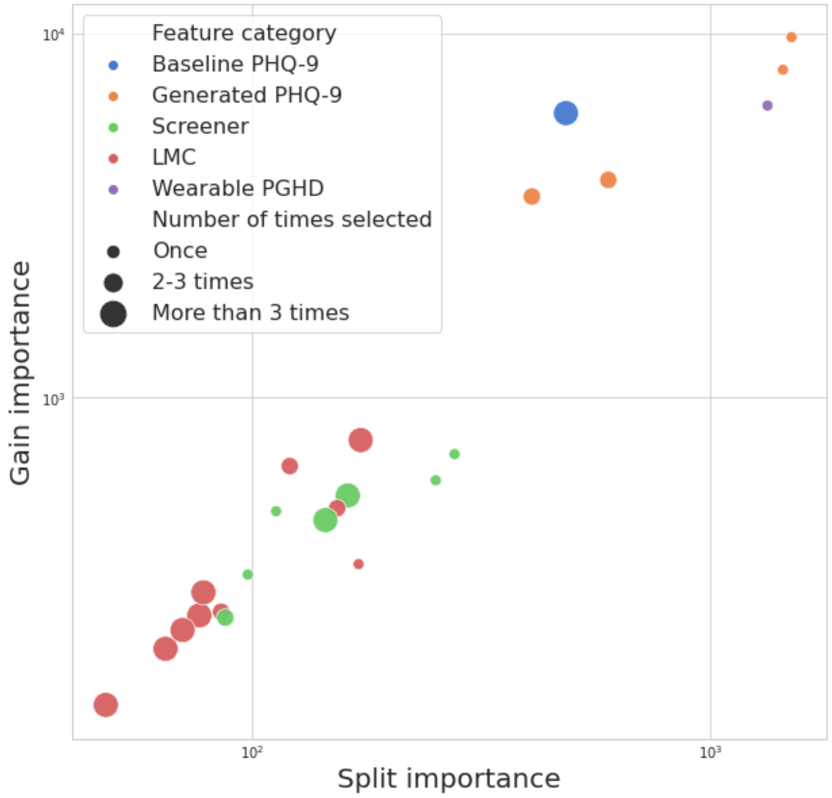


FIGURE 3.7: *Relative importance of selected features in the best performing phase 2c model in the combined pipeline. Gain importance is plotted on the vertical axis and split importance on the horizontal axis, each with a log scale. Each point represents the mean gain and split importance for a single feature, observed across 5-fold cross-validation. The size of the points reflects the number of times a feature has been selected for different splits (larger = more frequently selected), and color represents the type of feature (screener features = green, LMC features = red, wearable PGHD = purple, baseline PHQ-9 = blue, generated PHQ-9 features = orange).*

Chapter 4

Discussion

Person-generated health data represents a low burden, direct connection to the patient journey and such data has already been demonstrated to be a valuable component of models that predict health-relevant outcomes [59, 60]. In this thesis, we present a two-phase approach for predicting longitudinal deterioration in depression status. Starting from PHQ-9 category labels collected at 3-month intervals, in phase 1, we increased the label density by generating intermediate PHQ-9 category labels using wearable PGHD and lifestyle and medication change information. In a second phase, we combined self-reported and generated PHQ-9 category labels, plus additional recent wearable PGHD and lifestyle and medication change information to predict deterioration of depression status three months after the initial self-report. The final presented two-phase approach is very low burden and requires very little interaction from participants. The information we used as input consists of simple self-reports and passively-collected data from consumer-grade wearables.

Due to the initial combining approach of the two phases, we faced data leakage issues that had significant implications for the generalization capabilities of the phase 1 and 2 models. An important part of our work includes identifying and eliminating this data leakage. Having removed data leakage, we reran the model construction for both phases of PSYCHE-D, updating the selected features and parameters to the new data, thus creating the phase 1c and phase 2c models in the combined pipeline. We nonetheless discuss the initial phase 1 and 2 results, comparing them to the final phase 1c and 2c results, as they still provide valuable insights. We also remind the reader that prior to the implementation of the combined pipeline, analyses are based on randomized splitting, whereas in the combined pipeline we only perform participant splitting, as that is our end goal — to use our approach to generalize to new participants. Thus, the initial implementations nevertheless provide useful insights, but only regarding the potential of the PSYCHE-D approach on existing participants.

The phase 1 model in our approach was used to generate intermediate PHQ-9 category labels. Using random splitting, i.e. samples generated from the same participant can appear in both the train and test sets, our initial model achieved a Kappa score of 0.656 (±0.016 95%CI), with a high adjacent accuracy (89.0%). In the combined pipeline with no data leakage, we optimized the first phase for better performance on a participant-based split, i.e. all samples generated from a given participant appear either in the train or test set. Thus, the phase 1c model achieved a Kappa score of 0.476 (± 0.017 95%CI), and an adjacent accuracy of 77.6%. We remarked that the phase 1c model was best at correctly classifying participants with minimal and severe depression, the rest of the participants not achieving such high accuracy scores, as visualized in Figure 3.4.

The best implementation of the phase 2 model of PSYCHE-D, in which we predicted deterioration in PHQ-9 category three months after the initial PHQ-9 was completed, achieved a sensitivity of 88.2% using randomized splitting and a sensitivity of 87.3% using participant-based splitting. Unfortunately, this strong performance was caused by data leakage between the first and second phase models, and thus had to be amended in the combined pipeline with no data leakage between the two phases, using participant-based splitting. The final adapted phase 2c model achieved a sensitivity of 55.4%. We prioritized sensitivity as the primary key metric because the potential consequences of false negatives (i.e. not identifying a person with deteriorating depression) is much higher than the cost of false positives (i.e. incorrectly suspecting someone of deteriorating depression). We envision a consumer-facing application, where individuals suspected of worsening symptoms would be engaged directly, and thus individuals incorrectly labeled could simply dismiss the notification. Nevertheless, despite the data imbalance, the phase 2c model also achieves an average specificity of 65.3%, showing that using the PSYCHE-D approach in its current state still has significant targeting capabilities.

The second phase used generated intermediate PHQ-9 labels from the first phase as input features to determine increase in depression severity. We note that due to data sparsity, intermediate labeling was not always available, and thus some samples did not have two intermediate PHQ-9 category labels, but sometimes one or none. Nonetheless, as LightGBM handles missing values, the lack of intermediate labeling for previously unseen participants still allowed for adequate second phase model predictions. This highlights that the approach described in our work is indeed low-burden and robust to missing data, which is crucial in long-term healthcare monitoring systems.

Another strength of our approach is the use of interpretable classification algorithms. We can thus infer which factors are most important in determining whether an individual experiences worsening depression severity, and test these hypotheses in future work and begin open discussions with healthcare professionals regarding our findings, to leverage the use of consumer-grade wearables to monitor mental health in a way that is especially helpful to healthcare professionals to determine diagnoses and treatment options.

We examined the features that were selected as most important to determine absolute depression severity levels in both phase 1 and 1c models. The selected features in both models imply that, in general, socio-economic factors, demographic factors and life events (e.g. trauma, birth) are stronger predictors of depression severity than wearable PGHD, and lifestyle and medication changes. Large-scale studies have indeed shown that gender, traumatic experiences and comorbidity burden all have an influence on depression [61]. Age has also been shown to be strongly correlated with absolute depression levels according to previous studies [62].

Both models have also identified sleep and inactivity-related factors to be associated with absolute depression severity. The phase 1 model found hypersomnia and reduced activity to be associated with depression severity, which is consistent with previous studies [19, 61]. The phase 1c model found the mean ratio of time spent asleep to the time spent in bed to be a highly important feature. This feature could be a proxy to diagnose a participant with insomnia, which is strongly correlated with worsening depression [19, 63].

In the phase 2c model, features from all input sources were included in the best performing models, generated from different train:test splits, but with different relative importance. Various static features (i.e. those defined at enrollment, which don't change afterwards) were selected, but were of relatively low importance, and lower frequency. In particular, only two static features were selected as most important in more than half of the models: sex and whether a participant has insurance. This implies that generally in our cohort there was an observation of worsening depression for one sex, and that having insurance is an indicator of whether a participant's depression level worsens or not. We have two possible explanations for the latter, which could be explored in a follow-up study involving DiSCover participants. One reason as to why having insurance could be an indicator of increase in depressive symptoms, is that often people who experience depressive symptoms do not seek treatment for financial reasons [8]. Another possible correlation between depression status, having insurance and reporting medication changes is that if a participant feels like their depression is worsening, they may increase their healthcare engagement and try different medication strategies.

While LMC features were of low importance on average, they were frequently selected, implying that regardless of the participant split (i.e. regardless of socio-demographic characteristics and life events), self-awareness of lifestyle and medication changes is essential in predicting changes in depression status. In particular, starting to practice meditation and reducing stress-inducing activities were both frequently selected features, which is consistent with recent research on the effects of meditation and mindfulness on depressive symptoms [64, 65]. An additional frequently selected lifestyle change feature was reporting reducing or stopping alcohol consumption. This is consistent with previous research on the relationship between alcohol use and depression [66].

The only wearable PGHD feature selected by the phase 2c model was the slope of a 7-day linear regression of the total number of minutes awake during sleep, i.e. the average change in the number of minutes spent awake during sleep. Like the sleep feature selected by phase 1c, this could also be an indicator of insomnia, which has been found to be associated with depression [19, 63]. The feature was only selected once, but it was identified as highly important by the model when it was selected. A possible reason as to why only one wearable PGHD feature was selected once, compared to the majority of LMC features being selected frequently, is because wearable PGHD is significantly more noisy than self-reported binary LMC features, e.g. having reduced alcohol consumption vs not having reduced alcohol consumption. Baseline and generated PHQ-9 features were relatively the most important and frequently selected features. The generated

PHQ-9 features from the first phase model also summarize features from across all input sources. The intermediate generated PHQ-9 category labels are analogous to "weak labeling", which can help reduce large-scale, noisy data to a signal useful for supervised learning, for example Zhan et al [67].

Although the initial phase 2 model results were not accurate due to data leakage, it is still interesting to note the differences between the selected wearable PGHD features, specifically sleep features, using randomized vs participant-based splitting. In particular, as the data leakage primarily affected the generated PHQ-9 feature quality, with significantly higher accuracy for previously seen participants, and not wearable PGHD features. Relative activity trends, such as variability and change in sleep were selected as best determinants of increased depression severity using randomized splitting, whereas more absolute trends, in particular measures of central tendency, were selected using participant-based splitting. The randomized splitting strategy particularly selected features representing average sleep onset time, variability in sleep onset time, and changes in the percentage of sleep time spent awake, whereas the participant-based splitting strategy only selected features representing average sleep onset time, and average sleep time. From this, we can deduce that average sleep onset time is a good determinant of increasing depression severity, which is consistent with previous research [19], and that variability in sleep is participant-specific and not necessarily a good predictor for generalizing to other participants. However, general sleep and step-related trends are good indicators of inter-participant prediction of worsening depression.

4.1 Limitations

The work presented in this thesis demonstrates the power of a PGHD-based model for predicting long-term changes in depression status. The presented approach nevertheless has multiple limitations, some of which could be eliminated in a future work.

A key limitation of our work was to employ only a randomized splitting strategy throughout the development and testing of the first phase, and to only test participant-based splitting after the implementation of the second phase of PSYCHE-D. Initially, we focused on a randomized splitting strategy, in which samples generated from the same participants could be found in both the train and test sets, as described in Figure 2.7. This splitting strategy is valid, i.e. we would ideally use the presented approach to monitor participants' depression statuses over longer periods of time with continuous model refitting, so the same participants would eventually be part of the training and prediction sets. Nonetheless, our main goal was to be able to generalize to new participants, and, likely due to the initial focus on this splitting method, the model's generalization properties were lower than they could have been. Because of the randomized splitting, it is likely that the model did not focus on behavioural features as much as on the static socio-demographic characteristics to learn participant depression levels. This resulted in lower generalization properties, as seen by participant-based splitting, even in this very large and diverse population. If we had noticed this flaw sooner, the architecture of the model could have been further adapted to mitigate this issue.

Another major limitation is the model's reliance on the completion of several self-reported surveys over time. Participants were moderately engaged with the year-long research study, but to lower the barrier to participation, the number of surveys could be reduced or replaced with alternative sources of data. For example, instead of LMC surveys, medication changes could be assessed through electronic health records [68] or through other consumer-wearables which incorporate engagement, such as the Oura ring which allows participants to annotate days with a number of tags like medication [69]. Additionally, prescription apps [70, 71] or population health programs [72, 73] could provide a way to reach genuinely disengaged individuals who would not otherwise engage with tools which help them manage their mental health. Similar solutions could be used to track changes in what we defined to be "static" screener features, so that participants can make changes to the socio-demographic, socio-economic and life event-related features used to determine their depression status.

A further limitation of our work is that the data comes from a socio-demographically biased population. As with many studies that focus on using consumer wearable devices for the analysis of mental health [19, 22], our cohort over-represents female participants and college graduates, compared to US Census population statistics [74]. The socio-demographic bias could also be worsened by the assumption that scores from different participants are independent. This bias could however be reduced in a future work by running a larger study with participants of a wider socio-demographic distribution, and providing participants with consumer-grade wearable devices, to build a model that is able to learn the trends in depression changes based on PGHD from more diverse participants.

The data collection process of the DiSCover study design has also limited the time window for making predictions in depression status change to three months. A possible future work could include testing predictions beyond this time horizon. Such a prospective study could also incorporate dynamic (i.e. non-scheduled) self-reports, in order to further reduce participant burden. For instance, using predicted change as a trigger for interaction with the individual, who can then confirm/deny, or provide further context through a chatbot [75, 76, 77]. Such a system would also lend itself to active learning, which could further improve individual predictions. Further future work could focus on the application of PSYCHE to other aspects of mental health, such as anxiety, fatigue and stress.

Chapter 5

Conclusion

In this thesis we present PSYCHE-D, a two-phase approach to predict longitudinal deterioration in depression severity using person-generated health data. Using PHQ-9 category labels collected at 3-month intervals, combined with monthly survey responses and consumer-grade wearable data, in the first phase, we generated intermediate PHQ-9 category labels. In phase two, we used these labels in combination with PGHD to predict an increase in depression severity over a 3-month period.

One of the key challenges in our work was handling sparse data: due to the low-burden nature of the data collection process, we had at most monthly survey responses for lifestyle and medical changes, and quarterly PHQ-9 category labels. Previous studies have attempted to use consumer-grade wearable data, in combination with socio-demographic data, to study patterns in depression levels [19, 20, 21]. Such work usually consists of distinguishing between control and depressed participants, instead of using multi-class classification, and more frequent, often daily, depression severity labels, instead of quarterly ones. To our knowledge, this is the first work to use machine learning to predict depression levels beyond distinguishing controls from depressed participants. It is also one of the rare examples of low-burden data collection in the field of mental health, as depression severity labels were collected at such sparse intervals. We were able to achieve a model that can generalize to new participants to some extent, particularly given the sparsity of our dataset.

The final presented approach shows promise in its generalization capabilities given the variety in our participants and the sparsity of our data. Integrating such a model into consumer-grade wearable device applications to provide participants with a "check-in" on their mental health has the potential to help many individuals at risk of developing worsening depression symptoms by increasing engagement and awareness. The development of such a system has become even more necessary since the COVID-19 pandemic, which has been impacting the mental health of people worldwide for over a year at the time of writing.

Chapter 6

Outlook

The work presented in this thesis was an initial effort to use large-scale person-generated health data collected using low-burden methods to monitor changes in depressive status. The presented approach gave us an indication of what can be achieved and what the are limits of objective low-burden data collection methods for mental health monitoring. We reflect on how some of the approaches could be altered for future work, and how we can proceed with the knowledge acquired from completing this project.

One of the crucial design choices of the phase 1 model was deciding to use the five predefined PHQ-9 categories to define different levels of depression severity. We chose to use the five predefined categories, as opposed to defining participants using binary control/depressed labels. This choice was made because we hoped that the five-category split was sufficiently broad to be able to make distinctions between participants in different categories based on their PGHD, but also granular enough to detect changes between multiple depression categories, and observe longitudinal evolution. An alternative approach that we believe could result in a better performance would be to pose the first phase as a binary classification problem, instead of a multi-class classification problem, and use the second phase to detect participants that move from the "control" to the "depressed" category. Simplifying the first task in this way would also reduce the data imbalance, and thus hopefully result in better intermediate depression severity labels for the second phase to learn from.

A limitation of our work is that despite using a large population to develop the approach, our dataset is socio-demographically biased. This could be reduced in future work by performing independent validation using another low-burden longitudinal PGHD study focusing on mental health, for instance the mental health study presented in Kumar et al [78].

The use of objective wearable PGHD in combination with sparse subjective self-reported data has provided challenges when developing our approach. We were able to conclude that self-reported lifestyle and medication changes are essential to determining changes in depression severity, but further research is required to develop a low-burden model that could reliably monitor patients, for instance by integrating the use of objective health records into the PSYCHE-D approach.

The approach presented in this work shows promise in developing low-burden monitoring systems for depressive symptoms. The objectivity of our approach provides a non-stigmatizing environment to engage people about depression [8]. We hope that this demonstration of the ability to predict long-term changes in depression using a low-burden, PGHD-based approach will have great potential to deliver value to patients.

Bibliography

[1] P. Thera, "Dhammacakkappavattana sutta: Setting in motion the wheel of truth," 1999.

[2] "Depression." `https://www.who.int/news-room/fact-sheets/detail/depression`, Jan. 2020. Accessed: 2021-4-22.

[3] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, "Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the global burden of disease study 2017," *Lancet*, vol. 392, pp. 1789–1858, Nov. 2018.

[4] M. É. Czeisler, R. I. Lane, E. Petrosky, J. F. Wiley, A. Christensen, R. Njai, M. D. Weaver, R. Robbins, E. R. Facer-Childs, L. K. Barger, C. A. Czeisler, M. E. Howard, and S. M. W. Rajaratnam, "Mental health, substance use, and suicidal ideation during the COVID-19 pandemic - united states, june 24-30, 2020," *MMWR Morb. Mortal. Wkly. Rep.*, vol. 69, pp. 1049–1057, Aug. 2020.

[5] C. K. Ettman, S. M. Abdalla, G. H. Cohen, L. Sampson, P. M. Vivier, and S. Galea, "Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic," *JAMA Netw Open*, vol. 3, p. e2019686, Sept. 2020.

[6] "Major depression." `https://www.nimh.nih.gov/health/statistics/major-depression.shtml`, Feb. 2019. Accessed: 2021-4-22.

[7] B. Waitzfelder, C. Stewart, K. J. Coleman, R. Rossom, B. K. Ahmedani, A. Beck, J. E. Zeber, Y. G. Daida, C. Trinacty, S. Hubley, and G. E. Simon, "Treatment initiation for new episodes of depression in primary care settings," *J. Gen. Intern. Med.*, vol. 33, pp. 1283–1291, Aug. 2018.

[8] A. M. Chekroud, D. Foster, A. B. Zheutlin, D. M. Gerhard, B. Roy, N. Koutsouleris, A. Chandra, M. D. Esposti, G. Subramanyan, R. Gueorguieva, M. Paulus, and J. H. Krystal, "Predicting barriers to treatment for depression in a U.S. national sample: A Cross-Sectional, Proof-of-Concept study," *Psychiatr. Serv.*, vol. 69, pp. 927–934, Aug. 2018.

[9] L. J. Barney, K. M. Griffiths, A. F. Jorm, and H. Christensen, "Stigma about depression and its impact on help-seeking intentions," *Aust. N. Z. J. Psychiatry*, vol. 40, pp. 51–54, Jan. 2006.

[10] J.-P. Lépine and M. Briley, "The increasing burden of depression," *Neuropsychiatr. Dis. Treat.*, vol. 7, pp. 3–7, May 2011.

[11] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler, "The economic burden of adults with major depressive disorder in the united states (2005 and 2010)," *J. Clin. Psychiatry*, vol. 76, pp. 155–162, Feb. 2015.

[12] "Determining real-world data's fitness for use and the role of reliability," tech. rep., Duke-Margolis Center for Health Policy, Sept. 2019.

[13] P. Pedrelli, S. Fedor, A. Ghandeharioun, E. Howe, D. F. Ionescu, D. Bhathena, L. B. Fisher, C. Cusin, M. Nyer, A. Yeung, L. Sangermano, D. Mischoulon, J. E. Alpert, and R. W. Picard, "Monitoring changes in depression severity using wearable and mobile sensors," *Front. Psychiatry*, vol. 11, p. 584711, Dec. 2020.

[14] F. Cormack, M. McCue, N. Taptiklis, C. Skirrow, E. Glazer, E. Panagopoulos, T. A. van Schaik, B. Fehnert, J. King, and J. H. Barnett, "Wearable technology for High-Frequency cognitive and mood assessment in major depressive disorder: Longitudinal observational study," *JMIR Ment Health*, vol. 6, p. e12814, Nov. 2019.

[15] B. Kisliuk, "UCLA launches major mental health study to discover insights about depression." `https://newsroom.ucla.edu/releases/ucla-launches-major-mental-health-study-to-discover-insights-about-depression`, Aug. 2020. Accessed: 2021-8-15.

[16] F. Hardeveld, J. Spijker, R. De Graaf, W. A. Nolen, and A. T. F. Beekman, "Prevalence and predictors of recurrence of major depressive disorder in the adult population," *Acta Psychiatr. Scand.*, vol. 122, pp. 184–191, Sept. 2010.

[17] C. F. Reynolds, 3rd, P. Cuijpers, V. Patel, A. Cohen, A. Dias, N. Chowdhary, O. I. Okereke, M. A. Dew, S. J. Anderson, S. Mazumdar, F. Lotrich, and S. M. Albert, "Early intervention to reduce the global health and economic burden of major depression in older adults," *Annu. Rev. Public Health*, vol. 33, pp. 123–135, Apr. 2012.

[18] C. Schley, N. Pace, R. Mann, C. McKenzie, A. McRoberts, and A. Parker, "The headspace brief interventions clinic: Increasing timely access to effective treatments for young people with early signs of mental health problems," *Early Interv. Psychiatry*, vol. 13, pp. 1073–1082, Oct. 2019.

[19] Y. Zhang, A. A. Folarin, S. Sun, N. Cummins, R. Bendayan, Y. Ranjan, Z. Rashid, P. Conde, C. Stewart, P. Laiou, F. Matcham, K. M. White, F. Lamers, S. Siddi, S. Simblett, I. Myin-Germeys, A. Rintala, T. Wykes, J. M. Haro, B. W. Penninx,

V. A. Narayan, M. Hotopf, R. J. Dobson, and RADAR-CNS Consortium, “Relationship between major depression symptom severity and sleep collected using a wristband wearable device: Multicenter longitudinal observational study,” *JMIR Mhealth Uhealth*, vol. 9, p. e24604, Apr. 2021.

[20] J. T. O’Brien, P. Gallagher, D. Stow, N. Hammerla, T. Ploetz, M. Firbank, C. Ladha, K. Ladha, D. Jackson, R. McNaney, I. N. Ferrier, and P. Olivier, “A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression,” *Psychol. Med.*, vol. 47, pp. 93–102, Jan. 2017.

[21] P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu, S. Cohen, K. G. Creswell, J. Mankoff, J. D. Creswell, M. Goel, and A. K. Dey, “Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection,” *ACM Trans. Comput.-Hum. Interact.*, vol. 28, pp. 1–41, Jan. 2021.

[22] I. Moshe, Y. Terhorst, K. Opoku Asare, L. B. Sander, D. Ferreira, H. Baumeister, D. C. Mohr, and L. Pulkki-Råback, “Predicting symptoms of depression and anxiety using smartphone and wearable data,” *Front. Psychiatry*, vol. 12, p. 625247, Jan. 2021.

[23] S. Chancellor and M. De Choudhury, “Methods in predictive techniques for mental health status on social media: a critical review,” *NPJ Digit Med*, vol. 3, p. 43, Mar. 2020.

[24] Z. Vahedi and L. Zannella, “The association between self-reported depressive symptoms and the use of social networking sites (SNS): A meta-analysis,” *Curr. Psychol.*, Jan. 2019.

[25] B. N. Renn, A. Pratap, D. C. Atkins, S. D. Mooney, and P. A. Areán, “Smartphone-Based passive assessment of mobility in depression: Challenges and opportunities,” *Ment. Health Phys. Act.*, vol. 14, pp. 136–139, Mar. 2018.

[26] “Digital signals in chronic pain (DiSCover project).” `https://clinicaltrials.gov/ct2/show/NCT03421223`. Accessed: 2021-8-3.

[27] M. Makhmutova, R. Kainkaryam, M. Ferreira, J. Min, M. Jaggi, and I. Clay, “PSYCHE-D: predicting change in depression severity using person-generated health data (DATASET),” 2021.

[28] M. Makhmutova, R. Kainkaryam, M. Ferreira, J. Min, M. Jaggi, and I. Clay, “Prediction of self-reported depression scores using person-generated health data from a virtual 1-year mental health observational study,” in *Proceedings of the 2021 Workshop on Future of Digital Biomarkers*, DigiBiom ’21, (New York, NY, USA), pp. 4–11, Association for Computing Machinery, June 2021.

[29] J. L. Lee, C. J. Cerrada, M. K. Ying Vang, K. Scherer, C. Tai, J. L. A. Tran, J. L. Juusola, and C. N. Sang, "The DiSCover project: Protocol and baseline characteristics of a decentralized digital study assessing chronic pain outcomes and behavioral data." July 2021.

[30] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The PHQ-9: validity of a brief depression severity measure," *J. Gen. Intern. Med.*, vol. 16, pp. 606–613, Sept. 2001.

[31] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *J. Affect. Disord.*, vol. 114, Apr. 2009.

[32] J. McCambridge, J. Witton, and D. R. Elbourne, "Systematic review of the hawthorne effect: new concepts are needed to study research participation effects," *J. Clin. Epidemiol.*, vol. 67, pp. 267–277, Mar. 2014.

[33] R. Chen, F. Jankovic, N. Marinsek, L. Foschini, L. Kourtis, A. Signorini, M. Pugh, J. Shen, R. Yaari, V. Maljkovic, M. Sunga, H. H. Song, H. J. Jung, B. Tseng, and A. Trister, "Developing measures of cognitive impairment in the real world from Consumer-Grade multimodal sensor streams," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, (New York, NY, USA), pp. 2145–2155, Association for Computing Machinery, July 2019.

[34] J. H. Migueles, C. Cadenas-Sanchez, U. Ekelund, C. Delisle Nyström, J. Mora-Gonzalez, M. Löf, I. Labayen, J. R. Ruiz, and F. B. Ortega, "Accelerometer data collection and processing criteria to assess physical activity and other outcomes: A systematic review and practical considerations," *Sports Med.*, vol. 47, pp. 1821–1845, Sept. 2017.

[35] C. Tudor-Locke, S. M. Camhi, and R. P. Troiano, "A catalog of rules, variables, and definitions applied to accelerometer data in the national health and nutrition examination survey, 2003-2006," *Prev. Chronic Dis.*, vol. 9, p. E113, June 2012.

[36] M. M. Ohayon, "Epidemiology of insomnia: what we know and what we still need to learn," *Sleep Med. Rev.*, vol. 6, pp. 97–111, Apr. 2002.

[37] N. F. Watson, K. P. Harden, D. Buchwald, M. V. Vitiello, A. I. Pack, E. Strachan, and J. Goldberg, "Sleep duration and depressive symptoms: a gene-environment interaction," *Sleep*, vol. 37, pp. 351–358, Feb. 2014.

[38] A. Kandola, G. Ashdown-Franks, J. Hendrikse, C. M. Sabiston, and B. Stubbs, "Physical activity and depression: Towards understanding the antidepressant mechanisms of physical activity," *Neurosci. Biobehav. Rev.*, vol. 107, pp. 525–539, Dec. 2019.

[39] P. C. Dinas, Y. Koutedakis, and A. D. Flouris, "Effects of exercise and physical activity on depression," *Ir. J. Med. Sci.*, vol. 180, pp. 319–325, June 2011.

[40] W. E. Kraus, K. F. Janz, K. E. Powell, W. W. Campbell, J. M. Jakicic, R. P. Troiano, K. Sprow, A. Torres, K. L. Piercy, and 2018 PHYSICAL ACTIVITY GUIDELINES ADVISORY COMMITTEE*, "Daily step counts for measuring physical activity exposure and its relation to health," *Med. Sci. Sports Exerc.*, vol. 51, pp. 1206–1212, June 2019.

[41] "All about heart rate (pulse)." `https://www.heart.org/en/health-topics/high-blood-pressure/the-facts-about-high-blood-pressure/all-about-heart-rate-pulse`. Accessed: 2021-7-20.

[42] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, pp. 389–422, Jan. 2002.

[43] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (New York, NY, USA), ACM, Aug. 2016.

[44] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.*, vol. 33, pp. 613–619, Oct. 1973.

[45] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS One*, vol. 10, p. e0118432, Mar. 2015.

[46] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning from Data: Artificial Intelligence and Statistics V* (D. Fisher and H.-J. Lenz, eds.), pp. 199–206, New York, NY: Springer New York, 1996.

[47] A. Sano, S. Taylor, A. W. McHill, A. J. K. Phillips, L. K. Barger, E. Klerman, and R. Picard, "Identifying objective physiological markers and modifiable behaviors for Self-Reported stress and mental health status using wearable sensors and mobile phones: Observational study," *J. Med. Internet Res.*, vol. 20, June 2018.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. 85, pp. 2825–2830, 2011.

[49] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: a highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), pp. 3149–3157, Curran Associates Inc., Dec. 2017.

[50] R. Korlakai Vinayak and R. Gilad-Bachrach, "DART: Dropouts meet multiple additive regression trees," in *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics* (G. Lebanon and S. V. N. Vishwanathan, eds.), vol. 38 of *Proceedings of Machine Learning Research*, (San Diego, California, USA), pp. 489–497, PMLR, 2015.

[51] J. Budzik, "Many heads are better than one: The case for ensemble learning," Sep 2019. Accessed: 2021-7-23.

[52] M. A. Ganaie, M. Hu, M. Tanveer*, and P. N. Suganthan*, "Ensemble deep learning: A review," Apr. 2021.

[53] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Front. Neurorobot.*, vol. 7, p. 21, Dec. 2013.

[54] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, pp. 1340–1347, May 2010.

[55] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[56] A. Perrier, "Feature importance in random forests." `https://alexisperrier.com/datascience/2015/08/27/feature-importance-random-forests-gini-accuracy.html`, Aug. 2015. Accessed: 2021-6-15.

[57] S. Raschka, "Feature importance permutation." `http://rasbt.github.io/mlxtend/user_guide/evaluate/feature_importance_permutation/`. Accessed: 2021-6-15.

[58] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, ICML '06, (New York, NY, USA), pp. 233–240, Association for Computing Machinery, June 2006.

[59] M. Karas, N. Marinsek, J. Goldhahn, L. Foschini, E. Ramirez, and I. Clay, "Predicting subjective recovery from lower limb surgery using consumer wearables," *Digit Biomark*, vol. 4, pp. 73–86, Nov. 2020.

[60] J. Dunn, L. Kidzinski, R. Runge, D. Witt, J. L. Hicks, S. M. Schüssler-Fiorenza Rose, X. Li, A. Bahmani, S. L. Delp, T. Hastie, and M. P. Snyder, "Wearable sensors enable personalized predictions of clinical laboratory measurements," *Nat. Med.*, May 2021.

[61] A. Brailean, J. Curtis, K. Davis, A. Dregan, and M. Hotopf, "Characteristics, comorbidities, and correlates of atypical depression: evidence from the UK biobank mental health survey," *Psychol. Med.*, vol. 50, pp. 1129–1138, May 2020.

[62] J. Mirowsky and C. E. Ross, "Age and depression," *J. Health Soc. Behav.*, vol. 33, pp. 187–205; discussion 206–12, Sept. 1992.

[63] P. L. Franzen and D. J. Buysse, "Sleep disturbances and depression: risk relationships for subsequent depression and therapeutic implications," *Dialogues Clin. Neurosci.*, vol. 10, no. 4, pp. 473–481, 2008.

[64] F. B. R. Parmentier, M. García-Toro, J. García-Campayo, A. M. Yañez, P. Andrés, and M. Gili, "Mindfulness and symptoms of depression and anxiety in the general population: The mediating roles of worry, rumination, reappraisal and suppression," *Front. Psychol.*, vol. 10, p. 506, Mar. 2019.

[65] C. Elder, S. Nidich, F. Moriarty, and R. Nidich, "Effect of transcendental meditation on employee stress, depression, and burnout: a randomized controlled study," *Perm. J.*, vol. 18, no. 1, pp. 19–23, 2014.

[66] S. Awaworyi Churchill and L. Farrell, "Alcohol and depression: Evidence from the 2014 health survey for england," *Drug Alcohol Depend.*, vol. 180, pp. 86–92, Nov. 2017.

[67] A. Zhan, S. Mohan, C. Tarolli, R. B. Schneider, J. L. Adams, S. Sharma, M. J. Elson, K. L. Spear, A. M. Glidden, M. A. Little, A. Terzis, E. R. Dorsey, and S. Saria, "Using smartphones and machine learning to quantify parkinson disease severity: The mobile parkinson disease score," *JAMA Neurol.*, vol. 75, pp. 876–880, July 2018.

[68] R. Kukafka, J. S. Ancker, C. Chan, J. Chelico, S. Khan, S. Mortoti, K. Natarajan, K. Presley, and K. Stephens, "Redesigning electronic health record systems to support public health," *J. Biomed. Inform.*, vol. 40, pp. 398–409, Aug. 2007.

[69] "Oura ring." `https://ouraring.com/`. Accessed: 2021-6-11.

[70] "BfArM - digital health applications (DiGA)." `https://www.bfarm.de/EN/Medical-devices/Tasks/Digital-Health-Applications/_node.html`. Accessed: 2021-7-1.

[71] "Osmind." `https://www.osmind.org/`. Accessed: 2021-7-7.

[72] "LumiHealth." `https://www.lumihealth.sg`. Accessed: 2021-7-1.

[73] D. V. Gunasekeran, R. M. W. W. Tseng, Y.-C. Tham, and T. Y. Wong, "Applications of digital health for public health responses to COVID-19: a systematic scoping review of artificial intelligence, telehealth and related technologies," *NPJ Digit Med*, vol. 4, p. 40, Feb. 2021.

[74] United States Census Bureau, "QuickFacts: United States."

[75] H. Gaffney, W. Mansell, and S. Tai, "Conversational agents in the treatment of mental health problems: Mixed-Method systematic review," *JMIR Mental Health*, vol. 6, p. e14166, Oct. 2019.

[76] A. A. Abd-Alrazaq, M. Alajlani, N. Ali, K. Denecke, B. M. Bewick, and M. Househ, "Perceptions and opinions of patients about mental health chatbots: Scoping review," *J. Med. Internet Res.*, vol. 23, p. e17828, Jan. 2021.

[77] L. Tudor Car, D. A. Dhinagaran, B. M. Kyaw, T. Kowatsch, S. Joty, Y.-L. Theng, and R. Atun, "Conversational agents in health care: Scoping review and conceptual analysis," *J. Med. Internet Res.*, vol. 22, p. e17158, Aug. 2020.

[78] S. Kumar, J. L. A. Tran, E. Ramirez, W.-N. Lee, L. Foschini, and J. L. Juusola, "Design, recruitment, and baseline characteristics of a virtual 1-year mental health study on behavioral data and health outcomes: Observational study," *JMIR Mental Health*, vol. 7, p. e17075, July 2020.

Appendices

Appendix A

Patient Health Questionnaire-9 (PHQ-9)

Over the last 2 weeks, how often have you been bothered by any of the following problems?

	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself — or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed. Or the opposite — being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead, or of hurting yourself	0	1	2	3

Appendix B

Survey response features

The self-reported screener survey and monthly self-reported lifestyle and monthly changes surveys produced a set of static and dynamic survey response features, respectively.

Static survey features are not tracked for changes throughout the study, and report the participant's socio-economic, demographic and medical status at baseline. These features are the following:

- Sex
- Race/ethnicity (White, Black, Asian, Hispanic, other; multiple selection is possible)
- Birth year
- Education level (seven levels, from not completing high school to doctorate degree)
- Height, in inches
- Weight, in lbs
- BMI (calculated from height and weight values)
- Pregnant
- Has given birth in the past year
- Experienced trauma resulting in injury (e.g. accident, fall) in the past year
- Has health insurance
- Has had sufficient money in the past month to meet basic living needs over the past month (never, rarely, some of the time, most of the time, all the time, don't know, prefer not to answer)
- Has received money or assistance from the government (e.g. food stamps, disability payments) in the past month
- Annual income in the past year, before taxes
- Number of individuals in the household

- Diagnosed comorbidities:
  - Cancer
  - Type 1 diabetes
  - Type 2 diabetes
  - Gout
  - Migraines
  - Multiple sclerosis
  - Osteoporosis
  - Fibromyalgia, peripheral neuropathic pain or central neuropathic pain (sum of diagnoses)
  - Osteoarthritis, rheumatoid arthritis (sum of diagnoses)
- For participants with migraines:
  - Number of migraine days per month, on average
  - Takes daily prescription migraine medication

The self-reported lifestyle and monthly changes survey responses produced the following features, regarding changes that took place during the month prior to survey completion:

- Started a new medication
- Stopped a medication
- Changed the dosage of one of my medications
- Started a new non-medication therapy (e.g. physical therapy, acupuncture, meditation)
- Stopped a non-medication therapy
- Do not take any medications and do not use any non-medication wellness therapies
- Increased physical activity or improved eating habits
- Began meditation or other relaxation techniques
- Reduced stress inducing activities
- Reduced alcohol consumption or stopped drinking alcohol
- Quit smoking

Appendix C

Wearable PGHD features

A total of 124 features were generated from wearable PGHD. Three sets of features were generated from step and sleep data: statistical trends, linear regression model features, and threshold-based features.

Statistical trends include the mean, median, maximum, IQR, sum, and range. These trends were observed for the following day-level aggregated features:

- Number of steps taken while awake
- Number of minutes with light physical activity range steps
- Number of minutes with moderate to vigorous physical activity range steps
- Number of minutes asleep
- Number of minutes in bed
- Sleep start time

Linear regression models were fit over 7-day and 14-day windows to observe changes in participants' behaviour for the following day-level aggregated features:

- Number of steps taken while awake
- Number of minutes with ”not moving” range steps
- Number of minutes with light physical activity range steps
- Number of minutes with moderate to vigorous physical activity range steps
- Number of distinct step streaks (continuous physical activity minutes, lasting at least 10 minutes)
- Number of steps over a rolling 6 minute window
- Number of minutes asleep
- Number of minutes in bed
- Number of times in an awake state during sleep

- Sleep efficiency
- Sleep start time

Threshold-based features are aggregated over 7-day and 14-day windows, as counts and percentages, based on the following day-level thresholds:

- Hypersomnia: over 10 hours of sleep
- Hyposomnia: less than 5 hours of sleep
- Active: over 10,000 steps
- Sedentary: fewer than 5,000 steps

Appendix D

Phase 1 selected survey response features

The following list contains the features that are part of the "selected" subset of survey response features, contributing to the best performing phase 1 model:

- Sex
- Birth year
- Education level
- Height, in inches
- Weight, in lbs
- BMI (calculated from height and weight values)
- Has given birth in the past year
- Experienced trauma resulting in injury (e.g. accident, fall) in the past year
- Has health insurance
- Has had sufficient money in the past month to meet basic living needs over the past month
- Has received money or assistance from the government (e.g. food stamps, disability payments) in the past month
- Number of individuals in the household
- Diagnosed comorbidities:
  - Cancer
  - Type 2 diabetes
  - Gout
  - Migraines
  - Osteoporosis
  - Fibromyalgia, peripheral neuropathic pain or central neuropathic pain (sum of diagnoses)

  - Osteoarthritis, rheumatoid arthritis (sum of diagnoses)

- For participants with migraines:
  - Number of migraine days per month, on average
  - Takes daily prescription migraine medication
- Started a new medication
- Changed the dosage of one of my medications
- Do not take any medications and do not use any non-medication wellness therapies

Appendix E

Initial phase 1 experiments

The following table summarizes the results of initial phase 1 experiments, in which we tested the performance of tuned XGBoost models on wearable PGHD input features, combined with survey responses, using a randomized splitting strategy. One can see that using only step or sleep features does not generate a well performing model. As these were initial experiments, we did not perform multiple splits to obtain confidence intervals.

There is a significant increase in performance metric scores when survey responses are added as input features. The selected survey response features can be referred to in Appendix D. Extensive testing with confidence intervals was completed using input feature sets that include survey responses.

Input features	**Adjacent accuracy**	**Balanced accuracy**	**Kappa**	**F1-score**
Steps (statistical trends)	0.676	0.250	0.150	0.326
Steps (all)	0.733	0.257	0.227	0.355
Steps (selected using RFECV)	0.725	0.248	0.214	0.344
Steps (highly correlated with PHQ-9 based on training set)	0.717	0.252	0.213	0.338
Sleep (statistical trends)	0.720	0.261	0.207	0.346
Sleep (all)	0.745	0.259	0.201	0.356
Sleep (selected using RFECV)	0.736	0.241	0.204	0.330
Sleep (highly correlated with PHQ-9 based on training set)	0.703	0.241	0.133	0.325
Survey responses (static)	0.897	0.468	0.678	0.540
Survey responses (variable)	0.656	0.261	0.189	0.302
Survey responses (all)	0.900	0.454	0.672	0.536
Survey responses (selected)	0.886	0.463	0.666	0.546
Survey responses (selected), steps (all)	0.838	0.333	0.488	0.448
Survey responses (selected), steps (highly correlated)	0.855	0.392	0.558	0.479
Survey responses (selected), sleep (all)	0.840	0.356	0.490	0.454
Survey responses (selected), sleep (highly correlated)	0.882	0.417	0.617	0.513

Appendix F

Phase 1 experiments

The following table summarizes the performance of phase 1 predictive models using a combination of survey response and wearable PGHD input features, with 95% confidence intervals, based on five randomized train:test splits.

Input features	**Model**	**Adjacent accuracy (±95 CI)**	**Balanced accuracy (±95 CI)**	**Kappa (±95 CI)**	**F1-score (±95 CI)**
Survey responses (demographic)	Logistic regression	0.625 (± 0.032)	0.247 (± 0.018)	0.161 (± 0.024)	0.308 (± 0.018)
	XGBoost	0.868 (± 0.010)	0.465 (± 0.024)	0.623 (± 0.027)	0.525 (± 0.015)
Survey responses (static)	Logistic regression	0.723 (± 0.003)	0.349 (± 0.026)	0.423 (± 0.017)	0.379 (± 0.021)
	XGBoost	0.884 (± 0.009)	0.482 (± 0.026)	0.668 (± 0.023)	0.546 (± 0.019)
Survey responses (all)	Logistic regression	0.729 (± 0.004)	0.347 (± 0.022)	0.427 (± 0.015)	0.380 (± 0.017)
	XGBoost	0.887 (± 0.007)	0.463 (± 0.023)	0.656 (± 0.018)	0.539 (± 0.014)
Survey responses (selected)	Logistic regression	0.733 (± 0.004)	0.354 (± 0.024)	0.429 (± 0.017)	0.387 (± 0.017)
	XGBoost	0.887 (± 0.005)	0.470 (± 0.020)	0.659 (± 0.017)	0.542 (± 0.013)
Survey responses (all), wearable PGHD (all)	Logistic regression	0.722 (± 0.008)	0.348 (± 0.017)	0.420 (± 0.024)	0.380 (± 0.007)
	XGBoost	0.829 (± 0.013)	0.319 (± 0.007)	0.440 (± 0.023)	0.432 (± 0.008)
Survey responses (selected), steps (HC)	Logistic regression	0.729 (± 0.007)	0.347 (± 0.018)	0.429 (± 0.015)	0.377 (± 0.012)
	XGBoost	0.854 (± 0.008)	0.373 (± 0.004)	0.526 (± 0.009)	0.474 (± 0.007)
Survey responses (selected), sleep (HC)	Logistic regression	0.734 (± 0.006)	0.356 (± 0.028)	0.431 (± 0.019)	0.391 (± 0.021)
	XGBoost	0.873 (± 0.005)	0.417 (± 0.019)	0.591 (± 0.027)	0.511 (± 0.014)
Survey responses (selected), wearable PGHD (threshold-based, selected)	Logistic regression	0.734 (± 0.008)	0.355 (± 0.022)	0.433 (± 0.019)	0.386 (± 0.015)
	XGBoost	0.887 (± 0.006)	0.463 (± 0.020)	0.652 (± 0.015)	0.540 (± 0.013)

Appendix G

Phase 2 experiments

The following table summarizes the performance of various phase 2 classification algorithms for a range of numbers of selected features in forward sequential feature selection, from 5 to 30 selected features, using randomized splitting. The best performing models were selected according to the primary performance metric — sensitivity. We also compute specificity, AUROC, precision and AUPRC, where specificity and AUPRC are secondary performance metrics. AUPRC values were not computed for every experiment, as this was added as a secondary performance metric as the experiments were being conducted. AUPRC was added because we found it to be more appropriate than AUROC due to class imbalance [58]. Thus, AURPC was computed for the best performing models, to help select the final best performing models, based on the primary metric. We did not repeat experiments to compute AUPRC for models where the sensitivity was significantly lower in comparison.

Model	**Sensitivity (±95 CI)**	**AUROC (±95 CI)**	**Precision (±95 CI)**	**Specificity (±95 CI)**	**AUPRC (±95 CI)**	**Number of selected features**
Logistic Regression	0.780 (± 0.029)	0.841 (± 0.016)	0.440 (± 0.017)	0.740 (± 0.009)	-	5
LGBM (GBDT)	0.834 (± 0.016)	0.906 (± 0.006)	0.544 (± 0.012)	0.817 (± 0.010)	-	5
LGBM (DART)	0.857 (± 0.028)	0.911 (± 0.008)	0.540 (± 0.016)	0.808 (± 0.015)	-	5
XGBoost	0.792 (± 0.034)	0.896 (± 0.005)	0.551 (± 0.009)	0.831 (± 0.014)	-	5
Random Forest	0.733 (± 0.012)	0.854 (± 0.009)	0.607 (± 0.030)	0.875 (± 0.016)	0.644 (± 0.020)	5
Logistic Regression	0.799 (± 0.020)	0.850 (± 0.013)	0.470 (± 0.011)	0.764 (± 0.011)	0.609 (± 0.035)	11
LGBM (GBDT)	0.832 (± 0.022)	0.914 (± 0.014)	0.560 (± 0.016)	0.828 (± 0.012)	-	11
LGBM (DART)	0.879 (± 0.015)	0.922 (± 0.010)	0.547 (± 0.019)	0.809 (± 0.014)	-	11
XGBoost	0.778 (± 0.034)	0.907 (± 0.009)	0.579 (± 0.012)	0.852 (± 0.007)	-	11
Random Forest	0.657 (± 0.048)	0.859 (± 0.044)	0.591 (± 0.058)	0.880 (± 0.027)	-	11
Logistic Regression	0.791 (± 0.020)	0.850 (± 0.013)	0.469 (± 0.012)	0.766 (± 0.013)	-	13
LGBM (GBDT)	0.840 (± 0.024)	0.914 (± 0.012)	0.565 (± 0.018)	0.830 (± 0.009)	0.723 (± 0.041)	13
LGBM (DART)	0.881 (± 0.021)	0.922 (± 0.010)	0.544 (± 0.019)	0.807 (± 0.015)	-	13
XGBoost	0.781 (± 0.025)	0.907 (± 0.007)	0.579 (± 0.008)	0.851 (± 0.007)	0.705 (± 0.027)	13
Random Forest	0.640 (± 0.039)	0.876 (± 0.039)	0.610 (± 0.028)	0.893 (± 0.016)	-	13
LGBM (DART)	0.882 (± 0.018)	0.922 (± 0.010)	0.546 (± 0.016)	0.808 (± 0.014)	0.739 (± 0.043)	14
Logistic Regression	0.791 (± 0.019)	0.850 (± 0.014)	0.469 (± 0.014)	0.766 (± 0.011)	-	15
LGBM (GBDT)	0.833 (± 0.027)	0.915 (± 0.014)	0.568 (± 0.026)	0.834 (± 0.015)	0.722 (± 0.044)	15
LGBM (DART)	0.878 (± 0.015)	0.922 (± 0.010)	0.544 (± 0.017)	0.807 (± 0.014)	0.742 (± 0.042)	15
XGBoost	0.769 (± 0.026)	0.907 (± 0.010)	0.580 (± 0.016)	0.854 (± 0.007)	-	15
Random Forest	0.633 (± 0.037)	0.881 (± 0.044)	0.614 (± 0.036)	0.895 (± 0.017)	-	15
LGBM (DART)	0.882 (± 0.018)	0.922 (± 0.010)	0.545 (± 0.017)	0.807 (± 0.013)	0.740 (± 0.041)	16
Logistic Regression	0.792 (± 0.024)	0.849 (± 0.014)	0.466 (± 0.017)	0.762 (± 0.013)	-	17
LGBM (GBDT)	0.838 (± 0.019)	0.915 (± 0.012)	0.564 (± 0.011)	0.831 (± 0.009)	0.725 (± 0.036)	17
LGBM (DART)	0.880 (± 0.009)	0.921 (± 0.010)	0.547 (± 0.014)	0.809 (± 0.011)	0.738 (± 0.043)	17
XGBoost	0.770 (± 0.022)	0.906 (± 0.007)	0.587 (± 0.011)	0.858 (± 0.007)	-	17
Random Forest	0.615 (± 0.018)	0.889 (± 0.040)	0.608 (± 0.052)	0.896 (± 0.021)	-	17
Logistic Regression	0.792 (± 0.033)	0.849 (± 0.014)	0.472 (± 0.017)	0.768 (± 0.012)	-	30
LGBM (GBDT)	0.826 (± 0.017)	0.917 (± 0.012)	0.576 (± 0.014)	0.841 (± 0.008)	-	30
LGBM (DART)	0.874 (± 0.018)	0.922 (± 0.009)	0.553 (± 0.017)	0.815 (± 0.011)	-	30
XGBoost	0.756 (± 0.019)	0.906 (± 0.007)	0.605 (± 0.014)	0.871 (± 0.008)	-	30
Random Forest	0.561 (± 0.040)	0.898 (± 0.009)	0.661 (± 0.048)	0.924 (± 0.016)	-	30

Appendix H

Phase 2 participant-based splitting experiments

The following table summarizes the results obtained when performing phase 2 model construction, including feature selection, model fitting and testing of the performance using participant-based splitting. The classification algorithm used was LightGBM DART.

Sensitivity (±95 CI)	**Specificity (±95 CI)**	**AUPRC (±95 CI)**	**Number of features selected**
0.873 (± 0.017)	0.831 (± 0.020)	0.751 (± 0.044)	13
0.872 (± 0.018)	0.831 (± 0.019)	0.753 (± 0.042)	14
0.870 (± 0.021)	0.831 (± 0.019)	0.753 (± 0.042)	15
0.870 (± 0.021)	0.830 (± 0.019)	0.754 (± 0.042)	16
0.870 (± 0.021)	0.830 (± 0.019)	0.754 (± 0.042)	17

Appendix I

Phase 1c selected features

The following list contains the PGHD features that were selected as the most important over various train:test splits of the phase 1c model in the combined pipeline:

- Sex
- Race/ethnicity (White, Black, Asian, Hispanic)
- Birth year
- Education level
- Height, in inches
- Weight, in lbs
- BMI
- Pregnant
- Experienced trauma resulting in injury (e.g. accident, fall) in the past year
- Has health insurance
- Has had sufficient money in the past month to meet basic living needs over the past month
- Has received money or assistance from the government (e.g. food stamps, disability payments) in the past month
- Number of individuals in the household
- Diagnosed comorbidities:
  - Cancer
  - Type 1 diabetes
  - Gout
  - Migraines
  - Osteoporosis

  - Fibromyalgia, peripheral neuropathic pain or central neuropathic pain
  - Osteoarthritis or rheumatoid arthritis

- For participants with migraines:
  - Number of migraine days per month, on average
  - Takes daily prescription migraine medication
- Started a new medication
- Stopped a medication
- Changed the dosage of one of my medications
- Does not take any medications and do not use any non-medication wellness therapies
- Began meditation or other relaxation techniques
- Number of days the participant has slept over 10 hours in the past 7 days
- Range of the sleep start hour over the past 14 days
- Mean ratio of the time spent asleep to the time spent in bed over the past 4 days