Prediction of self-reported depression scores using Person-Generated Health Data from a Virtual 1-Year Mental Health Observational Study

Mariko Makhmutova
Evidation Health
San Mateo, CA, USA
EPFL
Lausanne, Switzerland
mariko.makhmutova@epfl.ch

Raghu Kainkaryam
Evidation Health
San Mateo, CA, USA
rkainkaryam@evidation.com

Marta Ferreira
Evidation Health
San Mateo, CA, USA
mferreira@evidation.com

Jae Min
Evidation Health
San Mateo, CA, USA
jmin@evidation.com

Martin Jaggi
EPFL
Lausanne, Switzerland
martin.jaggi@epfl.ch

Ieuan Clay
Evidation Health
San Mateo, CA, USA
iclay@evidation.com

ABSTRACT

Background: Undiagnosed mental illnesses represent one of the biggest challenges in our society. Due to stigma surrounding mental health, many people experience symptoms years before diagnosis and often never receive active management. ***Objectives:*** We use person-generated health data, consisting of self-reports and data from consumer wearable devices to predict an individual's depression severity level. ***Methods:*** Reference labels and input feature sets were derived from a 1-year long longitudinal cohort study consisting of 10,036 individuals. Participant-reported PHQ-9 scores were used as reference labels for depression severity, and input feature sets consisted of self-reported socio-demographic information, lifestyle and medication change surveys, and objective behavioral data collected using consumer wearables. ***Results:*** Our best performing model achieved an adjacent accuracy of 0.889 (CI ±0.006) and a Kappa of 0.655 (CI ±0.015). We observe that socio-demographic features contribute strongly to model performance, and that although good performance can be achieved with self-reported features, the addition of a small number of threshold-based features, derived from objective wearable data, improves model robustness. ***Conclusions:*** To our knowledge, the presented classification model is developed using the largest longitudinal cohort study ever considered for depression diagnosis, and one of the first attempts to predict granular depression severity, beyond binary classification of depressed individuals versus healthy controls. We demonstrate the feasibility of our approach for this non-trivial problem. Future work will focus on combining the output labels of this model with self-reports in order to attempt to predict changes in individual, longitudinal mental health status.

CCS CONCEPTS

• **Applied computing → Life and medical sciences**; **Consumer health**; • **Computing methodologies → Machine learning**.

KEYWORDS

datasets, self-reported depression, person-generated health data, machine learning

ACM Reference Format:
Mariko Makhmutova, Raghu Kainkaryam, Marta Ferreira, Jae Min, Martin Jaggi, and Ieuan Clay. 2021. Prediction of self-reported depression scores using Person-Generated Health Data from a Virtual 1-Year Mental Health Observational Study. In *Future of Digital Biomarkers (DigiBiom'21), June 25, 2021, Virtual, WI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3469266.3469878

DigiBiom'21, June 25, 2021, Virtual, WI, USA

ACM ISBN 978-1-4503-8606-7/21/06…$15.00
https://doi.org/10.1145/3469266.3469878

1 INTRODUCTION

Major depressive disorder (MDD) is a leading cause of disability worldwide, with more than 264 million people worldwide suffering from this mental illness, a problem further exacerbated by the COVID-19 pandemic, according to the World Health Organization [18, 19]. In 2020, 19% of American adults reported worsened mental health compared to the same time the year before, likely linked to the pandemic [1]. In 2017, an estimated 17.3 million adults in the US experienced at least one major depressive episode in the last year, with 35% of them not receiving any treatment [17].

Under-diagnosis of depression has been attributed to many reasons including stigma surrounding mental health, limited access to medical care or barriers due to cost [4]. In addition to impacting the quality of life of many individuals, undiagnosed and untreated depression has significant economic consequences, adding an economic burden of over $200 billion annually in the US alone [9].

Thus, it is essential to make the detection and monitoring of depression symptoms easier and more affordable. An increasingly explored and promising way to accomplish this is through the use of person-generated health data (PGHD) in the form of self-reports, geolocation, consumer wearables, biometrics, exercise and social interaction data [6, 20].

Multiple studies have shown that PGHD can provide an early indicator of changes in mental health via social media use [3, 24] or physical activity patterns [21]. For example, a recent study used consumer wearable devices to track the sleep of 368 participants and used a linear regression approach to find associations between sleep features and self-reported depression [25], identifying several individual features with strong associations (Z-scores up to 6.19). Another study showed that activity features collected over a two-week period for 23 participants could accurately (Kappa=0.773) classify individuals with depression from controls, and predict changes in depressive status over the 2 week period [11]. These studies are limited in sample size and time duration to generalize across larger populations, yet demonstrate the potential of this approach versus more burdensome active assessments [22].

In this work, we present a classification model that leverages socio-demographic data, medical history, and PGHD derived from consumer-grade wearables, in order to predict an individual's depression severity level. To our knowledge, the presented model is developed using the largest 1-year long longitudinal cohort study ever considered for depression diagnosis using wearable PGHD at the time of publication [14].

Acquiring PGHD on a large scale requires a low-burden data collection approach. However, this produces sparse data, as participants are not obliged to provide data at frequent intervals. Consequently, we are limited to a relatively small set of reference labels: on average 2.07 labels per enrolled participant over the course of one year. Thus, this work will focus on generation of more frequent intermediate prediction labels which can be used in combination with reference labels to reduce the sparsity of the dataset, and support our ultimate objective of creating generalizable models that predict changes in mental health status on an individual level.

2 METHODS

2.1 Data collection

The data used in this work are part of the DiSCover Project developed by Evidation Health ([7]; ClinicalTrials.gov ID: NCT03421223). The DiSCover Project is a 1-year long longitudinal study consisting of 10,036 individuals who wore consumer wearable devices throughout the study and regularly completed surveys about their mental health and lifestyle changes.

More specifically, the data subset used in this work comprises the following:

- *Wearable PGHD*: 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 complete a screening/baseline survey regarding socio-demographic information, as well as comorbidities
- *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, starting new medication)
- *Patient Health Questionnaire (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 [13], in order to track changes in their mental health status

Figure 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.

As our goal is to predict depression severity levels, we used the five predefined PHQ-9 score categories, ranging from minimal to severe [13]. The categories are described in the following section. As the PHQ-9 score aims to summarize depression severity over the past two weeks, we only considered wearable PGHD for participants over the 14 days prior to the PHQ-9 completion date.

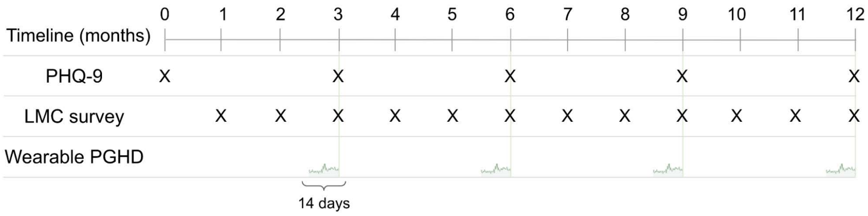

Figure 1: Data collection timeline. Crosses represent data collection for a given month and data type. For each participant, we extract the PHQ-9 score for a given quarter, the same month's lifestyle and medication (LMC) survey responses, and wearable PGHD collected in the 14 days prior to the PHQ-9 completion date.

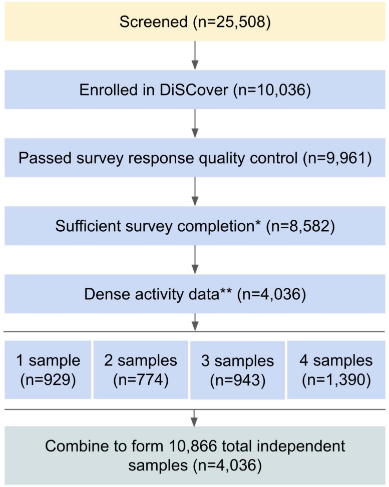
Screened (n=25,508)
Enrolled in DiSCover (n=10,036)
Passed survey response quality control (n=9,961)
Sufficient survey completion* (n=8,582)
Dense activity data** (n=4,036)
1 sample (n=929)
2 samples (n=774)
3 samples (n=943)
4 samples (n=1,390)
Combine to form 10,866 total independent samples (n=4,036)

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

2.2 Data processing

2.2.1 Data filtering process. The participant filtering process is illustrated in Figure 2. 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, participants were filtered based on their survey completion: we kept participants who completed the PHQ-9 for two contiguous quarters, as well as the LMC survey for the same month as the second PHQ-9. The selection criteria are illustrated in Figure 3. These requirements were set as we aim to study the evolution of mental health status, and recent lifestyle changes, such as changes in medication, are important factors in mental health.

Next, the participants were filtered based on the density of their available activity data. For each quarter, as defined by the PHQ-9 completion dates, we considered activity data in the two weeks prior to this completion date. We filtered participants based on whether they had a sufficient amount of activity data in these two weeks, according to standards proposed in literature [15, 23]. Participants were kept if they had at least 4 valid days of activity data in each of the two weeks. A day was considered "valid" if the participant has worn the device for at least 10 hours that day.

Each sample in our dataset was 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. At this stage, we had samples from 4,036 unique participants, with 929 of the participants providing only one sample. Initial data exploration showed that the overall evolution of PHQ-9 scores was stable throughout the year when grouping by demographic variables, such as sex, age, race, and geographic location. Based on this observation, we treated each of the samples (scores) from the participants as independent from each other. Thus, we obtained a total of 10,866 samples from 4,036 unique participants.

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

2.2.2 Survey responses. We split the survey response data into two categories: static and variable features. Features originating from the screener survey were considered as static as changes in these features were not tracked throughout the study. This included features such as sex, race and comorbidities.

Lifestyle and medication changes (LMC) features were considered as variable, as we observed self-reported monthly changes in lifestyle and medication throughout the study. LMC features included changes in medication dosage or alcohol consumption habits in the past month.

2.2.3 Wearable PGHD. Wearable PGHD was collected using participants' Fitbit devices throughout the study. In this work,

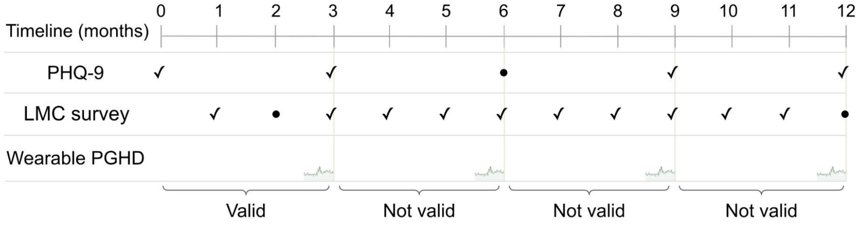

Figure 3: Illustration of the survey completion criteria. Check marks indicate completed surveys, dots indicate non-completed surveys. The first quarter (months 0-3) fulfills the selection criteria: the initial (month 0) and final (month 3) PHQ-9 were completed, and the LMC survey was completed for month 3. The second quarter is not valid, as the final PHQ-9 (month 6) was not completed. The third quarter is not valid as the initial PHQ-9 (month 6) was not completed. The fourth quarter is not valid as the final month LMC survey was not completed.

we focused specifically on the impact of trends in step and sleep on depression severity.

Step and sleep data, initially provided at a minute-level granularity, was aggregated at day-level. Channels taken into consideration include the number of steps taken, number of active minutes (determined by step count per minute), number of minutes slept, number of minutes spent in bed, and the sleep efficiency score.

Based on these channels, we used three different approaches to generate feature sets. The first set was based on general statistical trends aggregated over three time windows, relative to PHQ-9 completion date (most recent): 4, 7, and 14 days. The statistical trends considered include mean, median, IQR, and range.

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

The third approach to constructing features from day-level wearable data 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.

PHQ-9 score	Depression severity	Number of samples	% 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 1: Description of the PHQ-9 score categories and the dataset category distribution

2.2.4 PHQ-9 score categories. We aggregated PHQ-9 scores to the five predefined categories used to measure depression severity. As the study cohort was not screened based on their depression severity, the distribution of depression levels was imbalanced, as expected in a random population sample. Table 1 presents PHQ-9 score categories and the category distribution in our final dataset.

2.3 Modeling

The goal of this work is to develop a model that correctly predicts participants' PHQ-9 score categories from socio-demographic, medical and wearable PGHD.

A common problem with wearable PGHD is inconsistent and missing data. Participants may choose not to wear their devices for various reasons, so we do not want to impose assumptions on their activity through imputation. This decision motivated our classification algorithm selection – the Extreme Gradient Boosting (XGBoost) algorithm [5]. XGBoost efficiently constructs ensembles of decision trees, and it is able to handle sparse data. We used a logistic regression model with zero imputation as the baseline to compare to the performance of XGBoost.

As identified in the previous section, we are posed with an imbalanced multi-label classification problem. To mitigate the effects of imbalanced classes, we performed sample stratification during training, hyperparameter tuning, and testing, so as to optimize performance across all categories. In addition to this, we used XGBoost's sampled weighting functionality 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

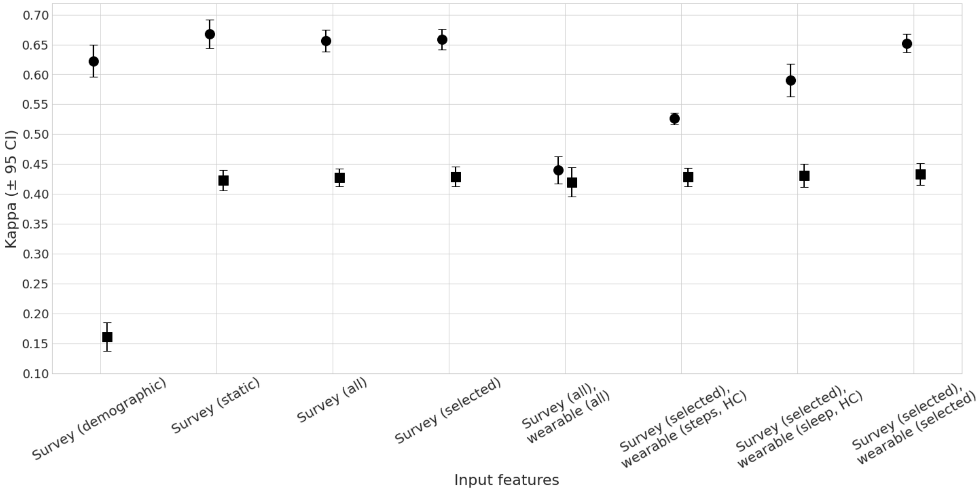

Figure 4: Performance comparison of models predicting PHQ-9 score category using various input feature sets, with 95% confidence intervals. Circular markers represent XGBoost models, and square markers represent logistic regression models.

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 2: Performance comparison of the best predictive XGBoost models after extended hyperparameter tuning, with 95% confidence intervals

feature elimination and cross-validated selection (RFECV; [10]) in order to eliminate features that had lower contributions to model performance.

Model performance was primarily measured using quadratic weighted Cohen's Kappa. Weighted Kappa is used to calculate the level of agreement between the predicted and target values, using a distance-based weight penalty [8]. Quadratic weighted Cohen's Kappa uses a quadratic penalty in lieu of the traditional linear penalty based on distance from the target value. The Kappa score ranges from -1 to 1, with 1 signifying full agreement, and -1 signifying maximal distance between predicted and target values.

We also use adjacent accuracy (i.e. fraction of samples predicted at most one off from the target value), balanced accuracy and weighted F1-scores as secondary performance metrics.

2.4 Hyperparameter tuning

We performed randomized search 5-fold cross validation 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 discuss the performance of the tuned models in the following section, and we report the performance metrics of the best tuned models with 95% confidence intervals across 5 training runs (5 outer shuffle splits).

3 RESULTS

In this section we present the performance of baseline models compared to tuned XGBoost classification models, for various input feature sets with a confidence interval (CI) of 95% for each model.

We observed 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 feature elimination on the remaining features, based on model performance, we obtained a final set of selected survey response features, used in further testing.

When performing feature selection on wearable PGHD features, we observed similar model performance when using RFECV and when selecting a small subset of features that are highly correlated with the PHQ-9 score categories based on the training set. To reduce noise from additional features, we proceeded using only these highly correlated (HC) features when fitting models with both survey and wearable data.

We performed an initial set of experiments using survey and wearable data separately. Based on the results, we proceeded with the best selected subset of survey response features in combination with wearable features to find the optimal model to predict the PHQ-9 score category. We present the summarized results in Figure 4, with a 95% CI.

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

The survey responses (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.

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 reduces the CI. We observe a comparable performance for the model that uses the same survey response features in combination with wearable features, but with an even narrower 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 2.

Table 2 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, it 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.

The average category accuracy is also displayed in Figure 5. We note that although the model performance is significantly better for the more prevalent categories representative of milder depression levels, the adjacent accuracy is consistently high across all classes. This implies that 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 level.

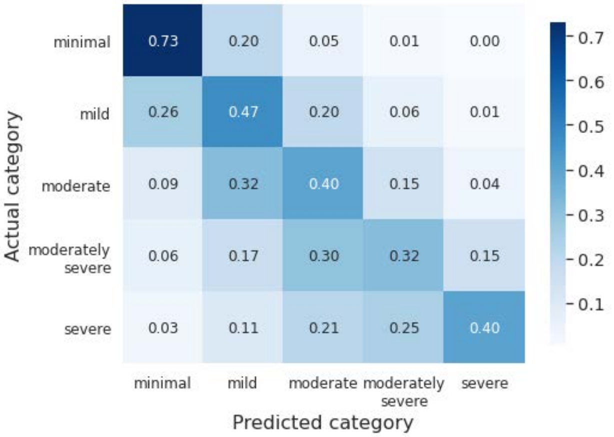

Figure 5: Confusion matrix showing the best model's PHQ-9 score category accuracy distribution.

4 DISCUSSION

Using survey responses and wearable PGHD from a study cohort of 10,036 participants, we were able to construct a robust model that predicts depression levels. Despite the sparsity of the survey response data and missingness of wearable data, the model is able to effectively predict the five PHQ-9 score categories with a quadratic weighted Cohen's Kappa score of 0.655. We also achieved an adjacent accuracy of 89%, meaning that participants are almost always predicted to be at most one category off from their true self-reported PHQ-9 score category.

Initial experiments showed that we can obtain a decent model with Kappa=0.623 using only basic demographic features, such as sex, age, pregnancy and BMI. However, the addition of socioeconomic features and medical history render the model more robust and give an overall better performance, with Kappa=0.661 after extensive hyperparameter tuning. Adding wearable features further improves the robustness of the model, narrowing the 95% CI.

We are able to conclude that demographic attributes are crucial to evaluate depression levels. Additionally, sociodemographics and medical history are important considerations for mental health status, and adding behavioural data in the form of wearable PGHD features adds more robustness to predictions.

Our results are consistent with a recent study that used consumer wearable devices to track participants' sleep, exploring associations between sleep features and depression [25]. The study found that the percentage of nights with hypersomnia is significantly associated with PHQ-8 scores. Likewise, we found the number of hypersomnia nights in the past 7 days to be one of the two most predictive wearable features for the PHQ-9 score category. The study adjusted models for socio-demographic factors, in the form of sex, age, education level and income. The focus of the study was directed towards more elaborate sleep features, such as the amount of Rapid Eye Movement (REM) sleep participants had. We did not have this data available at the time of study, but this could be further explored in a future work.

Another recent study used mixed linear models in an attempt to predict depression and anxiety levels from smartphone and wearable data [16]. This study also found that there is a significant relationship between total sleep time and depression, and no significant association between physical activity measures and depression.

Multiple studies have attempted to use consumer wearable data, in combination with socio-demographic data to study patterns in depression levels. However, to our knowledge, this is the first work to use machine learning to predict depression levels beyond identifying controls from depressed participants.

We acknowledge two important limitations of this work. The first limitation is that we are using self-reported PHQ-9 scores, as well as self-reported lifestyle and medication changes. Self-reported results can be highly biased [2]; thus, at the cost of non-invasive observation, we are faced with the problem of potentially inaccurate reference labels that we use to fit our models.

The second limitation of our work is in the diversity of the cohort population. As with many studies that focus on using consumer wearable devices for the analysis of mental health [16, 25], our cohort comprises a majority of Non-Hispanic White and female participants. This demographic bias could also be worsened by the initial assumption that scores from different participants are independent. Nevertheless, based on a preliminary error analysis, we observe no significant difference in model performance when comparing across various demographic attributes.

The long term aim of our work is to predict changes in mental health status, but the sparsity of labeling in our dataset has posed challenges for the task. We therefore decided to implement an intermediate weak labeling step – the outputs generated by the model presented in the current work – which, in a future work, will allow us to make inferences about individuals' mental health statuses without rendering the frequency of mental health questionnaires too frequent.

5 OUTLOOK

We likely underestimate the burden of mental health due to a lack of reporting caused by stigmatization, poor mental health literacy and other societal or environmental factors. Tools which enable patient activation through low burden and unobtrusive monitoring can be a part of the solution. The work presented here is a first step in developing proof of principle for a PGHD-based warning or alert system which could be used to tailor the delivery of behavioral interventions [12]. This might just start with simply asking someone who appears to be in a depressive state 'are you ok?'.

ACKNOWLEDGMENTS

The authors would like to thank Nikki Marinsek, Allison Shapiro, and Piyusha Gade for their advice, as well as Raul Eulogio and Filip Jankovic for technical input. The authors would also like to thank the study participants, and the DiSCover Project team for allowing us to work with the participant datasets.

6 ETHICS STATEMENT

Data used in this work was derived from the Digital Signals in Chronic Pain (DiSCover) Project ([7]; ClinicalTrials.gov identifier: NCT03421223). This study received expedited review and IRB approval from WCG IRB (IRB Study #: 1181760; Protocol #: 20172916; Initial Approved Date: December 21, 2017).

REFERENCES

[1] American Psychological Association. 2020. *Stress in America™ 2020.* Technical Report.

[2] Philip S Brenner and John DeLamater. 2016. Lies, Damned Lies, and Survey Self-Reports? Identity as a Cause of Measurement Bias. *Soc. Psychol. Q.* 79, 4 (Dec. 2016), 333–354.

[3] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digit Med* 3 (March 2020), 43.

[4] Adam M Chekroud, David Foster, Amanda B Zheutlin, Danielle M Gerhard, Brita Roy, Nikolaos Koutsouleris, Abhishek Chandra, Michelle Degli Esposti, Girish Subramanyan, Ralitza Gueorguieva, Martin Paulus, and John H Krystal. 2018. Predicting Barriers to Treatment for Depression in a U.S. National Sample: A Cross-Sectional, Proof-of-Concept Study. *Psychiatr. Serv.* 69, 8 (Aug. 2018), 927–934.

[5] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[6] Francesca Cormack, Maggie McCue, Nick Taptiklis, Caroline Skirrow, Emilie Glazer, Elli Panagopoulos, Tempest A van Schaik, Ben Fehnert, James King, and Jennifer H Barnett. 2019. Wearable Technology for High-Frequency Cognitive and Mood Assessment in Major Depressive Disorder: Longitudinal Observational Study. *JMIR Ment Health* 6, 11 (Nov. 2019), e12814.

[7] Evidation Health. 2019. Evidation Health's DiSCover Program Completes Initial Enrollment, Releases Data from Largest U.S. Study of Chronic Pain. https://evidation.com/news/evidation-healths-discover-program-completes-initial-enrollment-releases-data-from-largest-u-s-study-of-chronic-pain/. Accessed: 2021-6-2.

[8] Joseph L Fleiss and Jacob Cohen. 1973. The Equivalence of Weighted Kappa and the Intraclass Correlation Coefficient as Measures of Reliability. *Educ. Psychol. Meas.* 33, 3 (Oct. 1973), 613–619.

[9] Paul E Greenberg, Andree-Anne Fournier, Tammy Sisitsky, Crystal T Pike, and Ronald C Kessler. 2015. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *J. Clin. Psychiatry* 76, 2 (Feb. 2015), 155–162.

[10] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. 2002. Gene Selection for Cancer Classification using Support Vector Machines. *Mach. Learn.* 46, 1 (Jan. 2002), 389–422.

[11] Nicholas C Jacobson, Hilary Weingarden, and Sabine Wilhelm. 2019. Digital biomarkers of mood disorders and symptom change. *NPJ Digit Med* 2 (Feb. 2019), 3.

[12] Jan-Niklas Kramer, Florian Künzler, Varun Mishra, Bastien Presset, David Kotz, Shawna Smith, Urte Scholz, and Tobias Kowatsch. 2019. Investigating Intervention Components and Exploring States of Receptivity for a Smartphone App to Promote Physical Activity: Protocol of a Microrandomized Trial. *JMIR Res. Protoc.* 8, 1 (Jan. 2019), e11540.

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

[14] Shefali Kumar, Jennifer L A Tran, Ernesto Ramirez, Wei-Nchih Lee, Luca Foschini, and Jessie L Juusola. 2020. Design, Recruitment, and Baseline Characteristics of a Virtual 1-Year Mental Health Study on Behavioral Data and Health Outcomes: Observational Study. *JMIR Ment Health* 7, 7 (July 2020), e17075.

[15] Jairo H Migueles, Cristina Cadenas-Sanchez, Ulf Ekelund, Christine Delisle Nyström, Jose Mora-Gonzalez, Marie Löf, Idoia Labayen, Jonatan R Ruiz, and Francisco B Ortega. 2017. Accelerometer Data Collection and Processing Criteria to Assess Physical Activity and Other Outcomes: A Systematic Review and Practical Considerations. *Sports Med.* 47, 9 (Sept. 2017), 1821–1845.

[16] Isaac Moshe, Yannik Terhorst, Kennedy Opoku Asare, Lasse Bosse Sander, Denzil Ferreira, Harald Baumeister, David C Mohr, and Laura Pulkki-Råback. 2021. Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. *Front. Psychiatry* 12 (Jan. 2021), 625247.

[17] National Institute of Mental Health. 2019. Major Depression. https://www.nimh.nih.gov/health/statistics/major-depression.shtml. Accessed: 2021-4-22.

[18] World Health Organization. 2020. COVID-19 disrupting mental health services in most countries, WHO survey. https://www.who.int/news/item/05-10-2020-covid-19-disrupting-mental-health-services-in-most-countries-who-survey

[19] World Health Organization. 2020. Depression. https://www.who.int/news-room/fact-sheets/detail/depression. Accessed: 2021-4-22.

[20] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F Ionescu, Darian Bhathena, Lauren B Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, Lisa Sangermano, David Mischoulon, Johnathan E Alpert, and Rosalind W Picard. 2020. Monitoring Changes in Depression Severity Using Wearable and Mobile Sensors. *Front. Psychiatry* 11 (Dec. 2020), 584711.

[21] Brenna N Renn, Abhishek Pratap, David C Atkins, Sean D Mooney, and Patricia A Areán. 2018. Smartphone-Based Passive Assessment of Mobility in Depression: Challenges and Opportunities. *Ment. Health Phys. Act.* 14 (March 2018), 136–139.

[22] Oleksandr Sverdlov, Jelena Curcic, Kristin Hannesdottir, Liangke Gou, Valeria Luca, Francesco Ambrosetti, Bingsong Zhang, Jens Praestgaard, Vanessa Vallejo, Andrew Dolman, Baltazar Gomez-Mancilla, Konstantinos Biliouris, Mark Deurinck, Jang-Ho Cha, Francesca Cormack, John Anderson, Nicholas Bott, Ziv Peremen, Gil Issachar, and Gabriel Jacobs. 2021. A Study of Novel Exploratory Tools, Digital Technologies, and Central Nervous System Biomarkers to Characterize Unipolar Depression. *Frontiers in Psychiatry* 12 (04 2021). https://doi.org/10.3389/fpsyt.2021.640741

[23] Catrine Tudor-Locke, Sarah M Camhi, and Richard P Troiano. 2012. A catalog of rules, variables, and definitions applied to accelerometer data in the National Health and Nutrition Examination Survey, 2003-2006. *Prev. Chronic Dis.* 9 (June 2012), E113.

[24] Zahra Vahedi and Lesley Zannella. 2019. The association between self-reported depressive symptoms and the use of social networking sites (SNS): A meta-analysis. *Current Psychology* (26 Jan 2019). https://doi.org/10.1007/s12144-019-0150-6

[25] Yuezhou Zhang, Amos A Folarin, Shaoxiong Sun, Nicholas Cummins, Rebecca Bendayan, Yatharth Ranjan, Zulqarnain Rashid, Pauline Conde, Callum Stewart, Petroula Laiou, Faith Matcham, Katie M White, Femke Lamers, Sara Siddi, Sara Simblett, Inez Myin-Germeys, Aki Rintala, Til Wykes, Josep Maria Haro, Brenda Wjh Penninx, Vaibhav A Narayan, Matthew Hotopf, Richard Jb Dobson, and RADAR-CNS Consortium. 2021. Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study. *JMIR Mhealth Uhealth* 9, 4 (April 2021), e24604.