Associations between depression symptom severity and individuals' behaviors measured by smartphones and wearable devices

Zhang, Yuezhou

Awarding institution:
King's College London

Associations Between Depression Symptom Severity and Individuals' Behaviors Measured by Smartphones and Wearable Devices

Yuezhou Zhang

Department of Biostatistics & Health Informatics
Institute of Psychiatry, Psychology and Neuroscience
King's College London

A thesis submitted for the degree of
Doctor of Philosophy,
November 2022

Abstract

Depression is a prevalent and severe mental health disorder that is one of the leading causes of disability worldwide. It can cause various physical and psychological problems, leading to loss of productivity, increased social burden, and even suicide. The current diagnosis of depression relies on skilled clinicians and self-reported questionnaires, which have limitations including subjective recall bias and loss of day-to-day fluctuation information. As a result, the majority of individuals with depression did not receive timely and effective treatment. Therefore, there is a need for more effective auxiliary techniques for recognizing and monitoring depression.

With the development and widespread use of sensors, mobile technology provides a cost-effective and convenient means for gathering individuals' behavioral data related to depression symptoms. Several past studies have attempted to monitor depression using mobile phones and wearable devices. However, the majority of these studies were conducted on relatively small and homogeneous cohorts with short follow-up periods, which may have limited the generalizability of their findings. Furthermore, the impact of participant attrition and engagement, the direction of relationships over time, and individual differences need further exploration.

To address these limitations, this thesis extracts a variety of behavioral features from multiple data streams of mobile phone and wearable data and explores their associations with depression symptom severity using a large, longitudinal, multi-center data set. Specifically, Chapter 1 provides an overview of the background of depression, motivations for using mobile technology for depression monitoring, and existing related studies.

Chapter 2 performs a novel investigation into long-term participant retention and engagement from a European longitudinal observational program, the RADAR-MDD study, which is used throughout the whole thesis. A significantly higher participant retention rate is found in the RADAR-MDD study than in previous remote digital health studies. According to the data-driven method, lower participant engagement is found to be associated with higher depression symptom severity, younger age, and longer questionnaire response/completion time in the study app. Finally, the strategies for increasing participant engagement in future digital health research are also discussed in this chapter.

Next, the associations between depression symptom severity and various categories of behaviors are explored separately in the following chapters: sleep (Chapter 3), sociability as measured by Bluetooth device counts (Chapter 4), mobility (Chapter 5), daily walking (Chapter 6), and circadian rhythms (Chapter 7). These associations are examined using multilevel models that incorporate demographics as between-participant covariates. A number of significant associations between behavioral characteristics and depression symptom severity are found in these chapters. For example, higher depression severity is significantly associated with worse sleep, lower sociability, lower mobility, slower cadence of daily walking, and weaker circadian rhythmicity. Notably, the longitudinal association between mobility and depression over time is assessed using dynamic structural equation models in Chapter 5. Changes in several mobility features are found to significantly affect subsequent changes in depression severity. Furthermore, daily-life gait patterns are found to provide extra information for recognizing depression relative to laboratory gait patterns in Chapter 6.

Taken together, the findings in this thesis demonstrate that depression is closely associated with individuals' daily-life behaviors, which can be captured by mobile technology in real-world settings. Despite challenges of data quality and participant attrition, the evidence may provide support for the development of future clinical tools to passively monitor mental health status and trajectory with minimal burden on the participant.

Author Declaration

I, Yuezhou Zhang, declare that this thesis is my own work. Chapters 2–6 of this thesis consist of four publications and one preprint paper (under peer review). I am the first author of these five papers and did all the data analysis, coding, and writing up work. Other coauthors' contributions include data collection, platform development, data storage and maintenance, analysis planning, and critical review of manuscripts, which are stated in the “Authors’ Contributions” sections of related chapters in detail.

Acknowledgements

First and foremost, I would like to express my deepest thanks to my supervisors, Prof. Richard Dobson and Dr. Amos Folarin. I always feel lucky under their supervision. Despite their hectic schedules, they always made time to discuss the research plans with me, provide suggestions for problems I encountered, and critically review my papers. I enjoyed and benefited greatly from every regular meeting with them. During these years, they taught me creative thinking, academic writing, and team collaboration skills that will benefit me for the rest of my life.

I am especially grateful to my colleague and friend, Shaoxiong Sun. He gave me lots of advice about my research and personal life during the past 3 years. When I didn’t know much about academic writing in the first year of my PhD, he gently gave me lots of suggestions on my manuscripts. My first paper could not have been published without his help.

Next, I want to thank all members of our mobile health group, Yatharth Ranjan, Zulqarnain Rashid, Pauline Conde, Callum Stewart, Petroula Laiou, Heet Sankesara, and Xi Bai. It's been a pleasure to work with this group over the past 3 years. I'd also like to thank our group for their great work on the RADAR-base platform, which provides a means to gather active and passive data from participants of digital health studies. My thesis would not be finished without the data gathered by this platform.

I want to thank my postgraduate coordinator, Daniel Stahl. When I contacted him 3 years ago for a potential PhD opportunity, he thought I would be a better fit for the mobile health group and kindly introduced me to Richard and Amos. I also enjoyed the PGCert courses in “Applied Statistical Modelling and Health Informatics” organized by him, which helped me a lot and gave me lots of essential knowledge for my subsequent research.

I would also want to thank one of my collaborators, Abhishek Pratap, in particular. During our collaboration, I learned lots of skills for both data analysis and academic writing from him.

I want to express my gratitude for the amazing RADAR-CNS project, which undoubtedly made a significant contribution to the field of digital mental health research. I'd like to thank Prof. Matthew Hotopf in particular for his support and helpful comments on my papers. Also, I would like to thank all members of the RADAR-CNS consortium and participants for their contributions to this project.

I would also like to thank all my friends I met in London, as well as those who are linked remotely in China. Without their encouragement and company, I wouldn't be able to keep going through my hard times, especially during the COVID-19 pandemic.

Last but not least, I am extremely grateful to my father and mother, Guifa Zhang and Wenjun Yang, for their unconditional support and love. I also owe them an apology because there is an old Chinese saying, “父母在，不远游”, which means that “a son shouldn't go far away from home while his parents are still alive.” Since I am their only child, I was supposed to be at my parents’ side when they were old. But they made sacrifices to let me pursue my dream.

List of the Author’s Publications During the Period of This Thesis

First author publications:

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Bendayan, R., Ranjan, Y., … Dobson, R. J. B. (2021). Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study. *JMIR MHealth and UHealth, 9*(4).

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., … Dobson, R. J. B. (2021). Predicting depressive symptom severity through individuals’ nearby bluetooth device count data collected by mobile phones: Preliminary longitudinal study. *JMIR MHealth and UHealth, 9*(7).

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Vairavan, S., Bendayan, R., … Dobson, R. J. B. (2022). Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study. *JMIR Mental Health, 9*(3).

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Vairavan, S., Qian, L., … Dobson, R. J. B. (2022). Associations Between Depression Symptom Severity and Daily-Life Gait Characteristics Derived From Long-Term Acceleration Signals in Real-World Settings: Retrospective Analysis. *JMIR MHealth and UHealth, 10*(10).

Zhang, Y., Pratap, A., Folarin, A. A., Sun, S., Cummins, N., Matcham, F., … RADAR-CNS consortium. (2023). Long-term participant retention and engagement patterns in an app and wearable-based multinational remote digital depression study. *Npj Digital Medicine, 6*(1), 25.

Non-first author publications:

De Angel, V., Adeleye, F., Zhang, Y., Cummins, N., Munir, S., Lewis, S., Laporta Puyal, E., Matcham, F., Sun, S., Folarin, A. A., Ranjan, Y., Conde, P., Rashid, Z., Dobson, R., & Hotopf, M. (2023). The Feasibility of Implementing Remote Measurement Technologies in Psychological Treatment for Depression: Mixed Methods Study on Engagement. *JMIR Mental Health, 10,* e42866.

Laiou, P., Biondi, A., Bruno, E., Viana, P., Winston, J., Rashid, Z., Ranjan, Y., Conde, P., Stewart, C., Sun, S., Zhang, Y., Folarin, A., Dobson, R., Schulze-Bonhage, A., Dümpelmann, M., Richardson, M., & RADAR-CNS Consortium. (2022). Temporal Evolution of Multiday, Epileptic Functional Networks Prior to Seizure Occurrence. *Biomedicines, 10*(10), 2662.

Laiou, P., Kaliukhovich, D. A., Folarin, A. A., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Sun, S., Zhang, Y., Matcham, F., Ivan, A., Lavelle, G., Siddi, S., Lamers, F., Penninx, B. W. J. H., Haro, J. M., Annas, P., Cummins, N., Vairavan, S., … Hotopf, M. (2022). The Association between Home Stay and Symptom Severity in Major Depressive Disorder: Preliminary Findings from a

Multicenter Observational Study Using Geolocation Data from Smartphones. *JMIR MHealth and UHealth, 10*(1).

Sun, S., Folarin, A. A., Zhang, Y., Cummins, N., Liu, S., Stewart, C., Ranjan, Y., Rashid, Z., Conde, P., Laiou, P., Sankesara, H., Costa, G. D., Leocani, L., Sørensen, P. S., Magyari, M., Guerrero, A. I., Zabalza, A., Vairavan, S., Bailon, R., … Dobson, R. J. (2022). The utility of wearable devices in assessing ambulatory impairments of people with multiple sclerosis in free-living conditions. *Computer Methods and Programs in Biomedicine, 107204.*

List of Abbreviations

ACC	Acceleration
AIDS	Acquired Immunodeficiency Syndrome
ANS	Autonomic nervous system
CIBER	Centro de Investigación Biomédican en Red
CoxPH	Cox Proportional-Hazard
CR	Circadian rhythm
CRHR	Circadian rhythm in heart rate
DSEM	Dynamic structural equation modeling
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition
ECG	Electrocardiography
EDA	Electrodermal activity
EFPIA	European Federation of Pharmaceutical Industries and Associations
EU	European Union
FAST-R	Feasibility and Acceptability Support Team for Researchers
FD	Frequency domain
GDS-15	15-item Geriatric Depression Scale
GPBoost	Gaussian process boosting
GPS	Global Positioning System
HF	High frequency
HIV	Human Immunodeficiency Virus
HR	Hazard ratio (Chapter 2)
HR	Heart rate (Chapters 7)
HRV	Heart rate variability
IDS	Inventory of Depression Symptomatology
IDS-SR	Inventory of Depression Symptomatology – Self Report
IMI	Innovative Medicines Initiative
IQR	Interquartile range
KCL	King’s College London
LAO	Leave-all-out
LF	Low frequency
LOO	Leave-one-out
LR	Likelihood ratio
LSTM	Long short-term memory
LTMM	Long Term Movement Monitoring
MAC	Media Access Control
MF	Middle frequency

mHealth	Mobile health
MSE	Multiscale entropy
NBDC	Nearby Bluetooth device count
NHS	National Health Service
NIHR	National Institute for Health Research
NREM	Non-rapid eye movement
PHQ-8	8-item Patient Health Questionnaire
PPG	Photoplethysmography
PSD	Power spectral density
PSG	Polysomnography
PSQI	Pittsburgh Sleep Quality Index
RADAR-CNS	Remote Assessment of Disease and Relapse–Central Nervous System
RADAR-MDD	Remote Assessment of Disease and Relapse–Major Depressive Disorder
REDCap	Research electronic data capture
REM	Rapid eye movement
RMSE	Root mean squared error
RMT	Remote measurement technology
RSES	Rosenberg Self-Esteem Scale
SD	Standard deviation
SE	Standard error
SHAP	SHapley Additive exPlanations
SOL	Sleep onset latency
VUmc	Vrije Universiteit Medisch Centrum
WHO	World Health Organization
XGBoost	Extreme gradient boosting

Contents

Chapter 1
Introduction

Chapter 2

Chapter 1

Introduction

1.1 Overview

People with depressive symptoms frequently report abnormalities in daily behaviors, including sleep disturbances, diminished social connections, and decreased activity levels. With the advancement of mobile technology, individuals' daily behaviors can be tracked without additional user input. Therefore, the present thesis aims to explore associations between depression symptom severity and individuals' behaviors measured by mobile phones and wearable devices. In the introduction chapter, I first introduce the background of depression and the limitations of conventional depression diagnosis as well as other types of depression research (such as genetic and environmental research). Then, I explain the motivations for using mobile technology in depression research. Next, I review the existing related digital depression studies and summarize their findings and limitations. Finally, I list the objectives and outline of this thesis.

1.2 Background

1.2.1 Depression - Definitions and Symptoms

Depression disorder is one of the most prevalent mental diseases, characterized by sadness, loss of interest or pleasure, feelings of guilt or low self-esteem, disturbed sleep, feelings of fatigue, and poor concentration (World Health Organization, 2017).

Depression is now diagnosed using the Structured Interview for Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), which identifies nine common symptoms linked with depression disorder: 1) depressed mood, 2) loss of interest or pleasure in activities, 3) changes in sleep, 4) changes in weight, 5) fatigue or loss of energy, 6) restlessness or feeling slow, 7) diminished interest or pleasure in activities, 8) diminished ability to concentrate, 9) feelings of worthlessness, and thoughts of death and suicide (American Psychiatric Association, 2013).

1.2.2 Prevalence, Adverse Outcomes, and Treatment

More than 300 million individuals worldwide suffer from depression (Friedrich, 2017). The 12-month prevalence estimates for major depressive disorders range from 6.6% to 10.3%, whereas the estimated lifetime risk is between 16.6% and 17.1% (Kessler et al., 2003; Kessler et al., 2005; Kessler et al., 1994). There has been a growing trend in the prevalence of depression, with an estimated 18.4% increase in the number of people living with depression between 2005 and 2015 (World Health Organization, 2017).

Depression is one of the leading causes of disability worldwide (Ferrari et al., 2013). It is associated with a range of negative outcomes, including premature mortality (Miloyan & Fried, 2017), decreased quality of life (Saragoussi et al., 2018), loss of professional function (Cambridge, Knight, Mills, & Baune, 2018), and even suicide (World Health Organization, 2017). Further, depression has been found to be comorbid with many other medical diseases, such as cardiac disease, cancer, neurologic disorders, and HIV/AIDS (Krishnan et al., 2002). As a result of its chronic development, poor prognosis (Verhoeven et al., 2018), and comorbidities, depression can also impose a substantial burden on society in terms of cost, lost productivity, morbidity, and mortality (P. S. Wang, Simon, & Kessler, 2003; Wells et al., 2002).

Antidepressants and psychotherapy are two of the available treatments for depression (Frank, Novick, & Kupfer, 2022). Recognizing depression at an early stage would improve treatment results and prognosis (Kraus, Kadriu, Lanzenberger, Zarate Jr, & Kasper, 2019). However, the majority of depressed individuals did not receive adequate treatment (Evans-Lacko et al., 2018). The barriers to effective care include the lack of experienced healthcare providers and resources, the societal stigma of mental diseases, and the limitations of traditional depression assessments (Evans-Lacko et al., 2018).

1.2.3 Limitations of Traditional Depression Assessments

Traditional depression assessments, such as clinical interviews and self-reported questionnaires, have some limitations. First, these assessments rely on individuals' retrospective reports of their behaviors over weeks, months, or even years, which are known to be unreliable and susceptible to recall bias (Ben-Zeev & Young, 2010; De Beurs, Lange, & Van Dyck, 1992). Second, it is difficult to quantify the day-to-day fluctuations in mood and behavior using questionnaires (Bradshaw, Saling, Hopwood, Anderson, & Brodtmann, 2004; Snyder & Zhou, 2019). Third, physicians in primary care (where depression is managed (Simon & VonKorff, 1995)) may fail to recognize patients with depressive symptoms (Schulberg et al., 1985; Wells et al., 1989) due to the expertise and experience required for identifying depression. Last but not least, conventional evaluations often take place when the patient's mental health issues or functional impairments have progressed to a more severe, difficult-to-treat stage (Ben-Zeev, Scherer, Wang, Xie, & Campbell, 2015). Therefore, effective auxiliary methodologies for recognizing depression at an early stage are needed to enable preventative strategies.

1.2.4 Genetic and Environmental Factors of Depression

In order to understand the etiology of depression, a number of studies have investigated the genetic and environmental determinants of depression which may help prevent depression and guide therapy (Dunn et al., 2015).

There is substantial evidence linking genetic factors to depression and other mental diseases (Dunn et al., 2015). In comparison to individuals without MDD, people with MDD were three times more likely to have a first-degree relative who was depressed (Sullivan, Neale, & Kendler, 2000). Moreover, twin studies indicate a modest heritability of depression (Rice, Harold, & Thapar, 2002). Several existing large genetic studies, such as Human Genome Project (Sawicki, Samara, Hurwitz, & Passaro Jr, 1993); HapMap Project (Gibbs et al., 2003); 1,000 Genomes Project (Siva, 2008), offer researchers opportunities to explore the associations between candidate genes and depression etiology. However, the majority of candidate gene studies had insufficient power and results replication was rare (Dunn et al., 2015).

Numerous risk factors for depression have been identified as environmental factors, such as poverty (Brooks-Gunn & Duncan, 1997), bad family relations (Repetti, Taylor, & Seeman, 2002), family disruption (Gilman, Kawachi, Fitzmaurice, & Buka, 2003), childhood abuse (Slopen, Koenen, & Kubzansky, 2014), and other stressful circumstances in life (Kessler, 1997). While the risk of depression is heightened immediately after encountering these environmental risk events, the consequences of adversity can remain throughout the course of a person's lifetime (Dunn, Gilman, Willett, Slopen, & Molnar, 2012; Dunn, McLaughlin, Slopen, Rosand, & Smoller, 2013).

Nonetheless, both genetic and environmental approaches are time-consuming and expensive, and it is challenging to detect depression detection in a timely manner and capture the day-to-day mental states. Therefore, there is a need for more cost-effective methods, such as Mobile Health (mHealth), to gather complementary information on the day-to-day changes in mental states and related behaviors.

1.2.5 Mobile Health (mHealth)

With the development and global availability of mobile technologies (e.g., mobile phones and consumer wearables) (Vailshery, 2021), we now have an additional way to monitor people's daily behavior and health. Mobile phones, with embedded sensors, provide us with a cost-effective way to continually collect users' everyday context, including physical activity, location, and surroundings (Donker et al., 2013). Consumer wearables are generally unobtrusive, cost-effective, and comfortable to wear (Jia et al., 2018; Shin et al., 2019). In recent years, the quality and precision of wearable data have increased (Fuller et al., 2020), enabling long-term monitoring of individuals' behaviors, and environmental and physiological parameters including sleep stages, heart rate, and blood oxygen levels in real-world settings (Zapata-Lamana, Lalanza, Losilla, Parrado, & Capdevila, 2020).

1.2.6 Digital Phenotyping for Depression Monitoring

Depression is strongly related to a number of behavioral abnormalities, including decreases in activity and social interactions, sleep disturbances, and abnormal changes in circadian rhythm, walking patterns, and heart rate (Prigerson et al., 1995; Vallée, Cadot, Roustit, Parizot, & Chauvin, 2011). Individual behavioral patterns captured by mobile phones and wearables, often known as "digital phenotyping" (Mohr, Shilton, & Hotopf, 2020), provide a new way of detecting depression for both clinicians and depressed people (Donker et al., 2013). Combined with phone-based questionnaires, mobile technologies could evaluate the evolution and changes in the individual's depressive state and related behaviors more frequently than conventional clinical depression assessments (De Angel et al., 2022). The day-to-day fluctuations in individuals' behaviors and moods could also be captured by mobile technologies with minimal user burden (Marzano, Hollis, Cipriani, & Malhi, 2017). Moreover, large and diverse cohorts could be recruited remotely, cost-efficiently, and swiftly using mobile technologies (Moore, Tassé, Thorogood, Winship, & Doerr, 2017), which could aid in the knowledge of the etiology, pathophysiology, and effective treatments for depression.

1.3 Related mHealth Studies for Depression

In recent decades, numerous mHealth studies for depression have been conducted and the details of sample size and digital technologies of these studies are summarized in several recent systematic reviews (De Angel et al., 2022; Firth et al., 2017; Rohani, Faurholt-Jepsen, Kessing, & Bardram, 2018). A considerable proportion of studies were conducted on relatively small cohorts (De Angel et al., 2022). For instance, Wang et al. performed the StudentLife study, which gathered passive phone data from 48 Dartmouth College students for 10 weeks, along with the pre-post evaluations of depression severity (PHQ-9) (Wang et al., 2014). Ben-Zeev et al. reported a similar study that monitored the speech, sleep, and activity behaviors of 47 college students for 10 weeks (Ben-Zeev, Scherer, Wang, Xie, & Campbell, 2015). Saeb et al. recruited 40 adults from the general community to collect their GPS and phone usage data via an app for two weeks with the PHQ-9 assessments completed at enrollment (Saeb et al., 2015).

Several mHealth studies were conducted on relatively large cohorts but only collected one or two types of behaviors for short study periods. For example, Luik et al. collected sleep and activity actigraphy recordings from 1714 participants for 7 days, along with one depression assessment for each participant (Luik et al., 2015). Similarly, Li et al. presented a study that contains 375 participants' measurements of depressive symptom severity and 1-week accelerometer data (Li, Kearney, & Fitzgerald, 2018).

Some large longitudinal mHealth datasets for depression (such as GLOBEM (Xu et al., 2023) and BRIGHTEN (Pratap et al., 2022) datasets) were recently released online. GLOBEM dataset consists of 4 college student datasets from 2018 to 2021, containing behavioral data from 705 participants (Xu et al., 2023). Each sub-dataset gathered students' phone passive (e.g., phone usage, calls, Bluetooth, and GPS) and wearable data streams (e.g., sleep and steps) in the background 24×7 for 10 weeks, along with weekly mental health questionnaires (e.g., PHQ-9) (Xu et al., 2023).

BRIGHTEN (Bridging Research Innovations for Greater Health in Technology, Emotion, and Neuroscience) study recruited a total of 2193 adult participants at the baseline, of whom approximately 900 participants agreed to share their daily phone passive data (GPS and phone usage) and weekly remote surveys (e.g., PHQ-9 and sleep questionnaires) for 12 weeks (Pratap et al., 2022). However, approximately 50% of participants left the study between week 1 to week 4, and only 15% remained in the study at the end of 12 weeks (Pratap et al., 2018), indicating that participant retention and engagement are challenges for remote mHealth studies (Pratap et al., 2020).

1.4 Findings in Previous mHealth Studies

A number of significant associations between depression and individuals' behaviors have been identified by previous mHealth studies. The following is a summary of previous findings, broken down into seven categories of features, including sleep, mobility, sociability, circadian rhythm, environment, phone usage, and physiological parameters.

1.4.1 Sleep

Mental health research has linked sleep pathologies to depression (Alvaro, Roberts, & Harris, 2013; Mendelson, 2012). Depression is commonly accompanied by sleep disorders such as insomnia, hypersomnia, and sleep rhythm disturbances (Mendelson, 2012). However, several conventional sleep assessments, such as polysomnography (PSG) and sleep questionnaires, are unsuitable for long-term sleep monitoring in real-world settings (Sánchez-Ortuño, Edinger, Means, & Almirall, 2010). Therefore, recent digital studies have attempted to use mobile technologies for sleep monitoring in home settings (Chen et al., 2013; Rébecca Robillard et al., 2015; Van De Water, Holmes, & Hurley, 2011; R. Wang et al., 2014; Zhang et al., 2019).

Several studies used mobile phones to estimate sleep duration through the embedded light, microphone, and acceleration sensors, combined with the phone usage information (Chen et al., 2013; R. Wang et al., 2014). Some wearable devices could classify sleep into specific sleep stages (such as awake, light, deep, and REM sleep) and provide several sleep parameters (such as sleep quality and awakening counts) using heart rate (extracted from PPG or ECG signals) and accelerometry signals (Van De Water et al., 2011; Zhang et al., 2019).

Previous digital research indicated that higher depression severity was correlated with lower sleep quality, large sleep variance, and later sleep offset time (Chen et al., 2013; Rébecca Robillard et al., 2015; Van De Water et al., 2011; R. Wang et al., 2014). Notably, the total sleep duration showed opposite associations with depression symptom severity across different studies (positive: (R Robillard et al., 2013) and negative: (Kawada, Katsumata, Suzuki, & Shimizu, 2007; R. Wang et al., 2014)) which may be due to that insomnia and hypersomnia are both symptoms of depression (Alvaro et al., 2013; Kaplan & Harvey, 2009).

1.4.2 Mobility

A bidirectional link between depression and individuals' mobility was found in past depression studies (Roshanaei-Moghaddam, Katon, & Russo, 2009). Specifically, many depression studies reported that depressed people were more sedentary than healthy controls (Weyerer & Kupfer, 1994). On the other hand, several therapy trials have demonstrated that regular exercise could help lessen depression symptoms and the risk of developing depression (Mead et al., 2008; Teychenne, Ball, & Salmon, 2008).

Recent digital depression studies have attempted to measure individuals' mobility via acceleration and location data gathered via smartphones and wearable devices (Difrancesco et al., 2019; Farhan et al., 2016; Lu et al., 2018; Moukaddam, Truong, Cao, Shah, & Sabharwal, 2019; Saeb, Lattie, Schueller, Kording, & Mohr, 2016; Saeb et al., 2015; Yue et al., 2018). Embedded accelerometers allow smartphones and wearables to measure many activity characteristics, including step count, movement speed, and activity levels. According to these parameters, higher depression symptom severity was found to be associated with less time spent in activity (Lu et al., 2018), fewer daily step counts (Moukaddam et al., 2019), slower moving speed (Yue et al., 2018), and more sedentary behaviors (Difrancesco et al., 2019).

Several other studies tried to evaluate individuals' mobility patterns based on location features, including homestay (time at home), location entropy (time distribution across different locations), the number of location clusters, and moving distance (Farhan et al., 2016; Lu et al., 2018; Saeb et al., 2016; Saeb et al., 2015). In these studies, higher depression severity was found to be linked with lower location entropy, fewer location clusters, shorter moving distances, and longer homestay time (Farhan et al., 2016; Lu et al., 2018; Saeb et al., 2016; Saeb et al., 2015).

1.4.3 Sociability

The relationships between depression and sociability have been well-documented over the years (Cohen, 2004; Rook, 1984). In conventional survey-based studies, those who report fewer social network connections or less social support are more likely to have higher depressive symptom severity, and a substantial proportion of suicides have a history of social isolation (Burgess, Pirkis, Morton, & Croke, 2000; Cacioppo, Hughes, Waite, Hawkley, & Thisted, 2006). However, traditional survey-based sociability assessments are qualitative and susceptible to subjective bias (Boonstra, Werner-Seidler, O'Dea, Larsen, & Christensen, 2017). To address this limitation, several recent digital studies have approximated individuals' sociability using smartphone data, including call and message logs (frequency and length of communications) (Ben-Zeev, Schueller, et al., 2015; Doryab, Min, Wiese, Zimmerman, & Hong, 2014), Bluetooth sensor data (nearby devices) (Boonstra et al., 2017; R. Wang et al., 2014), and microphone sensor data (voice detection) (Ben-Zeev, Scherer, et al., 2015). Boonstra et al illustrated the feasibility of using Bluetooth data to approximate the social network of participants for depression analysis (Boonstra et al., 2017). Wang et al found the number of nearby Bluetooth devices is negatively correlated with depression severity (R. Wang et al., 2014). Ben-Zeev et al found that longer nearby human speech is correlated with lower depression severity (Ben-Zeev, Scherer, et al., 2015). The duration and frequency of communications (phone calls and text messages) were found to have inverse relationships with depression across genders (Cho et al., 2016; Doryab et al., 2014; Rohani, Faurholt-Jepsen, Kessing, & Bardram, 2018). In a past study, higher depression severity was found to be correlated with increased outgoing communications in female participants; in contrast, male participants who felt more depressed tended to make fewer calls than usual (Doryab et al., 2014). Therefore, sociability should be regarded as a highly individualized characteristic in the depression study (Rohani et al., 2018).

1.4.4 Circadian Rhythm

The circadian rhythm is an internal clock related to endogenous oscillations of about a 24-hour period, which affects and regulates the timing of almost all behavioral and physiological activities and has extensive associations with individuals' physical and mental health (Partch, Green, & Takahashi, 2014). Depression may lead to a misalignment of the circadian rhythm and make individuals' behaviors more irregular (Walker, Walton, DeVries, & Nelson, 2020). Dim-light melatonin onset detected from blood or saliva samples can be used to track the circadian rhythm (Bowman et al., 2021). However, this conventional method is not appropriate for long-term real-world monitoring (Bowman et al., 2021). Therefore, several recent digital studies have attempted to approximate the circadian rhythm using behavioral data collected by smartphones and wearable devices (such as heart rate, step, and GPS data), mainly based on two methodologies: the Cosinor model and spectrum analysis (Refinetti, Cornélissen, & Halberg, 2007; Saeb et al., 2015). The first method assumes individual behaviors follow a Cosinor function (or extended cosine function) (Refinetti et al., 2007). The circadian rhythms of individuals were represented by parameters of fitted Cosinor functions, such as acrophase (a phase marker showing the time when the fitted signal reaches its peak) and coefficient of determination of the model (R^2; reflecting the strength of circadian rhythm) (Refinetti et al., 2007). The second method for approximating the strength of the circadian rhythm is to calculate the power of the 24-hour frequency bands of an individual's spectrum generated from behavioral data (Saeb et al., 2015). In these past studies, higher depression severity was found to be associated with lower daytime activity, later acrophase of activity (Rébecca Robillard et al., 2015; Smagula, Krafty, Thayer, Buysse, & Hall, 2018; White, Rumble, & Benca, 2017), longer desynchronized phase between heart rate and activity (Carr et al., 2018), and lower strength of 24-hour periods in GPS data (Saeb et al., 2015).

1.4.5 Environment

It has been demonstrated that environmental factors, such as humidity, ambient temperature, and sunlight, are associated with people's moods (Howarth & Hoffman, 1984). For example, as an adjunct to antidepressant treatment, bright light therapy has been shown to be effective (Even, Schröder, Friedman, & Rouillon, 2008; Reeves et al., 2012). Several past digital studies have explored associations between environmental variables and depression symptom severity utilizing weather information (obtained via smartphone location data) and light sensors (Doryab et al., 2014; Moraes et al., 2013). Ávila-Moraes et al discovered that the amplitude and stability of light exposure were lower in the depressed group compared to the healthy control group (Moraes et al., 2013). Similar to sociability features, humidity was also found to have contradictory relationships with depression across genders; that is, humidity had a significant positive correlation with depression severity in female participants but a negative correlation in male participants (Doryab et al., 2014).

1.4.6 Phone Usage

The widespread usage of mobile phones has had a significant impact on the communication and social interactions of people (Thomée, Härenstam, & Hagberg, 2011). Several survey-based studies discovered that prolonged phone use had negative effects on sleep and mental health (Thomée et al., 2011). Similar to other categories of behaviors, self-reported statistics on phone usage are unreliable (Boase & Ling, 2013). Therefore, some recent digital research extracted several features, such as unlock duration, unlock frequency, and app usage time, to quantify the characteristics of individuals' phone usage (Rohani et al., 2018; Saeb et al., 2015). Duration and frequency of phone usage were both found to be positively correlated with depression severity in most studies (Rohani et al., 2018; Saeb et al., 2015). In a student population, the use of specific apps, such as Instagram, maps, and photo and video apps, was connected with more severe depressive symptoms, whereas the use of book apps was associated with lower depression severity (David, Roberts, & Christenson, 2018). Moreover, Mehrotra et al. discovered that while participants' emotional states have a causal influence on several elements of interactions with mobile phones, the usage of certain apps has a causal impact on participants' levels of happiness and stress (Mehrotra, Tsapeli, Hendley, & Musolesi, 2017).

1.4.7 Physiological Parameters

Several previous clinical studies have revealed that depression severity is associated with some physiological parameters, including heart rate, body temperature, and skin conductance (Kemp et al., 2010; Souetre et al., 1988). For instance, abnormal variations in body temperature have been observed in depressed people, which may be the result of circadian rhythm disorders (Souetre et al., 1988). Depressive disorders have an effect on human autonomic nervous system (ANS) function, resulting in decreased parasympathetic and/or elevated sympathetic tone, which may raise the risk of cardiovascular diseases (Agelink, Boz, Ullrich, & Andrich, 2002). Heart rate variability (HRV) and electrodermal activity (EDA) are two basic indicators of the state of ANS that can be measured in non-invasive manners (Sarchiapone et al., 2018). In some previous studies, depressed people have been found to have lower HRV (Kemp et al., 2010) and reduced EDA (Sarchiapone et al., 2018) compared to healthy controls.

In recent digital depression studies, depressed people were found to have a longer time in an elevated temperature state compared to healthy controls (Moraes et al., 2013). Further, Jepsen et al found a significant positive correlation between heart rate during sleep and the severity of depressive symptoms (Faurholt-Jepsen et al., 2015). However, few digital studies have directly investigated associations between depression severity and individuals' physiological parameters measured by wearable devices (De Angel et al., 2022).

1.5 Limitations of Previous Digital Depression Studies

Although previous digital studies have shown the feasibility of mobile technologies for monitoring depression in naturalistic settings, there are also some limitations. First, the primary limitation is that the majority of prior studies were conducted on relatively small, homogeneous (e.g., university students), and short-term cohorts. In a recent review of 51 digital depression studies, the median sample size was 58 participants and the median duration of follow-up was 9 days (De Angel et al., 2022). Small and homogeneous cohorts may restrict the generalizability of their findings, which may be the reason for some contradictory results among various studies (Rohani et al., 2018) and some unsuccessful replication work (Saeb et al., 2016).

Second, participant attrition and missing data may result in an unbalanced cohort, which may have a substantial impact on the generalizability of real-world findings (De Angel et al., 2022; Druce, Dixon, & McBeth, 2019). According to a recent engagement study, participant retention and engagement in digital studies were connected with various real-world factors, such as age, compensation, referral by clinicians, and clinical conditions (Pratap et al., 2020). However, most previous digital depression studies did not discuss the impact of participant attrition and missing data on their findings. Further, the comparison of participant participation in data collection of different data streams (e.g., remote questionnaires, passive phone data, and passive wearable data) has not yet been conducted.

Third, some studies only applied basic statistical features to characterize the behaviors of participants, which may lead to some information loss. For example, in a previous study, only the total number of nearby Bluetooth devices was calculated to approximate sociability (R. Wang et al., 2014), thus the periodicity, complexity, and variance of individuals' behaviors contained in Bluetooth data may be lost. In addition, several categories of behavioral characteristics, such as daily-life gait patterns and physiological parameters, have not been fully explored in past studies (De Angel et al., 2022).

Fourth, the longitudinal relationships (cross-lagged effects) between behavioral features and depression severity over time were not fully assessed in most past digital depression studies, possibly due to their short study durations. Cross-lagged effects

(McNeish & Hamaker, 2020) refer to the impact of present depression status on subsequent behaviors and the influence of current behaviors on the depression status at subsequent time points. If mobile technology could detect abnormal behavioral changes prior to the development of depression, early intervention may be used to avoid depression relapse and deterioration. Therefore, there is a need to further investigate the longitudinal links between depression and behavioral features derived from mobile technology.

Last but not least, the lack of reporting confounding variables in the analysis is also one of the limitations of past digital studies (Rohani et al., 2018, De Angel et al., 2022). Some demographic characteristics are associated with both depression and behavioral patterns. For example, older age is known to be significantly correlated with lower depression severity and decreased mobility (Akhtar-Danesh & Landeen, 2007; Kessler et al., 2003). Therefore, demographics must be considered as confounding variables when examining the association between depression and behavioral features.

1.6 Research Questions and Objectives of This Thesis

The main research questions investigated in the present thesis are: a) Can mobile technology capture the associations between depression and behaviors? b) Whether these associations found in previous studies can be observed consistently in a large multicenter (multinational) data set collected over an extended period? c) Can depression symptom severity be predicted using behavioral features extracted from mobile phones and wearable devices? d) What factors influence participant retention and engagement in a longitudinal digital depression study? To answer these research questions and address the limitations of previous digital studies, the objectives of this thesis are shown below.

Feature engineering This thesis aims to design and extract novel features for characterizing individuals' behaviors based on prior clinical findings and commonly used features in other research fields (e.g., signal processing). As an example, we compute the difference in sleep duration between weekends and weekdays as a novel feature to reflect sleep efficiency, based on a finding in previous sleep research: longer sleep during weekends than weekdays (weekend catch-up sleep) reflects insufficient weekday sleep (Kang et al., 2014; Liu, Zhao, Jia, & Buysse, 2008). On the other hand, frequency-domain and multiscale entropy features are widely used in the field of signal processing, and they are used in this thesis to describe the periodicity and complexity of individuals' behaviors.

Association analysis This thesis aims to explore associations between depression symptom severity and individuals' behavioral features derived from smartphones and wearables on a large, multicenter, and longitudinal dataset. In the association analysis, participants' demographics are regarded as confounding variables.

Cross-lagged effects analysis This thesis aims to explore the longitudinal associations (cross-lagged effects) between depression and behavioral features over time, testing if the current behavioral patterns impact the subsequent depression status and vice versa.

Long-term participant retention and engagement analysis This thesis aims to investigate the participant retention and engagement patterns in a large multinational digital study for major depressive disorder, as well as explore potential links between real-world factors (such as demographics, clinical characteristics, and types of devices) and participant engagement in digital health research.

1.7 Outline of Chapters

Chapter 2 investigates the long-term participant retention and engagement patterns in 3 data streams (Phone-Active, Phone-Passive, and Fitbit-Passive) of a large multicenter depression study. The survival analysis is used to present participant retention in this study and explore its connection with real-world factors. The longitudinal engagement patterns (levels of data density over time) are identified using unsupervised clustering methods. To identify factors substantially influencing engagement levels, participants' characteristics across different clusters are compared. Furthermore, this chapter discusses strategies for increasing participant engagement in future digital health research.

In chapter 3, we explore relationships between depression symptom severity and sleep measured by the Fitbit device. We extract a number of sleep features from Fitbit recordings to characterize the sleep of participants in 5 aspects: sleep architecture, sleep stability, sleep quality, insomnia, and hypersomnia. Associations between these sleep features and depression symptom severity are evaluated by linear mixed-effects regression models. The findings of this chapter are compared to those of previous sleep studies with conventional sleep measures (e.g., PSG and sleep surveys).

In chapter 4, we attempt to predict depression symptom severity using Bluetooth data gathered from mobile phones. We design and propose several statistical-based and nonlinear Bluetooth features to measure the distribution, regularity, and periodicity of participants' nearby Bluetooth device count data. Likewise, associations between Bluetooth features and depression are measured using linear mixed-effects regression models. We then predict depression symptom severity using extracted Bluetooth features via the hierarchical Bayesian linear regression model, which can capture the cohort's characteristics with individual differences.

Chapter 5 explores the longitudinal relationships (cross-lagged effects) between depression symptom severity and phone-measured mobility over time. Several mobility features are extracted from the location data provided by mobile phones' GPS sensors and networks. The framework of dynamic structural equation modeling is utilized to examine if current mobility features significantly affect the subsequent depression status and vice versa. The impact of individual differences on these longitudinal relationships is also investigated and discussed in this chapter.

Chapter 6 examines associations between depression symptom severity and daily-life gait characteristics derived from long-term acceleration data in real-world settings. Although gait patterns have been shown to be closely correlated with depression in laboratory settings, daily-life walking patterns and their relationships with depression have yet to be fully explored. To fill this gap, in this chapter, we extract several daily-life gait features from raw accelerometry data to describe the gait cadence and gait force over a long term and then examine their associations with depression symptom severity. To test whether these relationships can be captured by different devices, we perform our analysis on two ambulatory data sets containing acceleration data from wearable devices and mobile phones, respectively.

Chapter 7 explores the associations between depression symptom severity and circadian rhythm patterns estimated from wearable data. Since traditional laboratory-based measures for the circadian rhythm are unsuitable for long-term monitoring in naturalistic settings, this chapter attempts to approximate individuals' circadian rhythms using passive behavioral data gathered from Fitbit devices. I utilize the parameters of Cosinor models fitted using wearable data to reflect the characteristics of circadian rhythm patterns and then examine their associations with depression symptom severity using linear mixed-effects regression models. Additionally, seasonal influences on circadian rhythm patterns are also examined and discussed in this chapter.

The last chapter (Chapter 8) summarizes and discusses the findings of this thesis as well as our analysis plans and suggestions for future digital health studies.

Chapters 2–6 of this thesis consist of five publications. I am the first author of these five papers and did all the data analysis, coding, and writing up work. Other coauthors' contributions include data collection, platform development, data storage and maintenance, analysis planning, and critical review of manuscripts, which are stated in the "Authors' Contributions" sections of related chapters in detail. In due course, the work described in Chapters 7 will be refined and written up for publication. Supplementary material to Chapters 2-7 are provided in Appendixes.

References

Agelink, M. W., Boz, C., Ullrich, H., & Andrich, J. (2002). Relationship between major depression and heart rate variability.: Clinical consequences and implications for antidepressive treatment. *Psychiatry Research, 113*(1-2), 139-149.

Akhtar-Danesh, N., & Landeen, J. (2007). Relation between depression and sociodemographic factors. *International journal of mental health systems, 1*(1), 1-9.

Alvaro, P. K., Roberts, R. M., & Harris, J. K. (2013). A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. *Sleep, 36*(7), 1059-1068.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association.

Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal, 38*(3), 218.

Ben-Zeev, D., Schueller, S. M., Begale, M., Duffecy, J., Kane, J. M., & Mohr, D. C. (2015). Strategies for mHealth research: lessons from 3 mobile intervention studies. *Administration and Policy in Mental Health and Mental Health Services Research, 42*(2), 157-167.

Ben-Zeev, D., & Young, M. A. (2010). Accuracy of hospitalized depressed patients' and healthy controls' retrospective symptom reports: an experience sampling study.

The Journal of nervous and mental disease, 198(4), 280-285.

Boase, J., & Ling, R. (2013). Measuring mobile phone use: Self-report versus log data. *Journal of Computer-Mediated Communication, 18*(4), 508-519.

Brooks-Gunn, J., & Duncan, G. J. (1997). The effects of poverty on children. The future of children, 55-71.

Boonstra, T. W., Werner-Seidler, A., O'Dea, B., Larsen, M. E., & Christensen, H. (2017). *Smartphone app to investigate the relationship between social connectivity and mental health.* Paper presented at the 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC).

Bowman, C., Huang, Y., Walch, O. J., Fang, Y., Frank, E., Tyler, J., . . . Sen, S. (2021). A method for characterizing daily physiology from widely used wearables. *Cell reports methods, 1*(4), 100058.

Bradshaw, J., Saling, M., Hopwood, M., Anderson, V., & Brodtmann, A. (2004). Fluctuating cognition in dementia with Lewy bodies and Alzheimer’s disease is qualitatively distinct. *Journal of Neurology, Neurosurgery & Psychiatry, 75*(3), 382-387.

Burgess, P., Pirkis, J., Morton, J., & Croke, E. (2000). Lessons from a comprehensive clinical audit of users of psychiatric services who committed suicide. *Psychiatric services, 51*(12), 1555-1560.

Cacioppo, J. T., Hughes, M. E., Waite, L. J., Hawkley, L. C., & Thisted, R. A. (2006). Loneliness as a specific risk factor for depressive symptoms: cross-sectional and longitudinal analyses. *Psychology and aging, 21*(1), 140.

Cambridge, O. R., Knight, M. J., Mills, N., & Baune, B. T. (2018). The clinical relationship between cognitive impairment and psychosocial functioning in major depressive disorder: A systematic review. *Psychiatry Research, 269*, 157-171.

Carr, O., Saunders, K. E., Bilderbeck, A. C., Tsanas, A., Palmius, N., Geddes, J. R., . . . Goodwin, G. M. (2018). Desynchronization of diurnal rhythms in bipolar disorder and borderline personality disorder. *Translational psychiatry, 8*(1), 1-9.

Chen, Z., Lin, M., Chen, F., Lane, N. D., Cardone, G., Wang, R., . . . Campbell, A. T. (2013). *Unobtrusive sleep monitoring using smartphones.* Paper presented at the 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops.

Cho, Y. M., Lim, H. J., Jang, H., Kim, K., Choi, J. W., Shin, C., . . . Kim, N. (2016). A cross-sectional study of the association between mobile phone use and symptoms of ill health. *Environmental health and toxicology, 31*.

Cohen, S. (2004). Social relationships and health. *American psychologist, 59*(8), 676.

David, M. E., Roberts, J. A., & Christenson, B. (2018). Too much of a good thing: Investigating the association between actual smartphone use and individual well-being. *International Journal of Human–Computer Interaction, 34*(3), 265-275.

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., . . . Mohr, D. C. (2022). Digital health tools for the passive monitoring of depression: a systematic review of methods. *NPJ Digital Medicine, 5*(1), 1-14.

De Beurs, E., Lange, A., & Van Dyck, R. (1992). Self-monitoring of panic attacks and retrospective estimates of panic: Discordant findings. *Behaviour research and therapy, 30*(4), 411-413.

Difrancesco, S., Lamers, F., Riese, H., Merikangas, K. R., Beekman, A. T., van Hemert, A. M., . . . Penninx, B. W. (2019). Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: A 2-week ambulatory assessment study. *Depression and anxiety, 36*(10), 975-986.

Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M.-R., & Christensen, H. (2013). Smartphones for smarter delivery of mental health programs: a systematic review. *Journal of medical Internet research, 15*(11), e2791.

Doryab, A., Min, J. K., Wiese, J., Zimmerman, J., & Hong, J. (2014). *Detection of behavior change in people with depression.* Paper presented at the Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence.

Druce, K. L., Dixon, W. G., & McBeth, J. (2019). Maximizing engagement in mobile health studies: lessons learned and future directions. *Rheumatic diseases clinics of North America, 45*(2), 159.

Dunn, E. C., Brown, R. C., Dai, Y., Rosand, J., Nugent, N. R., Amstadter, A. B., & Smoller, J. W. (2015). Genetic determinants of depression: recent findings and future directions. Harvard review of psychiatry, 23(1), 1.

Dunn, E. C., Gilman, S. E., Willett, J. B., Slopen, N. B., & Molnar, B. E. (2012). THE IMPACT OF EXPOSURE TO INTERPERSONAL VIOLENCE ON GENDER DIFFERENCES IN ADOLESCENT-ONSET MAJOR DEPRESSION: RESULTS FROM THE N ATIONAL C OMORBIDITY S URVEY R EPLICATION (NCS-R). Depression and anxiety, 29(5), 392-399.

Dunn, E. C., McLaughlin, K. A., Slopen, N., Rosand, J., & Smoller, J. W. (2013). Developmental timing of child maltreatment and symptoms of depression and suicidal ideation in young adulthood: results from the National Longitudinal Study of Adolescent Health. Depression and anxiety, 30(10), 955-964.

Evans-Lacko, S., Aguilar-Gaxiola, S., Al-Hamzawi, A., Alonso, J., Benjet, C., Bruffaerts, R., . . . Gureje, O. (2018). Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: results from the WHO World Mental Health (WMH) surveys. *Psychological medicine, 48*(9), 1560-1571.

Even, C., Schröder, C. M., Friedman, S., & Rouillon, F. (2008). Efficacy of light therapy in nonseasonal depression: a systematic review. *Journal of affective disorders, 108*(1-2), 11-23.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., . . . Wang, B. (2016). *Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data.* Paper presented at the 2016 IEEE Wireless Health (WH).

Faurholt-Jepsen, M., Brage, S., Vinberg, M., Jensen, H. M., Christensen, E. M., Knorr, U., & Kessing, L. V. (2015). Electronic monitoring of psychomotor activity as a supplementary objective measure of depression severity. *Nordic Journal of Psychiatry, 69*(2), 118-125.

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J., . . . Whiteford, H. A. (2013). Burden of depressive disorders by country, sex, age, and year: findings from the global burden of disease study 2010. *PLoS medicine, 10*(11), e1001547.

Firth, J., Torous, J., Nicholas, J., Carney, R., Pratap, A., Rosenbaum, S., & Sarris, J. (2017). The efficacy of smartphone-based mental health interventions for depressive symptoms: a meta-analysis of randomized controlled trials. World Psychiatry, 16(3), 287-298.

Frank, E., Novick, D., & Kupfer, D. J. (2022). Antidepressants and psychotherapy: a clinical research review. *Dialogues in clinical neuroscience*.

Friedrich, M. J. (2017). Depression is the leading cause of disability around the world.

Jama, 317(15), 1517-1517.

Fuller, D., Colwell, E., Low, J., Orychock, K., Tobin, M. A., Simango, B., . . . Cullen, K. (2020). Reliability and validity of commercially available wearable devices for measuring steps, energy expenditure, and heart rate: systematic review. *JMIR mHealth and uHealth, 8*(9), e18694.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., . . . Shen, Y. (2003). The international HapMap project.

Gilman, S. E., Kawachi, I., Fitzmaurice, G. M., & Buka, S. L. (2003). Family disruption in childhood and risk of adult depression. American Journal of Psychiatry, 160(5), 939-946.

Howarth, E., & Hoffman, M. S. (1984). A multidimensional approach to the relationship between mood and weather. *British Journal of Psychology, 75*(1), 15-23.

Jia, Y., Wang, W., Wen, D., Liang, L., Gao, L., & Lei, J. (2018). Perceived user preferences and usability evaluation of mainstream wearable devices for health monitoring. *PeerJ, 6*, e5350.

Kang, S.-G., Lee, Y. J., Kim, S. J., Lim, W., Lee, H.-J., Park, Y.-M., . . . Hong, J. P. (2014). Weekend catch-up sleep is independently associated with suicide attempts and self-injury in Korean adolescents. *Comprehensive psychiatry, 55*(2), 319-325.

Kaplan, K. A., & Harvey, A. G. (2009). Hypersomnia across mood disorders: a review and synthesis. *Sleep medicine reviews, 13*(4), 275-285.

Kawada, T., Katsumata, M., Suzuki, H., & Shimizu, T. (2007). Actigraphic predictors of the depressive state in students with no psychiatric disorders. *Journal of affective disorders, 98*(1-2), 117-120.

Kemp, A. H., Quintana, D. S., Gray, M. A., Felmingham, K. L., Brown, K., & Gatt, J. M. (2010). Impact of depression and antidepressant treatment on heart rate variability: a review and meta-analysis. *Biological psychiatry, 67*(11), 1067-1074.

Kessler, R. C. (1997). The effects of stressful life events on depression. Annual review of psychology, 48(1), 191-214.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., . . . Wang, P. S. (2003). The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *Jama, 289*(23), 3095-3105.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 593-602.

Kessler, R. C., McGonagle, K. A., Zhao, S., Nelson, C. B., Hughes, M., Eshleman, S., . . . Kendler, K. S. (1994). Lifetime and 12-month prevalence of DSM-III-R psychiatric disorders in the United States: results from the National Comorbidity Survey. *Archives of General Psychiatry, 51*(1), 8-19.

Kraus, C., Kadriu, B., Lanzenberger, R., Zarate Jr, C. A., & Kasper, S. (2019). Prognosis and improved outcomes in major depression: a review. *Translational psychiatry, 9*(1), 1-17.

Krishnan, K. R. R., Delong, M., Kraemer, H., Carney, R., Spiegel, D., Gordon, C., . . . Buckwalter, K. (2002). Comorbidity of depression with other medical diseases in the elderly. Biological psychiatry, 52(6), 559-588.

Li, X., Kearney, P. M., & Fitzgerald, A. P. (2018). Accelerometer-based physical activity patterns and correlates of depressive symptoms. Paper presented at the

Health Information Science: 7th International Conference, HIS 2018, Cairns, QLD, Australia, October 5–7, 2018, Proceedings 7.

Liu, X., Zhao, Z., Jia, C., & Buysse, D. J. (2008). Sleep patterns and problems among Chinese adolescents. *Pediatrics, 121*(6), 1165-1173.

Lu, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., . . . Bi, J. (2018). Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(1), 1-21.

Luik, A. I., Zuurbier, L. A., Direk, N., Hofman, A., Van Someren, E. J., & Tiemeier, H. (2015). 24-hour activity rhythm and sleep disturbances in depression and anxiety: A population-based study of middle-aged and older persons. Depression and anxiety, 32(9), 684-692.

Marzano, L., Hollis, C., Cipriani, A., & Malhi, G. S. (2017). Digital technology: coming of age? (Vol. 20, pp. 97-97): Royal College of Psychiatrists.

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological methods, 25*(5), 610.

Mead, G. E., Morley, W., Campbell, P., Greig, C. A., McMurdo, M., & Lawlor, D. A. (2008). Exercise for depression. *Cochrane database of systematic reviews*(4).

Mehrotra, A., Tsapeli, F., Hendley, R., & Musolesi, M. (2017). MyTraces: Investigating correlation and causation between users' emotional states and mobile phone interaction. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 1(3), 1-21.

Mendelson, W. (2012). *Human sleep and its disorders*: Springer Science & Business Media.

Miloyan, B., & Fried, E. (2017). A reassessment of the relationship between depression and all-cause mortality in 3,604,005 participants from 293 studies. *World Psychiatry, 16*(2), 219.

Mohr, D. C., Shilton, K., & Hotopf, M. (2020). Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. *NPJ Digital Medicine, 3*(1), 1-2.

Moore, S., Tassé, A.-M., Thorogood, A., Winship, I., & Doerr, M. (2017). Consent processes for mobile app mediated research: systematic review. *JMIR mHealth and uHealth, 5*(8), e7014.

Moraes, C. Á., Cambras, T., Diez-Noguera, A., Schimitt, R., Dantas, G., Levandovski, R., & Hidalgo, M. P. (2013). A new chronobiological approach to discriminate between acute and chronic depression using peripheral temperature, rest-activity, and light exposure parameters. *BMC psychiatry, 13*(1), 1-10.

Moukaddam, N., Truong, A., Cao, J., Shah, A., & Sabharwal, A. (2019). Findings from a trial of the Smartphone and OnLine Usage-based eValuation for Depression (SOLVD) application: what do apps really tell us about patients with depression? Concordance between app-generated data and standard psychiatric questionnaires for depression and anxiety. *Journal of Psychiatric Practice®, 25*(5), 365-373.

Partch, C. L., Green, C. B., & Takahashi, J. S. (2014). Molecular architecture of the mammalian circadian clock. *Trends in cell biology, 24*(2), 90-99.

Pratap, A., Renn, B. N., Volponi, J., Mooney, S. D., Gazzaley, A., Arean, P. A., & Anguera, J. A. (2018). Using mobile apps to assess and treat depression in Hispanic and Latino populations: fully remote randomized clinical trial. Journal of medical Internet research, 20(8), e10130.

Pratap, A., Neto, E. C., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., . . . Wilbanks, J. (2020). Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *NPJ digital medicine, 3*(1), 1-10.

Pratap, A., Homiar, A., Waninger, L., Herd, C., Suver, C., Volponi, J., . . . Areán, P. (2022). Real-world behavioral dataset from two fully remote smartphone-based randomized clinical trials for depression. Scientific data, 9(1), 522.

Prigerson, H. G., Monk, T. H., Reynolds III, C. F., Begley, A., Houck, P. R., Bierhals, A. J., & Kupfer, D. J. (1995). Lifestyle regularity and activity level as protective factors against bereavement-related depression in late-life. *Depression, 3*(6), 297-302.

Reeves, G. M., Nijjar, G. V., Langenberg, P., Johnson, M. A., Khabazghazvini, B., Sleemi, A., . . . Tariq, M. (2012). Improvement in depression scores after 1 hour of light therapy treatment in patients with seasonal affective disorder. *The Journal of nervous and mental disease, 200*(1), 51.

Refinetti, R., Cornélissen, G., & Halberg, F. (2007). Procedures for numerical analysis of circadian rhythms. *Biological rhythm research, 38*(4), 275-325.

Repetti, R. L., Taylor, S. E., & Seeman, T. E. (2002). Risky families: family social environments and the mental and physical health of offspring. Psychological bulletin, 128(2), 330.

Rice, F., Harold, G., & Thapar, A. (2002). The genetic aetiology of childhood depression: a review. Journal of child Psychology and Psychiatry, 43(1), 65-79.

Robillard, R., Hermens, D. F., Naismith, S. L., White, D., Rogers, N. L., Ip, T. K., . . . Smith, K. L. (2015). Ambulatory sleep-wake patterns and variability in young people with emerging mental disorders. *Journal of Psychiatry and Neuroscience, 40*(1), 28-37.

Robillard, R., Naismith, S., Rogers, N., Scott, E., Ip, T., Hermens, D. F., & Hickie, I. (2013). Sleep-wake cycle and melatonin rhythms in adolescents and young adults with mood disorders: comparison of unipolar and bipolar phenotypes. *European Psychiatry, 28*(7), 412-416.

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018). Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. *JMIR mHealth and uHealth, 6*(8), e9691.

Rook, K. S. (1984). The negative side of social interaction: impact on psychological well-being. *Journal of personality and social psychology, 46*(5), 1097.

Roshanaei-Moghaddam, B., Katon, W. J., & Russo, J. (2009). The longitudinal effects of depression on physical activity. *General hospital psychiatry, 31*(4), 306-315.

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

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research, 17*(7), e4273.

Sánchez-Ortuño, M. M., Edinger, J. D., Means, M. K., & Almirall, D. (2010). Home is where sleep is: an ecological approach to test the validity of actigraphy for the assessment of insomnia. *Journal of Clinical Sleep Medicine, 6*(1), 21-29.

Saragoussi, D., Christensen, M. C., Hammer-Helmich, L., Rive, B., Touya, M., & Haro, J. M. (2018). Long-term follow-up on health-related quality of life in major depressive disorder: a 2-year European cohort study. *Neuropsychiatric disease* *and treatment.*

Sarchiapone, M., Gramaglia, C., Iosue, M., Carli, V., Mandelli, L., Serretti, A., . . . Zeppegno, P. (2018). The association between electrodermal activity (EDA), depression and suicidal behaviour: A systematic review and narrative synthesis. *BMC psychiatry, 18*(1), 1-27.

Sawicki, M. P., Samara, G., Hurwitz, M., & Passaro Jr, E. (1993). Human genome project. The American journal of surgery, 165(2), 258-264.

Schulberg, H. C., Saul, M., McClelland, M., Ganguli, M., Christy, W., & Frank, R. (1985). Assessing depression in primary medical and psychiatric practices. *Archives of General Psychiatry, 42*(12), 1164-1170.

Shin, G., Jarrahi, M. H., Fei, Y., Karami, A., Gafinowitz, N., Byun, A., & Lu, X. (2019). Wearable activity trackers, accuracy, adoption, acceptance and health impact: A systematic literature review. *Journal of biomedical informatics, 93*, 103153.

Simon, G. E., & VonKorff, M. (1995). Recognition, management, and outcomes of depression in primary care. *Archives of family medicine, 4*(2), 99-105.

Siva, N. (2008). 1000 Genomes project. Nature biotechnology, 26(3), 256-257.

Slopen, N., Koenen, K. C., & Kubzansky, L. D. (2014). Cumulative adversity in childhood and emergent risk factors for long-term health. The Journal of pediatrics, 164(3), 631-638. e632.

Smagula, S. F., Krafty, R. T., Thayer, J. F., Buysse, D. J., & Hall, M. H. (2018). Rest-activity rhythm profiles associated with manic-hypomanic and depressive symptoms. *Journal of psychiatric research, 102*, 238-244.

Snyder, M., & Zhou, W. (2019). Big data and health. *The Lancet Digital Health, 1*(6), e252-e254.

Souetre, E., Salvati, E., Wehr, T. A., Sack, D. A., Krebs, B., & Darcourt, G. (1988). Twenty-four-hour profiles of body temperature and plasma TSH in bipolar patients during depression and during remission and in normal control subjects. *The American journal of psychiatry, 145*(9), 1133-1137.

Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: review and meta-analysis. American Journal of Psychiatry, 157(10), 1552-1562.

Teychenne, M., Ball, K., & Salmon, J. (2008). Physical activity and likelihood of depression in adults: a review. *Preventive medicine, 46*(5), 397-411.

Thomée, S., Härenstam, A., & Hagberg, M. (2011). Mobile phone use and stress, sleep disturbances, and symptoms of depression among young adults-a prospective cohort study. *BMC public health, 11*(1), 1-11.

Vailshery, L. S. (2021). Ownership of smartphones in the UK 2020. Statista. Retrieved from https://www. statista.com/statistics/956297/ownership-of-smartphones-uk/

Vallée, J., Cadot, E., Roustit, C., Parizot, I., & Chauvin, P. (2011). The role of daily mobility in mental health inequalities: the interactive influence of activity space and neighbourhood of residence on depression. *Social science & medicine, 73*(8), 1133-1144.

Van De Water, A. T., Holmes, A., & Hurley, D. A. (2011). Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography–a systematic review. *Journal of sleep research, 20*(1pt2), 183-200.

Verhoeven, J. E., Verduijn, J., Schoevers, R. A., van Hemert, A. M., Beekman, A. T., & Penninx, B. W. (2018). Complete recovery from depression is the exception rather than the rule: prognosis of depression beyond diagnostic boundaries. *Nederlands Tijdschrift Voor Geneeskunde, 162*.

Walker, W. H., Walton, J. C., DeVries, A. C., & Nelson, R. J. (2020). Circadian rhythm disruption and mental health. *Translational psychiatry, 10*(1), 1-13.

Wang, P. S., Simon, G., & Kessler, R. C. (2003). The economic burden of depression and the cost-effectiveness of treatment. *International journal of methods in psychiatric research, 12*(1), 22-33.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., . . . Campbell, A. T. (2014). *StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones.* Paper presented at the Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing.

Wells, K. B., Hays, R. D., Burnam, M. A., Rogers, W., Greenfield, S., & Ware, J. E. (1989). Detection of depressive disorder for patients receiving prepaid or fee-for-service care: results from the Medical Outcomes Study. *Jama, 262*(23), 3298-3302.

Wells, K. B., Miranda, J., Bauer, M. S., Bruce, M. L., Durham, M., Escobar, J., . . . Horwitz, S. M. (2002). Overcoming barriers to reducing the burden of affective disorders. *Biological psychiatry, 52*(6), 655-675.

Weyerer, S., & Kupfer, B. (1994). Physical exercise and psychological health. *Sports Medicine, 17*(2), 108-116.

White, K. H., Rumble, M. E., & Benca, R. M. (2017). Sex differences in the relationship between depressive symptoms and actigraphic assessments of sleep and rest-activity rhythms in a population-based sample. *Psychosomatic medicine, 79*(4), 479.

World Health Organization. (2017). Depression and other common mental disorders: Global health estimates. World Health Organization. https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf

Xu, X., Liu, X., Zhang, H., Wang, W., Nepal, S., Sefidgar, Y., . . . Morris, M. E. (2023). GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 6(4), 1-34.

Yue, C., Ware, S., Morillo, R., Lu, J., Shang, C., Bi, J., . . . Wang, B. (2018). Fusing location data for depression prediction. *IEEE Transactions on Big Data, 7*(2), 355-370.

Zapata-Lamana, R., Lalanza, J. F., Losilla, J.-M., Parrado, E., & Capdevila, L. (2020). mHealth technology for ecological momentary assessment in physical activity research: a systematic review. *PeerJ, 8*, e8848.

Zhang, Y., Yang, Z., Lan, K., Liu, X., Zhang, Z., Li, P., . . . Pan, J. (2019). *Sleep stage classification using bidirectional lstm in wearable multi-sensor systems.* Paper presented at the IEEE INFOCOM 2019-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS).

Chapter 2

Long-term Participant Retention and Engagement Patterns in an App and Wearable-based Multinational Remote Digital Depression Study

This chapter is adapted from:

Zhang Y, Pratap A, Folarin AA, Sun S, Cummins N, Matcham F, Vairavan S, Dineley J, Ranjan Y, Rashid Z, Conde P, Stewart C, White KM, Oetzmann C, Ivan A, Lamers F, Siddi S, Rambla CH, Simblett S, Nica R, Mohr DC, Myin-Germeys I, Wykes T, Haro JM, Penninx BWJH, Narayan VA, Annas P, Hotopf M, Dobson RJB, and RADAR-CNS Consortium. Long-term participant retention and engagement patterns in an app and wearable-based multinational remote digital depression study. *Npj Digital Medicine, 6*(1), 25.

Background: Recent growth in remote studies has shown the effectiveness of digital health technologies in recruiting and monitoring the health and behavior of large and diverse populations of interest in real-world settings. However, retaining and engaging participants to monitor their long-term health trajectories has remained a significant challenge. Uneven participant engagement combined with attrition over the course of the study could lead to an imbalanced study cohort and data collection, which may severely impact the generalizability of real-world evidence.

Objective: To investigate long-term participant retention and engagement patterns, we performed a secondary analysis of a large multinational real-world dataset from an observational study for major depressive disorder.

Methods: The data was collected from 614 participants using Android smartphones and Fitbit devices, including three data streams: Phone-Active (surveys), Phone-Passive, and Fitbit-Passive data streams. Survival analyses (Kaplan-Meier curves and Cox Proportional-Hazards models) and unsupervised clustering (K-means) were used to explore participant retention and longitudinal engagement patterns, respectively.

Results: Retention analysis revealed that a considerable proportion of participants (54.6%, 47.7%, and 67.6% for three data streams, respectively) were retained during the first 43 weeks of the study. The unsupervised clustering identified three distinct subgroups with different engagement levels (most, medium, and least) for each data stream. Notable findings comparing participants' characteristics across these subgroups were: 1) Participants in the least engaged group had the highest depression symptom severity (up to 4 points higher PHQ-8 score, p < .01) compared to participants in the other two subgroups across all three data streams. 2) For the Phone-Active data, participants in the least engaged group (N=204; 33.2%) on average completed 4 bi-weekly surveys in comparison to participants in the most engaged group (N=231; 37.6%) who on average completed 20 bi-weekly surveys. The least engaged group also took significantly longer to respond to (3.8 hours more, p < .001) and complete (11.3 seconds more, p < .001) surveys and were younger (age difference = 5 years, p < .01) compared to the most engaged group. 3) A considerable proportion of participants (44.6%) in the least engaged group (completed 4 bi-weekly surveys) of the Phone-Active data stream still contributed the Fitbit-Passive data for an average of 42 weeks.

Conclusions: Our findings show various factors, such as sociodemographics,

depression severity, and day-to-day study app usage behavior, could be linked to participant retention and engagement in fully remote research studies. In particular, passive data gathered from wearables without additional participant burden showed great data contiguity over the long term. However, an assessment of daily data collection patterns revealed that participants with higher depression severity are less likely to engage with a study app in a remote study. Further research is needed to understand the sociotechnical and human factors for digital mental health tools to engage research participants, particularly those with higher disease severity. The data-driven findings related to participant engagement in real-world settings could inform the design of future remote digital health research studies to enable equitable and balanced health data collection from diverse target populations.

Please refer to Appendix A for supplementary material.

2.1 Introduction

To gain valuable insights into the etiology of depression and identify effective treatments tailored to individuals, large diverse cohort-based studies are required to assess the underlying temporal patterns in real-world risk and protective factors of depression in individuals (Cai, Choi, & Fried, 2020; Klasen et al., 2015). However, dynamic day-to-day changes in behavior in naturalistic settings are not captured effectively by conventional clinical assessments that rely on infrequent in-person assessments and subjective retrospective reporting of symptoms (Snyder & Zhou, 2019). Additionally, reaching and recruiting a large and diverse cohort in a cost-effective and timely manner continues to be challenging for conventional clinical studies (Gilchrist & Gunn, 2007).

Due to increasing ubiquity and cost-effectiveness, smartphones and wearable devices, compared to medical devices, allow researchers to monitor personalized daily behaviors and physiology over time for a large and diverse population (Bailon et al., 2019; Bardram & Matic, 2020; Liew, Wah, Shuja, & Daghighi, 2015). Combined with scalable data collection platforms, these technologies provide high-fidelity multimodal behavior sensing capabilities (Ranjan et al., 2019). Several recent large-scale remote digital depression studies have shown the feasibility of technology-based remote data collection with real-world behaviors (Cho et al., 2016; Luik et al., 2015; Matcham et al., 2019; Pratap et al., 2018). For example, sleep (Zhang et al., 2021a), social interactions (Zhang et al., 2021b), and mobility (Laiou et al., 2022; Zhang et al., 2022) features derived from digital apps, smartphones, or wearable devices have been demonstrated to be significantly associated with depressive symptoms. Remote digital studies also offer an effective medium to reach and recruit from larger and more diverse populations (Moore, Tassé, Thorogood, Winship, & Doerr, 2017) thereby considerably lowering the costs and time for creating cohorts of interest than conventional clinical studies (Pratap et al., 2018).

Although previous remote digital studies have shown the feasibility and utility of leveraging smartphones and wearable technology for assessing behavioral changes in naturalistic settings, long-term participant retention and engagement remain significant challenges (De Angel et al., 2022; Druce, Dixon, & McBeth, 2019). Moreover, differential recruitment and retention of participants can lead to imbalanced cohorts and biased data collection that can severely impact the generalizability of real-world evidence (O'connor et al., 2016; Pratap et al., 2020; Quisel, Foschini, Zbikowski, & Juusola, 2019; Simblett et al., 2018a). For example, Pratap et al. found that four specific indicators (referral by clinicians, older age of participants, compensation of participants, and having a clinical condition [as opposed to being healthy]) were significantly associated with participant retention, and participant demographics were also associated with long-term engagement patterns in a cross-study evaluation of eight observational digital health studies conducted between 2014–2019 (Pratap et al., 2020).

However, past studies investigated participant behavior and retention in the study for short follow-up periods and were primarily based on active tasks (surveys) completed by participants using a limited set of variables of interest (O'connor et al., 2016; Pratap et al., 2020; Quisel et al., 2019; Simblett et al., 2018a). To leverage digital health technology for assessing and managing complex chronic conditions (e.g., psychiatric and neurological disorders), gathering long-term real-world behavior is necessary. And to remotely engage large populations effectively and equitably, there is a further need to understand key risk factors that impact long-term participant engagement (months to years) in remote digital studies, including the feasibility of collecting active and passive data streams. Participants' behaviors of answering surveys via the study app, such as time spent responding to surveys and completing surveys in naturalistic settings, may also reflect the participants' interest early (Bassili, 1996; Fazio, Powell, & Herr, 1983; Heerwegh, 2003), which could be indicative of long-term engagement in the study. Furthermore, there is a need to understand the feasibility of collecting passive data via smartphones (e.g., Bluetooth and GPS data) and wearables (e.g., heart rate and sleep data) in comparison to active task-based data (e.g., surveys) requiring active participation and with additional user burden.

Here we present findings from a secondary analysis of data collected from the Remote Assessment of Disease and Relapse-Major Depressive Disorder (RADAR-MDD) study (Matcham et al., 2019; Matcham et al., 2021) to evaluate the real-world factors impacting long-term participant retention and engagement in a large, multinational cohort. Specifically, we assessed three specific key questions using participant-level usage data of study apps and wearables: A) Is participant retention associated with real-world factors (such as sociodemographics, medium of data collection [smartphones and wearables], and severity of depressive symptoms)? B) Are there potential patterns in participants' long-term engagement, including differences between active and passive data streams collected via the study apps and wearables? C) And if there are significant differences in participants' characteristics in the study across different long-term engagement patterns?

2.2 Methods

2.2.1 The RADAR-MDD Study Design

Data used in this study was collected from the EU research program, RADAR-MDD, which aimed to investigate the utility of smartphones and wearable devices to monitor depression in real-world settings and understand factors that could help predict relapse in MDD (Matcham et al., 2019). The study recruited 623 participants from 3 sites across 3 European countries (United Kingdom - King's College London [KCL]; Spain - Centro de Investigación Biomédican en Red [CIBER]; Netherlands - Vrije Universiteit Medisch Centrum [VUmc]) and followed participants for up to 2 years (Matcham et al., 2019). Nine participants recruited from a second site in Spain were not included in the present analysis due to the small sample size. All participants in this study were over 18 years old and had a history of recurrent MDD with at least one episode within the last 2 years that met DSM-5 diagnostic criteria for the diagnosis of MDD. Additionally, in order to be enrolled in the study, the participants were asked to use an Android smartphone as their primary phone if they had one, or were provided with one to use if they did not.

The study used the RADAR-base, an open-source platform, for smartphone-based health data collection via two Android study apps (active and passive monitoring apps) (Ranjan et al., 2019). Participants were asked to regularly complete self-reported surveys via the active app (Matcham et al., 2019). Additionally, participants' real-world behavior was gathered passively using the Android passive monitoring app and a Fitbit wearable (details below). The participants were also required to complete some clinical assessments via research electronic data capture (REDCap) surveys every 3 months. Participants' sociodemographics, medical history, lifetime depression history, and baseline mental health status were also collected during the participant enrollment session (Matcham et al., 2019). Although participants were not financially reimbursed for providing data via study apps and the Fitbit device, participants received £15/€20 for enrollment, £5/€10 for clinical assessments (REDCap surveys) every 3 months, and £10/€10 for every additional qualitative interview completed (Matcham et al., 2019). Furthermore, the "Human-in-the-loop" (Goodday et al., 2021) approach was used during the observation period. The research team contacted participants for various reasons, such as reminding clinical assessments, technical issues (e.g., Fitbit broken, problems in study apps, and phone issues), and congratulating participants on reaching key study milestones (e.g., one year in the study). The detailed study protocol and descriptions of the dataset have been reported by Matcham et al. (Matcham et al., 2019; Matcham et al., 2021).

The first participant was enrolled in November 2017 and the last participant was enrolled in June 2020, and the data collection was finished in April 2021 (Matcham et al., 2021). As a result of this rolling enrollment, the time in study for RADAR-MDD participants varies from 11 months to 24 months. There were temporal differences in participant recruitment across the three sites. The KCL site started participant recruitment first (November 2017), followed by the CIBER site (September 2018), and the VUmc enrolled participants later again (February 2019) (Matcham et al., 2021).

The RADAR-MDD protocol was co-developed together with a patient advisory board who shared their opinions on several user-facing aspects of the study, including the choice and frequency of survey measures, the usability of the study app, participant-facing documents, selection of optimal participation incentives, selection and deployment of wearable devices as well as the data analysis plan (Matcham et al., 2019; Simblett et al., 2019). All participants signed informed consent and the study had been approved by all local Ethics committees (Matcham et al., 2019).

2.2.2 Primary Data Streams

For evaluating long-term participant retention and engagement in the study, we classified the data collected by the study apps and Fitbit devices into three distinct categories: i) Phone active data - representing active tasks completed by participants via the study app, ii) Phone passive data - continuous data streams gathered by the smartphones without active input from participants, and iii) Fitbit passive data - continual physiological monitoring data collected through a wrist-worn Fitbit device during the observation period.

Phone Active Data

A variety of episodic surveys were administered via the study app. The complete list of surveys and deployment details are covered in the study protocol (Matcham et al., 2019). However, with the focus on present research evaluating long-term engagement, we considered the two longitudinal surveys, the 8-item Patient Health Questionnaire (PHQ-8 (Kroenke et al., 2009)) and the Rosenberg Self-Esteem Scale (RSES (Greenberger, Chen, Dmitrieva, & Farruggia, 2003)), which were conducted via smartphones remotely once every two weeks. The completion windows for PHQ-8 and RSES are both 3 days. Surveys could not be completed once the window expired. If the participants finished at least one of these two surveys, we considered they were engaging in the active assessment part of the study for the corresponding 2 weeks.

Phone Passive Data

The passive monitoring app unobtrusively and continuously collected information on participants' phone usage (e.g., battery level logs, app usage logs, and phone interaction data) and surrounding information (e.g., ambient light, nearby Bluetooth device count, and GPS location data) (Matcham et al., 2019). Following the phone passive data availability definition presented in (Matcham et al., 2021), we considered a participant to be using their study phone and sharing the phone passive data on a given day if at least one passive data point was collected from their smartphone during the day.

Fitbit Passive Data

Participants were also required to wear a Fitbit Charge 2 or 3 wrist-worn during the follow-up time to provide passive measures of their sleep stages, steps, calorie consumption, and heart rate. Matching the phone passive data availability definition, if at least one data point from the Fitbit-based data stream was available, we considered the participant to be wearing the Fitbit at least once during that given day.

2.2.3 Primary Outcomes

Metrics of Engagement

We defined two key metrics to assess the participant's engagement. (1) Duration in the study: the number of days between the first and last day of data contributed by the participant in a selected engagement observation period. (2) Longitudinal data-availability vector: a binary-encoded vector representing the density of the participant's contributed data in an engagement observation period, where the i-th element of the vector represents the i-th day in the study and is set to 1 if a data point is contributed by the participant on that day or is set to 0 otherwise. To align the frequency of passive data streams (daily), for the Phone-Active data, we set the 14 elements (2-week period) of the data-availability vector to 1 in which a survey was completed by a participant. We calculated these two metrics of engagement for each of the three data streams (Phone-Active, Phone-Passive, and Fitbit-Passive data), respectively.

Variables of Interest

A variety of factors may affect the duration and density of participants' engagement in remote digital studies (Baumel, Muench, Edan, & Kane, 2019; Pratap et al., 2020; Torous, Lipschitz, Ng, & Firth, 2020). In this engagement study, we considered a variety of factors, including participants' sociodemographics, recruitment study site, smartphone brand, baseline depression symptom severity, comorbidity, and depression medication, as well as app usage behavior (survey response time and survey completion time) as variables of interest. These are briefly described below.

Sociodemographics: Age, gender, ethnicity (not collected at the CIBER site), education, marital status, income, and accommodation type were recorded in the enrollment session.

Study site: Participant recruitment site (KCL, CIBER, and VUmc).

Smartphone brand: The brand of the participant's smartphone used in the study was also recorded in the enrollment session.

Baseline depression symptom severity: Depressive symptom severity was estimated by the PHQ-8 survey administered through the study app at the time of enrollment and every subsequent two weeks. The PHQ-8 contains 8 questions, and the total score of the PHQ-8 ranges from 0 to 24 with increasing severity of depressive symptoms (Kroenke et al., 2009). We considered the PHQ-8 surveys completed at enrollment to represent the participants' baseline depression severity.

Comorbidity and Medication: The participant’s comorbidity information related to 19 types of common comorbidities (listed in Supplementary Table 11, Appendix A) was recorded in the enrolment section. Also, participant use of depression medication was recorded at enrollment. For the present analysis, we used a binary variable to indicate whether the participant had comorbidities and whether they were taking depression medication at the time of enrollment.

Survey response and completion time: Survey response time is calculated as the time that elapsed between the notification arrival time in the study app and the time at which the participant started responding to the survey. Survey completion time was the total time the participant spent completing the survey. Several studies suggested that the response time and the speed of answering questions could reflect the participants’ attitude to the survey (Bassili, 1996; Fazio et al., 1983; Heerwegh, 2003). Therefore, we used these two metrics to reflect participants’ interests and enthusiasm about the study and test whether they are linked to long-term engagement patterns. Both metrics were calculated for the two surveys (PHQ-8 and RSES).

2.2.4 Survival Analysis

We used a survival modeling approach (Bewick, Cheek, & Ball, 2004) to assess participants' overall duration in the study (retention). Survival models are commonly used in medical research for exploring associations between the time passed before some events occur and one or more predictor variables (Singer & Willett, 1991). The survival models were also used in a recent participant retention study (Pratap et al., 2020). In our participant retention analysis, the event is the participant disengaging from the study app (stopping contributing data to the study) and the elapsed time is the duration in the study (described above).

Data Harmonization

We conducted our survival analysis on two separate observation periods. The first, referred to as our primary cohort, has an observation period of 43 weeks. This period matches the number of weeks between the last patient enrolled in the RADAR-MDD (June 2020) and the end of data collection in the RADAR-MDD (April 2021). Therefore, it represents the common maximum theoretical survival observation period for all participants enrolled in the RADAR-MDD study. We used this cohort for the presented primary analysis. We also defined a secondary cohort with a survival observation period of 94 weeks. This longer period of observation represents the maximum survival observation period for 50% of participants enrolled in the RADAR-MDD. Using this secondary cohort, we aimed to investigate even longer-term participant behavior patterns in remote studies.

Participant Retention Analysis

We first used Kaplan-Meier curves (Rich et al., 2010) to measure the overall participant retention rates over the two observation periods for three data streams, respectively. To further assess the joint effect of multiple variables of interest on participants' retention in the study, we used the Cox Proportional-Hazard (CoxPH) model (Kumar & Klefsjö, 1994). We considered the baseline PHQ-8 score, comorbidity, depression medication, sociodemographics (age, gender, marital status, children, years in education, annual income, and accommodation type), study site, and phone brand as predictor variables. If the duration in the study of a participant is equal to the cutoff observation period, we consider the participant to be engaged in the study (no event). To minimize undue influence associated with periodic disengagement (i.e., some participants stop engaging for a while, then re-engage), the right-censoring method (Rich et al., 2010) was used for participants whose duration in the study was less than the observation period. We relaxed the determination of the event by considering 4 more weeks after the cut-off day. For example, if a participant's last active survey was completed on Week 30 within the first 43 weeks (using the primary cutoff observation period), but if they completed more active surveys between Week 44–Week 47 (4-week extension), we still considered this participant engaged in contributing active data to the study (no event). Otherwise, if there was no completed survey during the 4 weeks after the cut-off day, we considered this participant stopped contributing active data to the study (the event happened i.e participant stopped contributing Phone-Active data to the study). Note, the same methodology was used to counter periodic disengagement in the Phone-Passive and Fitbit-Passive data. To assess the joint effect of multiple variables of interest on retention, we used separate CoxPH models for Phone-Active, Phone-Passive, and Fitbit-Passive data across the two observation periods (43 weeks and 94 weeks). The assumptions of proportional hazards (Kleinbaum & Klein, 2012), which means that relative hazard remains constant over time with different predictors, of all CoxPH models were tested using the scaled Schoenfeld residuals (Grambsch & Therneau, 1994).

2.2.5 Clustering Analysis

Long-term Participant Engagement Pattern Modeling

We used an unsupervised K-means clustering method (Syakur, Khotimah, Rochman, & Satoto, 2018) to explore potential latent patterns of participant long-term engagement in the study using the longitudinal data-availability vector (defined above). The elbow method was used to determine the optimal number of clusters (Syakur et al., 2018). The Kruskal-Wallis test was used to assess any potential enrichment of variables of interest (described above) across the clusters (Ostertagova, Ostertag, & Kováč, 2014). The same approach was applied to the three data streams and across the two observation periods. Transitions of participants in clusters across the three data streams were recorded and visualized by Sankey diagrams (Schmidt, 2008).

2.3 Results

2.3.1 Cohort Characteristics

In total, we analyzed data from 614 participants recruited from three recruitment sites (350, 146, and 118 participants from KCL, CIBER, and VUmc, respectively) between November 2017 and April 2021. The cohort's median (range) age was 49 (18 – 80) years; Supplementary Figure 1 of Appendix A shows the age distribution. The majority of the cohort are females (75.7%, N=465), which is expected because all enrolled participants had a current or prior history of depression, and the prevalence of depression is known to be higher in females than in males (Albert, 2015; Noble, 2005; Salk, Hyde, & Abramson, 2017; Van de Velde, Bracke, & Levecque, 2010). Differences in participant characteristics across study sites were assessed by Kruskal-Wallis tests (Ostertagova et al., 2014). Participants recruited at the CIBER site had the highest median age (54.0 [49.0, 61.0] years) across the three sites (KCL: 45.0 [30.0, 56.0] years and VUmc: 40.0 [26.0, 57.8] years) (p < .001). In addition, the CIBER site cohort also had a significantly higher median baseline PHQ-8 score (15.5 [10.0, 19.0]) than the KCL (9.0 [6.0, 13.0] scores) and VUmc (8.0 [6.0, 14.0] scores) sites (p < .001). For ethnicity, the majority of recruited participants were white across KCL (84.3%) and VUmc (92.4%) sites. Ethnicity data was not collected for participants recruited at the CIBER site. Table 2.1 summarizes sociodemographic and clinical characteristics for the overall cohort with comparisons stratified by sites. Briefly, the subcohort (secondary cohort) with a longer observation period (94 weeks) (See Methods) had 313 participants with a median age of 51.0 [37.0, 59.0] years, with the majority being females (75.1%, N=235). The full set of secondary cohort descriptive statistics is summarized in Supplementary Table 1 of Appendix A.

2.3.2 Participant Retention

For the primary cohort analysis, the participant retention (survival rate) at the end of the common maximum observation period of 43 weeks (described in Methods) as quantified using Phone-Active, Phone-Passive, and Fitbit-Passive data streams were

54.6% (N=335), 47.7% (N=293), and 67.6% (N=415), respectively. Similarly, for the secondary cohort, the participant retention rates in the 94 weeks measured by the three data streams were 48.2% (N=151), 39.3% (N=123), and 54.0% (N=169), respectively. Figure 2.1 displays the Kaplan-Meier survival curves that show participant retention across two observation periods stratified by three data streams.

Figure 2.1. The Kaplan-Meier survival curves for (a) the primary cohort (N=614) with an observation period of 43 weeks, and (b) the secondary cohort (N=313) with a longer observation period of 94 weeks stratified by Phone-Active, Phone-Passive, and Fitbit-Passive data streams.

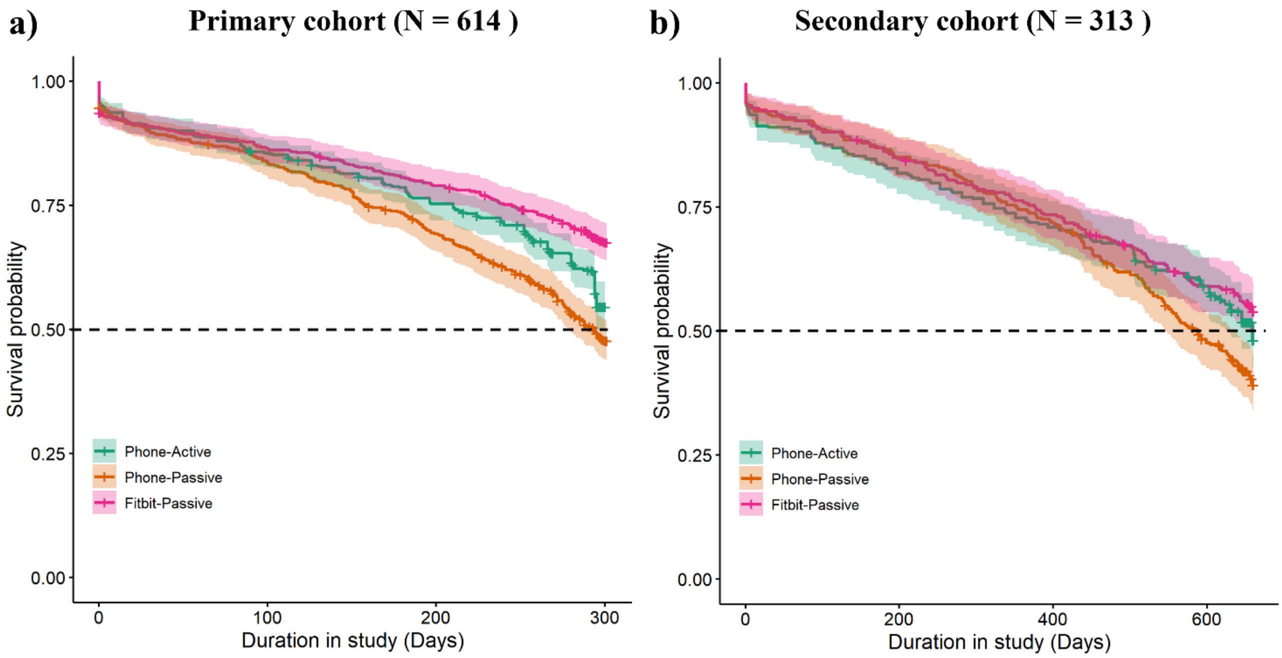


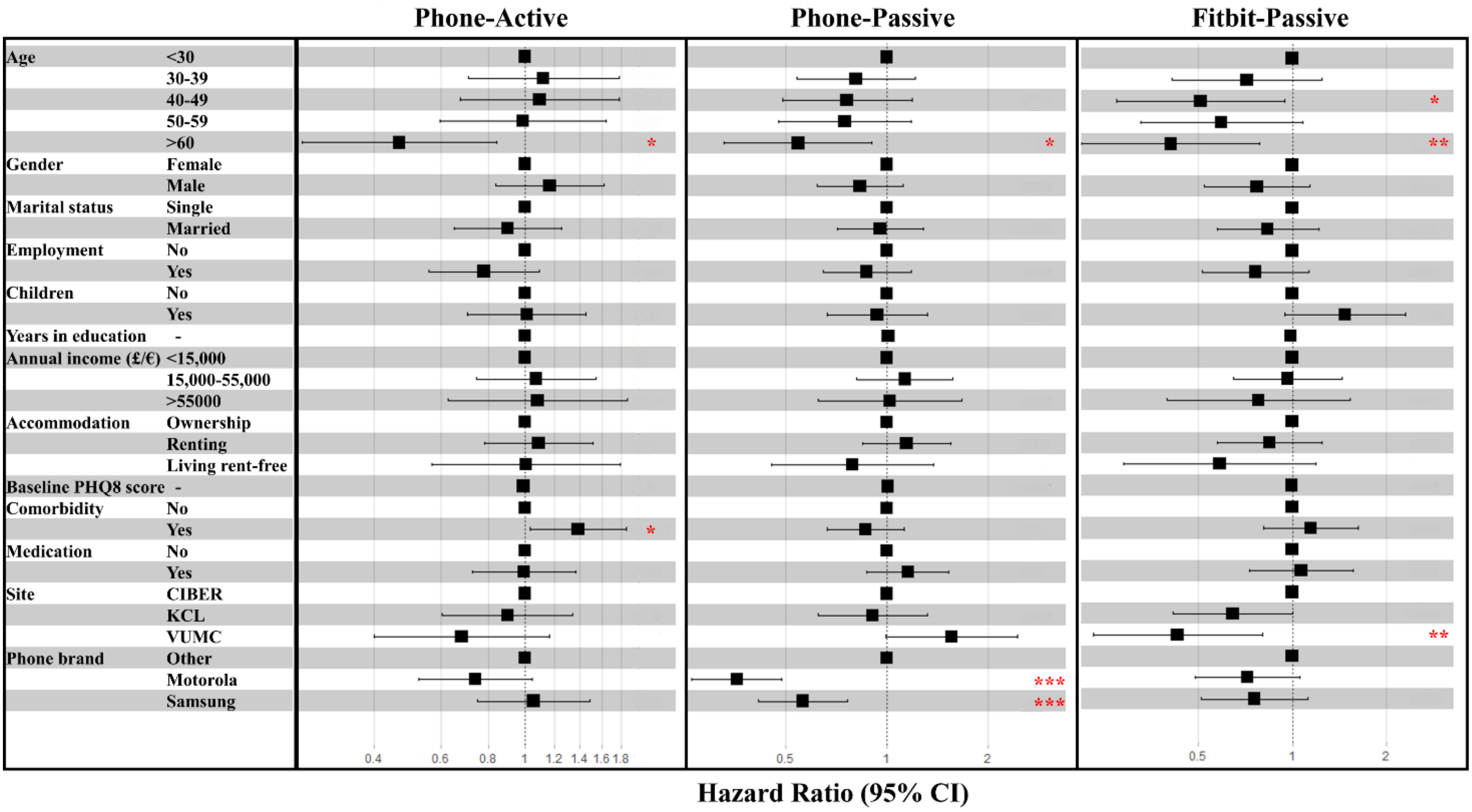
Table 2.1. A summary of characteristics of 614 participants in the RADAR-MDD study, with comparisons across the three study sites.

Characteristics	Total	KCL	CIBER	VUmc	P value
Number of participants, n	614	350	146	118	
Age (median [IQR])	49.00 [32.00, 58.75]	45.00 [30.00, 56.00]	54.00 [49.00, 61.00]	40.00 [26.00, 57.75]	<.001
Female, n (%)	465 (75.7)	267 (76.3)	106 (72.6)	92 (78.0)	.56
Marital status, n (%)					.005
Single/separated/divorced/widowed	328(53.4)	185(52.9)	66 (45.2)	77 (65.3)	
Married/cohabiting/LTR	286 (46.6)	165 (47.1)	80 (54.8)	41 (34.8)	
Ethnicity, n (%)					<.001
White	404 (86.3)	295 (84.3)	-	109 (92.4)	
Black	14 (3.0)	11 (3.1)	-	3 (2.5)	
Asian	16 (3.4)	16 (4.6)	-	0 (0)	
Other	34 (7.3)	28 (8.0)	-	6 (5.1)	
Employed, n (%)	258 (42.0)	186 (53.1)	33 (22.6)	39 (33.1)	<.001
Having children, n (%)	304(49.5)	152 (43.4)	111(76.0)	41 (34.8)	<.001
Years in education (median [IQR])	16.00 [13.00, 19.00]	17.00 [14.00, 19.00]	11.00 [9.00, 15.75]	16.50 [14.00, 20.00]	<.001
Annual income, n (%)					<.001
<15,000 (£/€)	152 (24.8)	74 (21.1)	47 (32.2)	31 (26.3)	
15,000-55,000 (£/€)	348 (56.7)	203 (58.0)	92 (63.0)	53 (44.9)	
>55000 (£/€)	98 (16.0)	72 (20.6)	7 (4.8)	19 (16.1)	
Accommodation, n (%)					<.001
Own outright/with mortgage	323 (52.6)	169 (48.3)	105 (71.9)	49 (41.5)	
Renting	236 (38.4)	151 (43.1)	27 (18.5)	58 (49.2)	
Living rent-free	46 (7.5)	29 (8.3)	10 (6.8)	7 (5.9)	
Baseline PHQ-8 score (median [IQR])	10.00 [7.00, 16.00]	9.00 [6.00, 13.00]	15.50 [10.00, 19.00]	8.00 [6.00, 14.00]	<.001
Having comorbidities, n (%)	311 (50.7)	176 (50.3)	96 (65.8)	39 (33.1)	
Taking depression medication, n (%)	400 (65.1)	206 (58.9)	133 (91.1)	61 (51.7)	
Number of contact logs (median [IQR])	4.00 [2.00, 7.00]	5.00 [3.00, 8.00]	3.00 [2.00, 5.00]	2.00 [1.00, 3.75]	<.001
Brand of smartphone, n (%)					<.001
Motorola	240 (39.7)	171 (49.7)	30 (20.8)	39 (33.3)	
Samsung	194 (32.1)	94 (27.3)	44 (30.6)	56 (47.9)	
Other	171 (28.3)	79 (23.0)	70 (48.6)	22 (18.8)	

To further assess the impact of multiple variables of interest (age, gender, marital status, employment, children, education, income, accommodation, the baseline PHQ-8 score, comorbidity, depression medication, smartphone brand, and study site), we used multivariate Cox Proportional-Hazards models (Kumar & Klefsjö, 1994). For each covariate, a hazard ratio (HR) greater than 1 indicates the variable is associated with a higher risk of participants not contributing data to the study, thus negatively impacting participant retention in the study. Across the three data streams, age was found to significantly affect participant retention in the study. Compared with the youngest group (18 – 30 years old), participants in older age groups tend to stay in the study for a longer time. Participants in the oldest group (>60 years old) had the lowest risks of stopping contributing data for all three data streams (Phone-Active: HR = 0.47, p < .05; Phone-Passive: HR = 0.54, p < .05; Fitbit-Passive: HR = 0.41, p < .01). Further, for Phone-Active data, participants with comorbidities had the higher risk (HR = 1.38, p < .05) for leaving the study early than participants with no comorbidities. Of note, participants using Motorola (HR = 0.36, p<.001) and Samsung (HR = 0.56, p < .001) branded phones contributed Phone-Passive data for significantly longer durations compared with other brands of smartphones. Further, participants in the VUmc site had the lowest risk of stopping sharing the Fitbit-Passive data (HR = 0.43, p < .01).

All three Cox Proportional-Hazards models met the global proportional hazards assumption tested using the scaled Schoenfeld residuals (Grambsch & Therneau, 1994). See Supplementary Tables 2 and 3 of Appendix A for proportional hazard assumption tests for each of the variables and the global models. Figure 2.2 shows the hazard ratio plots of the three data streams for the primary cohort. Participants' age also continued to significantly impact retention in the extended observation period (94 weeks) across all three data streams assessed in the secondary cohort (Supplementary Figure 2, Appendix A).

Figure 2.2. The hazard ratio plots of Cox Proportional-Hazards models for assessing the impact of multiple variables of interest on the participant retention time in the study of the primary cohort (43-week observation period) for Phone-Active, Phone-Passive, and Fitbit-Passive data streams, respectively. For each, a hazard ratio greater than 1, indicates the variable is positively associated with the risk of an event (disengaging from the study, i.e., stopping contributing data), thus negatively associated with the length of participant retention time in the study. Significance levels: < .05 *, p < .01 **, and p < .001 ***.

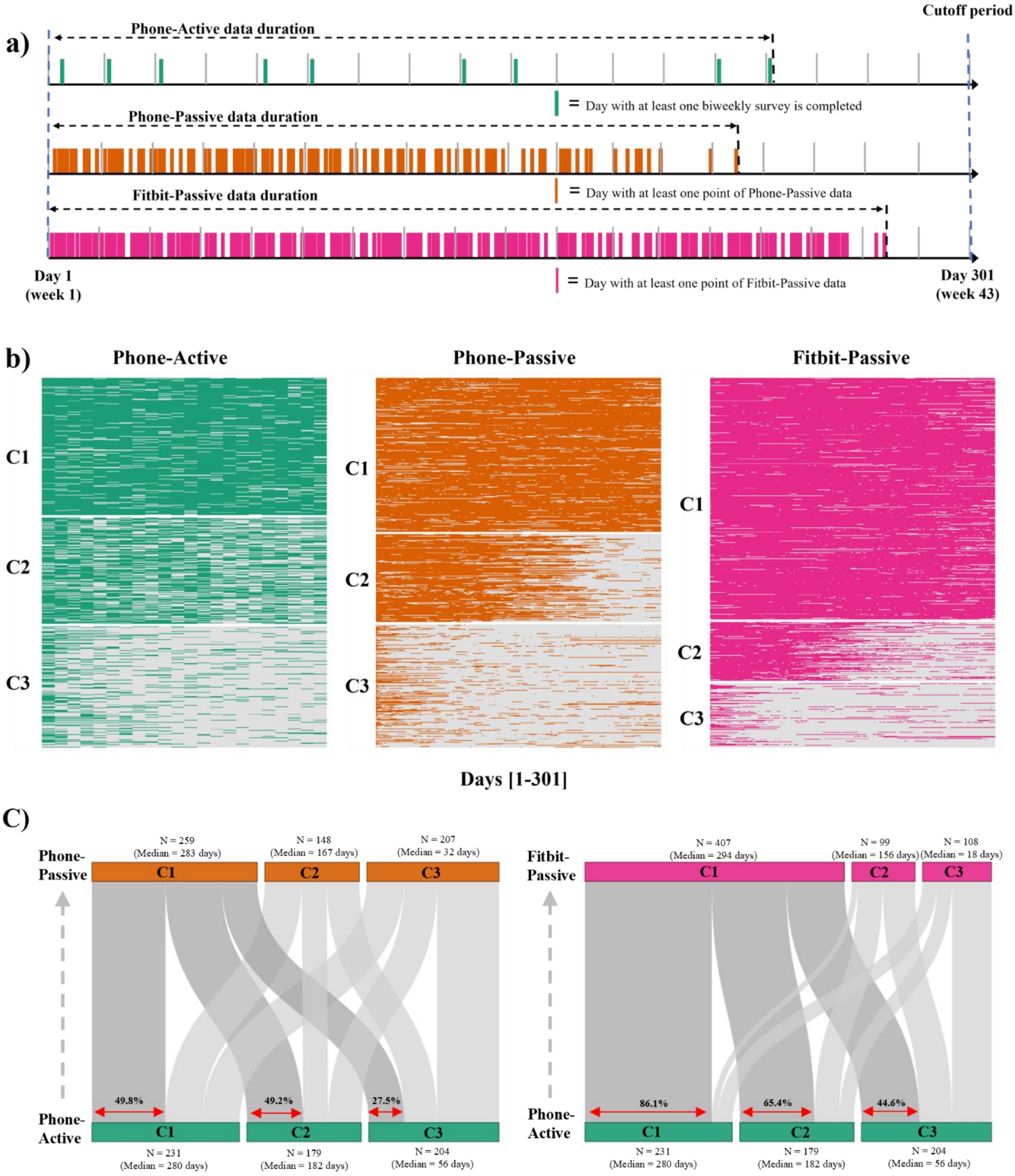


2.3.3 Participants' Long-term Engagement Patterns in the Study

Patterns in participants' day-to-day data sharing were assessed using an unsupervised K-means method (Wu, 2012) across the three data streams separately (Figure 2.3a). In the primary observation period, three subgroups showing distinct participant engagement patterns (C1: most engaged, C2: medium engaged, and C3: least engaged) emerged across each data stream (Figure 2.3b). Across the three engagement clusters in each data stream (Phone-Active, Phone-Passive, and Fitbit-Passive), we found notable differences in participants' behavior (survey response and completion times), baseline depression symptom severity, and age (Figure 2.4). Supplementary Tables 4-6 of Appendix A provide further details for comparisons of all variables across three clusters for all three data streams using the Kruskal-Wallis tests.

a) Long-term engagement patterns across active, passive, and wearable data streams: Participants in the most engaged C1 cluster (37.6% of the cohort; N=231) completed a median (IQR) of 20.0 (18.0, 21.0) bi-weekly surveys as opposed to 4.0 (1.0, 6.0) for those in the least engaged cluster (C3; 33.2% of the cohort; N=204). Similarly, the data sharing patterns for passive data streams showed significant differences as well. Participants (42.2% of the cohort; N=259) in the most engaged C1 cluster of the Phone-Passive data stream, shared phone-based passive data for a median (IQR) of 283 (257.0, 298.0) days as opposed to 32 (4.0, 67.5) days for the participants in the least engaged C3 cluster (33.7%; N=207). Similarly, for the Fitbit-based data gathered passively, the most engaged C1 cluster with 66.29% participants (N=407) shared the data for median (IQR) 294 (274.0, 301.0) days compared to just 18 (0, 67.0) days for the 17.6% participants (N=108) in the least engaged cluster (C3). Of note, we found a considerable proportion of participants in the medium (C2) and least (C3) engaged clusters of the Phone-Active data stream, despite completing a lesser number of active surveys (13 and 4 bi-weekly surveys, respectively), continued contributing passive data from Fitbit for an average of 42 weeks. Figure 2.3c shows this marked transition where 65.4% of participants (N=151) from the C2 cluster and 44.6% of participants (N=91) from the C3 cluster, based on the Phone-Active data stream, transitioned to the most engaged C1 cluster of the Fitbit-Passive data stream.

Figure 2.3. Participant longitudinal engagement patterns in the RADAR-MDD Dataset. a) Schematic representation of a participant’s 3 data streams in the study. b) Heatmaps of participant longitudinal engagement patterns, clustered using K-means clustering. In each heatmap, each row represents a data-availability vector of one participant (described in Methods), and subgroups were arranged from the most engaged cluster to the least engaged cluster (C1-C3). c) Sankey plots showing the proportion of common participants between clusters determined from Phone-Active (green), Phone-Passive (brown), and Fitbit-Passive (pink) data streams. To match passive data streams, if a survey that was due every two weeks was completed by a participant, 14 elements of the participant’s data-availability vector of Phone-Active corresponding to these two weeks are set to 1 (representing the participant was contributing active data).



b) Survey response and completion times: We also observed prominent linkages between long-term engagement and the survey response time (the time to respond to survey notifications) and completion time (total survey completion time). Participants in the most engaged C1 cluster of the Phone-Active data stream had significantly shorter survey response time (73.7 [31.3, 215.8] minutes) for the PHQ-8 survey compared to 302.4 (122.3, 527.1) minutes for the least engaged C3 cluster (Figure 2.4a) (p<.001). This finding is also consistent for subgroups in the Fitbit-Passive data stream (Figure 2.4a) and the RSES survey (Supplementary Table 4 and Supplementary Table 6, Appendix A). In terms of survey completion time, participants in the least engaged cluster (C3) of the Phone-Active data stream took significantly longer (61.6 [46.1, 83.0] seconds) to complete surveys than those in C1 (50.3 [37.9, 69.0] seconds) and C2 (49.4 [40.0, 67.0] seconds) clusters (Figure 2.4b) (p < .001). Likewise, the finding of survey completion time is consistent for the Fitbit-Passive data stream (Figure 2.4b) and the RSES survey (Supplementary Table 4 and Supplementary Table 6, Appendix A).

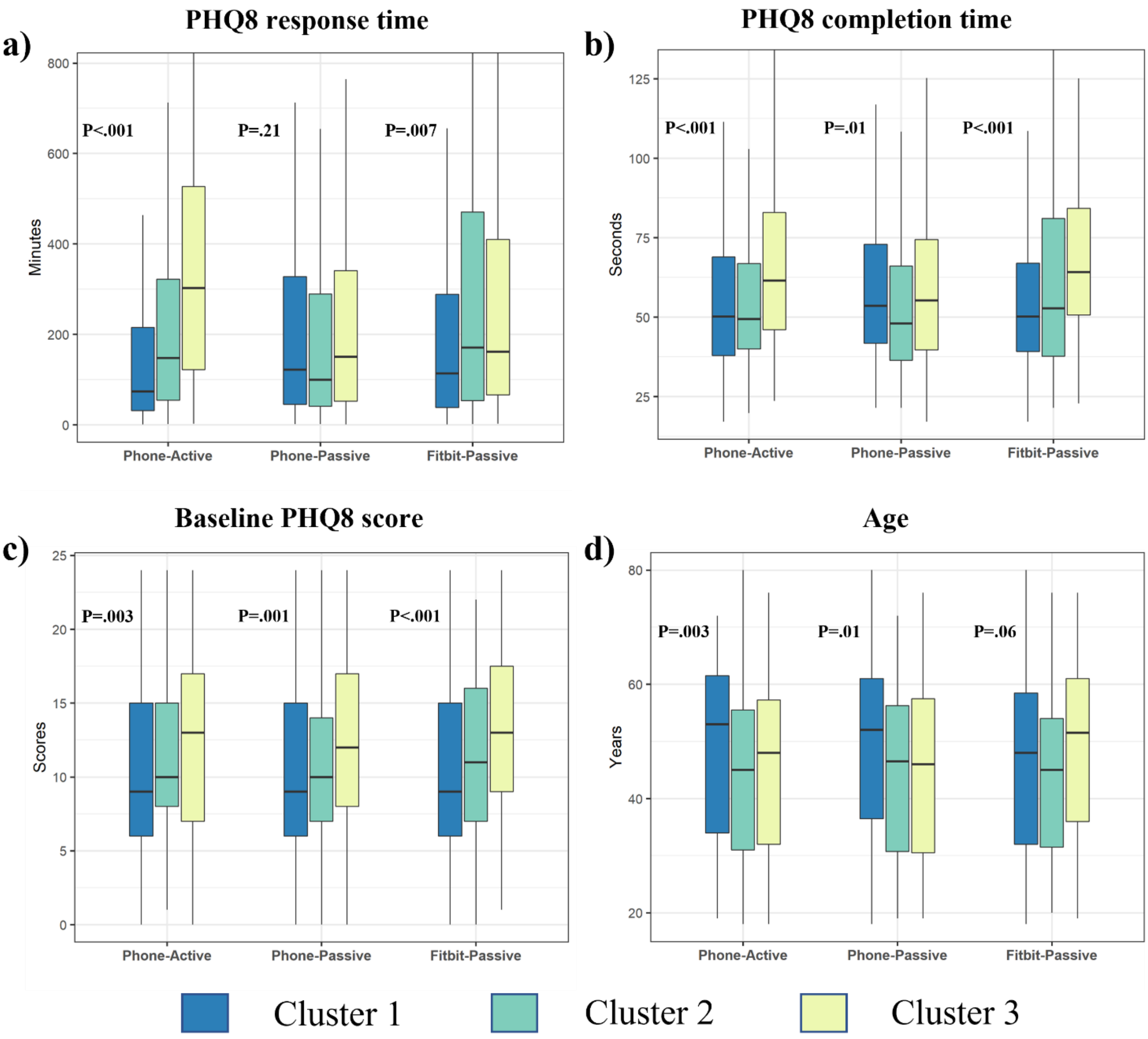
c) Baseline depression symptom severity: The baseline PHQ-8 scores of participants were significantly different across three subgroups (C1, C2, C3) for all three data streams. Overall, participants in the least engaged cluster (C3) had a significantly higher severity of depressive symptoms at enrollment (Figure 2.4c). For example, participants in C3 for the Phone-Active data stream had a 4 points difference in the median baseline PHQ-8 score (13.0 [7.0, 17.0]) compared to participants in the most engaged cluster (C1) with a median baseline PHQ-8 score of 9.0 (6.0, 15.0) (p < .01). Similarly, in participants in cluster C3 of Phone-Passive and Fitbit-Passive data streams showed a statistically significant difference in the baseline PHQ-8 scores compared with the most engaged cluster (C1) (Phone-Passive - C1: 9 [6.0, 15.0]; C3: 12 [8.0,17.0] and Fitbit-Passive - C1: 9 [6.0, 15.0]; C3: 13 [9.0, 17.5]) (p < .001).

d) Sociodemographics: The age of participants was significantly different across the 3 clusters of Phone-Active and Phone-Passive data streams. For the Phone-Active data stream, participants in C1 cluster had a significantly higher median (IQR) age of 53.0 (34.0, 61.5) years than participants in C2 (45.0 [31.0, 55.5]) and C3 (48.0 [32.0, 57.3]) clusters (p < .01). Similarly, for the Phone-Passive data stream, participants in the most active C1 cluster had the significantly highest median (IQR) age of 52.0 (36.5, 61.0) years across the 3 clusters (C2: 46.5 [30.8, 56.3] years and C3: 46.0 [30.5, 57.5] years) (p < .05). For ethnicity (available for KCL and VUmc sites), we found the proportion of white participants was significantly lower in the least engaged C3 group (77.8%) than C1 (95.1%) and C2 (84.0%) clusters for the Phone-Active data (p <.001). Likewise, Phone-Passive and Fitbit-Passive data had similar findings (Supplementary Table 10, Appendix A).

e) Phone brand and “human-in-the-loop” (research team contacting participants): Similar to the results of participant retention analysis, we found the Phone-Passive data collection to be significantly different across the smartphone brands. In the Phone-Passive data stream, the proportion of participants with Motorola brand phones in the least engaged C3 cluster (15%) was significantly lower than C1 (57.0%) and C2 (42.9%) (p < .001) (Supplementary Table 5, Appendix A). Further, for Phone-Active data stream, we found participants in the most engaged C1 cluster were contacted less frequently

(3.0 [2.0, 5.0]) than those in the C2 (5.0 [3.0, 7.0]) and C3 (5.0 [2.0, 9.0]) clusters (p < .001) (Supplementary Table 4, Appendix A).

Figure 2.4. Significant differences in participants' survey response time and survey completion time) baseline depression symptom severity (PHQ-8 score), and age across three long-term engagement patterns (Cluster 1, Cluster 2, and Cluster 3) for Phone-Active, Phone-Passive, and Fitbit-Passive data stream, respectively. Note: Cluster 1, Cluster 2, and Cluster 3 represent the most engaged, medium engaged, and least engaged patterns shown in Figure 2.3b.



For the secondary cohort with a longer observation period, unsupervised clustering of 94 weeks of individual-level engagement data showed 4 clusters (C1-C4) shown in Supplementary Figure 2.4 of Appendix A. The results of the participant characteristics enriched in the 4 engagement clusters for the secondary cohort are similar to the results of the primary cohort and are summarized in Supplementary Tables 7-9 of Appendix A for the three data streams, respectively.

2.4 Discussion

2.4.1 Principal Findings

We report findings from a novel investigation into long-term participant retention and engagement patterns from a large European multinational remote digital study for depression (Matcham et al., 2019; Matcham et al., 2021). Our findings show a significantly higher long-term participant retention than in past remote digital health studies (Druce et al., 2019; O'connor et al., 2016; Pratap et al., 2020; Quisel et al., 2019; Simblett et al., 2018a). However, we show several factors, that can significantly impact long-term participant retention and the density of real-world data collection. These range from participants' sociodemographics, and depression symptom severity, to study app usage behavior e.g., survey response, and completion times. Here we contextualize our key findings in the broader digital medicine context that may help inform the design and development of remote digital studies. We also compare the utility of using active and passive data collection for long-term remote monitoring of behavior and health outcomes. Finally, we share some of the participant engagement strategies deployed by the RADAR-MDD consortium (Matcham et al., 2019) and data-driven insights to help improve long-term participant engagement in future remote digital studies.

Key Learnings from Long-term Participant Engagement and Retention in the RADAR-MDD Study

One of the notable findings was that participants with higher severity of depression at the time of enrollment contributed less data both actively and passively. For example, participants in the least engaged cluster (C3) had the highest depression severity at the baseline and were up to 16 times less likely to share active or passive data from smartphones and wearables. The finding indicates that participants with higher depression symptom severity may be at a higher risk of not engaging in fully remote studies. A similar finding that the lowest engaged group had the highest depression and anxiety scores was observed in a previous web-based mental health intervention study (Chien et al., 2020). Non-uniform engagement in depression study apps, particularly by participants with higher depression severity, could bias the real-world data collection, impacting the generalizability and robustness of generated evidence. There is an urgent need for future research to develop solutions that alleviate non-uniform data collection. First, mixed methods research that aims to uncover the context behind quantitative findings by using qualitative methods (Tariq & Woodman, 2013) is needed to understand issues that impact the engagement of people with high depression severity. Second, the use of methods (Papoutsi, Wherton, Shaw, Morrison, & Greenhalgh, 2021; Shaw et al., 2018) to co-design study protocols and apps with representative patient advisory boards can help optimize the acceptability of the technology in real-world settings. Third, applying “Human-in-the-loop” (Awais et al., 2020; Goodday et al., 2021) approaches can help the timely resolution of problems that are encountered by participants and may reduce the risk of disengaging from the study. Finally, the present study showed that passive data gathered from wearables has greater contiguity and participant retention over the long term. Focusing efforts on collecting multimodal passive data streams without additional user burden may be a more effective and acceptable marker of individual behavior in real-world settings (Pratap et al., 2019). We discuss these strategies in further detail below.

We also observed that participants’ time in responding to and completing surveys is significantly associated with their long-term engagement patterns. Participants with shorter survey response and completion times tend to engage for the longer term, completing more surveys and wearing Fitbit for a significantly longer period. Past studies have also reported that if participants are more interested in the study, they are quicker to respond and complete study-related assessments (Bassili, 1996; Fazio et al., 1983; Heerwegh, 2003). Further, survey response and completion time may also be correlated with several other factors, such as participants’ familiarity with smartphones and study apps, life behaviors, and smartphone latency (battery and memory). To our best knowledge, this paper is the first to quantitatively link the survey response and completion times to participants’ longitudinal engagement patterns in a remote digital health study. Such real-world objective metrics on participants’ app-usage behavior may be potentially useful for passively assessing the quality of the active data and predicting long-term engagement early.

Finally, age is a significant indicator of participant retention and engagement. We found that older participants have a lower risk of disengaging from the study app (Figure 2.2) and tend to contribute more surveys and phone passive data (Figure 2.4d) than younger participants. This finding is consistent with several previous engagement studies (Dineley et al., 2021; Li S, 2022; Pratap et al., 2020).

Feasibility of Collecting Active Data and Passive Data Streams for Long-term Remote Behavior Monitoring in Major Depression

While there is growing interest amongst researchers in gathering real-world behavioral data without having to rely on episodic in-clinic assessments that may be subject to recall bias (Althubaiti, 2016), there is limited empirical research quantifying the long-term participant engagement differences between active (surveys) and passive data streams (smartphones and wearables). We compared the long-term differences in the density of active and passive data collected from surveys, smartphones, and wearable devices.

Passively gathered data from wearable devices showed the highest long-term engagement (C1 in Figure 2.3b and Supplementary Figure 4 of Appendix A) and the highest participant retention rates (Figure 2.1) over both observation periods (43 weeks and 94 weeks). The finding clearly shows that wearable devices with minimal participant burden could help researchers collect high-density data over a longer period. Another potential reason may be that the Fitbit app provides participants with timely feedback about their sleep quality and physical activity, which may increase their interest in wearing Fitbit devices. We found a significant proportion of the participants who completed fewer longitudinal surveys (C2 and C3 of the Phone-Active data stream) but contributed passive Fitbit data for significantly longer (Figure 2.3c). This illustrates the value of wearable devices for long-term monitoring of participants who cannot routinely actively engage in completing frequent health surveys.

On the other hand, we found that the passive data gathered from participants' phones had the lowest retention rate in both observation periods (Figure 2.1). A potential reason for lower compliance in passive data collection from smartphones could be the relatively high consumption of battery and users’ data plans. The study app collected high-resolution passive data frequently (e.g., GPS [every 10 minutes], Bluetooth [hourly], battery levels [every 10 minutes], and phone usage [event trigger]). The collection of highly granular passive data could have made some participants stop the app from collecting passive data or uninstall it. Future research is needed to understand the suitable balance between passive data collection and phone battery consumption that is acceptable to participants in real-world settings. Notably, we also found that smartphone brands significantly affected the retention and density of phone-passive data collection. Smartphone brands may have different policies on the duration for which an app can collect granular passive data continuously. However, the sample sizes of several phone brands were limited in our cohort. Additional research is needed to investigate intra-device/brand differences within and across Android and iOS phones to enable the robust and equitable collection of passive data. Finally, a small but significant group of participants were not contributing either active or passive data (Figure 2.3c). Further research is needed to understand the concerns of this subgroup to avoid the collection of unbalanced data.

Potential Reasons for High Participant Retention in the RADAR-MDD Study

We discuss four strategies developed and adopted by the RADAR-MDD consortium (Matcham et al., 2019), which may have helped increase long-term participant retention and engagement.

a) "Human-in-the-loop" (Goodday et al., 2021)**.** The RADAR-MDD research team contacted participants for various reasons, such as reminding 3-month assessments, any malfunctions in the Fitbit device, problems in study apps, and congratulating participants for completing the 1-year milestone. Timely resolution of technical issues and feedback and encouragement from the research team may help keep participants in the study (Simblett et al., 2018b).

b) Monetary incentives. Compensation for participant time and monetary incentives are known to enhance engagement (Bentley & Thacker, 2004; Simblett et al., 2018b). Although participants were not offered compensation for completing surveys remotely and sharing behavior data passively, existing monetary incentives could increase participants' willingness to remain engaged in the study. For example, participants were given monetary incentives for enrolling in the study, taking part in clinical assessments (every 3 months), and additional interviews (e.g., 1-year interview) (see Method section). This cyclical compensation (every 3 months and 1 year) could have indirectly incentivized participants to remain in the study. Furthermore, participants were allowed to keep the Fitbit device after the completion of the study, which could have impacted their motivation to join and remain engaged in the study for a longer term.

c) Participant-centric design. Participants' lack of familiarity with how to use digital technologies (study apps) and lack of intrinsic motivation (not familiar with the value of the study) are two key barriers to long-term engagement (Simblett et al., 2018b). Therefore, participants and patients were invited to provide input at all stages of the study process (Matcham et al., 2019). A patient advisory board comprising service users guided the early study protocol and study app design stages to the implementation and analysis phases. They contributed to improving the study design and engagement motivation strategy and shaped how the technology was used (Birnbaum, Lewis, Rosen, & Ranney, 2015). This approach, called "participant-centered initiative" (Anderson, Bragg, Hartzler, & Edwards, 2012; Kaye et al., 2012), treats participants as partners in the entire research cycle, which could provide a means to improve participant retention and engagement in long-term digital health studies. A recent study demonstrated that "participant-centric design" played an essential role in maximizing engagement in remote app-based studies (Druce et al., 2019).

d) Recruiting participants with the target disease of interest. The inclusion criteria in the present study required all participants to have at least one depressive episode in the last two years. Therefore, the study contains an enriched population with a specific clinical condition. Prior research has shown that participants with clinical conditions of interest in the study tend to remain engaged for significantly longer (Pratap et al., 2020; Simblett et al., 2018b). Experiences of having depression may make participants aware of the benefits of regularly completing the self-assessment and getting feedback from clinical teams to realize their status of mental health (Simblett et al., 2018b).

Potential Solutions to Improve Participant Retention and Engagement in Remote Digital Research

Although the incentives and recruitment strategies discussed above increased participant retention in the present cohort, a notable proportion of the cohort (17.59% – 33.71%) across active and passive data streams did not remain engaged in the study over the long-term (C3 clusters in Figure 2.3b). Long-term participant retention and engagement in remote digital studies, therefore, remains an active area of research. Several potential solutions could be learned from our findings. Participant characteristics, such as younger age, more depressive symptoms at baseline, and delayed responses to remote surveys, could act as early indicators of a subgroup of participants at a higher risk of disengagement from the study. Targeted engagement strategies including tailored communication and increased "Human-in-loop" interactions could be deployed to this subgroup. An alternative approach is to recruit more heavily from participants matching the characteristics of the low-engaged subgroup, which may help reduce the overall data imbalance.

Also, a near-real-time analytical framework could be deployed to monitor the incoming real-world data for known socio-technical biases continually. The system could triage participants who are falling below an acceptable level of compliance with the study team in terms of data quantity and quality. This could help just-in-time identify potential causes of unbalanced data collection in real-world settings and allow for timely and targeted interventions to re-engage participants at the highest risk of drop.

2.4.2 Limitations

Our findings should be viewed in the context of certain limitations related to the collection of real-world data in a fully remote European multinational remote digital study for depression. First, the RADAR-MDD study used an open enrollment model to gather real-world data and did not stratify or randomize participant recruitment based on sociodemographic characteristics, enrollment sites, etc. For example, the overall cohort had significantly fewer participants older than 70 years, which can be related to known barriers, e.g., lower use of digital technologies and health problems (Forsat, Palmowski, Palmowski, Boers, & Buttgereit, 2020; Mody et al., 2008; Pywell, Vijaykumar, Dodd, & Coventry, 2020). Further, the study population was predominantly white people, with the majority of females. While the higher proportion of females in the present study cohort is aligned with previous epidemiological and remote observational studies (Arean et al., 2016; Difrancesco et al., 2019; Lu et al., 2018) and a known higher prevalence of depression in females than in males (Albert, 2015; Noble, 2005; Salk et al., 2017; Van de Velde et al., 2010), the findings may not be generalizable to a more diverse or non-depressed population. Future studies should use randomized designs to investigate the causal impact of various demographic and socio-technical factors on participant retention using a representative target population linked to the condition of interest.

Second, although we accounted for many site-level differences, there could be several unmeasured and inconsistent factors across study sites. Third, there were some changes during the course of the study, such as changes in versions of some surveys, fixing technical bugs (e.g., missing notifications), and adding surveys as well as different study start times across three sites that could impact participant engagement. Fourth, the education system, language, income levels, and currency are also different across European countries and could lead to inconsistencies in the comparison of participant responses to sociodemographic questions across sites. These potential differences limited our interpretation of the different levels of participant engagement across the three sites. Fifth, the technical differences between the two versions of Fitbit devices deployed in the study (Charge 2 and Charge 3) were not tracked. Also, the present study was only based on the Android smartphone operating system. As a result, the impact of different versions of wearable devices and different smartphone operating systems on participant engagement is unclear.

Sixth, the specific impact of the number of contact logs on participant engagement may be bidirectional. For example, technical issues may decrease participant engagement with the study app despite the study team reaching out. On the other hand, reaching out to remind participants to complete an assessment or congratulating them for reaching the 1-year milestone may increase participant engagement. Also, participants in the most engaged clusters had the lowest number of contact logs, indicating that highly engaged participants did not need additional reminders to complete assessments and encountered fewer technical issues.

Seventh, while participants were not paid for completing remote surveys via smartphones or sharing passive data, compensation was given for clinical assessments every 3 months, which may also affect the generalizability of our findings in cohorts without any incentives.

2.4.3 Conclusions

This study demonstrated that participant retention in the RADAR-MDD study was significantly higher than in past digital studies. Higher retention is likely linked to the deployment of several engagement strategies such as “human-in-the-loop”, monetary incentives, participant-centric design, and a targeted clinical cohort. We found several notable indicators, such as age, depression severity, and survey response and competition times in the study app, significantly impacted the depth and density of our real-world data collection in fully remote research. Furthermore, passive data gathered from wearables without participant burden showed advantages in helping collect behavioral data with greater contiguity and over a longer duration. Combined, these objective engagement metrics could help identify and triage participants with the highest dropout risk for tailored and just-in-time engagement strategies to enable equitable and balanced health data collection from diverse target populations.

Acknowledgments

The RADAR-CNS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 115902. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA, www.imi.europa.eu. This communication reflects the views of the RADAR-CNS consortium and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding body has not been involved in the design of the study, the collection or analysis of data, or the interpretation of data. Participants in the CIBER site came from the following four clinical communities in Spain: Parc Sanitari Sant Joan de Déu Network services, Institut Català de la Salut, Institut Pere Mata, and Hospital Clínico San Carlos.

Participant recruitment in Amsterdam was partially accomplished through Hersenonderzoek.nl (www.hersenonderzoek.nl), a Dutch online registry that facilitates participant recruitment for neuroscience studies. Hersenonderzoek.nl is funded by ZonMw-Memorabel (project no 73305095003), a project in the context of the Dutch Deltaplan Dementie, Gieskes-Strijbis Foundation, the Alzheimer's Society in the Netherlands and Brain Foundation Netherlands.

This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design.

This research was reviewed by a team with experience of mental health problems and their careers who have been specially trained to advise on research proposals and documentation through the Feasibility and Acceptability Support Team for Researchers (FAST-R): a free, confidential service in England provided by the National Institute for Health Research Maudsley Biomedical Research Centre via King’s College London and South London and Maudsley NHS Foundation Trust.

The RADAR-MDD was conducted per the Declaration of Helsinki and Good Clinical Practice, adhering to principles outlined in the NHS Research Governance Framework for Health and Social Care (2nd edition). Ethical approval has been obtained in London from the Camberwell St Giles Research Ethics Committee (REC reference: 17/LO/1154), in London from the CEIC Fundacio Sant Joan de Deu (CI: PIC-128-17) and in The Netherlands from the Medische Ethische Toetsingscommissie VUmc (METc VUmc registratienummer: 2018.012 – NL63557.029.17)

Author’s contributions

YZ, AP, AF, and RD contributed to the design of the study. YZ and AP extracted and summarized the active survey and passive data from the Remote Assessment of Disease and Relapse–Major Depressive Disorder study, planned and performed the analysis, and drafted the manuscript. MH is the principal investigator for the RADAR-MDD study. RD, AF, YR, ZR, PC, and CS have contributed to the development of the RADAR-base platform for data collection. FM, KW, CO, AI, FL, S Siddi, CHR, S Simblett, JMH, BWJHP, and MH contributed to participant recruitment and data collection. AF, IMG, VAN, TW, PA, MH, and RD. contributed the administrative, technical, and clinical support of the study. All authors were involved in reviewing the manuscript, had access to the study data, and provided direction and comments on the manuscript.

Conflicts of Interest

SV and VAN are employees of Janssen Research and Development LLC. PA is employed by the pharmaceutical company H. Lundbeck A/S. DCM has accepted honoraria and consulting fees from Apple Inc, Otsuka Pharmaceuticals, Pear Therapeutics, and the One Mind Foundation; has received royalties from Oxford Press; and has an ownership interest in Adaptive Health Inc. MH is the principal investigator of the Remote Assessment of Disease and Relapse–Central Nervous System project, a private public precompetitive consortium that receives funding from Janssen, UCB, Lundbeck, MSD, and Biogen. All other authors declare no competing interests.

References

Albert, P. R. (2015). Why is depression more prevalent in women? *Journal of psychiatry & neuroscience: JPN, 40*(4), 219.

Althubaiti, A. (2016). Information bias in health research: definition, pitfalls, and adjustment methods. *Journal of multidisciplinary healthcare, 9*, 211.

Anderson, N., Bragg, C., Hartzler, A., & Edwards, K. (2012). Participant-centric initiatives: tools to facilitate engagement in research. *Applied & translational genomics, 1*, 25-29.

Arean, P. A., Hallgren, K. A., Jordan, J. T., Gazzaley, A., Atkins, D. C., Heagerty, P. J., & Anguera, J. A. (2016). The use and effectiveness of mobile apps for depression: results from a fully remote clinical trial. *Journal of medical Internet research, 18*(12), e6482.

Awais, M., Ghayvat, H., Krishnan Pandarathodiyil, A., Nabillah Ghani, W. M., Ramanathan, A., Pandya, S., . . . Faye, I. (2020). Healthcare professional in the loop (HPIL): classification of standard and oral cancer-causing anomalous regions of oral cavity using textural analysis technique in autofluorescence imaging. *Sensors, 20*(20), 5780.

Bailon, C., Damas, M., Pomares, H., Sanabria, D., Perakakis, P., Goicoechea, C., & Banos, O. (2019). Smartphone-based platform for affect monitoring through flexibly managed experience sampling methods. *Sensors, 19*(15), 3430.

Bardram, J. E., & Matic, A. (2020). A decade of ubiquitous computing research in mental health. *IEEE Pervasive Computing, 19*(1), 62-72.

Bassili, J. N. (1996). Meta-judgmental versus operative indexes of psychological attributes: The case of measures of attitude strength. *Journal of personality and social psychology, 71*(4), 637.

Baumel, A., Muench, F., Edan, S., & Kane, J. M. (2019). Objective User Engagement With Mental Health Apps: Systematic Search and Panel-Based Usage Analysis. *J Med Internet Res, 21*(9), e14567. doi:10.2196/14567

Bentley, J. P., & Thacker, P. G. (2004). The influence of risk and monetary payment on the research participation decision making process. *Journal of medical ethics, 30*(3), 293-298.

Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 12: survival analysis. *Critical care, 8*(5), 1-6.

Birnbaum, F., Lewis, D. M., Rosen, R., & Ranney, M. L. (2015). Patient engagement and the design of digital health. *Academic emergency medicine: official journal of the Society for Academic Emergency Medicine, 22*(6), 754.

Cai, N., Choi, K. W., & Fried, E. I. (2020). Reviewing the genetics of heterogeneity in depression: operationalizations, manifestations and etiologies. *Human molecular genetics, 29*(R1), R10-R18.

Chien, I., Enrique, A., Palacios, J., Regan, T., Keegan, D., Carter, D., . . . Richards, D. (2020). A machine learning approach to understanding patterns of engagement with internet-delivered mental health interventions. *JAMA network open, 3*(7), e2010791-e2010791.

Cho, Y. M., Lim, H. J., Jang, H., Kim, K., Choi, J. W., Shin, C., . . . Kim, N. (2016). A cross-sectional study of the association between mobile phone use and symptoms of ill health. *Environmental health and toxicology, 31*.

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., . . . Mohr, D. C. (2022). Digital health tools for the passive monitoring of depression: a systematic review of methods. *NPJ digital medicine, 5*(1), 1-14.

Difrancesco, S., Lamers, F., Riese, H., Merikangas, K. R., Beekman, A. T., van Hemert, A. M., . . . Penninx, B. W. (2019). Sleep, circadian rhythm, and physical activity patterns in depressive and anxiety disorders: A 2-week ambulatory assessment study. *Depression and anxiety, 36*(10), 975-986.

Dineley, J., Lavelle, G., Leightley, D., Matcham, F., Siddi, S., Peñarrubia-María, M. T., . . . Simblett, S. (2021). *Remote smartphone-based speech collection: Acceptance and barriers in individuals with major depressive disorder.* Paper presented at the 22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021.

Druce, K. L., Dixon, W. G., & McBeth, J. (2019). Maximizing engagement in mobile health studies: lessons learned and future directions. *Rheumatic diseases clinics of North America, 45*(2), 159.

Fazio, R. H., Powell, M. C., & Herr, P. M. (1983). Toward a process model of the attitude–behavior relation: Accessing one's attitude upon mere observation of the attitude object. *Journal of personality and social psychology, 44*(4), 723.

Forsat, N. D., Palmowski, A., Palmowski, Y., Boers, M., & Buttgereit, F. (2020). Recruitment and retention of older people in clinical research: a systematic literature review. *Journal of the American Geriatrics Society, 68*(12), 2955-2963.

Gilchrist, G., & Gunn, J. (2007). Observational studies of depression in primary care: what do we know? *BMC Family practice, 8*(1), 1-18.

Goodday, S. M., Karlin, E., Alfarano, A., Brooks, A., Chapman, C., Desille, R., . . . Boch, A. (2021). An Alternative to the Light Touch Digital Health Remote Study: The Stress and Recovery in Frontline COVID-19 Health Care Workers Study. *JMIR formative research, 5*(12), e32165.

Grambsch, P. M., & Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika, 81*(3), 515-526.

Greenberger, E., Chen, C., Dmitrieva, J., & Farruggia, S. P. (2003). Item-wording and the dimensionality of the Rosenberg Self-Esteem Scale: Do they matter? *Personality and individual differences, 35*(6), 1241-1254.

Heerwegh, D. (2003). Explaining response latencies and changing answers using client-side paradata from a web survey. *Social Science Computer Review, 21*(3), 360-373.

Kaye, J., Curren, L., Anderson, N., Edwards, K., Fullerton, S. M., Kanellopoulou, N., . . . Shepherd, J. (2012). From patients to partners: participant-centric initiatives in biomedical research. *Nature Reviews Genetics, 13*(5), 371-376.

Klasen, F., Otto, C., Kriston, L., Patalay, P., Schlack, R., & Ravens-Sieberer, U. (2015). Risk and protective factors for the development of depressive symptoms in children and adolescents: results of the longitudinal BELLA study. *European child & adolescent psychiatry, 24*(6), 695-703.

Kleinbaum, D. G., & Klein, M. (2012). Evaluating the proportional hazards assumption *Survival analysis* (pp. 161-200): Springer.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders, 114*(1-3), 163-173.

Kumar, D., & Klefsjö, B. (1994). Proportional hazards model: a review. *Reliability Engineering & System Safety, 44*(2), 177-188.

Laiou, P., Kaliukhovich, D. A., Folarin, A. A., Ranjan, Y., Rashid, Z., Conde, P., . . . Hotopf, M. (2022). The Association Between Home Stay and Symptom Severity in Major Depressive Disorder: Preliminary Findings From a Multicenter Observational Study Using Geolocation Data From Smartphones. *JMIR Mhealth Uhealth, 10*(1), e28095. doi:10.2196/28095

Li S, H. R., Selvarajan R, Woerner M, Fillipo IG, Banerjee S, Mosser B, Jain F, Arean P, Pratap A. (2022). Recruitment & Retention in Remote Research: Learnings from a Large Decentralized Real-World Study. *JMIR Preprints*.

Liew, C. S., Wah, T. Y., Shuja, J., & Daghighi, B. (2015). Mining personal data using smartphones and wearable devices: A survey. *Sensors, 15*(2), 4430-4469.

Lu, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., . . . Bi, J. (2018). Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(1), 1-21.

Luik, A. I., Zuurbier, L. A., Direk, N., Hofman, A., Van Someren, E. J., & Tiemeier, H. (2015). 24-hour activity rhythm and sleep disturbances in depression and anxiety: A population-based study of middle-aged and older persons. *Depression and anxiety, 32*(9), 684-692.

Matcham, F., di San Pietro, C. B., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., . . . Lamers, F. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. *BMC psychiatry, 19*(1), 1-11.

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K. M., Annas, P., . . . Horsfall, M. (2021). Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): Recruitment, retention, and data availability in a longitudinal remote measurement study.

Mody, L., Miller, D. K., McGloin, J. M., Freeman, M., Marcantonio, E. R., Magaziner, J., & Studenski, S. (2008). Recruitment and Retention of Older Adults in Aging Research: (See editorial comments by Dr. Stephanie Studenski, pp 2351–2352). *Journal of the American Geriatrics Society, 56*(12), 2340-2348.

Moore, S., Tassé, A.-M., Thorogood, A., Winship, I., & Doerr, M. (2017). Consent processes for mobile app mediated research: systematic review. *JMIR mHealth and uHealth, 5*(8), e126.

Noble, R. E. (2005). Depression in women. *Metabolism, 54*(5), 49-52.

O'connor, S., Hanlon, P., O'donnell, C. A., Garcia, S., Glanville, J., & Mair, F. S. (2016). Understanding factors affecting patient and public engagement and recruitment to digital health interventions: a systematic review of qualitative studies. *BMC medical informatics and decision making, 16*(1), 1-15.

Ostertagova, E., Ostertag, O., & Kováč, J. (2014). *Methodology and application of the Kruskal-Wallis test.* Paper presented at the Applied Mechanics and Materials.

Papoutsi, C., Wherton, J., Shaw, S., Morrison, C., & Greenhalgh, T. (2021). Putting the social back into sociotechnical: Case studies of co-design in digital health. *Journal of the American Medical Informatics Association, 28*(2), 284-293.

Pratap, A., Atkins, D. C., Renn, B. N., Tanana, M. J., Mooney, S. D., Anguera, J. A., & Areán, P. A. (2019). The accuracy of passive phone sensors in predicting daily mood. *Depression and anxiety, 36*(1), 72-81.

Pratap, A., Neto, E. C., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., . . . Wilbanks, J. (2020). Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *NPJ digital medicine, 3*(1), 1-10.

Pratap, A., Renn, B. N., Volponi, J., Mooney, S. D., Gazzaley, A., Arean, P. A., & Anguera, J. A. (2018). Using mobile apps to assess and treat depression in Hispanic and Latino populations: fully remote randomized clinical trial. *Journal of medical Internet research, 20*(8), e10130.

Pywell, J., Vijaykumar, S., Dodd, A., & Coventry, L. (2020). Barriers to older adults' uptake of mobile-based mental health interventions. *Digital health, 6*, 2055207620905422.

Quisel, T., Foschini, L., Zbikowski, S. M., & Juusola, J. L. (2019). The association between medication adherence for chronic conditions and digital health activity tracking: retrospective analysis. *Journal of medical Internet research, 21*(3), e11486.

Ranjan, Y., Rashid, Z., Stewart, C., Conde, P., Begale, M., Verbeeck, D., . . . Consortium, R.-C. (2019). RADAR-base: open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices. *JMIR mHealth and uHealth, 7*(8), e11734.

Rich, J. T., Neely, J. G., Paniello, R. C., Voelker, C. C., Nussenbaum, B., & Wang, E. W. (2010). A practical guide to understanding Kaplan-Meier curves. *Otolaryngology—Head and Neck Surgery, 143*(3), 331-336.

Salk, R. H., Hyde, J. S., & Abramson, L. Y. (2017). Gender differences in depression in representative national samples: Meta-analyses of diagnoses and symptoms. *Psychological bulletin, 143*(8), 783.

Schmidt, M. (2008). The Sankey diagram in energy and material flow management: part II: methodology and current applications. *Journal of industrial ecology, 12*(2), 173-185.

Shaw, J., Agarwal, P., Desveaux, L., Palma, D. C., Stamenova, V., Jamieson, T., . . . Bhattacharyya, O. (2018). Beyond "implementation": digital health innovation and service design. *NPJ Digital Medicine, 1*(1), 1-5.

Simblett, S., Greer, B., Matcham, F., Curtis, H., Polhemus, A., Ferrão, J., . . . Wykes, T. (2018a). Barriers to and facilitators of engagement with remote measurement technology for managing health: systematic review and content analysis of findings. *Journal of medical Internet research, 20*(7), e10480.

Simblett, S., Greer, B., Matcham, F., Curtis, H., Polhemus, A., Ferrão, J., . . . Wykes, T. (2018b). Barriers to and Facilitators of Engagement With Remote Measurement Technology for Managing Health: Systematic Review and Content Analysis of Findings. *J Med Internet Res, 20*(7), e10480. doi:10.2196/10480

Simblett, S., Matcham, F., Siddi, S., Bulgari, V., Barattieri di San Pietro, C., Hortas

López, J., . . . Wykes, T. (2019). Barriers to and Facilitators of Engagement With mHealth Technology for Remote Measurement and Management of Depression: Qualitative Analysis. *JMIR Mhealth Uhealth, 7*(1), e11325. doi:10.2196/11325

Singer, J. D., & Willett, J. B. (1991). Modeling the days of our lives: using survival analysis when designing and analyzing longitudinal studies of duration and the timing of events. *Psychological bulletin, 110*(2), 268.

Snyder, M., & Zhou, W. (2019). Big data and health. *The Lancet Digital Health, 1*(6), e252-e254.

Syakur, M., Khotimah, B., Rochman, E., & Satoto, B. D. (2018). *Integration k-means clustering method and elbow method for identification of the best customer profile cluster.* Paper presented at the IOP Conference Series: Materials Science and Engineering.

Tariq, S., & Woodman, J. (2013). Using mixed methods in health research. *JRSM short reports, 4*(6), 2042533313479197.

Torous, J., Lipschitz, J., Ng, M., & Firth, J. (2020). Dropout rates in clinical trials of smartphone apps for depressive symptoms: a systematic review and meta-analysis. *Journal of affective disorders, 263*, 413-419.

Van de Velde, S., Bracke, P., & Levecque, K. (2010). Gender differences in depression in 23 European countries. Cross-national variation in the gender gap in depression. *Social science & medicine, 71*(2), 305-313.

Wu, J. (2012). Cluster analysis and K-means clustering: an introduction *Advances in K-means Clustering* (pp. 1-16): Springer.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Bendayan, R., Ranjan, Y., . . . Dobson, R. J. B. (2021a). Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. *JMIR mHealth and uHealth, 9*(4), e24604.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., . . . Dobson, R. J. B. (2021b). Predicting Depressive Symptom Severity Through Individuals' Nearby Bluetooth Device Count Data Collected by Mobile Phones: Preliminary Longitudinal Study. *JMIR Mhealth Uhealth, 9*(7), e29840. doi:10.2196/29840

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Vairavan, S., Bendayan, R., . . . Dobson, R. J. (2022). Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study. *JMIR Ment Health, 9*(3), e34898. doi:10.2196/34898

Chapter 3

Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study

This chapter has been published as:

Zhang Y, Folarin AA, Sun S, Cummins N, Bendayan R, Ranjan Y, Rashid Z, Conde P, Stewart C, Laiou P, Matcham F, White KM, Lamers F, Siddi S, Simblett S, Myin-Germeys I, Rintala A, Wykes T, Haro JM, Penninx BW, Narayan VA, Hotopf M, Dobson RJ, 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 2021;9*(4): e24604.

Background: Sleep problems tend to vary according to the course of the disorder in individuals with mental health problems. Research in mental health has associated sleep pathologies with depression. However, the gold standard for sleep assessment, polysomnography (PSG), is not suitable for long-term, continuous monitoring of daily sleep, and methods such as sleep diaries rely on subjective recall, which is qualitative and inaccurate. Wearable devices, on the other hand, provide a low-cost and convenient means to monitor sleep in home settings.

Objective: The main aim of this study was to devise and extract sleep features from data collected using a wearable device and analyze their associations with depressive symptom severity and sleep quality as measured by the self-assessed 8-item Patient Health Questionnaire (PHQ-8).

Methods: Daily sleep data were collected passively by Fitbit wristband devices, and depressive symptom severity was self-reported every 2 weeks by the PHQ-8. The data used in this paper included 2812 PHQ-8 records from 368 participants recruited from 3 study sites in the Netherlands, Spain, and the United Kingdom. We extracted 18 sleep features from Fitbit data that describe participant sleep in the following 5 aspects: sleep architecture, sleep stability, sleep quality, insomnia, and hypersomnia. Linear mixed regression models were used to explore associations between sleep features and depressive symptom severity. The *z* score was used to evaluate the significance of the coefficient of each feature.

Results: We tested our models on the entire dataset and separately on the data of 3 different study sites. We identified 14 sleep features that were significantly (*P*<.05) associated with the PHQ-8 score on the entire dataset, among them awake time percentage (*z*=5.45, *P*<.001), awakening times (z=5.53, *P*<.001), insomnia (z=4.55, *P*<.001), mean sleep offset time (z=6.19, *P*<.001), and hypersomnia (z=5.30, *P*<.001) were the top 5 features ranked by *z* score statistics. Associations between sleep features and PHQ-8 scores varied across different sites, possibly due to differences in the populations. We observed that many of our findings were consistent with previous

studies, which used other measurements to assess sleep, such as PSG and sleep questionnaires.

Conclusions: We demonstrated that several derived sleep features extracted from consumer wearable devices show potential for the remote measurement of sleep as biomarkers of depression in real-world settings. These findings may provide the basis for the development of clinical tools to passively monitor disease state and trajectory, with minimal burden on the participant.

3.1 Introduction

According to the report of the World Health Organization, the total number of people with depression was estimated to exceed 300 million in 2015, equivalent to 4.4% of the world’s population (World Health Organization, 2017). There are several depression-related adverse outcomes, including premature mortality (Cuijpers & Schoevers, 2004), decline in quality of life (Lenox-Smith et al., 2013), and loss of occupational function (Lerner et al., 2004).

Sleep disturbances are prevalent among depression patients; more than 90% of patients with depression reported poor sleep quality (Mendelson, 2012). Sleep disturbances cover a wide range of different symptoms and disorders including insomnia, hypersomnia, excessive daytime sleepiness, and circadian rhythm disturbance (Alvaro et al., 2013). Insomnia and sleep quality have been observed to be bidirectionally related to depression in several longitudinal studies (Alvaro et al., 2013). Hypersomnia is more frequently present in depressive episodes of bipolar patients (Detre et al., 1972; Thase et al., 1989). Changes in sleep architecture, such as reduced deep sleep, increased rapid eye movement (REM) sleep, and shortened REM latency, are also significant predictors of depression (Palagini et al., 2013; Riemann et al., 2001).

The gold standard for sleep evaluation is polysomnography (PSG), which involves several physiological measurements including electroencephalogram, electrocardiogram, electromyogram, and accelerometers (Berry et al., 2012). Using PSG to assess sleep lacks ecological validity and is time-consuming, expensive, and labor-intensive, requiring dedicated equipment and separate laboratory rooms as well as experts to analyze the physiological signals. Since depression can affect patients for an extended period, long-term monitoring of sleep quality is essential. Due to the above shortcomings, PSG is not suitable for long-term sleep monitoring (Sánchez-Ortuño et al., 2010). A sleep questionnaire, such as the Pittsburgh Sleep Quality Index (PSQI) (Buysse et al., 1989), is another useful method to assess sleep. This method relies on the self-reporting of subjective factors, like low recall of sleep, that may affect the accuracy of the assessment (Moore et al., 2015).

Several recent studies have used wearable devices to estimate sleep quality and sleep-related parameters (Beattie et al., 2017; Van de Water et al., 2011; Zhang et al., 2018, 2019) and analyzed the relationship between sleep and depression (Demasi et al., 2016; Mark et al., 2016; Miwa et al., 2007). Miwa et al estimated sleep quality by detecting rollover movements during sleep and observed a significant difference in sleep quality between nondepressed and depressed people (Miwa et al., 2007). Mark et al estimated the sleep duration of 40 information workers for 12 days using a Fitbit wristband and found that sleep duration was positively correlated with mood (Mark et al., 2016). DeMasi et al found that sleep was significantly related to changes in depressive symptoms (Demasi et al., 2016). These studies have mostly been performed on single center and relatively small datasets (number of participants fewer than 100). Moreover, most of these studies only used basic sleep parameters, such as sleep duration; detailed information on sleep architecture, sleep patterns, and stability of sleep was not considered. The relationship between detailed sleep features, as estimated from data supplied by wearable devices, and depression is yet to be fully explored.

The first aim of this study was to design more sleep-related features, from wearable device data, that reflect the sleep architecture, sleep stability, sleep quality, and sleep disturbances (insomnia and hypersomnia) of the participant. The second aim was to explore associations between these sleep features and depressive symptom severity on a relatively large, multisite dataset. The third aim was to compare our findings with previous studies that used other measurements to assess sleep such as PSG and sleep questionnaires.

3.2 Methods

3.2.1 Data set

Study Participants and Settings

The data we used in this paper were collected from a major EU Innovative Medicines Initiative research project, Remote Assessment of Disease and Relapse–Central Nervous System (RADAR-CNS) (Khoulji et al., 2017). This project aims to investigate the use of remote measurement technologies to monitor people with depression, epilepsy, and multiple sclerosis in real-world settings. The study protocol for the depression component (Remote Assessment of Disease and Relapse–Major Depressive Disorder [RADAR-MDD]) is described in detail in Matcham et al (Matcham et al., 2019). The RADAR-MDD project aims to recruit 600 participants with a recent history of depression in 3 study sites (King’s College London [KCL], UK; Vrije Universiteit Medisch Centrum [VUmc], Amsterdam, The Netherlands; and Centro de Investigación Biomédican en Red [CIBER], Barcelona, Spain). Recruitment procedures vary slightly across sites and eligible participants are identified either through existing research cohorts (in KCL and VUmc) who had given consent to be contacted for research purposes; advertisements in general practices, psychologist practices, newspapers, and Hersenonderzoek.nl (https://hersenonderzoek.nl/), which is a Dutch online registry (VUmc); or through mental health services (in KCL and CIBER) (Matcham et al., 2019). Participants from KCL and VUmc are community-based, while the participants from CIBER come from a clinical population. As part of the study, participants are asked to install several remote monitoring technology apps and use an activity tracker for up to 2 years of follow-up. Many categories of passive and active data are being collected and uploaded to an open-source platform, RADAR-base (Ranjan et al., 2019). In this paper, we focus on the sleep and 8-item Patient Health Questionnaire (PHQ-8) data (Kroenke et al., 2009).

Sleep Data

According to the American Academy of Sleep Medicine manual for the scoring of sleep and associated events, sleep can be divided into 2 phases, REM sleep and non-REM (NREM) sleep, and NREM sleep can be subdivided into N1, N2, and N3 stages according to characteristic patterns of brain waves collected by PSG (Berry et al., 2012). In our project, the daily sleep records of participants were collected by the Charge 2 or

Charge 3 (Fitbit Inc). An entire night's sleep is divided into 4 stages: awake, light, deep, and REM. The light stage provides estimates for the N1 and N2 stages in PSG, while the deep stage provides estimates for the N3 stage in PSG. According to several validation studies of Fitbit, the Fitbit wristband had limited specificity in sleep stages estimation (de Zambotti et al., 2018; Haghayegh et al., 2019; Liang & Chapa-Martell, 2019). Therefore, in this study, we were not expecting the Fitbit devices to provide information as accurate as PSG would have provided. However, the Fitbit devices were deemed sensitive enough to detect changes in sleep-wake states (de Zambotti et al., 2018; Haghayegh et al., 2019; Liang & Chapa-Martell, 2019); therefore, the provided sleep stage information could be used to determine estimates for detailed sleep parameters based on known sleep pathology.

PHQ-8 Data

The variability of each participant's depressive symptom severity was measured via the PHQ-8, conducted by mobile phone every 2 weeks. The questionnaire contains 8 questions, with the score of each subitem ranging from 0 to 3. The total score (range 0 to 24) of all subitems is the PHQ-8 score, which can evaluate depressive symptom severity of the participant for the past 2 weeks. A PHQ-8 score ≥10 is the most commonly recommended cutpoint for clinically significant depressive symptoms (Kroenke et al., 2009) (i.e., if the PHQ-8 of a participant is ≥10, the participant is likely to have had depressive symptoms in the previous 2 weeks). In the PHQ-8, subitem 3 refers to sleep. The content of subitem 3 is "Trouble falling or staying asleep, or sleeping too much" (Kroenke et al., 2009). A higher score in subitem 3 indicates worse self-reported sleep in the past 2 weeks. For reading convenience, we denoted the score of subitem 3 as the sleep subscore in this paper.

Sociodemographics

Sociodemographic of participants were collected during the enrollment session. According to previous studies on the associations between depression and sociodemographic characteristics (Akhtar-Danesh & Landeen, 2007; Aluoja et al., 2004), we considered baseline age, gender, education level, and annual income as potential confounding variables in our analyses. Due to the different educational systems in different countries, we simply divided the education level into 2 levels: degree (or above) and below degree. The annual income levels of Spain and the Netherlands were transformed into equivalent British levels.

3.2.2 Feature Extraction

Feature Window Size

For each PHQ-8 record, we extracted sleep features from a 2-week time window prior to the PHQ-8 completion time, as the PHQ-8 score is used to represent the depressive symptom severity of the participant for the past 2 weeks. The feature window is denoted as Δt in this paper.

Sleep Features

According to known sleep pathology and our experience, 18 sleep features extracted in this paper were divided into the following 5 categories (Table 3.1): sleep architecture, representing the basic and cyclical patterns of sleep; sleep stability, representing the variance of sleep in the feature window; sleep quality, measures relating to total sleep and wake times; insomnia, trouble falling or staying asleep; and hypersomnia, excessive sleepiness.

a) Sleep Architecture

The features of sleep architecture were intended to describe the basic and cyclical patterns of sleep. Therefore, we extracted some features similar to those in the PSG report (total sleep time, time in bed, sleep onset time, sleep offset time, and REM latency) (Ohayon et al., 2017), and features of the percentages of all sleep stages. Total sleep time of one night is defined as the sum of all nonawake stages (light, deep, and REM) (Ohayon et al., 2017). The mean total sleep time in Δt was denoted as *Av_tst*. Time in bed of one night is defined as the sum of all sleep stages (awake, light, deep, and REM) of the entire night (Ohayon et al., 2017). The mean time in bed in Δt was denoted as *Av_time_bed*. Percentage of each sleep stage is defined as the percentage of the time in the sleep stage to the time in bed of the entire night. Different sleep stages have different functions and can reflect the quality of sleep. Deep sleep is considered essential for memory consolidation (Walker, 2008), and REM sleep favors the preservation of memory (Rasch & Born, 2013). A previous sleep report has shown that more deep sleep and fewer awakenings represent better sleep quality (Ohayon et al., 2017). Therefore, we extracted the mean percentages of these 4 sleep stages in Δt, and denoted them as *Deep_pct, Light_pct, REM_pct, Awake_pct,* respectively. The combination of deep and light sleep is NREM sleep. The mental activity that occurs in NREM and REM sleep is a result of 2 different mind generators, which also explains the difference in mental activity (Manni, 2005). So, we extracted the mean percentage of NREM sleep in Δt, which was denoted as *NREM_pct*. We calculated the mean sleep onset time (the first nonawake stage) in Δt, denoted as *Av_onset*. Mean sleep offset time (the last nonawake stage) in Δt was calculated and denoted as *Av_offset*. Previous literature has shown that shortened REM latency can be considered as a biological mark of depression relapse (Palagini et al., 2013). REM latency is defined as the interval between sleep onset and occurrence of the first REM stage. The mean REM latency in Δt was denoted as *REM_L*.

b) Sleep Stability

The features in this category were used to estimate the variance of sleep during Δt. We extracted the standard deviation of total sleep time, sleep onset time, and sleep offset time in Δt, which were denoted as *Std_tst*, *Std_onset,* and *Std_offset*, respectively.

c) Sleep Quality

In this paper, we used features of sleep efficiency, awakenings, and weekend catch-up sleep to describe sleep quality. The definition of sleep efficiency is the percentage of total sleep time to time in bed (Ohayon et al., 2017). Mean sleep efficiency in Δt was denoted as *Efficiency*. The definition of awakenings (>5 minutes) for one night is the number of episodes in which an individual is awake for more than 5 minutes (Ohayon et al., 2017). The average number of awakenings in Δt was denoted as *Awake_5*. Weekend catch-up sleep is an indicator of insufficient weekday sleep, which might be associated with depression level (Kang et al., 2014). A longer total sleep time during the weekend compared with weekdays may reflect the actual sleep needed (Liu et al., 2008). Therefore, we calculated the mean total sleep time difference between weekend and weekdays in Δt, which was denoted as *WKD_diff*.

d) Insomnia

A review of several longitudinal studies suggested that insomnia is bidirectionally related to depression (Alvaro et al., 2013). According to the diagnostic features provided in the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition* (American Psychiatric Association, 2013), insomnia manifests as initial insomnia (difficulty initiating sleep at bedtime), middle insomnia (frequent or prolonged awakening throughout the night), and late insomnia (early-morning awakening with an inability to return to sleep).

For initial insomnia and late insomnia, mean sleep onset time (*Av_onset*) and sleep offset time (*Av_offset*) can be used to partially reflect them, respectively. We define potential middle insomnia to be whether the total sleep time is less than 6 hours and there is at least one prolonged awakening (≥30 minutes) during the night. The percentage of days with potential middle insomnia in the feature window was denoted as *M_Insomnia*.

e) Hypersomnia

Hypersomnia can be another symptom of depression (Detre et al., 1972). The hypersomnia criteria used in Tam et al (Tam et al., 1997) is sleeping more than 10 hours per day, 3 days per week. In this paper, the percentage of days with total sleep time greater than 10 hours was extracted in Δt and denoted as *Dur_10*.

Table 3.1. A list of sleep features used in this study and their short descriptions.

Features	Description	Unit
Sleep architecture		
Av_tst	Mean total sleep time	Hour
Av_time_bed	Mean time in bed	Hour
Deep_pct	Mean percentage of deep sleep	%
Light_pct	Mean percentage of light sleep	%
REM_pct	Mean percentage of REM[a] sleep	%
NREM_pct	Mean percentage of NREM[b] sleep	%
Awake_pct	Mean percentage of awake time	%
Av_onset	Mean sleep onset time	Hour
Av_offset	Mean sleep offset time	Hour
REM_L	Mean REM latency time	Hour
Sleep stability		
Std_tst	Standard deviation of total sleep time	Hour
Std_onset	Standard deviation of sleep onset time	Hour
Std_offset	Standard deviation of sleep offset time	Hour
Sleep quality		
Efficiency	Mean sleep efficiency	%
Awake_5	Mean number of awakenings (>5 minutes) per night	Times
WKD_diff	Total sleep time difference between weekend and weekdays	Hour
Insomnia		
M_insomnia	Percentage of days with potential middle insomnia	%
Hypersomnia		
Dur_10	Percentage of days with total sleep time >10 hours	%

[a] REM: rapid eye movement.

[b] Non-REM: non–rapid eye movement.

3.2.3 Statistical Method

Data Inclusion Criteria

Sleep and PHQ-8 records were missing in our data cohort for a variety of expected reasons, including the participants not wearing the Fitbit wristband when they slept, participants forgetting to complete the PHQ-8, and the Fitbit wristband being damaged during follow-up. We, therefore, specified the following inclusion criteria: (1) PHQ-8 record should be completed (i.e., participant answered all 8 questions in the questionnaire); (2) number of days with sleep records in the feature window should be at least 12 days (approximately 85% of the feature window size) (Farhan et al., 2016); (3) number of PHQ-8 records for each participant should be greater than or equal to 3 (Singer & Willett, 2003); (4) date of PHQ-8 records should be before February 2020, because the impact of the COVID-19 pandemic on sleep needs to be excluded (Sun et al., 2020).

Statistical Analyses

In our study, each participant had multiple PHQ-8 records and repeated sleep measures. For this reason, we used linear mixed models, which allow for accounting of both within and between-individual variability over time (Laird & Ware, 1982). For each sleep feature, a 3-level linear mixed model with a participant-specific random intercept and a site-specific random intercept was built on the entire dataset to explore the association between this sleep feature and depressive symptom severity (PHQ-8) by bivariate analysis. We then used 2-level linear mixed models with participant-specific random intercepts to test these associations on the 3 subsets (KCL, CIBER, and VUmc) separately. We similarly analyzed the associations between sleep features and sleep subscore. All models were adjusted for baseline age, gender, education level, and annual income, which were specified as fixed effects. Model assumptions were checked by the histograms of residuals and Q-Q plots. If the residuals are not normally distributed, the Box-Cox transformation was performed (Box & Cox, 1964). The *z* score was used to evaluate the statistical significance of the coefficient of each model. All *P* values of these tests were corrected by using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995) for multiple comparisons, and the significance level of the corrected *P* value was set to .05. All linear mixed models were implemented by using the lme4 package for R software version 3.6.1 (R Foundation for Statistical Computing).

In order to identify and compare the relationship between self-reported sleep and self-reported depression among different study sites, Spearman correlations were calculated between the PHQ-8 score and sleep subscore on the 3 study sites separately.

An example of such a 3-level linear mixed model is as follows:

$$Sleep_{ijk} = \delta_{000} + V_{00k} + U_{0jk} + \beta_1(PHQ\text{-}8_{ijk}) + \beta_2(age_{\mathrm{jk}}) + \beta_3(gender_{jk}) + \beta_4(education_{jk}) + \beta_5(income_{jk}) + \varepsilon_{ijk},$$

where PHQ-8$_{\mathrm{ijk}}$ is the i$^{\mathrm{th}}$ PHQ-8 score of the participant j of the site k, Sleep$_{\mathrm{ijk}}$ is one sleep feature extracted in Δt before the i$^{\mathrm{th}}$ PHQ-8 record of the participant j of the site k, age$_{jk}$, gender$_{jk}$, education$_{jk}$, and income$_{jk}$ are potential confounding variables of the participant j of the site k, ε_{ijk} is the residual, δ_{000} is the fixed effect on intercept, U$_{0jk}$ is the random intercept of the participant j in the site k, and V$_{00k}$ is the random intercept of the site k.

3.3 Results

3.3.1 Data Summary

According to our data inclusion criteria, from June 2018 to February 2020, 2812 PHQ-8 records from 368 participants collected from 3 study sites were included for our analysis. A summary of the sociodemographic characteristics of these participants at baseline and scores of all PHQ-8 records is shown in Table 3.2. The Kruskal-Wallis test was used to determine whether there were any significant differences for these characteristics between the sites. These tests revealed that, except for gender, sociodemographic characteristics and distribution of PHQ-8 scores differed between the study sites. The histograms of PHQ-8 scores of the study sites and the entire dataset are shown in Figure 1. We can observe that the KCL site had the most PHQ-8 records among the sites. PHQ-8 scores from the CIBER site were relatively high, probably because participants in the CIBER site came from a clinical population. Figure 2 presents pairwise Spearman correlation coefficients between all 18 sleep features. Table 3.3 shows the results of Spearman correlation analysis; we can observe there was a strong positive correlation between the sleep subscore and PHQ-8 score (*r*=.73,

z=54.48, P<.001) on the entire dataset, but this correlation was relatively weaker on the VUmc data (r=.64, z=18.75, P<.001).

Table 3.2. A summary of sociodemographic characteristics and PHQ-8 records of participants from the 3 study sites and results of Kruskal-Wallis tests on these characteristics.

Characteristic	KCL[a]	CIBER[b]	VUmc[c]	*P* value[d]
Participants, n	189	96	83	—[e]
PHQ-8[f] records, n	1547	708	557	—
PHQ-8 scores, median (Q1, Q3)	8 (4, 12)	14 (8, 19)	9 (5, 13)	<.001
The PHQ-8 score ≥10, n (%)	599 (38.7)	492 (69.5)	248 (44.5)	<.001
Age at baseline, median (Q1, Q3)	46 (30.3, 59.0)	55 (49.3, 60.8)	42 (28.0, 57.0)	<.001
Female sex, n (%)	144 (76.2)	69 (71.9)	65 (81.9)	.62
Education[g], n (%)	—	—	—	<.001
Degree or above	116 (61.4)	21 (21.9)	40 (48.2)	—
Below degree	73 (38.6)	75 (78.1)	43 (51.8)	—
Annual income[h] (£), n (%)	—	—	—	.009
<15,000	40 (21.2)	28 (29.2)	24 (28.9)	—
15,000-40,000	80 (42.3)	53 (55.2)	34 (41.0)	—
>40,000	67 (35.5)	15 (15.6)	14 (16.9)	—
Not mentioned	2 (1.1)	0 (0)	11 (13.3)	—

[a]KCL: King's College London.

[b]CIBER: Centro de Investigación Biomédican en Red.

[c]VUmc: Vrije Universiteit Medisch Centrum.

[d]*P* value of Kruskal-Wallis test.

[e]Not applicable.

[f]PHQ-8: 8-item Patient Health Questionnaire.

[g]Education levels of Spain and the Netherlands transformed into equivalent British education levels.

[h]Annual income levels of Spain and the Netherlands transformed into equivalent British levels.

Figure 3.1. Histograms of the PHQ-8 scores of the three study sites and the entire dataset.

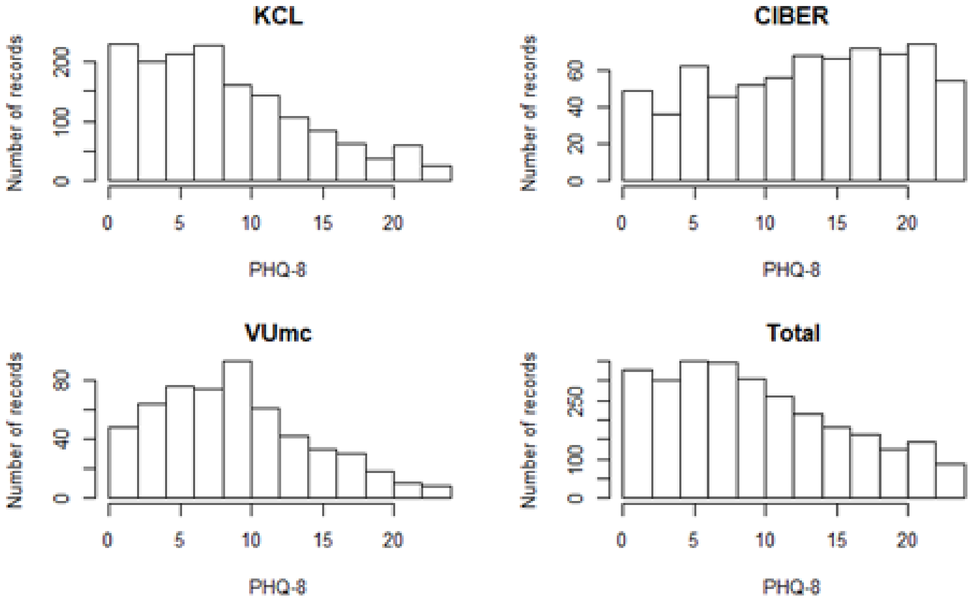


Figure 3.2. Correlation plot of pairwise Spearman correlations between all sleep features. Descriptions of abbreviations of sleep features are shown in Table 3.1.

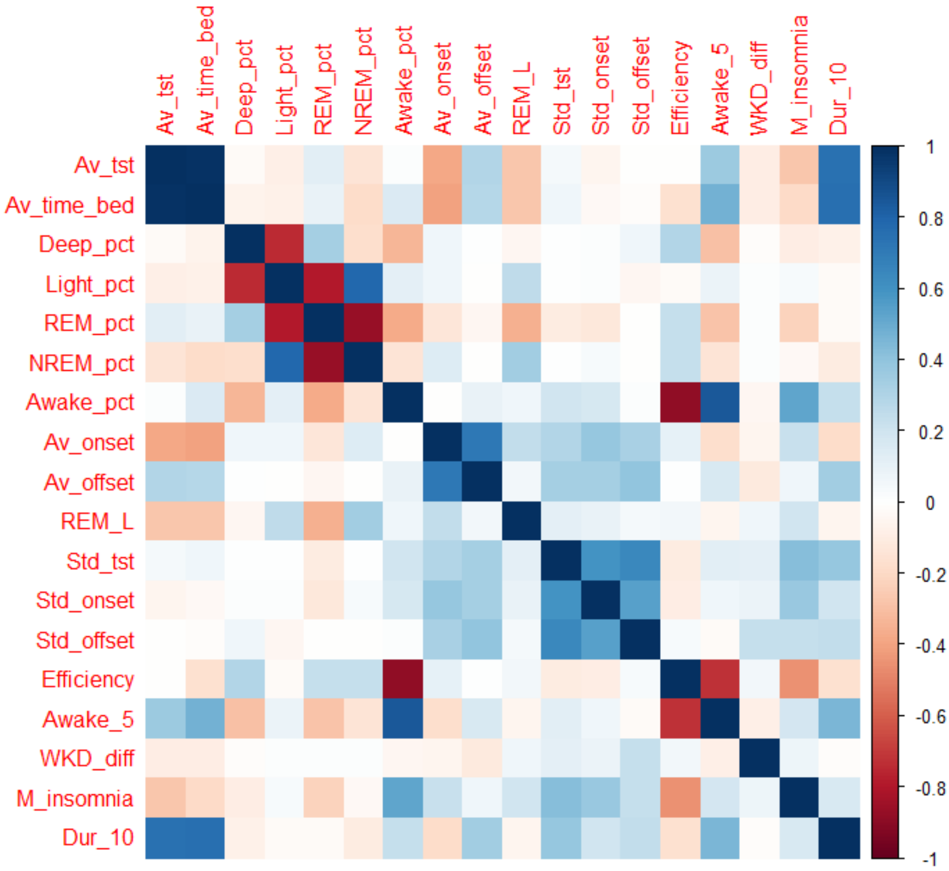


Table 3.3. Spearman correlation coefficients between the PHQ-8 score and sleep subscore[a] on the 3 study sites and their 95% confidence intervals, *z* score statistics, and *P* values.

Study site	*r*	95% CI	*z* score	*P* value
KCL[b]	.74	0.71, 0.76	41.99	<.001
CIBER[c]	.78	0.75, 0.81	32.09	<.001
VUmc[d]	.64	0.58, 0.69	18.75	<.001
Total	.73	0.71, 0.74	54.48	<.001

[a] Sleep subscore represents the score of subitem 3 in the PHQ-8.

[b] KCL: King’s College London.

[c] CIBER: Centro de Investigación Biomédican en Red.

[d] VUmc: Vrije Universiteit Medisch Centrum.

3.3.2 Three-Level Linear Mixed Models on the Entire Dataset

Table 3.4 shows the results from 3-level linear mixed regression models that reflect the associations between sleep features and the PHQ-8 score and sleep subscore, respectively. A total of 14 sleep features were found to be significantly associated with the PHQ-8 score, among them awake percentage (*z*=5.45, *P*<.001), awakening times (*z*=5.53, *P*<.001), insomnia (*z*=4.55, *P*<.001), mean sleep offset time (*z*=6.19, *P*<.001), and hypersomnia (*z*=5.30, *P*<.001) were the top 5 features ranked by *z* score statistics. The percentages of light sleep (*Light_pct*) and NREM sleep (*NREM_pct*) and sleep efficiency (*Efficiency*) were significantly and negatively associated with the PHQ-8 score, whereas the rest of the significant features were positively associated with the PHQ-8 score.

For sleep subscore, we can notice that deep sleep percentage (*Deep_pct*), REM sleep percentage (*REM_pct*), and sleep efficiency (*Efficiency*) were significantly and negatively associated with the sleep subscore, whereas features of the percentage of awake time (*Awake_pct*), unstable sleep (*Std_tst*, *Std_onset*, *Std_offset*), awakening times (*Awake_5*), weekend catch-up sleep (*WKD_diff*), sleep onset time (*Av_onset*), sleep offset time (*Av_offset*), insomnia (*M_insomnia*), and hypersomnia (*Dur_10*) were significantly and positively associated with the sleep subscore.

Table 3.4. Slope coefficient estimates, 95% confidence intervals, *z* score statistics, and *P* values from 3-level linear mixed models on the entire dataset for exploring associations between sleep features[a] and the PHQ-8 score and sleep subscore[b].

Features		PHQ-8[c] score				Sleep subscore		
	Coeff.[d]	95% CI	z score	*P* value	Coeff.	95% CI	z score	*P* value
Av_tst	0.013	0.006, 0.019	3.93	<.001	–0.004	–0.034, 0.025	–0.28	.78
Av_time_bed	0.016	0.009, 0.023	4.45	<.001	0.005	–0.028, 0.038	0.29	.77
Deep_pct	–0.007	–0.026, 0.011	–0.75	.45	–0.104	–0.191, –0.017	–2.34	.02
Light_pct	–0.032	–0.064, –0.001	–2.02	.04	0.090	–0.057, 0.237	1.20	.23
REM_pct	0.003	–0.021, 0.027	0.25	.80	–0.125	–0.238, –0.012	–2.17	.03
NREM_pct	–0.038	–0.062, –0.014	–3.12	.002	–0.014	–0.127, 0.098	–0.25	.80
Awake_pct	0.035	0.022, 0.048	5.45	<.001	0.139	0.079, 0.199	4.58	<.001
Av_onset	0.007	–0.001, 0.015	1.71	.09	0.078	0.040, 0.115	4.03	<.001
Av_offset	0.025	0.017, 0.033	6.19	<.001	0.097	0.060, 0.135	5.10	<.001
REM_L	0.034	–0.021, 0.088	1.21	.23	0.085	–0.178, 0.347	0.63	.53
Std_tst	0.008	0.004, 0.012	4.07	<.001	0.047	0.028, 0.067	4.77	<.001
Std_onset	0.012	0.004, 0.019	3.11	.002	0.060	0.022, 0.097	3.13	.002
Std_offset	0.012	0.005, 0.018	3.58	<.001	0.069	0.037, 0.100	4.26	<.001
Efficiency	–0.025	–0.037, –0.012	–3.91	<.001	–0.108	–0.167, –0.050	–3.65	<.001
Awake_5	0.016	0.010, 0.022	5.53	<.001	0.038	0.011, 0.065	2.77	.006
WKD_diff	0.134	0.039, 0.230	2.76	.006	0.747	0.255, 1.240	2.98	.003
M_insomnia	0.370	0.211, 0.530	4.55	<.001	2.373	1.595, 3.151	5.98	<.001
Dur_10	0.309	0.195, 0.423	5.30	<.001	0.909	0.357, 1.462	3.23	.001

[a]Definitions of sleep features in this table are shown in Table 3.1.

[b]Sleep subscore represents the score of subitem 3 in the PHQ-8.

[c]PHQ-8: 8-item Patient Health Questionnaire.

[d]Slope coefficient estimates for all sleep features.

3.3.3 Two-Level Linear Mixed Models on Different Research Sites

Table 3.5 provides the results from 2-level linear mixed models which show the associations between sleep features and the PHQ-8 score on different research sites separately. On the KCL data, most associations between sleep features and depression were consistent with the results on the entire dataset. On the CIBER data, some features were no longer significantly associated with the PHQ-8 score. However, on the VUmc data, most features lost their significance except features of total sleep time (*Av_tst*), time in bed (*Av_time_bed*), REM latency (*REM_L*), and awakenings (*Awake_5*).

Table 3.6 shows associations between sleep features and the sleep subscore on different research sites. The significance of associations between sleep features and the sleep subscore were different among the 3 study sites. Notably, the insomnia feature (*M_insomnia*) and at least one feature of sleep stability were significantly positively associated with sleep subscore on the data of all 3 sites.

Table 3.5. Coefficient estimates, 95% confidence intervals, and *P* values from 2-level linear mixed models on the 3 study sites for exploring associations between sleep features[a] and the PHQ-8 score.

Features	KCL[b]			CIBER[c]			VUmc[d]		
	Coeff.[e]	95% CI	*P*	Coeff.	95% CI	*P*	Coeff.	95% CI	*P*
Av_tst	0.013	0.005, 0.020	.001	0.016	–0.001, 0.033	.06	0.011	0, 0.022	.049
Av_time_bed	0.016	0.008, 0.024	<.001	0.021	0.002, 0.040	.03	0.013	0.001, 0.025	.04
Deep_pct	–0.005	–0.028, 0.018	.69	0.024	–0.022, 0.071	.31	–0.037	–0.074, 0.001	.06
Light_pct	–0.046	–0.087, –0.006	.03	–0.081	–0.155, –0.007	.03	0.019	–0.043, 0.082	.55
REM_pct	0.013	–0.018, 0.043	.43	0.015	–0.042, 0.071	.62	–0.007	–0.055, 0.041	.77
NREM_pct	–0.049	–0.080, –0.018	.002	–0.060	–0.116, –0.005	.04	–0.016	–0.062, 0.030	.50
Awake_pct	0.037	0.020, 0.054	<.001	0.043	0.015, 0.071	.003	0.022	–0.003, 0.047	.09
Av_onset	0.010	0.000, 0.020	.047	0.004	–0.018, 0.025	.74	–0.005	–0.021, 0.010	.52
Av_offset	0.029	0.018, 0.039	<.001	0.024	0.004, 0.043	.02	0.012	–0.004, 0.029	.14
REM_L	0.019	–0.049, 0.088	.58	0.106	–0.026, 0.237	.12	–0.126	–0.231, –0.020	.02
Std_tst	0.008	0.003, 0.013	.001	0.009	0, 0.019	.06	0.002	–0.006, 0.010	.62
Std_onset	0.007	–0.002, 0.016	.14	0.019	–0.001, 0.039	.06	0.001	–0.011, 0.013	.93
Std_offset	0.009	0.001, 0.017	.03	0.019	0.002, 0.036	.03	0.003	–0.008, 0.015	.56
Efficiency	–0.025	–0.041, –0.008	.004	–0.043	–0.071, –0.016	.002	–0.012	–0.037, 0.013	.34
Awake_5	0.014	0.006, 0.022	<.001	0.022	0.009, 0.035	.001	0.016	0.005, 0.027	.01
WKD_diff	0.211	0.084, 0.339	.001	0.071	–0.126, 0.268	.48	0.077	–0.144, 0.299	.49
M_insomnia	0.472	0.259, 0.685	<.001	0.381	0.028, 0.734	.04	–0.048	–0.385, 0.289	.78
Dur_10	0.331	0.191, 0.472	<.001	0.340	0.052, 0.627	.02	0.181	–0.051, 0.413	.13

[a] Definitions of sleep features in this table are shown in Table 3.1.

[b] KCL: King's College London.

[c] CIBER: Centro de Investigación Biomédican en Red.

[d] VUmc: Vrije Universiteit Medisch Centrum.

[e] Slope coefficient estimates for all sleep features.

Table 3.6. Coefficient estimates, 95% confidence intervals, and *P* values from 2-level linear mixed models on the 3 study sites for exploring associations between sleep features[a] and the sleep subscore[b].

Features	KCL[c]			CIBER[d]			VUmc[e]		
	Coeff.[f]	95% CI	*P*	Coeff.	95% CI	*P*	Coeff.	95% CI	*P*
Av_tst	0.015	–0.021, 0.050	.41	–0.035	–0.116, 0.047	.41	–0.017	–0.070, 0.035	.52
Av_time_bed	0.026	–0.013, 0.066	.19	–0.025	–0.116, 0.065	.58	–0.015	–0.074, 0.043	.61
Deep_pct	–0.027	–0.134, 0.081	.63	–0.196	–0.412, 0.020	.07	–0.191	–0.369, –0.014	.04
Light_pct	–0.024	–0.213, 0.166	.81	0.098	–0.250, 0.445	.58	0.312	0.016, 0.608	.04
REM_pct	–0.116	–0.260, 0.028	.12	–0.037	–0.304, 0.230	.79	–0.169	–0.398, 0.060	.15
NREM_pct	–0.048	–0.194, 0.098	.52	–0.123	–0.389, 0.143	.37	0.125	–0.096, 0.346	.27
Awake_pct	0.165	0.085, 0.245	<.001	0.150	0.020, 0.280	.02	0.049	–0.073, 0.170	.43
Av_onset	0.055	0.008, 0.101	.02	0.075	–0.023, 0.172	.13	0.128	0.054, 0.202	.001
Av_offset	0.102	0.053, 0.150	<.001	0.048	–0.040, 0.135	.29	0.133	0.056, 0.210	.001
REM_L	0.073	–0.255, 0.401	.66	0.146	–0.494, 0.787	.65	–0.171	–0.683, 0.340	.51
Std_tst	0.046	0.022, 0.071	<.001	0.046	–0.002, 0.094	.06	0.043	0.004, 0.082	.03
Std_onset	0.028	–0.015, 0.070	.21	0.089	–0.018, 0.195	.10	0.079	0.020, 0.139	.01
Std_offset	0.046	0.008, 0.084	.02	0.109	0.022, 0.195	.01	0.072	0.016, 0.127	.01
Efficiency	–0.118	–0.196, –0.041	.003	–0.152	–0.280, –0.024	.02	–0.044	–0.162, 0.074	.46
Awake_5	0.047	0.011, 0.083	.01	0.037	–0.022, 0.097	.22	0.013	–0.042, 0.067	.65
WKD_diff	1.169	0.534, 1.804	<.001	0.210	–0.864, 1.284	.70	0.283	–0.830, 1.395	.62
M_insomnia	2.302	1.274, 3.329	<.001	2.777	1.070, 4.485	.001	1.823	0.180, 3.465	.03
Dur_10	1.057	0.387, 1.728	.002	0.576	–0.844, 1.995	.43	0.706	–0.411, 1.823	.22

[a] The definitions of sleep features in this table are shown in Table 3.1.

[b] The sleep subscore represents the score of subitem 3 in the PHQ-8.

[c] KCL: King’s College London.

[d] CIBER: Centro de Investigación Biomédican en Red.

[e] VUmc: Vrije Universiteit Medisch Centrum.

[f] Slope coefficient estimates for all sleep features.

3.4 Discussion

3.4.1 Principal Findings

In this study, we extracted 18 sleep features through Fitbit data to quantitatively describe participant sleep characteristics in 5 categories (sleep architecture, sleep stability, sleep quality, insomnia, and hypersomnia) associated with the severity of depression. Along with the depressive status worsening, the following changes may be seen in the past 2 weeks: (1) percentage of light/NREM sleep decreased and the percentage of wakefulness during sleep increased (sleep architecture); (2) sleep duration/onset/offset were unstable (sleep stability); (3) reduced sleep efficiency, more awakenings during sleep, and longer weekend catch-up sleep were observed (sleep quality); (4) more days with insomnia were observed (insomnia); (5) more days with hypersomnia were observed (hypersomnia). Table 3.4 illustrated that our sleep features of these 5 categories could reflect both the participant sleep condition (sleep subscore) and depressive symptom severity (PHQ-8 score) of the past 2 weeks.

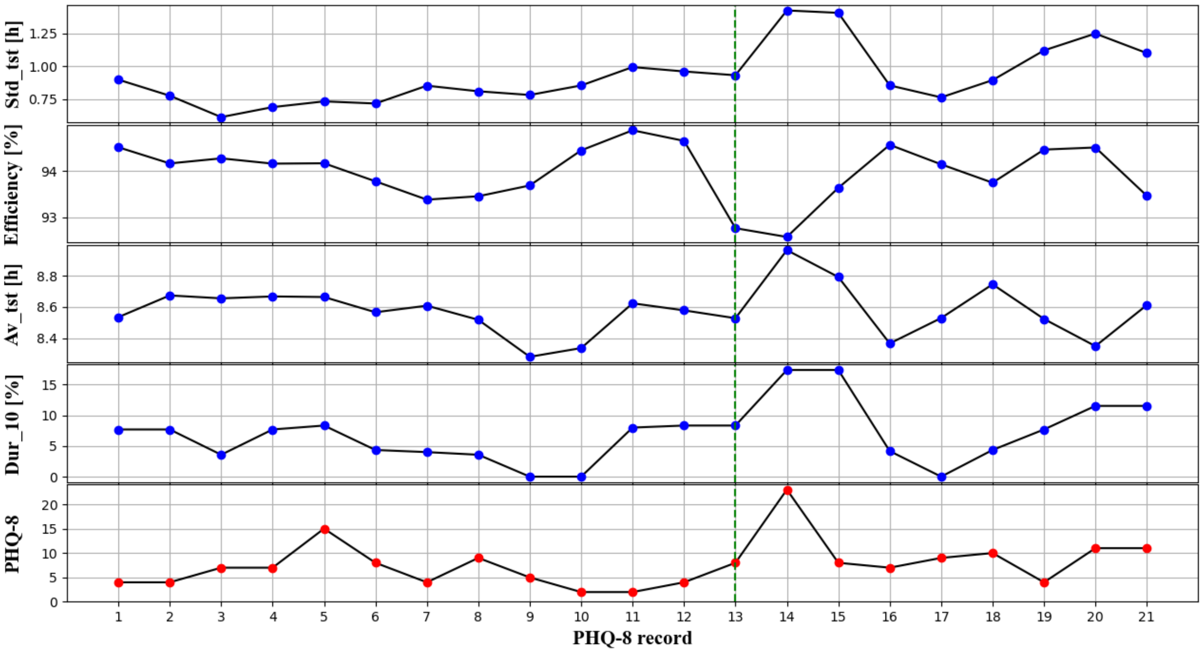
3.4.2 Potential Factors Affecting Associations

We evaluated our models on the research sites separately. From Table 3.5 and Table 3.6, we can notice that the associations between sleep features and PHQ-8 score/sleep subscore varied across different sites. Several factors may affect the associations. First, the populations of the 3 sites were significantly different (Table 3.2). For example, participants in the CIBER site came from a clinical population and their average age was oldest, so one speculation is that there was less difference between their weekday sleep and weekend sleep for inpatients or people in retirement. Therefore, this may be the reason why the feature of weekend catch-up sleep (*WKD_diff*) lost significance on the CIBER data. In addition, the reduced significance of features related to sleep onset and offset time on the CIBER site might be related to the regular sleep pattern in CIBER site favors going to bed later, as seen in our previous study (Sun et al., 2020).

The associations between sleep features and the sleep subscore on the VUmc data (Table 3.6) were similar to that in the entire dataset (Table 3.4), which demonstrated sleep features have the same ability to capture the sleep condition of participants on the VUmc data. However, the significance of associations between these sleep features and the PHQ-8 score was reduced in the VUmc data (Table 3.5). One possible reason is that, as seen on Table 3.3, the correlation between the sleep subscore and PHQ-8 score in the VUmc data (*r*=.64) was weaker than other 2 study sites (KCL: *r*=.74 and CIBER: *r*=.78), which may be caused by confounding variables that we did not consider or record in the VUmc population such as medication and occupational status.

Sample size and heterogeneity of the dataset were other possible factors that may affect results. Table 3.2 shows that the KCL site had the most PHQ-8 records, whereas VUmc had the least data. As depression manifests itself in distinctive symptoms on different people, it may be difficult to fully explore the associations between sleep and depression on a relatively smaller dataset (VUmc). For example, hypersomnia is specifically related to bipolar patients (Detre et al., 1972; Thase et al., 1989); therefore, if the dataset did not contain enough bipolar patients or bipolar patients were not in depressive episodes when they completed their PHQ-8 records, it would be hard to find the association between hypersomnia and depression.

Figure 3.3. The PHQ-8 scores and a select 4 sleep features of one participant with an obvious increasing trend in PHQ-8 score at 13th PHQ-8 record. Descriptions of abbreviations of sleep features in this figure are shown in Table 3.1.



3.4.3 Comparison with Prior Work

Our study has a relatively larger sample size and a longer follow-up duration than previous studies on monitoring depression by using wearable devices and mobile phones (Demasi et al., 2016; Mark et al., 2016; Miwa et al., 2007). Each participant has multiple PHQ-8 records and repeated measurements of sleep, so we can not only explore the relationships between sleep and depression between individuals but also find the associations within individuals by using the linear mixed model. Figure 3 is an example of a possible depression relapse of one participant, showing an obvious increasing trend in PHQ-8 scores at the 13th PHQ-8 record of this participant. We can observe the sleep features in Figure 3 are significantly associated with the PHQ-8 score.

This indicates that the sleep features extracted in this paper have the potential to be the biomarkers of depression.

We also compared our findings with previous studies that used other measurements to assess sleep, such as PSG and sleep questionnaires. Although the sample size, population, measurements, duration of these studies are different, the comparison may help to find more general associations between sleep and depression. Table 3.7 provides a summary of the comparison. Several longitudinal studies based on sleep questionnaires have shown that insomnia and hypersomnia are both symptoms of depression (Alvaro et al., 2013; Kaplan & Harvey, 2009), which we found in our research. Kang et al found the weekend catch-up sleep was significantly positively correlated with the severity of depression by analyzing the self-sleep questionnaires of 4553 Korean adolescents, and this is consistent with the finding in our paper (Kang et al., 2014). A sleep report has shown that higher sleep efficiency, more deep sleep, and fewer awakenings after sleep onset represent better sleep quality (Ohayon et al., 2017), which is also consistent with the relationships we found between deep sleep percentage, awake percentage, and awakenings (>5 minutes) with sleep subscore. A review showed that according to PSG research, the shortened REM latency and increased percentage of REM sleep are biological markers of depression relapse (Palagini et al., 2013); however, relationships between depressive symptom severity with REM sleep percentage and REM latency were not significant in our results.

Table 3.7. Summary of the comparisons with previous studies using other measurements to assess sleep.

Type of feature	Findings in previous studies	Consistent[a]	Measurement
Insomnia	Insomnia is significantly related to depression (Alvaro et al., 2013).	Yes	Questionnaire
Hypersomnia	Prevalence of hypersomnia is high in depressed patients (Kaplan & Harvey, 2009).	Yes	Questionnaire
Weekend catch-up sleep	Weekend catch-up sleep is significantly positively correlated with the severity of depression (Kang et al., 2014).	Yes	Questionnaire
Deep sleep percentage	More deep sleep represents higher sleep quality (Ohayon et al., 2017).	Yes	Questionnaire
Awake percentage, Awakenings (>5 mins)	Fewer awakenings after sleep onset represents better sleep quality (Ohayon et al., 2017).	Yes	Questionnaire
Sleep efficiency	Higher sleep efficiency represents better sleep quality (Ohayon et al., 2017).	Yes	Questionnaire
REM sleep percentage	Increased REM sleep percentage can be biomarkers of depression (Palagini et al., 2013).	No	PSG
REM[b] latency	Shortened REM latency can be biomarkers of depression (Palagini et al., 2013).	No	PSG

[a]Whether it is consistent with our findings.
[b]REM: rapid eye movement.

3.4.4 Limitations

Missing data is the major hindrance in our study. For various reasons, there were many missing records of sleep. We set the completion rate of sleep records greater than 85% (12 days) as one of the data inclusion criteria. However, the optimum threshold is unclear, which needs to be further studied in future research. Missingness could also be associated with depressive status and could be a useful marker of relapse of depression;

for example, participants may not feel like complying if they are feeling depressed. In future research, we will consider missingness as a potential feature.

Although we adjusted our models for age, gender, education level, and annual income, it is hard to consider all potential confounding variables. For example, some participants with sleep disorders may take sleep medications. Sleep medications have a significant influence on the features of sleep. Unfortunately, there was no daily record of whether the participant took medication. This confounding variable may affect the result.

The data of sleep stages used in this paper were provided by the Fitbit wristband. According to their validation studies, the Fitbit wristband showed promise in detecting sleep-wake states but limitations in other sleep stages estimation (de Zambotti et al., 2018; Haghayegh et al., 2019; Liang & Chapa-Martell, 2019). This may be the reason the features of REM percentage and REM latency in our paper did not show significant relationships with depressive symptoms. For detecting insomnia, the sleep onset latency (SOL) in the PSG report is a reliable indicator of insomnia, but the Charge 2 and 3 are not able to measure SOL directly. The features related to insomnia in our paper can partially reflect insomnia, but they may be affected by factors (such as work schedules or activities) other than insomnia. Therefore, in future research, we will combine multiple features (such as a late sleep onset time accompanied by a short total sleep time) to determine whether a participant has insomnia and try to use activity information (e.g., steps) provided by Fitbit to approximate SOL. Although there are some limitations of Fitbit data, it provides a means to investigate sleep characteristic in home settings.

In feature extraction, we did not consider the impact of individual circumstances on sleep features. For example, some participants may need to shift work at night, which our features are unable to capture. We will consider the impact of sleep habits and lifestyles on sleep features in the future. Further, we did not explore the impact of individual patterns of depression (Brailean et al., 2020)—for example, the distinction between people with typical and atypical depression who report reduced and increased sleep, respectively, during depressive episodes. In future work, we will explore whether including this dimension improves specificity of our findings.

In this paper, we focused on analyzing the manifestations of depression in sleep characteristics. We will investigate whether these relationships are bidirectional in future research. We only performed bivariate analysis (ie, separately analyzing the association between each feature and the PHQ-8 score). The combination of features and nonlinear relationships was not considered. We will try to apply machine/deep learning models to predict the severity of depression by using sleep features in future research.

3.4.5 Conclusions

Although consumer wearable devices may not be a substitute for PSG to assess sleep quality accurately, we demonstrated that some derived sleep features extracted from these wearable devices show potential for remote measurement of sleep and consequently can act as a biomarker of depression in real-world settings. These findings may provide the basis for the development of clinical tools that could be used to passively monitor disease state and trajectory with minimal burden on the participant.

Acknowledgments

The RADAR-CNS project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant 115902. This joint undertaking receives support from the European Union's Horizon 2020 research and innovation program and European Federation of Pharmaceutical Industries and Associations (EFPIA). This communication reflects the views of the RADAR-CNS consortium and neither the Innovative Medicines Initiative nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. Participants in the CIBER site came from the following 4 clinical communities in Spain: Parc Sanitari Sant Joan de Déu Network services, Institut Català de la Salut, Institut Pere Mata, and Hospital Clínico San Carlos. Participant recruitment in Amsterdam was partially accomplished through Hersenonderzoek.nl, a Dutch online registry that facilitates participant recruitment for neuroscience studies (https://hersenonderzoek.nl). Hersenonderzoek.nl is funded by grant 73305095003 from ZonMw-Memorabel, a project in the context of the Dutch Deltaplan Dementie, Gieskes-Strijbis Foundation, the Alzheimer's Society in the Netherlands, and Brain Foundation Netherlands. This paper represents independent research partially funded by the National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley National Health

Service (NHS) Foundation Trust and King’s College London. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or the Department of Health and Social Care. RB is funded in part by grant MR/R016372/1 from the King’s College London Medical Research Council Skills Development Fellowship program funded by the UK Medical Research Council and by grant IS-BRC-1215-20018 from the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London.

Authors’ Contributions

YZ extracted and integrated the depression questionnaires and Fitbit sleep data for the analysis, planned and performed the analysis, and drafted the manuscript. MH and VAN gained funding and co-led the Remote Assessment of Disease and Relapse–Central Nervous System program. MH is the principal investigator for the Remote Assessment of Disease and Relapse–Major Depressive Disorder study. RJBD, AAF, YR, ZR, PC, and CS have contributed to the development of the RADAR-base platform used for data collection and management across sites, data protection, security, and storage. YZ, AAF, S Sun, NC, RB, PL, MH, and RJBD contributed to the design of the study. FM, KMW, FL, S Siddi, S Simblett, JMH, BWJHP, MH contributed to data collection. AAF, IMG, AR, VAN, TW, MH, and RJBD contributed to the administrative, technical, and clinical support of the study. All authors were involved in reviewing the manuscript, had access to the study data, and provided direction and comments on the manuscript.

Conflicts of Interest

VAN is an employee of Janssen Research and Development LLC and may own equity in the company.

References

Akhtar-Danesh, N., & Landeen, J. (2007). Relation between depression and sociodemographic factors. *International Journal of Mental Health Systems, 1*(1), 4.

Aluoja, A., Leinsalu, M., Shlik, J., Vasar, V., & Luuk, K. (2004). Symptoms of depression in the Estonian population: Prevalence, sociodemographic correlates and social adjustment. *Journal of Affective Disorders, 78*(1), 27–35.

Alvaro, P. K., Roberts, R. M., & Harris, J. K. (2013). A Systematic Review Assessing Bidirectionality between Sleep Disturbances, Anxiety, and Depression. *Sleep, 36*(7), 1059–1068.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association.

Beattie, Z., Oyang, Y., Statan, A., Ghoreyshi, A., Pantelopoulos, A., Russell, A., & Heneghan, C. (2017). Estimation of sleep stages in a healthy adult population from optical plethysmography and accelerometer signals. *Physiological Measurement, 38*(11), 1968–1979.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300.

Berry, R. B., Brooks, R., Gamaldo, C. E., Harding, S. M., Marcus, C., & Vaughn, B. V. (2012). The AASM manual for the scoring of sleep and associated events. *Rules, Terminology and Technical Specifications, Darien, Illinois, American Academy of Sleep Medicine, 176*, 2012.

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological), 26*(2), 211–243.

Brailean, A., Curtis, J., Davis, K., Dregan, A., & Hotopf, M. (2020). Characteristics, comorbidities, and correlates of atypical depression: Evidence from the UK Biobank Mental Health Survey. *Psychological Medicine, 50*(7), 1129–1138.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research, 28*(2), 193–213.

Cuijpers, P., & Schoevers, R. A. (2004). Increased mortality in depressive disorders: A review. *Current Psychiatry Reports, 6*(6), 430–437.

de Zambotti, M., Goldstone, A., Claudatos, S., Colrain, I. M., & Baker, F. C. (2018). A validation study of Fitbit Charge 2^{TM} compared with polysomnography in adults. *Chronobiology International*, *35*(4), 465–476.

Demasi, Aguilera, & Recht. (2016). Detecting change in depressive symptoms from daily wellbeing questions, personality, and activity. *2016 IEEE Wireless Health (WH)*, 1–8.

Detre, T., Himmelhoch, J., Swartzburg, M., Anderson, C. M., Byck, R., & Kupfer, D. J. (1972). Hypersomnia and Manic-Depressive Disease. *American Journal of Psychiatry*, *128*(10), 1303–1305.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., Kamath, J., Russell, A., Bamis, A., & Wang, B. (2016). Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data. *2016 IEEE Wireless Health, WH 2016*, 30–37.

Haghayegh, S., Khoshnevis, S., Smolensky, M. H., Diller, K. R., & Castriotta, R. J. (2019). Accuracy of Wristband Fitbit Models in Assessing Sleep: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *21*(11), e16273.

Kang, S.-G., Lee, Y. J., Kim, S. J., Lim, W., Lee, H.-J., Park, Y.-M., Cho, I. H., Cho, S.-J., & Hong, J. P. (2014). Weekend catch-up sleep is independently associated with suicide attempts and self-injury in Korean adolescents. *Comprehensive Psychiatry*, *55*(2), 319–325.

Kaplan, K. A., & Harvey, A. G. (2009). Hypersomnia across mood disorders: A review and synthesis. *Sleep Medicine Reviews*, *13*(4), 275–285.

Khoulji, S., Garzón-Rey, J. M., & Aguilo, J. (2017). Remote Assessment of Disease and Relapse – Central Nervous System- RADAR-CNS. *Transactions on Machine Learning and Artificial Intelligence*, *5*(4).

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Lenox-Smith, A., Macdonald, M. T. B., Reed, C., Tylee, A., Peveler, R., Quail, D., & Wildgust, H. J. (2013). Quality of Life in Depressed Patients in UK Primary Care: The FINDER Study. *Neurology and Therapy*, *2*(1–2), 25–42.

Lerner, D., Adler, D. A., Chang, H., Berndt, E. R., Irish, J. T., Lapitsky, L., Hood, M. Y., Reed, J., & Rogers, W. H. (2004). The Clinical and Occupational Correlates of Work Productivity Loss Among Employed Patients With Depression. *Journal of Occupational & Environmental Medicine*, *46*(6), S46–S55.

Liang, Z., & Chapa-Martell, M. A. (2019). Accuracy of Fitbit Wristbands in Measuring Sleep Stage Transitions and the Effect of User-Specific Factors. *JMIR MHealth and UHealth*, *7*(6), e13384.

Liu, X., Zhao, Z., Jia, C., & Buysse, D. J. (2008). Sleep patterns and problems among chinese adolescents. *Pediatrics*, *121*(6), 1165–1173.

Manni, R. (2005). Rapid eye movement sleep, non-rapid eye movement sleep, dreams, and hallucinations. *Current Psychiatry Reports*, *7*(3), 196–200.

Mark, G., Czerwinski, M., Iqbal, S., & Johns, P. (2016). Workplace Indicators of Mood: Behavioral and Cognitive Correlates of Mood Among Information Workers. *Proceedings of the 6th International Conference on Digital Health Conference*, 29–36.

Matcham, F., Barattieri Di San Pietro, C., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., Folarin, A. A., Haro, J. M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V., Mohr, D. C., Myin-Germeys, I., Narayan, V., Bwjh, P., Ranjan, Y., Rashid, Z., Rintala, A., … Meyer, N. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry*, *19*(1), 1–11.

Mendelson, W. (2012). *Human Sleep and Its Disorders*. Springer Science & Business Media.

Miwa, H., Sasahara, S., & Matsui, T. (2007). Roll-over Detection and Sleep Quality Measurement using a Wearable Sensor. *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1507–1510.

Moore, C. M., Schmiege, S. J., & Matthews, E. E. (2015). Actigraphy and Sleep Diary Measurements in Breast Cancer Survivors: Discrepancy in Selected Sleep Parameters. *Behavioral Sleep Medicine*, *13*(6), 472–490.

Ohayon, M., Wickwire, E. M., Hirshkowitz, M., Albert, S. M., Avidan, A., Daly, F. J., Dauvilliers, Y., Ferri, R., Fung, C., Gozal, D., Hazen, N., Krystal, A., Lichstein, K., Mallampalli, M., Plazzi, G., Rawding, R., Scheer, F. A., Somers, V., & Vitiello, M. V. (2017). National Sleep Foundation’s sleep quality recommendations: First report. *Sleep Health*, *3*(1), 6–19.

Palagini, L., Baglioni, C., Ciapparelli, A., Gemignani, A., & Riemann, D. (2013). REM sleep dysregulation in depression: State of the art. *Sleep Medicine Reviews*, *17*(5), 377–390.

Ranjan, Y., Rashid, Z., Stewart, C., Conde, P., Begale, M., Verbeeck, D., Boettcher, S., Hyve, Dobson, R., Folarin, A., & RADAR-CNS Consortium. (2019). RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. *JMIR MHealth and UHealth*, *7*(8), e11734.

Rasch, B., & Born, J. (2013). About sleep’s role in memory. *Physiological Reviews*, *93*(2), 681–766.

Riemann, D., Berger, M., & Voderholzer, U. (2001). Sleep and depression--results from psychobiological studies: An overview. *Biological Psychology*, *57*(1–3), 67–103.

Sánchez-Ortuño, M. M., Edinger, J. D., Means, M. K., & Almirall, D. (2010). Home is where sleep is: An ecological approach to test the validity of actigraphy for the assessment of insomnia. *Journal of Clinical Sleep Medicine: JCSM: Official Publication of the American Academy of Sleep Medicine*, *6*(1), 21–29.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence* (1st ed.). Oxford University PressNew York.

Sun, S., Folarin, A. A., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Cummins, N., Matcham, F., Dalla Costa, G., Simblett, S., Leocani, L., Lamers, F., Sørensen, P. S., Buron, M., Zabalza, A., Guerrero Pérez, A. I., Penninx, B. W., Siddi, S., Haro, J. M., … RADAR-CNS Consortium. (2020). Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19. *Journal of Medical Internet Research*, *22*(9), e19992.

Tam, E. M., Lam, R. W., Robertson, H. A., Stewart, J. N., Yatham, L. N., & Zis, A. P. (1997). Atypical depressive symptoms in seasonal and non-seasonal mood disorders. *Journal of Affective Disorders*, *44*(1), 39–44.

Thase, M. E., Himmelhoch, J. M., Mallinger, A. G., Jarrett, D. B., & Kupfer, D. J. (1989). Sleep EEG and DST findings in anergic bipolar depression. *The American Journal of Psychiatry*, *146*(3), 329–333.

Van de Water, A. T. M., Holmes, A., & Hurley, D. A. (2011). Objective measurements of sleep for non-laboratory settings as alternatives to polysomnography—A systematic review. *Journal of Sleep Research*, *20*(1 Pt 2), 183–200.

Walker, M. P. (2008). Sleep-dependent memory processing. *Harvard Review of Psychiatry*, *16*(5), 287–298.

World Health Organization. (2017). Depression and other common mental disorders: Global health estimates. *World Health Organization*. https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf

Zhang, Y., Yang, Z., Lan, K., Liu, X., Zhang, Z., Li, P., Cao, D., Zheng, J., & Pan, J. (2019). Sleep Stage Classification Using Bidirectional LSTM in Wearable Multi-sensor Systems. *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 443–448.

Zhang, Y., Yang, Z., Zhang, Z., Li, P., Cao, D., Liu, X., Zheng, J., Yuan, Q., & Pan, J. (2018). Poster abstract: Breathing disorder detection using wearable electrocardiogram and oxygen saturation. *SenSys 2018 - Proceedings of the 16th Conference on Embedded Networked Sensor Systems*, 313–314.

Chapter 4

Predicting Depressive Symptom Severity Through Individuals' Nearby Bluetooth Device Count Data Collected by Mobile Phones: Preliminary Longitudinal Study

This chapter has been published as:

Zhang Y, Folarin AA, Sun S, Cummins N, Ranjan Y, Rashid Z, Conde P, Stewart C, Laiou P, Matcham F, Oetzmann C, Lamers F, Siddi S, Simblett S, Rintala A, Mohr DC, Myin-Germeys I, Wykes T, Haro JM, Penninx BWJH, Narayan VA, Annas P, Hotopf M, Dobson RJB, and RADAR-CNS Consortium. Predicting Depressive Symptom Severity Through Individuals' Nearby Bluetooth Device Count Data Collected by Mobile Phones: Preliminary Longitudinal Study. *JMIR Mhealth Uhealth 2021;9*(7): e29840

Background: Research in mental health has found associations between depression and individuals' behaviors and statuses, such as social connections and interactions, working status, mobility, and social isolation and loneliness. These behaviors and statuses can be approximated by the nearby Bluetooth device count (NBDC) detected by Bluetooth sensors in mobile phones.

Objective: This study aimed to explore the value of the NBDC data in predicting depressive symptom severity as measured via the 8-item Patient Health Questionnaire (PHQ-8).

Methods: The data used in this paper included 2886 biweekly PHQ-8 records collected from 316 participants recruited from three study sites in the Netherlands, Spain, and the United Kingdom as part of the EU Remote Assessment of Disease and Relapse-Central Nervous System (RADAR-CNS) study. From the NBDC data 2 weeks prior to each PHQ-8 score, we extracted 49 Bluetooth features, including statistical features and nonlinear features for measuring the periodicity and regularity of individuals' life rhythms. Linear mixed-effect models were used to explore associations between Bluetooth features and the PHQ-8 score. We then applied hierarchical Bayesian linear regression models to predict the PHQ-8 score from the extracted Bluetooth features.

Results: A number of significant associations were found between Bluetooth features and depressive symptom severity. Generally speaking, along with depressive symptom worsening, one or more of the following changes were found in the preceding 2 weeks of the NBDC data: (1) the amount decreased, (2) the variance decreased, (3) the periodicity (especially the circadian rhythm) decreased, and (4) the NBDC sequence became more irregular. Compared with commonly used machine learning models, the proposed hierarchical Bayesian linear regression model achieved the best prediction metrics (R^2=0.526) and a root mean squared error (RMSE) of 3.891. Bluetooth features can explain an extra 18.8% of the variance in the PHQ-8 score relative to the baseline model without Bluetooth features (R^2=0.338, RMSE=4.547).

Conclusions: Our statistical results indicate that the NBDC data have the potential to reflect changes in individuals’ behaviors and statuses concurrent with the changes in the depressive state. The prediction results demonstrate that the NBDC data have a significant value in predicting depressive symptom severity. These findings may have utility for the mental health monitoring practice in real-world settings.

4.1 Introduction

Existing studies have demonstrated that depression is significantly associated with individuals’ behaviors and statuses, such as social connections and interactions, working status, mobility, and social isolation and loneliness (Burgess et al., 2000; Cacioppo et al., 2006; Lampinen & Heikkinen, 2003; Rizvi et al., 2015). For example, individuals reporting fewer social network connections or less social support tend to have higher depressive symptomatology (Cacioppo et al., 2006). As the depressive mood and medical comorbidity can make people unable to work, the unemployment rate in depression is high (Rizvi et al., 2015). Reduced mobility and physical activity are associated with depressive symptoms (Lampinen & Heikkinen, 2003). Loneliness is a specific risk factor for depression, and a significant proportion of suicides have a history of social isolation (Burgess et al., 2000; Cacioppo et al., 2006). Although these findings have been replicated in different populations, these studies relied on participant self-report, which is susceptible to recall bias and typically does not capture dynamic information (Boonstra et al., 2017).

Mobile phone technology provides an unobtrusive, continuous, and cost-efficient means to capture individuals’ daily behaviors and statuses using a number of embedded sensors, such as accelerometers, GPS sensors, and Bluetooth sensors (Rohani et al., 2018). The embedded Bluetooth sensor can be used to record individuals’ local proximity information, such as the nearby Bluetooth device count (NBDC) that includes the Bluetooth signal of other phone users (Yan et al., 2013). The continuously recorded NBDC data represents a mixed signal that has been used to estimate individuals' behaviors and statuses, including face-to-face social interactions (Dissing et al., 2019; Eagle et al., 2009; Eagle & (Sandy) Pentland, 2006), working status (Clark et al., 2018), mobility (Nordström et al., 2007), and isolation and loneliness (Doryab et al., 2019; Wu et al., 2021). Therefore, the NBDC data have the potential to reflect changes in people's behaviors and statuses during the depressive state.

There have been a few studies exploring the relationship between the NBDC data and depression directly. Wang et al found a negative association (r=−0.362, P=.03) between the NBDC and self-reported depressive symptoms on the StudentLife data set, which contained mobile phone data from 48 students across a 10-week term at Dartmouth College (Wang et al., 2014). Boonstra et al illustrated the feasibility of collecting nearby Bluetooth device information for the depression recognition task, but they did not provide further findings (Boonstra et al., 2017).

Several recent studies have investigated the relationships between Bluetooth proximity data and mental health (Bogomolov et al., 2013, 2014; Moturu et al., 2011). Moturu et al found that individuals with lower sociability (estimated by the NBDC) tend to report lower mood more often (Moturu et al., 2011). Bogomolov et al established machine learning models to recognize happiness and stress with features of Bluetooth records, calls, and text messages, which obtained accuracy rates of 80.81% and 72.28%, respectively (Bogomolov et al., 2013, 2014). The above three studies were all performed on the "Friends and Family" data set, including 8 weeks of mobile phone data from 117 participants living in a major US university's married graduate student residency.

Previous studies (Bogomolov et al., 2013, 2014; Moturu et al., 2011; Wang et al., 2014) have been performed on relatively small (approximately 100 participants) homogeneous (e.g., university students) cohorts of participants over relatively short periods (8-10 weeks), which may limit their generalizability. Besides, Bluetooth features used in these studies have been limited to basic statistical features (e.g., sum, mean, and standard deviation), which are unable to characterize some nonlinear aspects (such as complexity, regularity, and periodicity) of the Bluetooth data. These nonlinear characteristics can reflect individuals' life rhythms, such as circadian and social rhythms, which are affected by depressive symptoms (Walker et al., 2020). Therefore, the associations between the NBDC data and depression are yet to be fully explored.

In this paper, we aimed to explore the value of the NBDC data in predicting self-reported depressive symptom severity in a relatively large cohort of individuals with a history of recurrent major depressive disorder. Our first objective was to explore the associations between statistical Bluetooth features and depressive symptom severity. Our second objective was to extract nonlinear features for quantifying complexity, regularity, and periodicity from the NBDC data and test their associations with depression. The third objective was to leverage appropriate machine learning models to predict the severity of depressive symptoms using extracted Bluetooth features.

4.2 Methods

4.2.1 Data Set

Study Participants and Settings

The data used in this study were collected from a major EU Innovative Medicines Initiative (IMI) research program Remote Assessment of Disease and Relapse-Central Nervous System (RADAR-CNS) (Khoulji et al., 2017). The project aimed to investigate the use of remote measurement technologies (RMTs) to monitor people with depression, epilepsy, and multiple sclerosis in real-world settings. The study protocol for the depression component (Remote Assessment of Disease and Relapse-Major Depressive Disorder; RADAR-MDD) has been described in detail by Matcham et al (Matcham et al., 2019). The RADAR-MDD project aimed to recruit 600 participants with a recent history of depression from three study sites in Spain (Centro de Investigación Biomédican en Red [CIBER], Barcelona), the Netherlands (Vrije Universiteit Medisch Centrum [VUmc], Amsterdam]), and the United Kingdom (King's College London [KCL]). Recruitment procedures varied slightly across sites with eligible participants identified through existing research infrastructures (in KCL and VUmc) where consent to be contacted for research purposes exists; advertisements in general practices, psychologist practices, and newspapers; Hersenonderzoek.nl (https://hersenonderzoek.nl), a Dutch online registry (VUmc); and mental health services (in KCL and CIBER) (Matcham et al., 2019).

Participants were asked to install passive and active remote monitoring technology (pRMT and aRMT, respectively) apps and use an activity tracker for up to 2 years of follow-up. Many categories of passive and active data were collected and uploaded to an open-source platform, RADAR-base (Ranjan et al., 2019).

As the purpose of this paper was to explore the value of the NBDC data in predicting self-reported depressive symptom severity, we focused on the NBDC data, 8-item Patient Health Questionnaire (PHQ-8) data (Kroenke et al., 2009), and baseline demographics. However, according to our previous research, the COVID-19 pandemic and related lockdown policies greatly impacted the behaviors (particularly mobility, social interactions, and working environment [working from home]) of European people (Sun et al., 2020). To exclude the impact of the COVID-19 pandemic, we performed a preliminary analysis with the data before February 2020.

PHQ-8 Data

The variability of each participant's depressive symptom severity was measured via the PHQ-8, conducted by mobile phones every 2 weeks. The PHQ-8 score ranges from 0 to 24 (increasing severity) (Kroenke et al., 2009). According to the PHQ-8 score, the severity of depression can usually be divided into the following five levels: asymptomatic (PHQ-8 <5), mild (5 ≤ PHQ-8 < 10), moderate (10 ≤ PHQ-8 < 15), moderately severe (15 ≤ PHQ-8 < 20), and severe (PHQ-8 ≥20) (Kroenke et al., 2009).

NBDC Data

The RADAR-base pRMT app scanned other Bluetooth devices in the participant's physical proximity once every hour. To avoid privacy leaks from participants and passers, the Media Access Control (MAC) address and types of Bluetooth devices were not recorded in this study. The NBDC was uploaded to the RADAR-base platform for further analyses.

Figure 4.1. A schematic diagram showing an individual's Nearby Bluetooth devices count (NBDC) in different scenarios in daily activities and life.

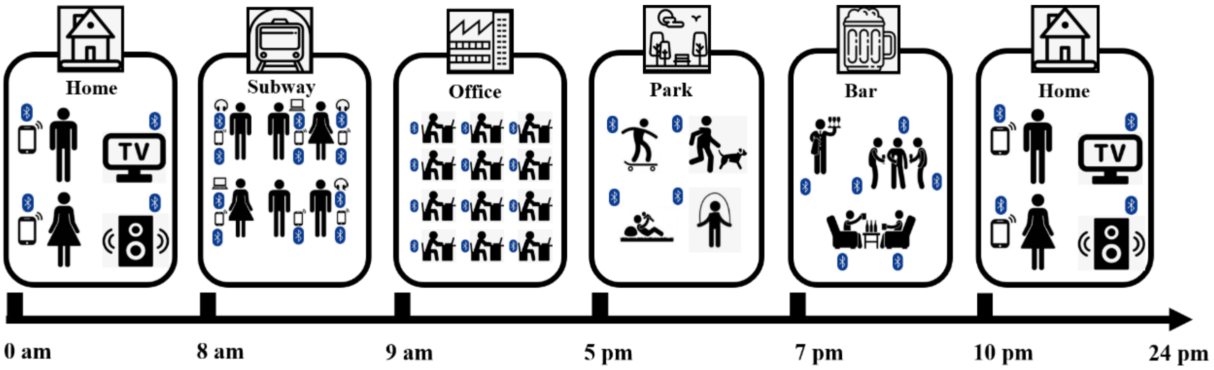
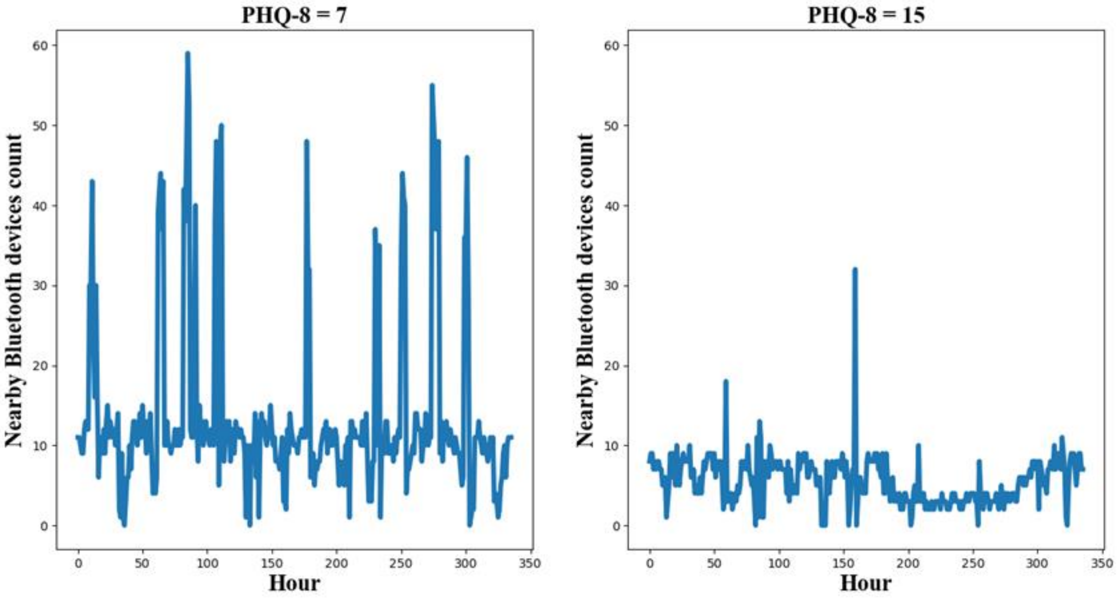


Figure 4.1 is a schematic diagram showing an individual's NBDC in different scenarios in daily activities and life. At home, the NBDC is related to the number of family members and Bluetooth devices in the house, reflecting the participant's connections with family (whether living alone) and the number of other Bluetooth devices. In public transportation (such as the train, subway, and bus), the NBDC is affected by the number of surrounding passengers' Bluetooth devices, reflecting the participant's social connections with strangers. Studies have shown that whether feeling comfortable in the presence of strangers is related to the intensity of social connections (Lee et al., 2001).

In the company, the NBDC can reflect the participant’s social connections and interactions with co-workers. After work, the NBDC can reflect whether the participant joins other social activities, such as going to the park or bar. Therefore, the NBDC data contain information about participants’ social connections and interactions with family, friends, co-workers, and strangers, and the data can also reflect participants’ time at home, mobility, social isolation, and working status, as well as the number of other Bluetooth devices in the house and working environment.

Figure 4.2 shows an example of two NBDC sequences collected over 14 days (336 hours) before two PHQ-8 records from one participant at two different depression severity levels (mild vs moderately severe).

Figure 4.2. An example of two 14-days nearby Bluetooth devices count (NBDC) sequences from the same participant at the mild depression level (left) and moderately severe level (right).



Demographics

Participants' demographics were recorded during the enrollment session. According to previous studies (Akhtar-Danesh & Landeen, 2007; Aluoja et al., 2004), baseline age, gender, and education level were considered as covariates in our analyses. Due to the different educational systems in the three countries in our data set, we used the number of years in education to represent education level.

Data Inclusion Criteria and Data Preprocessing

For each PHQ-8 record, we considered a "PHQ-8 interval" of 14 days before the day when the participant fills in the PHQ-8 questionnaire, as the PHQ-8 score is used to represent the depressive symptom severity of the participant for the past 2 weeks. To reduce the impact of the COVID-19 pandemic and missing data on our analysis, we specified the following two data inclusion criteria:

1. As mentioned in the data set section, to exclude the impact of the COVID-19 pandemic, we restricted our analysis to PHQ-8 records prior to February 2020.

2. Saeb et al (Saeb et al., 2015) and Farhan et al (Farhan et al., 2016) used 50% as each day's completeness threshold for passive data. In our data set, 89.62% of days have 50% (12 hours) or more of the NBDC data. We considered one day as a "valid day" if it contained at least 12 hours of the NBDC data. Then, we empirically selected PHQ-8 intervals with at least 10 valid days as valid PHQ-8 intervals to retain the majority (81.78%) of PHQ-8 intervals.

For the NBDC sequence in each selected PHQ-8 interval, we used linear interpolation to impute the missing hours in all valid days and discarded the NBDC data that did not belong to a valid day. The "NBDC sequence" in the rest of this paper refers to the preprocessed NBDC data in the 14-day PHQ-8 interval.

4.2.2 Feature Extraction

According to past Bluetooth-related research (Bogomolov et al., 2013, 2014; Moturu et al., 2011; Wang et al., 2014) and research on nonlinear features of signal processing (Broughton & Bryan, 2018; Costa et al., 2005), we extracted 49 Bluetooth features from the NBDC sequence in the PHQ-8 interval in the following three categories: second-order statistics, multiscale entropy (MSE), and frequency domain (FD). Table 4.1 summarizes all Bluetooth features extracted in this paper.

Second-Order Statistical Features

We first calculated four daily features (max, min, mean, and standard deviation) of daily NBDC data from all valid days in the PHQ-8 interval. For each daily feature, we calculated four second-order features (max, min, mean, and standard deviation) to reflect the amount and variance of the NBDC in the PHQ-8 interval. These features were denoted in the following format: [Second-order feature]_[Daily feature]. For example, the average value of the daily maximum number of the NBDC in the PHQ-8 interval was denoted as *Mean_Max*. A total of 16 second-order statistical features were extracted.

Nonlinear Bluetooth Features

The second-order statistical features can only reflect the amount (max, min, and mean) and variance (standard deviation) of the NBDC data. To exploit more information embedded in the NBDC data, we proposed MSE and FD features to measure the nonlinear characteristics, such as regularity, complexity, and periodicity, of the NBDC sequence.

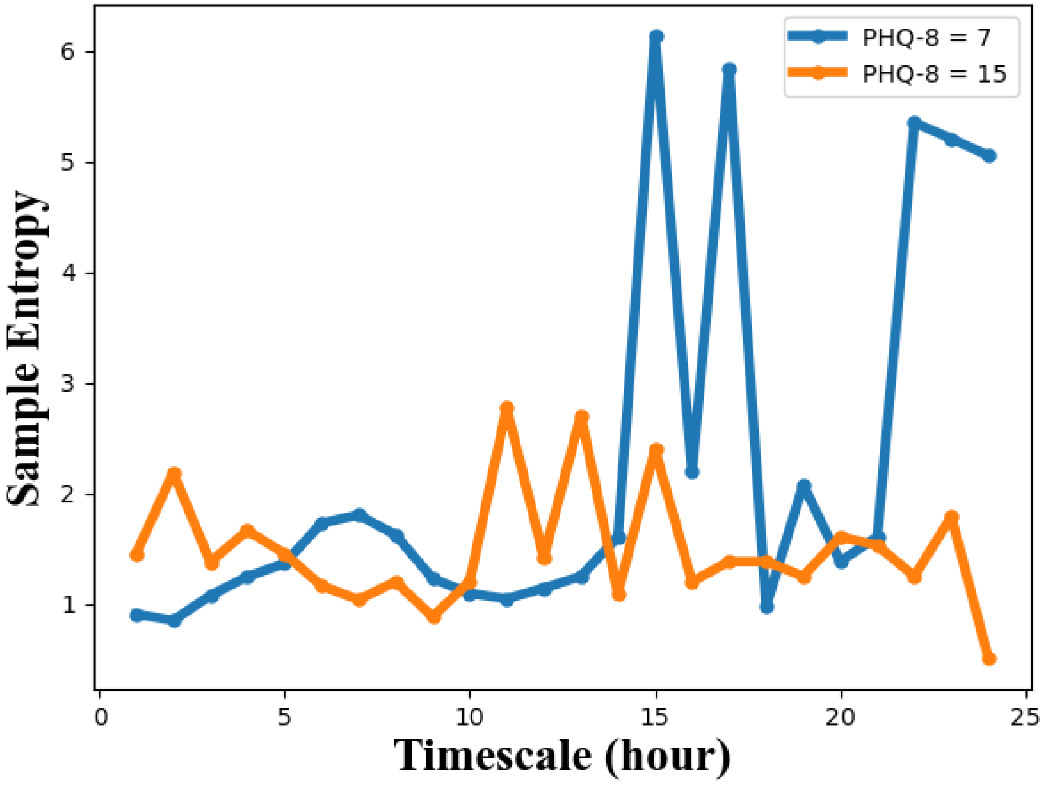
a) Multiscale Entropy Features

MSE analysis has been used to provide insights into the complexity and periodicity of signals over a range of timescales since the method was proposed by Costa et at (Costa et al., 2005). It has been widely used in the field of signal analysis, such as heart rate variability analysis (Silva et al., 2015), electroencephalogram analysis (T. Mizuno et al., 2010), and gait dynamics analysis (Costa et al., 2003). Compared with other entropy techniques (e.g., sample entropy and approximate entropy), the advantage of MSE analysis is that the assessments of complexity at shorter and longer timescales can be analyzed separately (Busa & van Emmerik, 2016). The MSE at short timescales reflects the complexity of the sequence. The larger the MSE at short timescales, the more chaotic and irregular the signal. The MSE at relatively long timescales assesses fluctuations occurring at a certain period, reflecting the periodicity of the signal.

To explore the complexity and periodicity of the NBDC sequence on different timescales (from 1 hour to 24 hours), we calculated MSE features of the NBDC sequences from scale 1 to scale 24, denoted as *MSE_1, MSE_2, …*, *MSE_24*. Figure 4.3

shows an example of MSE features calculated on two NBDC sequences at different depression severity levels from the same participant shown in Figure 4.2. In this example, the NBDC sequence at the mild depression level (PHQ-8=7) has lower MSE at relatively short timescales (scale 1-3) and higher MSE at relatively long timescales than the sequence at the moderately severe depression level (PHQ-8=15). This indicated that this participant's NBDC sequence at the mild depression level was more regular and periodic than the NBDC sequence at the moderately severe depression level.

Figure 4.3. An example of multiscale entropy (scale 1- 24) of two 14-days nearby Bluetooth device count (NBDC) sequences at the mild depression level (blue) and the moderately severe level (orange) from the same participant in Figure 4.2.



b) FD Features

FD analysis has been widely used in the signal processing field, especially for signals with periodic characteristics (Broughton & Bryan, 2018). People's behaviors follow a quasiperiodic routine, such as sleeping at night, working on weekdays, and gathering with friends on weekends (K. Mizuno, 2014; Walker et al., 2020). We therefore leveraged FD analysis to explore the periodic patterns in the NBDC data. Fast Fourier transformation (FFT) was performed to transform the NBDC sequence from the time domain to the FD. We set the sample rate to 24 hours, and then, the spectrum generated by FFT had the frequency axis scaled to reflect cycles per day.

Figure 4.4. An example of a 14-days NBDC sequence in the time domain (left) and its spectrum in the frequency domain (right).

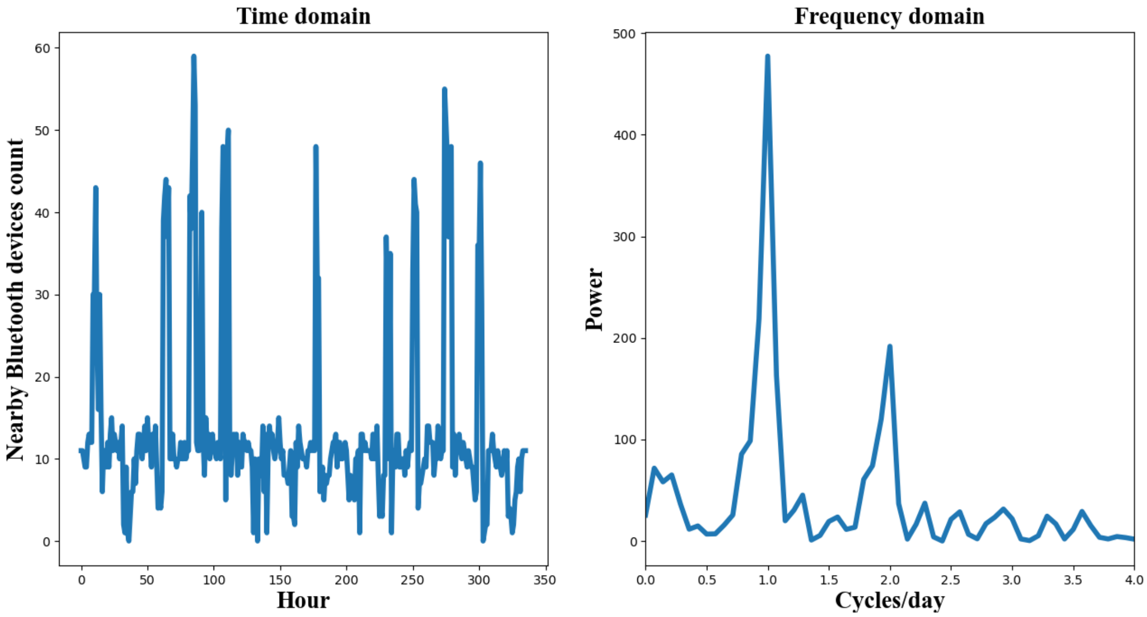


Figure 4.4 is an example of a NBDC sequence in the time domain and its spectrum in the FD. According to the spectrum's definition, spectrum power around 1 cycle per day reflects the participant's circadian rhythm (approximately 24-hour rhythm) (Walker et al., 2020). To explore the periodic rhythms of different period lengths, we empirically defined the following three frequency intervals: low frequency (LF) (0-0.75 cycles/day), middle frequency (MF) (0.75-1.25 cycles/day), and high frequency (HF) (>1.25

cycles/day). The power in MF represents the circadian rhythm. Similarly, the power in LF represents the long-term (>1 day) rhythm, while the power in HF represents the short-term (<1 day) rhythm.

The sums of spectrum power in these three frequency intervals were calculated and denoted as *LF_sum*, *MF_sum*, and *HF_sum*, respectively. The percentages of spectrum powers in these three frequency intervals to the total spectrum power were extracted and denoted as *LF_pct*, *MF_pct*, and *HF_pct*, respectively. To estimate the complexity and regularity of the spectrum, we calculated spectral entropy (SE) (Shannon, 1948) in these three intervals, denoted as *LF_se*, *MF_se*, and *HF_se*, respectively.

Table 4.1. Summary of 49 Bluetooth features used in this paper and their short descriptions.

Category	Abbreviation	Description	Number of features (N=49)
Statistical features	[Second-order feature]_[Daily feature], eg, Max_Mean	Second-order features (max, min, mean, and standard deviation) calculated in the PHQ-8[a] interval based on daily statistical Bluetooth features (max, min, mean, and standard deviation).	16
Multiscale entropy (MSE)	MSE_1, MSE_2, …, MSE_24	Multiscale entropy of the NBDC[b] sequences from scale 1 to scale 24.	24
Frequency domain[c]	LF_sum, MF_sum, HF_sum	The sums of spectrum power in LF, MF, and HF.	3
Frequency domain	LF_pct, MF_pct, HF_pct	The percentages of spectrum power in LF, MF, and HF to the total spectrum power.	3
Frequency domain	LF_se, MF_se, HF_se	Spectral entropy in LF, MF, and HF.	3

[a] PHQ-8: 8-item Patient Health Questionnaire.
[b] NBDC: nearby Bluetooth device count.
[c] LF: low frequency (0-0.75 cycles/day); MF: middle frequency (0.75-1.25 cycles/day); HF: high frequency (>1.25 cycles/day).

4.2.3 Statistical Methods

The linear mixed-effect model contains both fixed and random effects, allowing for both within-participant and between-participants variations over repeated measurements (Laird & Ware, 1982). Therefore, we used linear mixed-effect models in our statistical analyses.

Pairwise Association Analyses

To explore the association between each Bluetooth feature and depression severity, a series of pairwise linear mixed-effect models with random participant intercepts were performed to regress the PHQ-8 score with each of the Bluetooth features. All mixed-effect models, baseline age, gender, and years in education were considered as covariates. The z-test was used to evaluate the statistical significance of the coefficient of each model. The Benjamini-Hochberg method (Benjamini & Hochberg, 1995) was used for correction of multiple comparisons, and the significant level for the adjusted P value was set to .05. All linear mixed-effect models were implemented by using the R package "lmerTest," and the Benjamini-Hochberg method was performed by using the command "p.adjust" in R software (R Foundation for Statistical Computing).

Likelihood Ratio Test

One objective of this paper was to assess what value these Bluetooth features provide beyond other information that might be readily available, such as baseline demographics. The likelihood ratio test is a statistical test of goodness of fit between two nested models (Glover & Dixon, 2004). If the model with more parameters fits the data significantly better, it indicates that additional parameters provide more information and improve the model's fitness (Glover & Dixon, 2004). Therefore, we built three nested linear mixed-effect models with random participant intercepts (model A, model B, and model C). The predictors of model A were only demographics. The predictors of model B were demographics and 16 second-order statistical features. The predictors of model C were demographics and all 49 Bluetooth features. The likelihood ratio tests were performed to test whether these Bluetooth features have a significant value in fitting the PHQ-8 score regression model.

4.2.4 Prediction Models

Another objective of this paper was to examine whether it is possible to predict participants' depressive symptom severity using Bluetooth features combined with some known information (demographics and previous PHQ-8 scores). A subset of PHQ-8 intervals was selected for the prediction task based on the following two additional criteria:

1. To ensure that each participant had sufficient PHQ-8 intervals for the time-series cross-validation (described in the following model evaluation section), the number of valid PHQ-8 intervals for each participant should be at least 3.

2. To test whether the model can predict variability of depression severity, the difference of one participant's PHQ-8 scores should be more than or equal to 5 (clinically meaningful change) (Saeb et al., 2016).

Hierarchical Bayesian Linear Regression Model

The hierarchical Bayesian approach is an intermediate method compared to the completely pooled model and individualized model, capturing the whole population's characteristics while allowing individual differences (Gelman et al., 2013). We leveraged the hierarchical Bayesian linear regression model to predict participants' PHQ-8 scores using Bluetooth features, demographics (age, gender, and years in education), and the last observed PHQ-8 score. In this study, we implemented the hierarchical Bayesian linear regression using the "PyMC3" package (Salvatier et al., 2016) in Python. To compare the results with other commonly used machine learning models, we also implemented the LASSO regression model (Tibshirani, 1996) and XGBoost regression model (Chen & Guestrin, 2016) using the Scikit-learn machine learning library (Pedregosa et al., 2011) in Python. As depressive mood has a strong autocorrelation (Busk et al., 2020), we considered a baseline hierarchical Bayesian linear regression model with the last observed PHQ-8 score and demographics as predictors.

Model Evaluation

We selected root mean squared error (RMSE) and the predicted coefficient of determination (R^2) as two metrics for model discrimination evaluation. As we used the temporal data, "future data" should not predict "past data." Therefore, only the data observed before test data can be included in the training set. We applied leave-all-out (LAO) and leave-one-out (LOO) time-series cross-validation (Busk et al., 2020). As the number of PHQ-8 intervals of each participant in our data was different, we made some minor modifications to these two schemes (Figure 4.5).

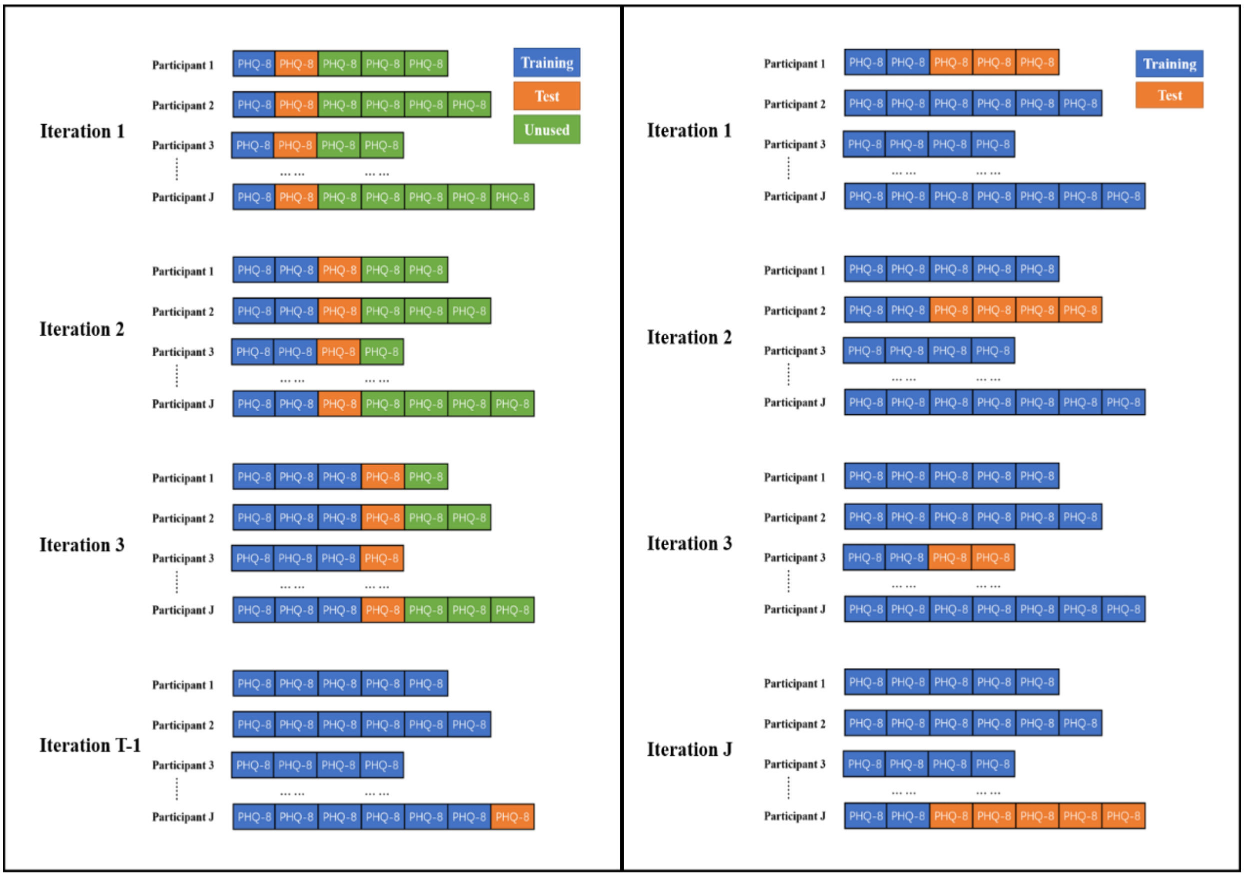
a) LAO Time-Series Cross-Validation

Each participant's data were divided into a sequence of t consecutive same-sized test sets, where the size of each test set is the length of one PHQ-8 interval (14 days) and t is the number of PHQ-8 intervals of this participant. The corresponding training set included all PHQ-8 intervals before each test set. Then, test sets and training sets were pooled across all participants. This process generated T-1 test and training set pairs (no prior data to predict the first PHQ-8 score), where T is the maximum number of PHQ-8 intervals of one participant in our data set (t≤T).

b) LOO Time-Series Cross-Validation

Each participant's data were divided into a training set and a test set. The training set was constructed using the first two PHQ-8 intervals of a participant, with the test set containing the rest of the participant's PHQ-8 intervals. Then, the training set was pooled with all data from all other participants. This scheme generated J training and test set pairs, where J is the number of participants in our data set.

Figure 4.5. Two schematic diagrams of leave-all-out time-series cross-validation (left) and leave-one-out time-series cross-validation (right), where T is the maximum number of PHQ-8 intervals of one participant, J is the number of participants, the training set is indicated by blue, the test set is indicated by orange, and unused data is indicated by green.



Leave-all-out time-series cross-validation **Leave-one-out time-series cross-validation**

4.3 Results

4.3.1 Data Summary

According to our date inclusion criteria, from June 2018 to February 2020, 2886 PHQ-8 intervals from 316 participants collected from three study sites were selected for our analysis. Table 4.2 presents a summary of the demographics and distribution of PHQ-8 records of all selected participants. Table 4.3 shows the descriptive statistics for all

49 Bluetooth features, and Figure 4.6 presents pairwise Spearman correlation coefficients between all features. Figure 4.7 presents boxplots of the NBDC for every hour in the whole population.

Table 4.2. Summary of the demographics and 8-item Patient Health Questionnaire (PHQ-8) record distribution of all selected participants.

Characteristic	Value
Number of participants	316
Demographics	
Age at baseline, median (Q1, Q3)	51.0 (35.0, 59.0)
Female sex, n (%)	234 (74.1%)
Number of years in education, median (Q1, Q3)	16.0 (14.0, 19.0)
PHQ-8[a] record distribution	
Number of PHQ-8 intervals	2886
Number of PHQ-8 intervals for each participant, median (Q1, Q3)	8.0 (3.0, 14.0)
PHQ-8 score, median (Q1, Q3)	9.0 (5.0, 15.0)

[a]PHQ-8: 8-item Patient Health Questionnaire.

Figure 4.6. A correlation plot of pairwise Spearman correlations between all 49 Bluetooth features. Definitions of Bluetooth features in this figure are shown in Table 4.1.

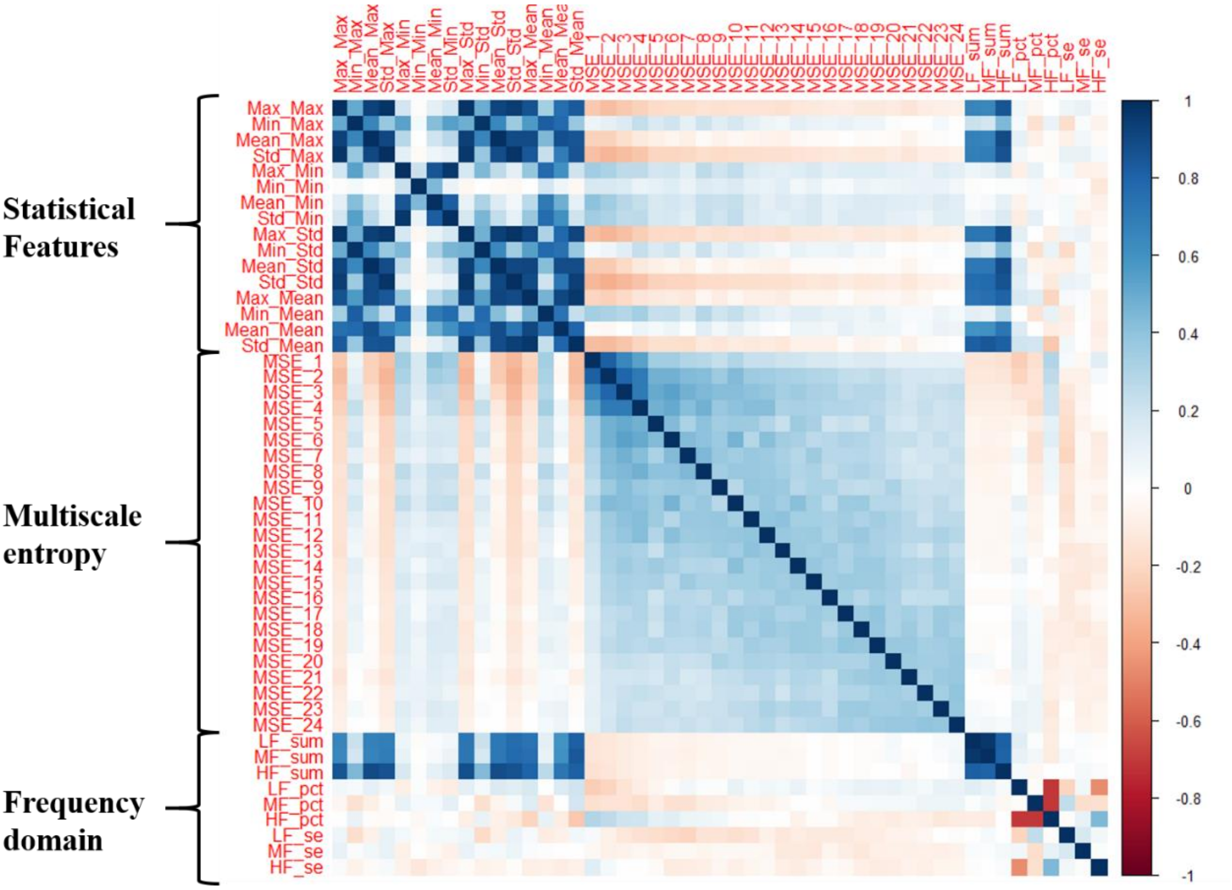


Figure 4.7. Boxplots of nearby Bluetooth devices count (NBDC) for every hour on the whole population. Boxes extend between 25th and 75th percentiles, and green solid lines inside the boxes are medians. Note the relative stationary NBDC during the night-time hours.

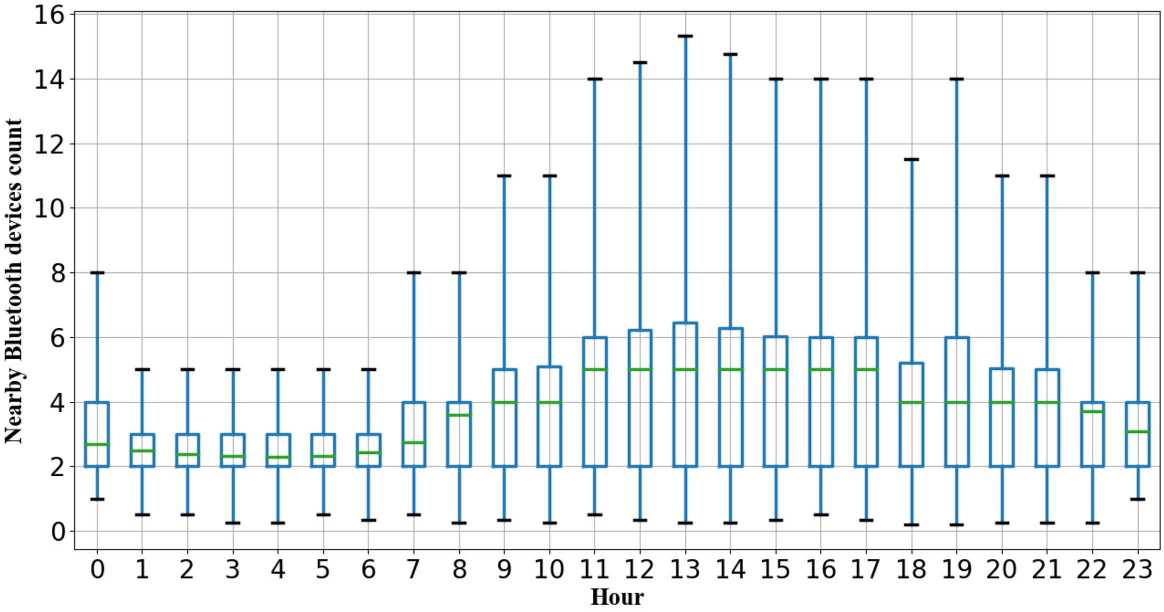


Table 4.3. Descriptive statistics for all 49 Bluetooth features.

Feature[a]	Mean	SD	Min	Q1	Median	Q3	Max
Second-order statistics							
Max_Max	49.79	48.48	1.00	25.00	40.00	60.00	621.00
Min_Max	5.09	6.22	0.00	2.00	4.00	6.00	90.00
Mean_Max	18.56	18.94	0.75	9.23	14.07	21.62	268.29
Std_Max	13.14	14.05	0.00	6.14	10.45	16.22	195.19
Max_Min	1.59	2.08	0.00	0.00	1.00	2.00	43.00
Min_Min	0.06	0.27	0.00	0.00	0.00	0.00	3.00
Mean_Min	0.58	0.88	0.00	0.00	0.21	0.79	13.71
Std_Min	0.50	0.62	0.00	0.00	0.42	0.70	11.94
Max_Std	12.31	12.76	0.34	5.60	9.51	15.39	185.98
Min_Std	1.20	1.45	0.00	0.56	0.87	1.32	21.61
Mean_Std	4.55	4.87	0.16	2.17	3.25	5.24	70.65
Std_Std	3.24	3.71	0.09	1.34	2.43	4.04	62.52
Max_Mean	9.32	9.34	0.17	4.38	6.88	11.04	136.10
Min_Mean	1.88	2.14	0.00	0.50	1.42	2.50	32.00
Mean_Mean	4.42	4.19	0.07	2.19	3.40	5.28	49.55
Std_Mean	2.13	2.59	0.05	0.84	1.45	2.54	49.37
Multiscale entropy (MSE)							
MSE_1	0.80	0.46	0.05	0.42	0.71	1.13	2.44
MSE_2	0.97	0.54	0.04	0.56	0.85	1.31	3.58
MSE_3	1.12	0.66	0.09	0.70	1.01	1.42	9.41
MSE_4	1.23	0.69	0.05	0.82	1.15	1.51	8.83
MSE_5	1.35	0.82	0.10	0.93	1.27	1.62	8.51
MSE_6	1.38	0.84	0.08	0.97	1.28	1.63	8.00
MSE_7	1.47	0.97	0.10	1.01	1.33	1.70	7.72
MSE_8	1.50	1.07	0.10	1.00	1.30	1.67	7.40
MSE_9	1.58	1.22	0.10	0.99	1.32	1.72	7.30
MSE_10	1.58	1.23	0.08	0.97	1.30	1.72	7.08
MSE_11	1.58	1.29	0.09	0.95	1.25	1.67	7.02
MSE_12	1.59	1.33	0.10	0.92	1.23	1.66	6.70
MSE_13	1.74	1.46	0.11	0.98	1.30	1.79	6.55
MSE_14	1.85	1.53	0.11	1.01	1.36	1.87	6.70
MSE_15	1.96	1.62	0.13	1.03	1.39	1.95	6.55
MSE_16	1.98	1.62	0.13	1.03	1.39	1.95	6.40
MSE_17	2.04	1.67	0.14	1.02	1.39	2.08	6.14
MSE_18	2.03	1.65	0.15	1.01	1.39	2.08	6.04
MSE_19	2.09	1.69	0.17	1.01	1.39	2.08	6.04
MSE_20	2.09	1.67	0.17	0.98	1.39	2.08	5.94
MSE_21	2.10	1.66	0.18	0.98	1.39	2.20	5.83
MSE_22	2.13	1.68	0.18	0.98	1.39	2.30	5.83
MSE_23	2.17	1.69	0.18	0.98	1.39	4.28	5.61
MSE_24	2.27	1.70	0.20	0.98	1.39	4.28	5.35

Feature	Mean	SD	Min	Q1	Median	Q3	Max
Frequency domain (FD)							
LF[b]_sum	330.66	2469.74	0.05	17.41	53.87	184.80	85956.16
MF[c]_sum	157.24	1166.32	0.02	8.16	25.77	83.05	34970.35
HF[d]_sum	602.22	3272.44	0.47	55.72	151.74	403.38	64127.16
LF_pct[e]	0.25	0.10	0.03	0.17	0.23	0.31	0.63
MF_pct	0.13	0.10	0.01	0.07	0.11	0.17	0.74
HF_pct	0.62	0.15	0.12	0.53	0.64	0.72	0.92
LF_se[f]	0.83	0.10	0.38	0.78	0.85	0.90	1.00
MF_se	0.82	0.09	0.40	0.77	0.83	0.88	0.99
HF_se	0.90	0.04	0.72	0.88	0.90	0.92	0.99

[a] Definitions of Bluetooth features in this table are shown in Table 4.1.
[b] LF: low frequency (0-0.75 cycles/day).
[c] MF: middle frequency (0.75-1.25 cycles/day).
[d] HF: high frequency (>1.25 cycles/day).
[e] pct: percentage of spectrum power.
[f] se: spectral entropy.

4.3.2 Association Analysis Results

The significant associations between depression severity (the PHQ-8 score) and Bluetooth features are presented in Table 4.4.

Associations Between the PHQ-8 Score and Second-Order Statistical Features

There were 10 second-order statistical features significantly associated with the PHQ-8 score. All these significant associations were negative, that is, the larger the value of these features, the lower the PHQ-8 score. Notably, *Min_Max* (the minimum value of daily maximum NBDC in the past 14 days) had the strongest association (z=−4.431, P<.001), which indicated that participants with a lower PHQ-8 score tended to have more daily social activities (such as social interactions and traveling) in the past 2 weeks.

In addition, four features related to daily variance (*Max_Std, Min_Std, Mean_Std,* and *Std_Std*) of the NBDC were all significantly and negatively associated with depression.

Table 4.4. Coefficient estimates, standard error, z-test statistics, and *P* values from pairwise linear mixed-effect models for exploring associations between Bluetooth features and the depressive symptom severity (8-item Patient Health Questionnaire).

Feature[a]	Estimate	SE	z score	Adjusted *P* value[b,c]
Second-order statistics				
Min_Max	−0.052	0.012	−4.431	<.001
Mean_max	−0.016	0.006	−2.809	.005
Max_Std	−0.015	0.006	−2.657	.008
Min_Std	−0.215	0.056	−3.838	<.001
Mean_Std	−0.065	0.023	−2.802	.005
Std_Std	−0.048	0.020	−2.385	.02
Max_Mean	−0.030	0.008	−3.498	<.001
Min_Mean	−0.093	0.046	−2.036	.04
Mean_Mean	−0.083	0.026	−3.225	.001
Std_Mean	−0.095	0.027	−3.464	.001
Multiscale entropy				
MSE_1	0.642	0.225	2.853	.005
MSE_2	0.433	0.192	2.255	.02
MSE_3	0.401	0.202	1.985	.04
MSE_16	−0.102	0.042	−2.429	.01
MSE_22	−0.123	0.043	−2.860	.005
Frequency domain (FD)				
LF[d]_sum	−0.021	0.005	−3.865	<.001
MF[e]_sum	−0.067	0.014	−4.766	<.001
HF[f]_sum	−0.027	0.010	−2.606	.009
MF_pct[g]	−1.834	0.812	−2.259	.02
HF_se[h]	3.821	1.820	2.099	.04

[a] Definitions of Bluetooth features in this table are shown in Table 4.1.

[b] Only significant associations (adjusted *P* value <.05) are reported.

[c] *P* values were adjusted by the Benjamini-Hochberg method for correction of multiple comparisons.

[d] LF: low frequency (0-0.75 cycles/day).

[e] MF: middle frequency (0.75-1.25 cycles/day).

[f] HF: high frequency (>1.25 cycles/day).

[g] pct: percentage of spectrum power.

[h] se: spectral entropy.

Associations Between the PHQ-8 Score and Multiscale Entropy Features

MSE at scale 1, scale 2, and scale 3 (*MSE_1*, *MSE_2*, and *MSE_3*) were significantly and positively associated with the PHQ-8 score, while MSE at scale 16 and scale 22 (*MSE_16* and *MSE_22*) were significantly and negatively associated with depressive symptom severity. According to the explanations of MSE we mentioned in the Methods section, these associations indicated that participants with more irregular and chaotic NBDC sequences were likely to have more severe depressive symptoms, while those with periodic and regular NBDC sequences may have lower PHQ-8 scores.

Associations Between the PHQ-8 Score and FD Features

There were five FD features significantly associated with the PHQ-8 score. The spectrum power was related to both the amount and frequency components of the NBDC sequence, so it had relatively strong correlations with second-order statistical features (Figure 4.6). Therefore, the spectrum power of three frequency intervals (*LF_sum*, *MF_sum*, and *HF_sum*) were all significantly and negatively associated with the PHQ-8 score. Among them, the *MF_sum* had the strongest association (z=−4.766, P<.001) with depression, which indicated that the circadian rhythm of the NBDC sequence is important to reflect the severity of depression. Likewise, the percentage of middle-frequency power (*MF_pct*) was significantly and negatively associated with depressive symptom severity. The spectral entropy of HF (*HF_se*) was significantly and positively associated with depression. This indicated that participants with irregular short-term (<1 day) rhythms were likely to have more severe depressive symptoms.

4.3.3 Results of Likelihood Ratio Tests

The results of the likelihood ratio tests are presented in Table 4.5. Model B (with second-order statistical Bluetooth features) and model C (with all Bluetooth features) fitted data significantly better than model A (without Bluetooth features), indicating that Bluetooth features could improve the statistical model significantly. The goodness of fit of model C was significantly better than that of model B, indicating that nonlinear Bluetooth features (MSE and FD features) provided additional information to the statistical model.

Table 4.5. Results of the likelihood ratio tests of the three nested linear mixed-effect models.

Model	Difference of parameters	Chi-square[a]	*P* value
Model B[b] vs model A[c]	16	31.04	.01
Model C[d] vs model A	49	135.19	<.001
Model C vs model B	33	104.15	<.001

[a]The critical values of the likelihood ratio statistic are as follows: $\chi^2_{0.05}(16)=26.296$, $\chi^2_{0.05}(33)=47.400$, and $\chi^2_{0.05}(49)=66.339$.
[b]Predictors of model B: demographics + 16 second-order statistical features.
[c]Predictors of model A: demographics.
[d]Predictors of model C: demographics + 16 second-order statistical features + 24 multiscale entropy features + 9 frequency domain features.

4.3.4 Performance of Prediction Models

A subset of 183 participants was selected for the prediction models. The results of the LAO and LOO time-series cross-validation are presented in Table 4.6. The R^2 score of the baseline model was 0.338 in LAO time-series cross-validation, which showed that more than 30% variance could be explained by the last observed PHQ-8 score and baseline demographics. In LOO time-series cross-validation, the R^2 score of the baseline model was negative, which indicated that the baseline model did not explain any variance in the LOO time-series cross-validation. To assess the improvement from nonlinear Bluetooth features, we tested the hierarchical Bayesian model with and without nonlinear Bluetooth features separately.

Table 4.6. Results of the leave-all-out time-series cross-validation and leave-one-out time-series cross-validation of the hierarchical Bayesian linear regression model, commonly used machine learning models, and the baseline model.

Model	Leave-all-out		Leave-one-out	
	R^2	RMSE[a]	R^2	RMSE
Baseline model[b]	0.338	4.547	−0.074	5.802
LASSO regression	0.458	4.114	0.144	5.178
XGBoost regression	0.464	4.092	0.346	4.523
Hierarchical Bayesian linear (second-order statistical features)	0.481	4.026	0.353	4.501
Hierarchical Bayesian linear (all Bluetooth features)	0.526	3.891	0.387	4.426

[a] RMSE: root mean squared error.
[b] The baseline model is the hierarchical Bayesian linear regression model with only the last observed 8-item Patient Health Questionnaire score and demographics as predictors.

In the subset, the maximum number of PHQ-8 intervals of one participant was 27, so the LAO time-series cross-validation went through T-1=26 iterations. The hierarchical Bayesian linear regression model with all Bluetooth features achieved the best result (R^2=0.526, RMSE=3.891), beating the LASSO and XGBoost regression models.

Compared with the result of the baseline model (R^2=0.338), the improvement in the R^2 score was 0.188, which means the Bluetooth features explained an additional 18.8% of data variance. The nonlinear Bluetooth features explained an additional 4.5% of data variance in the hierarchical Bayesian model.

The number of subset participants was 183, so J=183 iterations of the LOO time-series cross-validation were performed. The hierarchical Bayesian linear model with all Bluetooth features had the best performance (R^2=0.387, RMSE=4.426), but the result was close to that of the XGBoost regression model (R^2=0.346, RMSE=4.523).

The performance of the hierarchical Bayesian linear regression model evaluated by the LAO cross-validation was better than the LOO cross-validation performance. One potential reason is that only the first two PHQ-8 intervals of one participant were used for training in the LOO cross-validation, which may have caused the model to underfit the patterns at the participant level.

4.4 Discussion

4.4.1 Principal Findings

This paper explored the value of the NBDC data in predicting depression severity. Compared with previous Bluetooth-related studies (Bogomolov et al., 2013, 2014; Moturu et al., 2011; Wang et al., 2014), our study was performed on a larger (N=316) multicenter data set with a longer follow-up (median 4 months). We extracted 49 features from the NBDC sequences in the following three categories: second-order statistical features, MSE features, and FD features. To the best of our knowledge, this is the first time that MSE and FD features have been used in NBDC and depression data analyses. According to the results of association analyses (Table 4.4), when depression symptoms worsened (increase in the PHQ-8 score), one or more of the following changes were seen in the preceding 14 days of the NBDC sequence: (1) the amount decreased, which is consistent with the finding by Wang et al (Wang et al., 2014), (2) the variance decreased, (3) the periodicity (especially the circadian rhythm) decreased, and (4) the NBDC sequence became more irregular and chaotic.

These changes in the NBDC data can be explained by depression symptoms. The main manifestations of depression include negative feelings (such as sadness, guilt, stress, and tiredness) and loss of interest or pleasure (World Health Organization, 2017). This may lead to changes in behaviors, such as increased time at home (Chow et al., 2017; Saeb et al., 2015), decreased mobility (Lampinen & Heikkinen, 2003; Saeb et al., 2015), loss of the ability to work or study (Rizvi et al., 2015; World Health Organization, 2017), reduced intensity of social interactions (Cacioppo et al., 2006), unstable and irregular sleep (Zhang et al., 2021), and decreased engagement in activities (Goldberg et al., 2002). The increased time at home, inability to work or study, and diminished social interactions are reflected in the reduced amount of the NBDC sequence. The decreased mobility and engagement in activities may be possible reasons why participants with higher PHQ-8 scores have lower variance-related features (*Max_Std*, *Min_Std*, *Mean_Std*, and *Std_Std*). Depression also may lead to misalignment of the circadian rhythm and make people’s life rhythms (such as sleep rhythms and social rhythms) more irregular (Walker et al., 2020). This can be reflected in reduced periodicity and increased irregularity of the NBDC sequence. Saeb et al (Saeb et al., 2015) and Farhan et al (Farhan et al., 2016) found similar findings in GPS data, and showed that the circadian rhythm of the GPS signal was significantly and negatively correlated with depression.

From the perspective of the statistical model, Bluetooth features extracted in this paper significantly improved the goodness of fit for the PHQ-8 score, and nonlinear Bluetooth features (MSE and FD features) can provide additional information to second-order statistical features (Table 4.5). From the perspective of the prediction model, these 49 Bluetooth features explained an extra 18.8% of the variance in the PHQ-8 score relative to the baseline model, containing only the last PHQ-8 score and demographics, and MSE and FD features explained an extra 4.5% of data variance in the hierarchical Bayesian model (Table 4.6). From the perspective of the correlations between Bluetooth features (Figure 4.6), we can observe that, except for three FD features related to the spectrum power that had relatively strong correlations with second-order statistical features, the correlations between other nonlinear Bluetooth features and second-order statistical features were not obvious. This indicated that the MSE and FD features captured dimensions of information to second-order statistical features.

In our prediction model, the hierarchical Bayesian linear regression model achieved the best results in both the LAO and LOO time-series cross-validation. Compared with other models, one of the advantages of the hierarchical Bayesian model is that it performs individual predictions while considering the population's common characteristics (Gelman et al., 2013). Therefore, the hierarchical Bayesian model can be considered a suitable prediction modelling method for longitudinal data. The LOO time-series cross-validation results illustrated that the hierarchical Bayesian model could predict depression for participants with few observations (only two PHQ intervals in the training set) that overcomes the cold start problem. The hierarchical Bayesian linear model achieved a better result in the LAO time-series cross-validation, which indicated that the prediction results gradually became more accurate and individualized when each participant had more data available in the training set.

4.4.2 Limitations

The RADAR-MDD project was designed for long-term monitoring (up to 2 years) and collecting many other passive data, such as GPS data, acceleration data, app usage, and screen lightness, which need to be collected simultaneously through the mobile phone. Therefore, to avoid excessive battery consumption, nearby Bluetooth devices were scanned hourly in this study. However, some past studies suggested scanning nearby Bluetooth devices every 5 minutes to achieve high enough temporal resolution (Bogomolov et al., 2014; Eagle & (Sandy) Pentland, 2006). Although hourly NBDC data can also reflect individuals' behaviors and statuses, our lower data resolution may cause the loss of some dynamic information. On the other hand, using the relatively low resolution enabled us to collect multimodal data without excessive battery consumption. As the NBDC data are related to individuals' movement and location information, we will combine the NBDC data with GPS and acceleration data for future analysis to understand the context of the Bluetooth data.

As we mentioned in the Methods section, the MAC addresses and types of Bluetooth devices were not recorded for private issues. This made it impossible to distinguish between mobile phones and other Bluetooth devices (such as headphones, printers, and laptops), and between strangers' and acquaintances' devices. The advantage of the NBDC data is that the data contain mixed and rich information. The disadvantage is that it is difficult to explain the specific reasons for changes in the NBDC, that is, we cannot know whether the changes in the NBDC are caused by social interactions, working status, traveling, or isolation. Therefore, this paper did not explain in depth the actual meaning behind the Bluetooth features. For this limitation, we plan to use hashed MAC addresses in future research.

For the FD features, the division of the frequency intervals of the spectrum of the NBDC sequence in this paper was manually specified by our experience. The purpose of extracting these FD features was to prove that the NBDC sequence's FD has the potential to provide more information about individuals' behaviors and life rhythms. It is necessary to discuss the optimal boundaries of frequency intervals of the NBDC data in future research.

This paper applied the hierarchical Bayesian linear regression model to explore the linear relationships between Bluetooth features and depression. However, there may be nonlinear relationships between social connections and depressive symptom severity.

The Gaussian process (Dearmon & Smith, 2016), using the kernel method to find nonlinear relationships, will be considered in future research.

4.4.3 Conclusions

Our statistical results indicated that the NBDC data have the potential to reflect changes in individuals’ behaviors and statuses during a depressive state. The prediction results demonstrated that the NBDC data have significant value in predicting depressive symptom severity. The nonlinear Bluetooth features proposed in this paper provide additional information to statistical and prediction models. The hierarchical Bayesian model is an appropriate prediction model for predicting depression with longitudinal data, as both participant-level and population-level characteristics are considered in the model. These findings may support the mental health monitoring practice in real-world settings.

Acknowledgments

The Remote Assessment of Disease and Relapse-Central Nervous System (RADAR-CNS) project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking under grant agreement number 115902. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation program and European Federation of Pharmaceutical Industries and Associations (EFPIA) (www.imi.europa.eu). This communication reflects the views of the RADAR-CNS consortium, and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding body was not involved in the design of the study, the collection or analysis of the data, or the interpretation of the data. Participants in the CIBER site came from the following four clinical communities in Spain: Parc Sanitari Sant Joan de Déu Network services, Institut Català de la Salut, Institut Pere Mata, and Hospital Clínico San Carlos. Participant recruitment in Amsterdam was partially accomplished through Hersenonderzoek.nl, a Dutch online registry that facilitates participant recruitment for neuroscience studies. Hersenonderzoek.nl is funded by ZonMw-Memorabel (project number 73305095003), a project in the context of the Dutch Deltaplan Dementie, Gieskes-Strijbis Foundation, the Alzheimer's Society in the Netherlands, and Brain Foundation Netherlands. This paper represents independent research partly funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. The views expressed are those of the authors and not necessarily those of the NHS, NIHR, or Department of Health and Social Care. We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures and for their invaluable advice throughout the study protocol design. This research was reviewed by a team with experience in mental health problems and their careers who have been specially trained to advise on research proposals and documentation through the Feasibility and Acceptability Support Team for Researchers (FAST-R), a free confidential service in England provided by the National Institute for Health Research Maudsley Biomedical Research Centre via King’s College London and South London and Maudsley NHS Foundation Trust. Remote Assessment of Disease and Relapse-Major Depressive Disorder (RADAR-MDD) will be conducted according to the Declaration of Helsinki and Good Clinical Practice, adhering to principles outlined in the NHS Research Governance Framework for Health and Social Care (2nd edition). Ethical approval has been obtained in London from the Camberwell St Giles Research Ethics Committee (REC reference: 17/LO/1154), in London from the CEIC Fundacio Sant Joan de Deu (CI: PIC-128-17), and in the Netherlands from the Medische Ethische Toetsingscommissie VUms (METc VUmc registratienummer: 2018.012 – NL63557.029.17). RJBD is supported by the following: (1) NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London, London, UK; (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; (3) The BigData@Heart Consortium, funded by the Innovative Medicines Initiative-2 Joint Undertaking under grant agreement number 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation program and EFPIA; it is chaired by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and the European Society of Cardiology; (4) The National Institute for Health Research University College London

Hospitals Biomedical Research Centre; (5) The NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London; (6) The UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare; (7) The NIHR Applied Research Collaboration South London (NIHR ARC South London) at King’s College Hospital NHS Foundation Trust. This paper represents independent research partly funded by the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. MH and TW have been supported by NIHR Senior Investigator Awards.

Authors’ Contributions

YZ extracted and integrated the depression questionnaires and Bluetooth data for the analysis, planned and performed the analysis, and drafted the manuscript. MH and VAN gained funding and co-led the Remote Assessment of Disease and Relapse–Central Nervous System program. MH is the principal investigator for the Remote Assessment of Disease and Relapse–Major Depressive Disorder study. RJBD, AAF, YR, ZR, PC, and CS have contributed to the development of the RADAR-base platform used for data collection and management across sites, data protection, security, and storage. YZ, AAF, S Sun, NC, PL, DCM, MH, and RJBD contributed to the design of the study. FM, CO, FL, S Siddi, S Simblett, JMH, BWJHP, MH contributed to data collection. AAF, IMG, AR, VAN, TW, PA, MH, and RJBD contributed to the administrative, technical, and clinical support of the study. All authors were involved in reviewing the manuscript, had access to the study data, and provided direction and comments on the manuscript.

Conflicts of Interest

VAN is an employee of Janssen Research and Development LLC. PA is employed by the pharmaceutical company H. Lundbeck A/S. DCM has accepted honoraria and consulting fees from Apple, Inc, Otsuka Pharmaceuticals, Pear Therapeutics, and the One Mind Foundation; has received royalties from Oxford Press; and has an ownership interest in Adaptive Health, Inc.

References

Akhtar-Danesh, N., & Landeen, J. (2007). Relation between depression and sociodemographic factors. *International Journal of Mental Health Systems, 1*(1), 4.

Aluoja, A., Leinsalu, M., Shlik, J., Vasar, V., & Luuk, K. (2004). Symptoms of depression in the Estonian population: Prevalence, sociodemographic correlates and social adjustment. *Journal of Affective Disorders*, *78*(1), 27–35.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Bogomolov, A., Lepri, B., Ferron, M., Pianesi, F., & Pentland, A. (Sandy). (2014). Daily Stress Recognition from Mobile Phone Data, Weather Conditions and Individual Traits. *Proceedings of the 22nd ACM International Conference on Multimedia*, 477–486.

Bogomolov, A., Lepri, B., & Pianesi, F. (2013). Happiness Recognition from Mobile Phone Data. *2013 International Conference on Social Computing*, 790–795.

Boonstra, T. W., Werner-Seidler, A., O’Dea, B., Larsen, M. E., & Christensen, H. (2017). Smartphone app to investigate the relationship between social connectivity and mental health. *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 287–290.

Broughton, S. A., & Bryan, K. (2018). *Discrete Fourier analysis and wavelets: Applications to signal and image processing*. John Wiley & Sons.

Burgess, P., Pirkis, J., Morton, J., & Croke, E. (2000). Lessons From a Comprehensive Clinical Audit of Users of Psychiatric Services Who Committed Suicide. *Psychiatric Services*, *51*(12), 1555–1560.

Busa, M. A., & van Emmerik, R. E. A. (2016). Multiscale entropy: A tool for understanding the complexity of postural control. *Journal of Sport and Health Science*, *5*(1), 44–51.

Busk, J., Faurholt-Jepsen, M., Frost, M., Bardram, J. E., Vedel Kessing, L., & Winther, O. (2020). Forecasting Mood in Bipolar Disorder From Smartphone Self-assessments: Hierarchical Bayesian Approach. *JMIR MHealth and UHealth*, *8*(4), e15028.

Cacioppo, J. T., Hughes, M. E., Waite, L. J., Hawkley, L. C., & Thisted, R. A. (2006). Loneliness as a specific risk factor for depressive symptoms: Cross-sectional and longitudinal analyses. *Psychology and Aging*, *21*(1), 140–151.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chow, P. I., Fua, K., Huang, Y., Bonelli, W., Xiong, H., Barnes, L. E., & Teachman, B. A. (2017). Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students. *Journal of Medical Internet Research*, *19*(3), e62.

Clark, B. K., Winkler, E. A., Brakenridge, C. L., Trost, S. G., & Healy, G. N. (2018). Using Bluetooth proximity sensing to determine where office workers spend time at work. *PLOS ONE*, *13*(3), e0193971.

Costa, M., Goldberger, A. L., & Peng, C.-K. (2005). Multiscale entropy analysis of biological signals. *Physical Review E*, *71*(2), 021906.

Costa, M., Peng, C.-K., L. Goldberger, A., & Hausdorff, J. M. (2003). Multiscale entropy analysis of human gait dynamics. *Physica A: Statistical Mechanics and Its Applications*, *330*(1–2), 53–60.

Dearmon, J., & Smith, T. E. (2016). Gaussian Process Regression and Bayesian Model Averaging: An Alternative Approach to Modeling Spatial Phenomena: Gaussian Process Regression and BMA. *Geographical Analysis*, *48*(1), 82–111.

Dissing, A. S., Jørgensen, T. B., Gerds, T. A., Rod, N. H., & Lund, R. (2019). High perceived stress and social interaction behaviour among young adults. A study based on objective measures of face-to-face and smartphone interactions. *PLOS ONE*, *14*(7), e0218429.

Doryab, A., Villalba, D. K., Chikersal, P., Dutcher, J. M., Tumminia, M., Liu, X., Cohen, S., Creswell, K., Mankoff, J., Creswell, J. D., & Dey, A. K. (2019). Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: Statistical Analysis, Data Mining and Machine Learning of Smartphone and Fitbit Data. *JMIR MHealth and UHealth*, *7*(7), e13209.

Eagle, N., Pentland, A. (Sandy), & Lazer, D. (2009). Inferring friendship network structure by using mobile phone data. *Proceedings of the National Academy of Sciences*, *106*(36), 15274–15278.

Eagle, N., & (Sandy) Pentland, A. (2006). Reality mining: Sensing complex social systems. *Personal and Ubiquitous Computing*, *10*(4), 255–268.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., Kamath, J., Russell, A., Bamis, A., & Wang, B. (2016). Behavior vs. introspection: Refining prediction of clinical depression via smartphone sensing data. *2016 IEEE Wireless Health (WH)*, 1–8.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (0 ed.). Chapman and Hall/CRC.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*(5), 791–806.

Goldberg, B., Brintnell, E. S., & Goldberg, J. (2002). The Relationship Between Engagement in Meaningful Activities and Quality of Life in Persons Disabled by Mental Illness. *Occupational Therapy in Mental Health*, *18*(2), 17–44.

Khoulji, S., Garzón-Rey, J. M., & Aguilo, J. (2017). Remote Assessment of Disease and Relapse – Central Nervous System- RADAR-CNS. *Transactions on Machine Learning and Artificial Intelligence*, *5*(4).

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Lampinen, P., & Heikkinen, E. (2003). Reduced mobility and physical activity as predictors of depressive symptoms among community-dwelling older adults: An eight-year follow-up study. *Aging Clinical and Experimental Research*, *15*(3), 205–211.

Lee, R. M., Draper, M., & Lee, S. (2001). Social connectedness, dysfunctional interpersonal behaviors, and psychological distress: Testing a mediator model. *Journal of Counseling Psychology*, *48*(3), 310–318.

Matcham, F., Barattieri Di San Pietro, C., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., Folarin, A. A., Haro, J. M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V., Mohr, D. C., Myin-Germeys, I., Narayan, V., Bwjh, P., Ranjan, Y., Rashid, Z., Rintala, A., … Meyer, N. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry*, *19*(1), 1–11.

Mizuno, K. (2014). Human circadian rhythms and exercise: Significance and application in real-life situations. *The Journal of Physical Fitness and Sports Medicine*, *3*(3), 307–315.

Mizuno, T., Takahashi, T., Cho, R. Y., Kikuchi, M., Murata, T., Takahashi, K., & Wada, Y. (2010). Assessment of EEG dynamical complexity in Alzheimer’s disease using multiscale entropy. *Clinical Neurophysiology*, *121*(9), 1438–1446.

Moturu, S. T., Khayal, I., Aharony, N., Wei Pan, & Pentland, A. (2011). Sleep, mood and sociability in a healthy population. *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 5267–5270.

Nordström, E., Diot, C., Gass, R., & Gunningberg, P. (2007). Experiences from measuring human mobility using Bluetooth inquiring devices. *Proceedings of the 1st International Workshop on System Evaluation for Mobile Platforms - MobiEval '07*, 15–20.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Ranjan, Y., Rashid, Z., Stewart, C., Conde, P., Begale, M., Verbeeck, D., Boettcher, S., Hyve, Dobson, R., Folarin, A., & RADAR-CNS Consortium. (2019). RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. *JMIR MHealth and UHealth*, *7*(8), e11734.

Rizvi, S. J., Cyriac, A., Grima, E., Tan, M., Lin, P., Gallaugher, L. A., McIntyre, R. S., & Kennedy, S. H. (2015). Depression and Employment Status in Primary and Tertiary Care Settings. *The Canadian Journal of Psychiatry*, *60*(1), 14–22.

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018). Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: Systematic review. *JMIR MHealth and UHealth*, *6*(8).

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P., & Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*, *2016*(9), 1–15.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, *17*(7), 1–11.

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, *2*, e55.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Silva, L. E. V., Cabella, B. C. T., Neves, U. P. da C., & Murta Junior, L. O. (2015). Multiscale entropy-based methods for heart rate variability complexity analysis. *Physica A: Statistical Mechanics and Its Applications*, *422*, 143–152.

Sun, S., Folarin, A. A., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Cummins, N., Matcham, F., Dalla Costa, G., Simblett, S., Leocani, L., Lamers, F., Sørensen, P. S., Buron, M., Zabalza, A., Guerrero Pérez, A. I., Penninx, B. W., Siddi, S., Haro, J. M., … RADAR-CNS Consortium. (2020). Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19. *Journal of Medical Internet Research*, *22*(9), e19992.

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Walker, W. H., Walton, J. C., DeVries, A. C., & Nelson, R. J. (2020). Circadian rhythm disruption and mental health. *Translational Psychiatry*, *10*(1), 28.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 3–14.

World Health Organization. (2017). Depression and other common mental disorders: Global health estimates. *World Health Organization*. https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf

Wu, C., Barczyk, A. N., Craddock, R. C., Harari, G. M., Thomaz, E., Shumake, J. D., Beevers, C. G., Gosling, S. D., & Schnyer, D. M. (2021). Improving prediction of real-time loneliness and companionship type using geosocial features of personal smartphone data. *Smart Health*, *20*, 100180.

Yan, Z., Yang, J., & Tapia, E. M. (2013). Smartphone bluetooth based social sensing. *Proceedings of the 2013 ACM Conference on Pervasive and Ubiquitous Computing Adjunct Publication*, 95–98.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Bendayan, R., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Laiou, P., Matcham, F., White, K. M., Lamers, F., Siddi, S., Simblett, S., Myin-Germeys, I., Rintala, A., Wykes, T., Haro, J. M., … Dobson, R. J. B. (2021). Relationship Between Major Depression Symptom Severity and Sleep Collected Using a Wristband Wearable Device: Multicenter Longitudinal Observational Study. *JMIR MHealth and UHealth*, *9*(4), e24604.

Chapter 5

Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study

This chapter has been published as:

Zhang Y, Folarin AA, Sun S, Cummins N, Vairavan S, Bendayan R, Ranjan Y, Rashid Z, Conde P, Stewart C, Laiou P, Sankesara H, Matcham F, White KM, Oetzmann C, Ivan A, Lamers F, Siddi S, Vilella E, Simblett S, Rintala A, Bruce S, Mohr DC, Myin-Germeys I, Wykes T, Haro JM, Penninx BW, Narayan VA, Annas P, Hotopf M, Dobson RJ, and RADAR-CNS consortium. Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study. *JMIR Ment Health 2022;9*(3): e34898

Background: The mobility of an individual measured by phone-collected location data has been found to be associated with depression; however, the longitudinal relationships (the temporal direction of relationships) between depressive symptom severity and phone-measured mobility have yet to be fully explored.

Objective: We aimed to explore the relationships and the direction of the relationships between depressive symptom severity and phone-measured mobility over time.

Methods: Data used in this paper came from a major EU program, called the Remote Assessment of Disease and Relapse–Major Depressive Disorder, which was conducted in 3 European countries. Depressive symptom severity was measured with the 8-item Patient Health Questionnaire (PHQ-8) through mobile phones every 2 weeks. Participants' location data were recorded by GPS and network sensors in mobile phones every 10 minutes, and 11 mobility features were extracted from location data for the 2 weeks prior to the PHQ-8 assessment. Dynamic structural equation modeling was used to explore the longitudinal relationships between depressive symptom severity and phone-measured mobility.

Results: This study included 2341 PHQ-8 records and corresponding phone-collected location data from 290 participants (age: median 50.0 IQR 34.0, 59.0) years; of whom 215 (74.1%) were female, and 149 (51.4%) were employed. Significant negative correlations were found between depressive symptom severity and phone-measured mobility, and these correlations were more significant at the within-individual level than the between-individual level. For the direction of relationships over time, Homestay (time at home) ($\varphi=0.09$, $P=.01$), Location Entropy (time distribution on different locations) ($\varphi=-0.04$, $P=.02$), and Residential Location Count (reflecting traveling) ($\varphi=0.05$, $P=.02$) were significantly correlated with the subsequent changes in the PHQ-8 score, while changes in the PHQ-8 score significantly affected ($\varphi=-0.07$, $P<.001$) the subsequent periodicity of mobility.

Conclusions: Several phone-derived mobility features have the potential to predict future depression, which may provide support for future clinical applications, relapse prevention, and remote mental health monitoring practices in real-world settings.

5.1 Introduction

Depression is a prevalent and serious mental health disorder that is a leading cause of disability worldwide (Ferrari et al., 2013). It can cause physical health and psychological function problems, resulting in loss of productivity and a high social burden (Beck et al., 2011; Cuijpers & Schoevers, 2004; Katon & Ciechanowski, 2002; Simon, 2003). Currently, diagnosis of depression relies on skilled clinicians and self-report questionnaires, which have limitations that include subjective bias and dynamic information loss (Zhang et al., 2021). Consequently, many people with depression do not receive timely and effective treatment (Kessler et al., 2005), and more efficient methods for detecting and monitoring depression are needed. Recently, the use of mobile phones with embedded sensors for depression detection and monitoring, to provide new ways for supporting both depressed people and clinicians, has been investigated (Donker et al., 2013).

We focused on exploring how phone-collected location data could link individuals' mobility and depression. Past survey-based studies found that mobility is significantly and negatively associated with depression (Perrino et al., 2010; Roshanaei-Moghaddam et al., 2009; Weyerer & Kupfer, 1994). Several longitudinal survey–based studies reported a bidirectional relationship between depression and mobility over time, that is, decreased mobility worsened subsequent depressive symptoms and vice versa (Perrino et al., 2010; Roshanaei-Moghaddam et al., 2009). If the changes in mobility that occur before changes in depression can be captured by mobile phone technologies, early intervention can take place, which could prevent depression relapse or deterioration. Therefore, it is valuable to investigate relationships between depressive symptom severity and phone location data over time.

In recent years, there have been several studies (Ben-Zeev et al., 2015; Chow et al., 2017; Farhan et al., 2016; Laiou et al., 2022; Lu et al., 2018; Meyerhoff et al., 2021; Pratap et al., 2019; Saeb et al., 2015, 2016; Wang et al., 2014, 2018) exploring the associations between depressive symptom severity and mobility features extracted from phone-collected location data that have shown that mobility measured by phones is negatively associated with the severity of depressive symptoms which is consistent with past survey-based studies; however, not many have explored the direction of the relationships between depression and mobility over time. Meyerhoff et al recently found that phone-derived mobility features were correlated with subsequent changes in depression, but not vice versa (Meyerhoff et al., 2021). However, the autoregressive nature of depressive states and mobility levels (Gana et al., 2017; Rhodes & Courneya, 2003; Wichers, 2014) and the influence of individual differences may affect the results. In addition, the limitations of many previous phone-based studies included relatively small and homogeneous (e.g., university students) populations and the lack of comparison of between-individual and within-individual differences. To address these limitations, we aimed to explore the relationships and the direction of relationships over time between phone-derived mobility features and depressive symptom severity on a large multicenter data set.

5.2 Methods

5.2.1 Study Design

We used a large longitudinal data set of an EU research program called Remote Assessment of Disease and Relapse–Major Depressive Disorder, which explored the utility of remote measurement technologies in long-term (up to 2 years) depression monitoring (Matcham et al., 2019). We first used existing mobility features and then designed several new mobility features, which were extracted from this data set. Then, we assessed the relationships and direction of the relationships between depressive symptom severity and mobility features over time using dynamic structural equation models (Asparouhov et al., 2018). Furthermore, we investigated the effects of individual differences (such as demographics) on the models at the between-individual level.

5.2.2 Study Participants and Settings

All participants in the study had at least one diagnosis of depression in the most recent 2 years and were recruited from 3 countries (Netherlands, Spain, and the United Kingdom); additional details descriptions are reported in (Matcham et al., 2022). Participants' passive data (e.g., location, steps, and sleep) and active data (e.g., questionnaires) were respectively collected via passive remote measurement technologies and active remote measurement technologies apps provided by an open-source platform (RADAR-base) (Ranjan et al., 2019). A patient advisory board comprising service users co-developed the study and were involved in the choice of measures, the timing, and issues of engagement and in developing the analysis plan.

5.2.3 Ethics

Ethical approval was obtained from the Camberwell St. Giles Research Ethics Committee (17/LO/1154) in London, from the Fundacio Sant Joan de Deu Clinical Research Ethics Committee (CI: PIC-128-17) in London, and from the Medische Ethische Toetsingscommissie VUms (2018.012–NL63557.029.17) in the Netherlands.

5.2.4 Phone Location and Depression Questionnaire Data

We focused on phone location data and data from the 8-item Patient Health Questionnaire (PHQ-8) (Kroenke et al., 2009). The passive remote measurement technologies app measured participants' location coordinates (longitude and latitude) using 2 providers (GPS and network sensors) periodically every 10 minutes. To protect participants' private information, raw locations were obfuscated by adding a unique and random reference location which was assigned to each participant at the start of the study (RADAR-Base, 2022). The participant's self-reported depressive symptom severity was measured via the PHQ-8, with a score between 0 and 24 (Kroenke et al., 2009), which was assessed through the active remote measurement technologies app every 2 weeks (thus, the 2 weeks preceding each PHQ-8 record was the PHQ-8 interval).

5.2.5 Data Inclusion Criteria

Several factors may affect our analysis, such as the COVID-19 pandemic, location data accuracy, and missing data. Notably, the COVID-19 pandemic and related lockdown policies greatly impacted European people's mobility behaviors (Sun et al., 2020). Therefore, according to suggestions in previous studies (Farhan et al., 2016; Lu et al., 2018; Saeb et al., 2015) and our experiences, we selected a subset of the data set (Matcham et al., 2019) using the 3 criteria: (1) data from before February 2020 (prior to COVID-19 interventions in Europe) (Zhang et al., 2021) were included, (2) location records with an error larger than 165 meters were removed (Farhan et al., 2016; Lu et al., 2018), and (3) the amount of missing location data in a given PHQ-8 interval was limited to 50% (Farhan et al., 2016; Lu et al., 2018; Saeb et al., 2015).

5.2.6 Data Preprocessing

We calculated the distances between consecutive location records and the instantaneous speeds at all location records. The distance between 2 consecutive location records was computed by using the Haversine formula (Depp et al., 2019). The instantaneous speed was approximated by dividing the distance by the time between 2 consecutive location records. We regarded one location record as a stationary point if its instantaneous speed was less than 1 km/h; otherwise, we considered it a moving point (Farhan et al., 2016; Saeb et al., 2015).

The second procedure was location clustering. Since the density-based spatial clustering of applications with noise method (Ester et al., 1996) can treat low-density location points as outliers, avoiding overestimating the number of locations clusters (Farhan et al., 2016), we used this method for location clustering, using hyperparameters and the method for handling unequal sampling intervals from (Farhan et al., 2016).

5.2.7 Feature Extraction

We extracted 11 mobility features (Table 5.1) from location data in each PHQ-8 interval (14 days), of which 4 features (3 frequency-domain features to reflect periodic characteristics of mobility and 1 feature to represent the number of temporary residential locations during the past 14 days) are new.

Table 5.1. A list of mobility features used in this study and their short descriptions.

Feature	Description
Location Variance	Variance of longitude and latitude coordinates
Moving Time	Percentage of time spent in moving
Moving Distance	Distance between all location points weighted by available time
Number of Clusters	The number of location clusters found using density-based spatial clustering of applications with noise
Location Entropy	Entropy of time distribution over different locations
Normalized Entropy	Location Entropy normalized by the number of clusters
Homestay	Percentage of time spent at home
Residential Location Count	The number of temporary residential locations
Long-term Rhythm	Percentage of frequency bins within the long-term period (>1 day) of spectrum for longitude and latitude coordinates
Circadian Rhythm	Percentage of frequency bins within the circadian period (24 hours) of spectrum for longitude and latitude coordinates
Short-term Rhythm	Percentage of frequency bins within the short-term period (<1 day) of spectrum for longitude and latitude coordinates

5.2.8 Time-Domain Features

Location Variance

The Location Variance represented the variability of each participant's locations (Saeb et al., 2015) and is calculated as *log(Var*(*Lon*)+*Var*(*Lat*)), where *log* is the logarithm, and *Var*(*Lon*) and *Var*(*Lat*) represent the variances of the longitude and latitude coordinates, respectively, in one PHQ-8 interval.

Moving Time

The Moving Time represented the percentage of time that a participant spent in moving in one PHQ-8 interval (Saeb et al., 2015). The feature was computed by dividing the sum duration for all moving points by the sum of available time in one PHQ-8 interval.

Moving Distance

The Moving Distance was adjusted by dividing the total distance by the available time (in hours) in one PHQ-8 interval. In previous studies (Saeb et al., 2015, 2016), the total distance obtained by accumulating distances between all location records; however, this total distance was affected by the missing data rate.

Number of Clusters

The number of the unique location clusters that a participant visited in one PHQ-8 interval was calculated using density-based spatial clustering of applications with noise (Farhan et al., 2016).

Location Entropy

Location Entropy represented the distribution of time spent by a participant at different location clusters in one PHQ-8 interval (Saeb et al., 2015) and was calculated as

$$Location\ Entropy = -\sum_{i}^{Number\ of\ Clusters} p_i log\, p_i$$

where p_i is the percentage of time spent at location cluster *i*, thus the greater the average time, the higher the Location Entropy and vice versa (Saeb et al., 2015).

Normalized Entropy

Because the number of location clusters varies across participants and the number of clusters is positively correlated with Location Entropy (Farhan et al., 2016; Lu et al., 2018; Saeb et al., 2015), we also used Normalized Entropy which is given by *Normalized Entropy* = *Location Entropy* / *log* (*Number of Clusters*)

Homestay

In previous studies (Chow et al., 2017; Farhan et al., 2016; Lu et al., 2018; Saeb et al., 2015, 2016; Wang et al., 2018), each participant was assigned only one home location, which was the most visited location cluster between 12 AM to 6 AM; however, in our study, due to the long follow-up time and community-based population, participants may have more than one residential location in one PHQ-8 interval (for example, for reasons, such as traveling, business trips, or moving to a new house). Therefore, we adjusted the method of determining the residential locations. We first selected all location clusters visited at night (12 AM to 6 AM) in one PHQ-8 interval. Then, if multiple clusters were visited in the same night, the location cluster with the most location records was selected as the home location. This step partially excluded the impact of activities at night. The Homestay was the time spent at all stationary location points belonging to all home locations as the percentage of the available time in one PHQ-8 interval.

Residential Location Count

This new feature represented the number of residential locations. Since temporary home locations could reflect traveling (Isabelle et al., 2019), we used the number of residential locations in one PHQ-8 interval to reflect traveling.

5.2.9 Frequency-Domain Features

People’s life rhythms (such as circadian rhythm, sleep rhythm, and social rhythm) are related to depression (Walker et al., 2020). We propose 3 frequency-domain features to reflect the periodicity of participants’ mobility. To compute frequency-domain features, we used linear interpolation and the fast Fourier transformation to get the spectrums of longitude and latitude data, respectively (Figure 5.1). The frequency axis of the spectrum was scaled in cycles per day to reflect the number of periodic patterns that occurred daily. To explore the periodic rhythms of different period lengths, we used the same frequency-domain division as in our previous publication (Zhang et al., 2021), that is, frequency bands of low frequency (0 to 0.75 cycles per day), middle frequency

(0.75 to 1.25 cycles per day), and high frequency (>1.25 cycles per day). The power in the middle frequency was used to represent the strength of the circadian rhythm (around 1 cycle/day) of the participant's mobility. Likewise, the power in low frequency and high frequency represents the long-term (>1 day) periodic rhythm and short-term (<1 day) rhythm, respectively. We extracted 3 features to reflect the percentages of these 3 periodic rhythms (long-term, circadian, and short-term rhythms) in individuals' mobility. We summed the power in the same frequency band of longitude and latitude, then divided it by the sum of the total spectral power of longitude and latitude. The formulas of these 3 features are

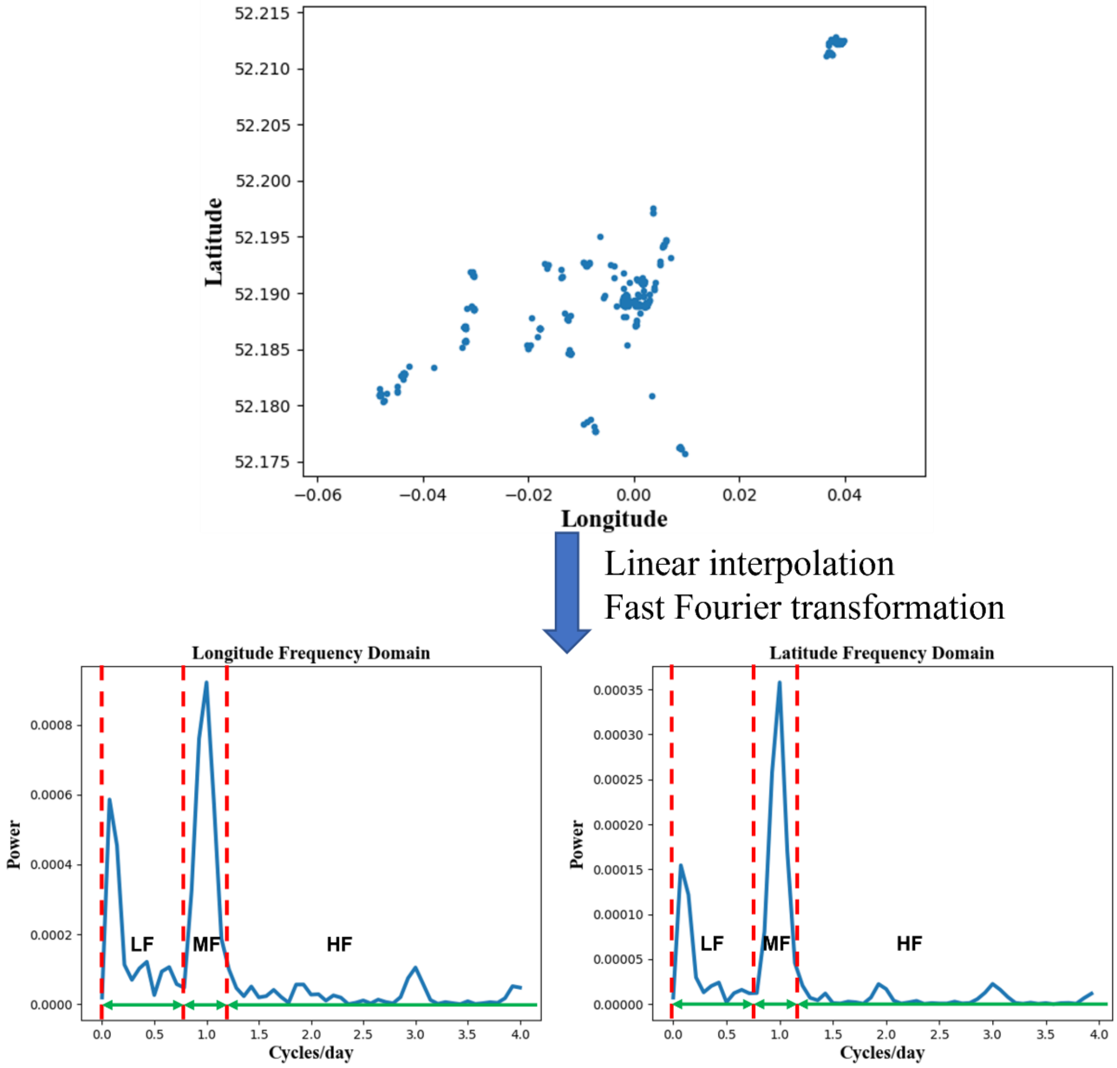
$$Long\text{-}term\ Rhythm=(PSD_{lon}(LF) + PSD_{lat}(LF)) / (PSD_{lon}(Total) + PSD_{lat}(Total))$$

$$Circadian\ Rhythm=(PSD_{lon}(MF) + PSD_{lat}(MF)) / (PSD_{lon}(Total) + PSD_{lat}(Total))$$

$$Short\text{-}term\ Rhythm=(PSD_{lon}(HF) + PSD_{lat}(HF)) / (PSD_{lon}(Total) + PSD_{lat}(Total))$$

where PSD_{lon} and PSD_{lat} represent the power spectral density of longitude and latitude, respectively, and *LF*, *MF*, *HF*, and *Total* are the low frequency, middle frequency, high frequency, and total spectral power, respectively. If the individuals' mobility is regular, the Long-term Rhythm or Circadian Rhythm will be high, otherwise, Short-term Rhythm will be high.

Figure 5.1. A schematic diagram showing the transformation of location data from time domain to frequency domain. (LF=low frequency (0-0.75 cycles/day), MF=middle frequency (0.75-1.25 cycles/day), and HF=high frequency (>1.25 cycles/day)).



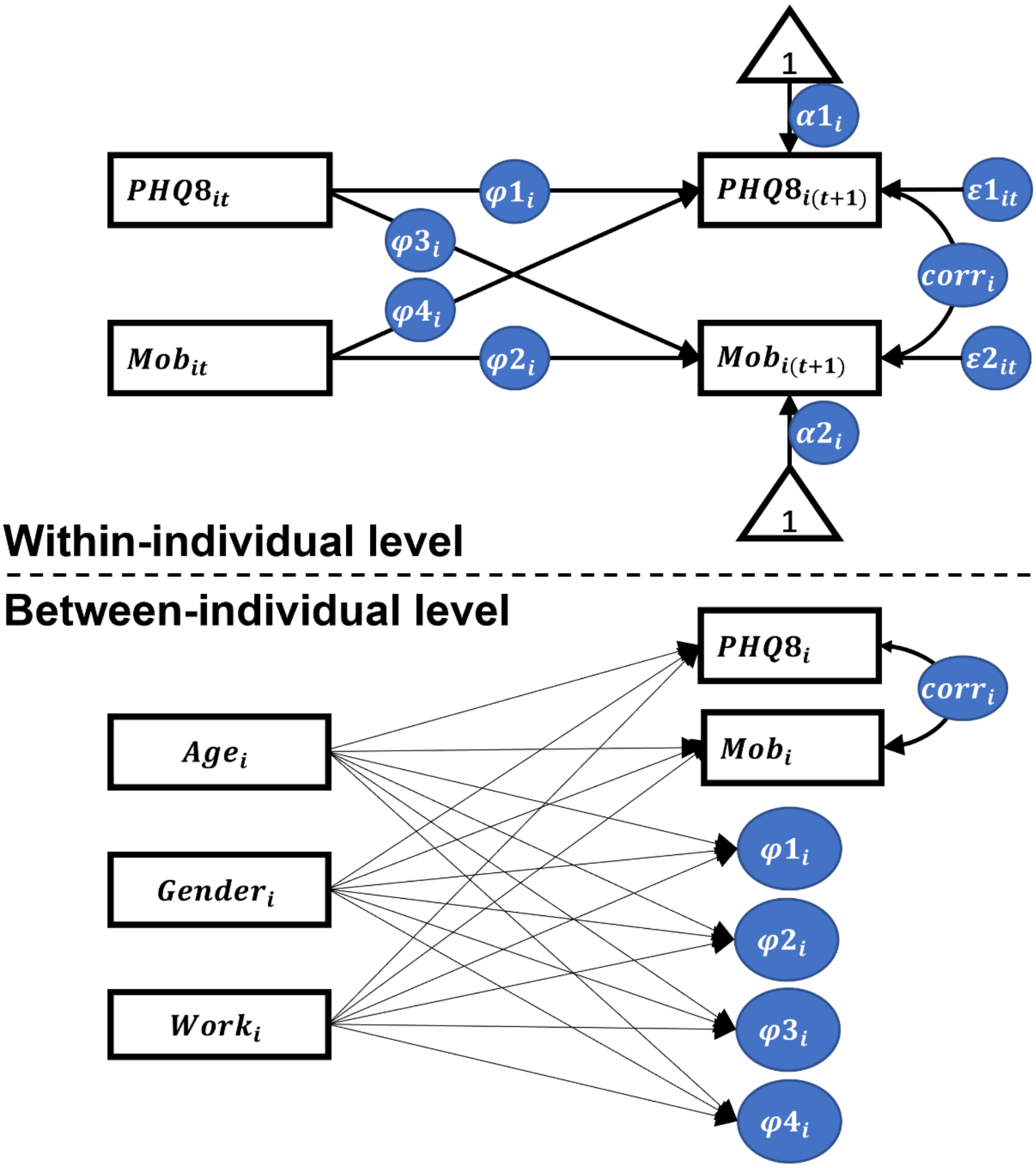
5.2.10 Data Analyses

We used dynamic structural equation modeling to explore the relationships and the direction of relationships between mobility features and PHQ-8 scores over time. Dynamic structural equation modeling is a broad integrated framework that blends multilevel, time-series, and structural equation modeling (Asparouhov et al., 2018; McNeish, 2019; McNeish & Hamaker, 2020) and which has shown to be particularly useful for intensive longitudinal data (McNeish, 2019; McNeish & Hamaker, 2020).

Specifically, the 2-level vector autoregressive model can estimate the lagged effects and cross-lagged effects between 2 outcome variables while considering the variability at both within-individual and between-individual levels (Asparouhov et al., 2018; McNeish & Hamaker, 2020). The lagged effect is the impact of one variable on itself over time, which was used to represent the autoregressive nature of depressive states and mobility levels (Gana et al., 2017; Rhodes & Courneya, 2003; Wichers, 2014). The cross-lagged effect is the impact of one variable on the other variable over time, which was used to explore the direction of relationships between mobility features and PHQ-8 score. In this study, we only considered the Lag-1 model (Figure 5.2), that is, the lagged effects and cross-lagged effects between a time point *t* and the immediately subsequent (2 weeks later) time point (*t* + 1).

We built a vector autoregressive model with each mobility feature and PHQ-8 score as outcome variables and used age, gender, and work status as covariates (Akhtar-Danesh & Landeen, 2007; Aluoja et al., 2004; Rizvi et al., 2015) at the between-individual level for adjusting individual differences. The correlations between the PHQ-8 score and the mobility feature (Figure 5.2) at both within-individual and between-individual levels were also estimated by the vector autoregressive model. We established a total of 11 vector autoregressive models for all mobility features. All *P* values of coefficients in vector autoregressive models and correlations were adjusted using the Benjamini-Hochberg method (Benjamini & Hochberg, 1995) for multiple comparisons. Findings were considered significant at adjusted *P* value<.05. Vector autoregressive models were implemented in Mplus (version 8) (Muthén et al., 2016) and multiple comparison corrections were performed in R software (version 3.6.3).

Figure 5.2. The path diagram of the VAR (1) model used in this paper. $PHQ8_{it}$ and Mob_{it} respectively represent the score of 8-item Patient Health Questionnaire and one of mobility features (Table 5.1) of participant i at time point t (the interval between two time points is 2 weeks), and age, gender, and work status were considered as covariates at the between-individual level.

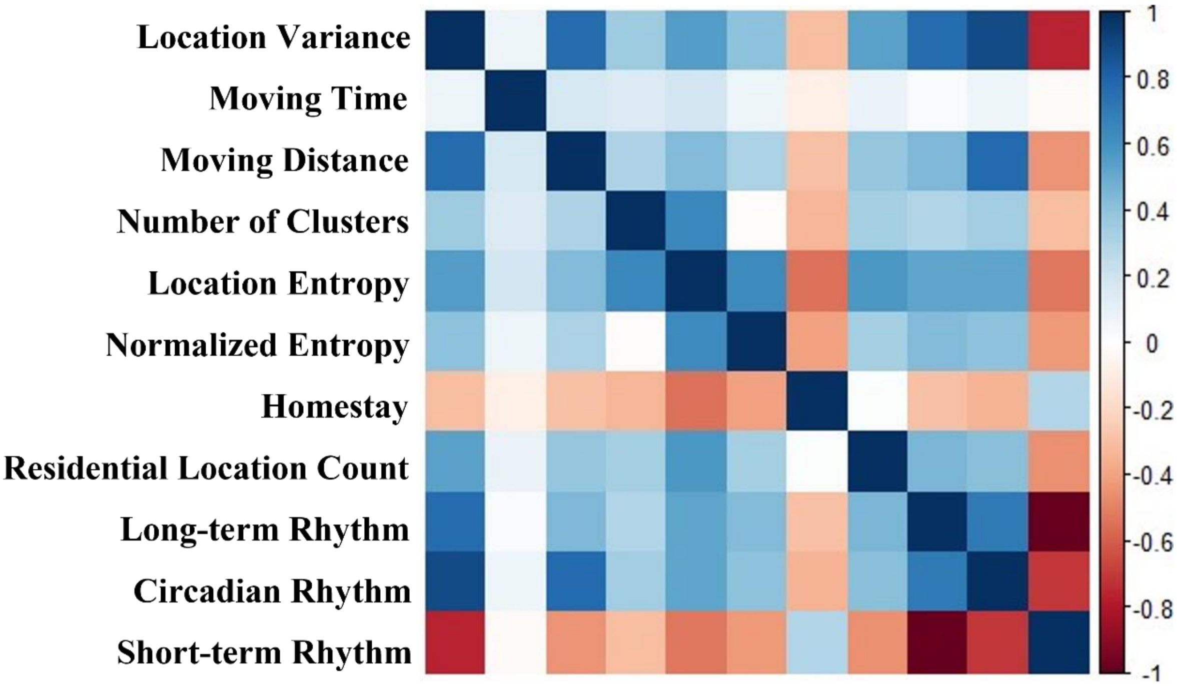


5.3 Results

5.3.1 Data Summary

The 2341 PHQ-8 intervals of 290 participants collected between November 2017 and February 2020 were included in our analysis. The sample had a median age of 50.0 (IQR 34.0, 59.0) years, with 215 (74.14%) female participants and 149 (51.38%) employed participants, with a median of 10 (IQR 5, 15) PHQ-8 scores and a median of 8.0 (IQR 3.0, 14.0) PHQ-8 intervals for each participant. The pairwise Spearman correlations between all 11 mobility features are presented in Figure 5.3.

Figure 5.3. A heatmap of pairwise Spearman correlations between all 11 mobility features extracted in this paper. Definitions of mobility features in this figure are shown in Table 5.1 and the Feature Extraction section.



5.3.2 Vector Autoregressive Models

Correlation

Except for Moving Time (*P*=.11), all mobility features were significantly correlated with the PHQ-8 score at the within-individual level (Table 5.2); Homestay (ρ=0.11, *P*<.001) and Short-term Rhythm (ρ=0.07; *P*=.004) were positively correlated, while other mobility features were negatively correlated. Between individuals, Location Variance (ρ=−0.22, *P*=.04) and Moving Distance (ρ=−0.26, *P*=.04) were significantly and negatively correlated with PHQ-8 scores.

Table 5.2. Mobility features' correlations with PHQ-8 scores at within- and between-individual levels.

Mobility feature	Within-individual level		Between-individual level	
	ρ	Adjusted *P* value	ρ	Adjusted *P* value
Location Variance	−0.10	<.001	−0.22	.04
Moving Time	0.03	.11	−0.09	.28
Moving Distance	−0.08	.002	−0.26	.04
Number of Clusters	−0.09	.001	−0.02	.44
Location Entropy	−0.15	<.001	−0.09	.22
Normalized Entropy	−0.05	.02	−0.14	.11
Homestay	0.11	<.001	0.10	.20
Residential Location Count	−0.09	.001	−0.09	.27
Long-term Rhythm	−0.07	.004	−0.17	.09
Circadian Rhythm	−0.12	<.001	−0.16	.11
Short-term Rhythm	0.07	.004	0.16	.09

Lagged and Cross-lagged Effects

There were significant and positive lagged effects exist in both PHQ-8 scores (φ1=0.45-0.51, *P*<.001) and mobility features (φ2=0.11-0.53, *P*<.001) (Table 5.3). For cross-lagged effects, PHQ-8 scores were significantly and negatively correlated with the subsequent Circadian Rhythm of mobility (φ3=−0.07, *P*<.001), while Location Entropy (φ4=−0.04, *P*=.02), Homestay (φ4=0.09, *P*=.01), and Residential Location Count (φ4=0.05, *P*=.02) were significantly correlated with subsequent PHQ-8 scores.

Table 5.3. Lagged and cross-lagged effects between mobility features and PHQ-8 scores estimated by vector autoregressive models.

Mobility feature	Lagged effects				Cross-lagged effects			
	φ1	Adjusted	φ2	Adjusted	φ3	Adjusted	φ4	Adjusted
Location Variance	0.49	<.001	0.2	<.001	−0.03	.22	0.02	.23
Moving Time	0.47	<.001	0.53	<.001	0.02	.22	0.02	.31
Moving Distance	0.48	<.001	0.38	<.001	0.03	.21	0.03	.21
Number of Clusters	0.49	<.001	0.3	<.001	0.005	.50	−0.01	.32
Location Entropy	0.47	<.001	0.22	<.001	−0.01	.33	−0.04	.02
Normalized Entropy	0.46	<.001	0.14	<.001	−0.004	.44	0.003	.45
Homestay	0.45	<.001	0.34	<.001	−0.01	.30	0.09	.01
Residential Location Count	0.51	<.001	0.11	<.001	−0.01	.34	0.05	.02
Long-term Rhythm	0.49	<.001	0.21	.001	−0.05	.06	0.001	.45
Circadian Rhythm	0.48	<.001	0.11	<.001	−0.07	<.001	0.03	.12
Short-term Rhythm	0.48	<.001	0.11	<.001	0.05	.06	−0.03	.34

The Influence of Individual Differences

Older and employed participants had significantly lower intercepts of the PHQ-8 score than younger and unemployed participants (Table 5.4). For mobility features, age was significantly and negatively correlated with Number of Clusters (γ=−0.12, *P*=.01), Location Entropy (γ=−0.18, *P*<.001), and Residential Location Count (γ=−0.16, *P*<.001), while work status was significantly correlated with most mobility features (except for Moving Time [*P*=.42] and Residential Location Count [*P*=.09]). For lagged effects, older participants had significantly lower lagged effects on Moving Distance (γ=−0.16, *P*=.02) and Homestay (γ=−0.14, *P*=.03) than younger participants. Female participants had significantly lower lagged effects on Location Entropy (γ=−0.15, *P*=.02) and Residential Location Count (γ=−0.24, *P*=.01) than male participants. Compared with unemployed participants, employed participants have significantly lower lagged effects on the PHQ-8 score (γ=−0.14, *P*=.03) and significantly higher lagged effects on Normalized Entropy (γ=0.25, *P*=.01). For cross-lagged effects, age was significantly and negatively correlated with the φ3 coefficient of Circadian Rhythm (γ=−0.49, *P*=.004) in the corresponding vector autoregressive model.

Table 5.4. Significant effects of individual difference at the between level of the vector autoregressive models. Only significant effects of at least one covariate are reported.

Characteristic	Age		Female		Employed	
	γ	Adjusted	γ	Adjusted	γ	Adjusted
Effects on the intercept of						
Patient Health Questionnaire–8	−0.21	<.001	0.07	.09	−0.10	.01
Location Variance	−0.08	.06	0.03	.29	0.12	.01
Moving Distance	0.01	.47	−0.01	.40	0.07	.01
Number of Clusters	−0.12	.01	0.02	.36	0.09	.03
Location Entropy	−0.18	<.001	0.01	.40	0.20	<.001
Normalized Entropy	−0.09	.09	−0.01	.45	0.26	<.001
Homestay	0.01	.32	0.03	.16	−0.15	<.001
Residential Location Count	−0.16	<.001	0.04	.17	0.06	.09
Long-term Rhythm	−0.07	.07	0.02	.34	0.14	.01
Circadian Rhythm	−0.07	.08	0.06	.10	0.13	<.001
Short-term Rhythm	0.10	.06	−0.06	.13	−0.16	<.001
Effects on the lagged effect of						
Patient Health Questionnaire–8	0.01	.47	−0.07	.13	−0.14	.03
Moving Distance	−0.16	.02	−0.04	.31	−0.08	.06
Location Entropy	−0.01	.46	−0.15	.02	0.02	.38
Normalized Entropy	0.09	.19	−0.19	.05	0.25	.01
Homestay	−0.14	.03	−0.09	.13	0.05	.27
Residential Location Count	0.01	.48	−0.24	.01	−0.04	.36
Effects on the cross-lagged effect of						
Circadian Rhythm (φ3)[a]	−0.49	.004	0.01	.48	0.164	.25

[a] φ3 represents the effect of the Patient Health Questionnaire–8 on the subsequent mobility feature.

5.4 Discussion

5.4.1 Principal Findings

This study provides a comprehensive understanding of the relationships and the direction of the relationships between depressive symptom severity and phone-measured mobility over time by using dynamic structural equation modeling on a large longitudinal data set and considering correlations at both individual and population levels, lagged effects (the autoregressive nature over time), cross-lagged effects (direction of the relationships over time), and the influences of individual differences (demographic characteristics).

Most mobility features extracted in this paper were significantly correlated with the PHQ-8 score at the within-individual level (Table 5.2), which indicated that, for a participant, the higher the severity of depressive symptoms, the lower mobility. This is consistent with both past survey-based (Weyerer & Kupfer, 1994) and phone-based studies (Saeb et al., 2015, 2016). These findings reaffirmed that the link between depressive symptom severity and mobility can be captured by mobile phones. However, many of the mobility features' correlations with PHQ-8 score were not significant at the between-individual level, possibly due to the significant effects of individual differences (age and work status) on both PHQ-8 score and mobility features (Table 5.4). Notably, features of Location Variance ($\rho=-0.22$, $P=.04$) and Moving Distance ($\rho=-0.26$, $P=.04$) were still significantly correlated with PHQ-8 score at the between-individual level, which indicated these features are relatively robust for reflecting depressive symptom severity in the whole population. Compared with the results of previous phone-based studies, our results showed that population diversity affects correlations between mobility features and the depression score. Most mobility features were significantly correlated with depression scores in student-based studies (Lu et al., 2018; Saeb et al., 2016), while several features lost their significance in a community-based population with a wide age distribution (Saeb et al., 2015). These findings indicated that individual differences need to be considered during exploring relationships between depression and mobility.

PHQ-8 score and mobility features both had significant and positive lagged effects (Table 5.3), indicating that the autoregressive nature of individuals' depressive states (Wichers, 2014) and movement habits (Rhodes & Courneya, 2003) could be captured by mobile phones. For the direction of relationships over time, we found 3 mobility features significantly correlated with the subsequent PHQ-8 score. Specifically, increases in PHQ-8 score are probably preceded by one or more following changes in the mobility: (1) lower average time spent at different places (Location Entropy), (2) more time at home (Homestay), and (3) more traveling (Residential Location Count). Conversely, change in PHQ-8 score was significantly and negatively correlated (φ3=−0.07, *P*<.001) with the subsequent circadian rhythm measured by location data. The findings of a recent study (Meyerhoff et al., 2021) showed changes in several mobility features were associated with subsequent depression changes, but not vice versa. The differences in populations and applied methods could be potential reasons for the slightly inconsistent results. Both our study and that study (Meyerhoff et al.,

2021) have shown that the changes in mobility prior to changes in depressive symptom severity can be captured by mobile phones. An interesting finding is that the number of residential locations was positively correlated (φ4=0.05, *P*=.02) with the subsequent PHQ-8 score (Table 5.3), which is opposite to their negative correlation (ρ=−0.09, *P*=.001) at the within-individual level (Table 5.2). As the number of temporary residential locations could reflect traveling (Isabelle et al., 2019), this finding indicated that traveling may reduce the current depressive symptoms but may worsen some existing depressive feelings. This finding may provide insight into a phenomenon called "post-travel depressed feelings (Jafari, 1987; *Post-Vacation Blues*, 2022)." The causes of "post-travel depressed feelings" are fatigue from trips, the shock of re-entry of ordinary life, and jet lag (Jafari, 1987; Katz et al., 2001).

For influences of individual differences on the levels of depressive symptom severity and mobility, we found that PHQ-8 scores tended to be lower in participants who are older or have jobs, which can be expected because previous survey-based studies have shown that depression is negatively correlated with age, and the unemployment rate in the depressed population is high (Akhtar-Danesh & Landeen, 2007; Aluoja et al., 2004; Rizvi et al., 2015). Gender was not significantly correlated with the PHQ-8 score (γ=0.07, *P*=.09) in our population, possibly due to all participants in our study having at least one diagnosis of depression in recent 2 years (Matcham et al., 2019), which may reduce the link between gender and depressive symptom severity. For the effects of demographic characteristics on mobility features, we found that the mobility in older participants or participants without jobs tended to be lower, which is also expected. For influences of individual differences on the lagged and cross-lagged effects, we found the participants with jobs had lower autocorrelation of the PHQ-8 score, indicating more depressive symptoms severity changes over time in employed participants than unemployed participants. Female participants, older participants, and unemployed participants tended to have lower autocorrelations of some mobility features than male participants, young participants, and employed participants, which indicated that variabilities of mobility over time were larger in these participants. For influences of age on cross-lagged effects, the impact of changes in PHQ-8 score on the subsequent circadian rhythm for older participants was significantly lower than that of young participants ($\gamma=-0.49$, *P*=.004), indicating that the mobility rhythm of the older participants is affected by depressive symptoms for a shorter period than the young participants.

We proposed 3 frequency-domain features to reflect the periodic characteristics of individuals' mobility (Figure 5.1). They were all significantly correlated with the PHQ-8 score at the within-individual level. Higher values of Long-term Rhythm and Circadian Rhythm represent more regular movement and activity, which were correlated with lower depressive symptom severity. Notably, Circadian Rhythm had the strongest correlation ($\rho=-0.12$, *P*<.001) among these 3 features, and it had significant cross-lagged effect ($\varphi3=-0.07$, *P*<.001) with the preceding PHQ-8 score. These findings demonstrated that the frequency-domain of location data can provide some additional information for evaluating depressive symptom severity in future research.

5.4.2 Limitations

We obfuscated the raw location data due to privacy issues. Therefore, we did not have access to contextual information, which may mean some information was lost. Another limitation is that we only used the Lag-1 vector autoregressive models. We did not use high-order vector autoregressive models because we wanted to make our preliminary model simple to allow easier explanation and to avoid convergence problems in the procedure of coefficient estimations. We will attempt high-order vector autoregressive models in future research when we have more data without the impact of the COVID-19.

We chose to build 11 dynamic structural equation modeling models, one for each mobility feature. Since each mobility feature has a specific meaning, the bivariate model can better explain changes of the feature before and after the changes in PHQ-8 scores indicating the longitudinal relationships. We attempted multivariate dynamic structural equation modeling with all mobility features, but the model failed to converge, possibly due to the multicollinearity between mobility features and complexity of the model. As all mobility features were devised for describing characteristics of individuals' mobility, there were high correlations between mobility features (Figure 5.3). In future research, we plan to solve the multicollinearity in the multivariate model through further feature engineering and feature selection methods or by using other multivariate time series models which are robust to multicollinearity (Garg & Tai, 2013).

5.4.3 Conclusions

This study provides initial evidence of the relationship and the direction of the relationship between depressive symptom severity and phone-measured mobility over time. We found several mobility features affected depressive symptom severity, while changes in the depression score were associated with the subsequent periodic rhythm of mobility. These mobility features have the potential to be used as indicators for assessing depression risk in future clinical applications, which could provide timely suggestions for both people with depression risk (eg, encouraging to attend more activities) and physicians (eg, early interventions). This work may provide support for remote mental health monitoring practice in real-world settings.

Acknowledgments

The Remote Assessment of Disease and Relapse–Major Depressive Disorder project is a component of the Remote Assessment of Disease and Relapse–Central Nervous System (RADAR-CNS) project, which has received funding from the Innovative Medicines Initiative 2 Joint Undertaking (grant 115902), which receives support from European Union Horizon 2020 and the European Federation of Pharmaceutical Industries and Associations. The study was conducted in accordance with the Declaration of Helsinki and Good Clinical Practice, adhering to principles outlined in the National Health Service Research Governance Framework for Health and Social Care (2nd edition).

Participant recruitment in Amsterdam was partially accomplished through Hersenonderzoek.nl (https://hersenonderzoek.nl), which is funded by ZonMw-Memorabel (project 73305095003).

This study was funded in part by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS (National Health Service) Foundation Trust and King's College London.

The funding bodies have not been involved in the design of the study, the collection or analysis of data, or the interpretation of data, and the views expressed are those of the authors.

We thank following 4 clinical communities in Spain: *Parc Sanitari Sant Joan de Déu* Network services, *Institut Català de la Salut, Institut Pere Mata,* and *Hospital Clínico San Carlos* for recruiting participants in the CIBER site.

We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design. We acknowledge that this research was reviewed by the Feasibility and Acceptability Support Team for Researchers (National Institute for Health Research Maudsley Biomedical Research Centre, King's College London and South London and Maudsley NHS Foundation Trust).

We thank all GLAD Study volunteers for their participation and gratefully acknowledge the NIHR BioResource, NIHR BioResource centers, NHS Trusts and staff for their contribution. We also acknowledge NIHR Biomedical Research Centre, King's College

London, South London and Maudsley NHS Trust and King's Health Partners. We thank the NIHR, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub program.

MH and TW were supported by NIHR Senior Investigator Awards.

RB is funded in part by the King’s College London Medical Research Council Skills Development Fellowship (grant MR/R016372/1), which is funded by the UK Medical Research Council (grant IS-BRC-1215-20018) for the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London.

RJBD is supported by (1) NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome Trust; (3) the BigData@Heart Consortium, which funded by the Innovative Medicines Initiative-2 Joint Undertaking (grant 116074), which receives support from the European Union Horizon 2020 research and innovation program and European Federation of Pharmaceutical Industries and Associations; (4) the NIHR University College London Hospitals Biomedical Research

Centre; (5) the UK Research and Innovation London Medical Imaging & Artificial Intelligence Centre for Value Based Health care; and (6) the NIHR Applied Research Collaboration South London at King’s College Hospital NHS Foundation Trust.

Authors’ Contributions

YZ extracted and integrated the questionnaire and location data for the analysis, planned and performed the analysis, and drafted the manuscript. MH and VAN gained funding and co-led the Remote Assessment of Disease and Relapse–Central Nervous System program. MH is the principal investigator for the Remote Assessment of Disease and Relapse–Major Depressive Disorder study. RJBD, AAF, YR, ZR, PC, HS, and CS have contributed to the development of the RADAR-base platform used for data collection and management across sites, data protection, security, and storage. YZ, AAF, S Sun, NC, SV, RB, PL, SB, DCM, MH, and RJBD contributed to the design of the study. FM, KMW, CO, AI, FL, S Siddi, EV, S Simblett, JMH, BWJHP, MH contributed to data collection. AAF, IMG, AR, VAN, TW, PA, MH, and RJBD contributed to the administrative, technical, and clinical support of the study. All authors were involved in reviewing the manuscript, had access to the study data, and provided direction and comments on the manuscript.

Conflicts of Interest

SV and VAN are employees of Janssen Research and Development LLC. PA is employed by the pharmaceutical company H. Lundbeck A/S. DCM has accepted honoraria and consulting fees from Apple Inc, Otsuka Pharmaceuticals, Pear Therapeutics, and the One Mind Foundation; has received royalties from Oxford Press; and has an ownership interest in Adaptive Health Inc. MH is the principal investigator of the Remote Assessment of Disease and Relapse–Central Nervous System project, a private public precompetitive consortium that receives funding from Janssen, UCB, Lundbeck, MSD, and Biogen.

References

Akhtar-Danesh, N., & Landeen, J. (2007). Relation between depression and sociodemographic factors. *International Journal of Mental Health Systems*, *1*(1), 4.

Aluoja, A., Leinsalu, M., Shlik, J., Vasar, V., & Luuk, K. (2004). Symptoms of depression in the Estonian population: Prevalence, sociodemographic correlates and social adjustment. *Journal of Affective Disorders*, *78*(1), 27–35.

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic Structural Equation Models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(3), 359–388.

Beck, A., Crain, A. L., Solberg, L. I., Unutzer, J., Glasgow, R. E., Maciosek, M. V., & Whitebird, R. (2011). Severity of Depression and Magnitude of Productivity Loss. *The Annals of Family Medicine*, *9*(4), 305–311.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Ben-Zeev, D., Scherer, E. A., Wang, R., Xie, H., & Campbell, A. T. (2015). Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal*, *38*(3), 218–226.

Chow, P. I., Fua, K., Huang, Y., Bonelli, W., Xiong, H., Barnes, L. E., & Teachman, B. A. (2017). Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students. *Journal of Medical Internet Research*, *19*(3), e62.

Cuijpers, P., & Schoevers, R. A. (2004). Increased mortality in depressive disorders: A review. *Current Psychiatry Reports*, *6*(6), 430–437.

Depp, C. A., Bashem, J., Moore, R. C., Holden, J. L., Mikhael, T., Swendsen, J., Harvey, P. D., & Granholm, E. L. (2019). GPS mobility as a digital biomarker of negative symptoms in schizophrenia: A case control study. *Npj Digital Medicine*, *2*(1), 108.

Donker, T., Petrie, K., Proudfoot, J., Clarke, J., Birch, M.-R., & Christensen, H. (2013). Smartphones for Smarter Delivery of Mental Health Programs: A Systematic Review. *Journal of Medical Internet Research*, *15*(11), e247.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, *96*(34), 226–231.

Farhan, A. A., Yue, C., Morillo, R., Ware, S., Lu, J., Bi, J., Kamath, J., Russell, A., Bamis, A., & Wang, B. (2016). Behavior vs. Introspection: Refining prediction of clinical depression via smartphone sensing data. *2016 IEEE Wireless Health, WH 2016*, 30–37.

Ferrari, A. J., Charlson, F. J., Norman, R. E., Patten, S. B., Freedman, G., Murray, C. J. L., Vos, T., & Whiteford, H. A. (2013). Burden of Depressive Disorders by Country, Sex, Age, and Year: Findings from the Global Burden of Disease Study 2010. *PLoS Medicine*, *10*(11), e1001547.

Gana, K., Bailly, N., Broc, G., Cazauvieilh, C., & Boudouda, N. E. (2017). The Geriatric Depression Scale: Does it measure depressive mood, depressive affect, or both?: Geriatric Depression Scale. *International Journal of Geriatric Psychiatry*, *32*(10), 1150–1157.

Garg, A., & Tai, K. (2013). Comparison of statistical and machine learning methods in modelling of data with multicollinearity. *International Journal of Modelling, Identification and Control*, *18*(4), 295.

Isabelle, F., Dominique, K., & Statia, E. (2019). Home away from home: A longitudinal study of the holiday appropriation process. *Tourism Management*, *71*, 327–336.

Jafari, J. (1987). Tourism models: The sociocultural aspects. *Tourism Management*, *8*(2), 151–159.

Katon, W., & Ciechanowski, P. (2002). Impact of major depression on chronic medical illness. *Journal of Psychosomatic Research*, *53*(4), 859–863.

Katz, G., Durst, R., Zislin, Y., Barel, Y., & Knobler, H. Y. (2001). Psychiatric aspects of jet lag: Review and hypothesis. *Medical Hypotheses*, *56*(1), 20–23.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, Severity, and Comorbidity of 12-Month DSM-IV Disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry*, *62*(6), 617.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173.

Laiou, P., Kaliukhovich, D. A., Folarin, A. A., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Sun, S., Zhang, Y., Matcham, F., Ivan, A., Lavelle, G., Siddi, S., Lamers, F., Penninx, B. W. J. H., Haro, J. M., Annas, P., Cummins, N., Vairavan, S., … Hotopf, M. (2022). The Association between Home Stay and Symptom Severity in Major Depressive Disorder: Preliminary Findings from a Multicenter Observational Study Using Geolocation Data from Smartphones. *JMIR MHealth and UHealth*, *10*(1).

Lu, J., Shang, C., Yue, C., Morillo, R., Ware, S., Kamath, J., Bamis, A., Russell, A., Wang, B., & Bi, J. (2018). Joint Modeling of Heterogeneous Sensing Data for

Depression Assessment via Multi-task Learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1–21.

Matcham, F., Barattieri Di San Pietro, C., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., Folarin, A. A., Haro, J. M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V., Mohr, D. C., Myin-Germeys, I., Narayan, V., Bwjh, P., Ranjan, Y., Rashid, Z., Rintala, A., … Meyer, N. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry*, *19*(1), 1–11.

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K. M., Annas, P., de Girolamo, G., Difrancesco, S., Haro, J. M., Horsfall, M., Ivan, A., Lavelle, G., Li, Q., Lombardini, F., Mohr, D. C., Narayan, V. A., Oetzmann, C., Penninx, B. W. J. H., Bruce, S., … Hotopf, M. (2022). Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): Recruitment, retention, and data availability in a longitudinal remote measurement study. *BMC Psychiatry*, *22*(1), 1–19.

McNeish, D. (2019). Two-Level Dynamic Structural Equation Models with Small Samples. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(6), 948–966.

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, *25*(5), 610–635.

Meyerhoff, J., Liu, T., Kording, K. P., Ungar, L. H., Kaiser, S. M., Karr, C. J., & Mohr, D. C. (2021). Evaluation of Changes in Depression, Anxiety, and Social Anxiety Using Smartphone Sensor Features: Longitudinal Cohort Study. *Journal of Medical Internet Research*, *23*(9), e22844.

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2016). *Regression and mediation analysis using Mplus*. Muthén & Muthén.

Perrino, T., Mason, C. A., Brown, S. C., & Szapocznik, J. (2010). The relationship between depressive symptoms and walking among Hispanic older adults: A longitudinal, cross-lagged panel analysis. *Aging & Mental Health*, *14*(2), 211–219.

Post-vacation blues. (2022). https://en.wikipedia.org/wiki/Post-vacation_blues#cite_note-Lillywhite_2017-5

Pratap, A., Atkins, D. C., Renn, B. N., Tanana, M. J., Mooney, S. D., Anguera, J. A., & Areán, P. A. (2019). The accuracy of passive phone sensors in predicting daily mood. *Depression and Anxiety*, *36*(1), 72–81.

RADAR-base. (2022). https://radar-base.atlassian.net/wiki/spaces/RAD/overview

Ranjan, Y., Rashid, Z., Stewart, C., Conde, P., Begale, M., Verbeeck, D., Boettcher, S., Hyve, Dobson, R., Folarin, A., & RADAR-CNS Consortium. (2019). RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. *JMIR MHealth and UHealth*, *7*(8), e11734.

Rhodes, R., & Courneya, K. (2003). Modelling the theory of planned behaviour and past behaviour. *Psychology, Health & Medicine*, *8*(1), 57–69.

Rizvi, S. J., Cyriac, A., Grima, E., Tan, M., Lin, P., Gallaugher, L. A., McIntyre, R. S., & Kennedy, S. H. (2015). Depression and Employment Status in Primary and Tertiary Care Settings. *The Canadian Journal of Psychiatry*, *60*(1), 14–22.

Roshanaei-Moghaddam, B., Katon, W. J., & Russo, J. (2009). The longitudinal effects of depression on physical activity. *General Hospital Psychiatry*, *31*(4), 306–315.

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P., & Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*, *2016*(9), 1–15.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research*, *17*(7), 1–11.

Simon, G. E. (2003). Social and economic burden of mood disorders. *Biological Psychiatry*, *54*(3), 208–215.

Sun, S., Folarin, A. A., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Cummins, N., Matcham, F., Dalla Costa, G., Simblett, S., Leocani, L., Lamers, F., Sørensen, P. S., Buron, M., Zabalza, A., Guerrero Pérez, A. I., Penninx, B. W., Siddi, S., Haro, J. M., … RADAR-CNS Consortium. (2020). Using Smartphones and Wearable Devices to Monitor Behavioral Changes During COVID-19. *Journal of Medical Internet Research*, *22*(9), e19992.

Walker, W. H., Walton, J. C., DeVries, A. C., & Nelson, R. J. (2020). Circadian rhythm disruption and mental health. *Translational Psychiatry*, *10*(1), 28.

Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). Studentlife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 3–14.

Wang, R., Wang, W., daSilva, A., Huckins, J. F., Kelley, W. M., Heatherton, T. F., & Campbell, A. T. (2018). Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *2*(1), 1–26.

Weyerer, S., & Kupfer, B. (1994). Physical Exercise and Psychological Healtha: *Sports Medicine*, *17*(2), 108–116.

Wichers, M. (2014). The dynamic nature of depression: A new micro-level perspective of mental disorder that meets current challenges. *Psychological Medicine*, *44*(7), 1349–1360.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Laiou, P., Matcham, F., Oetzmann, C., Lamers, F., Siddi, S., Simblett, S., Rintala, A., Mohr, D. C., Myin-Germeys, I., Wykes, T., Haro, J. M., … Dobson, R. J. B. (2021). Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: Preliminary longitudinal study. *JMIR MHealth and UHealth*, *9*(7), 1–19.

Chapter 6

Associations Between Depression Symptom Severity and Daily-Life Gait Characteristics Derived from Long-Term Acceleration Signals in Real-World Settings: Retrospective Analysis

This chapter has been published as:

Zhang Y, Folarin AA, Sun S, Cummins N, Vairavan S, Qian L, Ranjan Y, Rashid Z, Conde P, Stewart C, Laiou P, Sankesara H, Matcham F, White KM, Oetzmann C, Ivan A, Lamers F, Siddi S, Simblett S, Rintala A, Mohr DC, Myin-Germeys I, Wykes T, Haro JM, Penninx BWJH, Narayan VA, Annas P, Hotopf M, Dobson RJB, and RADAR-CNS Consortium. Associations Between Depression Symptom Severity and Daily-Life Gait Characteristics Derived from Long-Term Acceleration Signals in Real-World Settings: Retrospective Analysis. *JMIR Mhealth Uhealth 2022;10*(10): e40667

Background: Gait is an essential manifestation of depression. However, the gait characteristics of daily walking and their relationships with depression have yet to be fully explored.

Objective: The aim of this study was to explore associations between depression symptom severity and daily-life gait characteristics derived from acceleration signals in real-world settings.

Methods: We used two ambulatory data sets (N=71 and N=215) with acceleration signals collected by wearable devices and mobile phones, respectively. We extracted 12 daily-life gait features to describe the distribution and variance of gait cadence and force over a long-term period. Spearman coefficients and linear mixed-effects models were used to explore the associations between daily-life gait features and depression symptom severity measured by the 15-item Geriatric Depression Scale (GDS-15) and 8-item Patient Health Questionnaire (PHQ-8) self-reported questionnaires. The likelihood-ratio (LR) test was used to test whether daily-life gait features could provide additional information relative to the laboratory gait features.

Results: Higher depression symptom severity was significantly associated with lower gait cadence of high-performance walking (segments with faster walking speed) over a long-term period in both data sets. The linear regression model with long-term daily-life gait features (R^2=0.30) fitted depression scores significantly better (LR test *P*=.001) than the model with only laboratory gait features (R^2=0.06).

Conclusions: This study indicated that the significant links between daily-life walking characteristics and depression symptom severity could be captured by both wearable devices and mobile phones. The daily-life gait patterns could provide additional information for predicting depression symptom severity relative to laboratory walking. These findings may contribute to developing clinical tools to remotely monitor mental health in real-world settings.

Please refer to appendix B for supplementary material.

6.1 Introduction

Depression affects the lives of over 300 million people worldwide (Friedrich, 2017) and is associated with many adverse outcomes, including decreased quality of life, loss of occupational function, disability, premature mortality, and suicide (Hawton et al., 2013; Lenox-Smith et al., 2013; Lerner et al., 2004; Lewinsohn et al., 2000). While early treatment can be effective and prevent more serious adverse outcomes (Kamphuis et al., 2012), more than half of depressed people do not receive timely treatment (Harman et al., 2004; Young et al., 2001). Current questionnaire-based depression assessments may be affected by recall bias and may not be able to collect dynamic information (Althubaiti, 2016; Devaux & Sassi, 2016). Therefore, several recent studies have attempted to explore the associations between depression and changes in individuals' behaviors using mobile technologies (Rohani et al., 2018).

Changes in gait are essential manifestations of depression (Schrijvers et al., 2008; Sobin & Sackeim, 1997). The main hypothesis linking gait with depression is a bidirectional interaction between the brain motor system and cortical and subcortical structures, which are related to emotions and cognitive functions (Deligianni et al., 2019; R. D. Sanders & Gillig, n.d.; Yogev-Seligmann et al., 2008). Many studies have explored the relationships between depression and gait characteristics based on "gold-standard" laboratory walking tests. Longer gait cycles, reduced stride length, and slower gait cadence were observed in participants with depression compared with healthy controls, which have been consistently shown in several studies (Brandler et al., 2012; Lemke et al., 2000; Michalak et al., 2009, 2011; Pieruccini-Faria et al., 2018; Radovanović et al., 2014; J. B. Sanders et al., 2016; Sloman et al., 1987; van Iersel et al., 2005). Other gait abnormalities such as reduced gait force (Sloman et al., 1987), increased double support time (Radovanović et al., 2014), reduced swing time variability (Brandler et al., 2012), slumped postures (Michalak et al., 2009), and increased body sway (Pieruccini-Faria et al., 2018) have been reported, but with less consistency across studies.

Laboratory gait tests are hard to be applied in real-world settings because of the need for expensive equipment (e.g., video camera and force plates), specialized laboratories, and the inconvenience of wearing sensors on the knees and ankles, for example (Deligianni et al., 2019; Wang et al., 2021). Some researchers have suggested that people's daily-life activity characteristics should have stronger links to their health conditions than laboratory tests (Atrsaei et al., 2021; Notthoff et al., 2018; Rispens et al., 2015). Therefore, it is necessary to monitor and evaluate daily-life walking using efficient methods.

In recent years, several studies have used mobile technologies to measure daily-life walking patterns and explored their associations with depression. However, most of these studies only measured the number of cumulative steps of daily-life walking (Abedi et al., 2015; Große et al., 2021; McKercher et al., 2009), which is more related to individuals' mobility and physical activity than to gait patterns (e.g., gait cadence and gait force). To our knowledge, there have been only a few studies exploring the associations between daily-life gait patterns and depression directly. Adolph et al found that depressed participants had reduced walking speed, reduced vertical up-and-down movements, and more slumped postures compared with controls by placing two accelerometers on the participant’s trunk and right leg for 2 days (Adolph et al., 2021). However, wearing multiple sensors on the body may not be suitable for long-term monitoring. With the development of sensors, the mobile phone provides a cost-effective, continuous, and unobtrusive means to measure individuals’ behaviors, including daily walking. Therefore, the mobile phone may be a potential tool for long-term gait monitoring.

The aim of this study was to explore the value of daily-walking monitoring for improving the evaluation of depression symptom severity. Our first objective was to design and extract gait features from raw acceleration signals to describe the characteristics of daily walking. The second objective was to explore the associations between gait features and depression symptom severity, and to test whether these associations could be captured by different acceleration devices. The third objective was to test whether daily-life walking could provide additional information for predicting depression relative to laboratory walking. To achieve the second and third objectives, we performed our analyses on two ambulatory data sets, the Long Term Movement Monitoring (LTMM) and Remote Assessment of Disease and Relapse–Major Depressive Disorder (RADAR-MDD) data sets (Matcham et al., 2019; Weiss et al., 2013), with acceleration signals collected by a wearable device and mobile phone, respectively. Importantly, the LTMM data set contains data related to both laboratory and daily walking, which could address the third study objective.

6.2 Methods

6.2.1 Data Sets

LTMM Data Set

The LTMM data set includes demographics (age and gender), depression scores (15-item Geriatric Depression Scale [GDS-15] (D'Ath et al., 1994)), and raw acceleration signals (100 Hz) of laboratory walking tests and 3-day activities for 71 elderly adults (Weiss et al., 2013), which can be downloaded at PhysioNet (Goldberger et al., 2000). Participants were included if they did not have any cognitive or gait/balance disorders (Weiss et al., 2013). Participants were asked to walk at a self-selected and comfortable speed for 1 minute in the laboratory while wearing a 3-axis accelerometer on their lower back (Weiss et al., 2013). The GDS-15 questionnaire contains 15 easy-to-understand, yes/no format questions, which is suitable for depression screening in the older population (Williams & Wallace, 1993; Yesavage & Sheikh, 1986). After the laboratory walking test, all participants were asked to wear the accelerometer for the next 3 consecutive days to record daily activities (Weiss et al., 2013).

Ethics Considerations

RADAR-MDD was conducted per the Declaration of Helsinki and Good Clinical Practice, adhering to principles outlined in the National Health Service (NHS) Research Governance Framework for Health and Social Care (2nd edition). Ethical approval has been obtained in London from the Camberwell St Giles Research Ethics Committee

(REC reference 17/LO/1154), in Spain from the CEIC Fundació Sant Joan de Deu (CI PIC-128-17), and in the Netherlands from the Medische Ethische Toetsingscommissie VUms (METc VUmc registratienummer 2018.012–NL63557.029.17).

RADAR-MDD Data Set

The EU research program RADAR-MDD aimed to investigate the utility of mobile technologies for the long-term monitoring of participants with depression in real-world settings (Matcham et al., 2019, 2022). Adult participants with a depression history were included in the study if they did not meet the following criteria: (1) have other psychiatric disorders (eg, bipolar disorder, schizophrenia, and dementia), (2) have received treatment for drug or alcohol use in the past 6 months, (3) a major medical diagnosis that affects daily activities, and (4) pregnancy (Matcham et al., 2019). A detailed study protocol was published previously (Matcham et al., 2019). In this study, we used a subset of RADAR-MDD data collected from a study site in the United Kingdom (King's College London [KCL]) between November 2017 and April 2021, because the KCL site was the only site to acquire ethical approval for collecting the phone's acceleration signals. We hereafter denote this subset as the RADAR-MDD-KCL data set for convenience. The phone's acceleration signals were collected at 50 Hz and uploaded to an open-source platform, RADAR-base (Ranjan et al., 2019). The participants' depression symptom severity was assessed by the 8-item Patient Health Questionnaire (PHQ-8) (Kroenke et al., 2009) self-reported through mobile phones every 2 weeks. A patient advisory board comprising service users co-developed the study. They were involved in the choice of measures, timing, and issues of engagement, and have also been involved in developing the analysis plan.

6.2.2 Step Detection Algorithm

Since we needed to respectively detect steps on the acceleration signals collected by wearable devices and mobile phones, we chose to use the step detection algorithm (Marron et al., 2016), which was based on mobile phones (Figure 6.1). Given a segment of 3-axis acceleration signals (x_i, y_i, z_i), the magnitude of the acceleration of the segment of acceleration signals was calculated to combine 3D signals to a single series, r_i, where $r_i = \sqrt{x_i^2 + y_i^2 + z_i^2}$. The magnitude of the acceleration signals does not depend on the orientation and tilt of the mobile phone during walking (Marron et al., 2016). Subsequently, r_i was filtered by a weighted moving-average filter to remove noise (Equation 1, w=150 milliseconds) Next, the filtered r_i was subtracted by the mean of $\bar{r_i}$ to make $\bar{r_i}$ symmetric to the x-axis. We calculated two new series, $B1_\text{i}$ and $B2_\text{i}$, based on two thresholds to detect the walking swing phase and stance phase, respectively (see Equations 2 and 3). If a swing phase ends and a stance phase starts, we can identify a step that occurred. The formal detection rule of a step S_i at sample i is that the following two conditions must be satisfied: (1) a change from –0.5 to 0 in B1 ($B1_i$=0 and $B1_{i-1}$=0.5); (2) there is at least one detection of $B2$=–0.5 in a window of size w=150 milliseconds in sample i ($Min(B2_{i:i+w})$=–0.5).

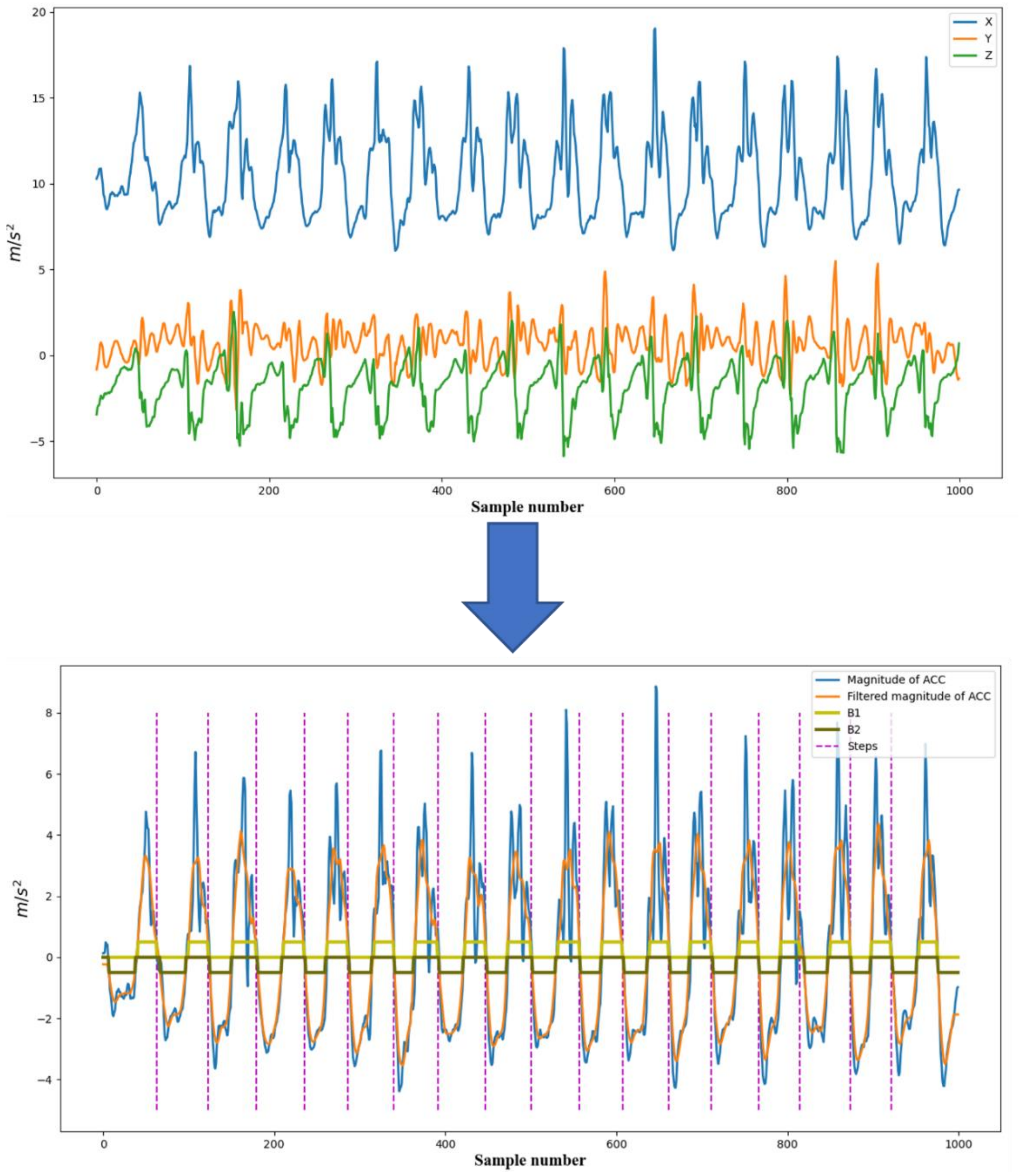
$$\overline{r_i} = \frac{1}{2\omega+1}\sum_{j=i-\omega}^{i+\omega} r_j \quad (1)$$

$$B1_i = \begin{cases} 0.5, & if\ \overline{r_i} \geq 0.5 \\ 0, & otherwise \end{cases} \quad (2)$$

$$B2_i = \begin{cases} -0.5, & if\ \overline{r_i} \leq -0.5 \\ 0, & otherwise \end{cases} \quad (3)$$

Then, the gait cycle series could be derived by calculating time intervals between consecutive steps, which was denoted as *Cycles.* During each gait cycle, the amplitude from the peak to the valley of the magnitude of the acceleration signals was used to reflect the gait force of each step. The force of all steps in the given acceleration signal was denoted as the series *Force.*

Figure 6.1. Step detection algorithm. ACC is the 3-axis acceleration signals, B1 and B2 are two series calculated by thresholds to detect walking swing and stance phase respectively, and pink dash lines represent detected steps.



6.2.3 Feature Extraction

Feature Window Size

Since the PHQ-8 score is used to estimate depression symptom severity for the past 2 weeks (Kroenke et al., 2009), we extracted gait features from a 14-day time window prior to each PHQ-8 record from the RADAR-MDD-KCL data set. For the LTMM data set, we extracted gait features from 3-day activities to link daily-life walking with the GDS-15 score.

Step Detection Window and the Continuous Walking Segment

Daily-life walking in real-world settings is complex and contains some intermittent walking segments (such as walking in a crowded environment or a walking-rest transition status). These intermittent walking segments may not fully reflect a participant's normal walking patterns. Therefore, to distinguish between continuous and intermittent walking, we used a 1-minute sliding window (Ihlen et al., 2015) to detect steps from the long-term raw acceleration signals. If the participant was walking most of the time in this minute, we considered this minute as the continuous walking segment. Based on our experience, we set 50 seconds as the threshold for selecting the continuous walking segment; that is, the segment with more than 50 seconds of walking time (sum of all gait cycles in the minute) was selected for further analysis (Figure 6.2b).

Gait Features

a) Overview

The performance of walking varies over time due to several factors such as mood, energy, and environment. Therefore, the long-term gait features need to represent the distribution and variance of walking patterns over the feature window. We first extracted three short-term gait features from every detected continuous walking segment in the feature window. Then, for each short-term gait feature, we calculated four statistical second-order features (long-term features) across all values of continuous walking segments. In total, 12 long-term gait features were extracted in this study, and a summary of these features is shown in Table 6.1. A schematic diagram of long-term gait feature extraction is shown in Figure 6.2.

b) Short-Term Gait Features from the 1-Minute Continuous Walking Segment

Gait cadence and gait force are essential characteristics of walking. Gait cadence is the rate at which the individual feet contact the ground (Levine et al., 2012). Gait force reflects the ground reaction force during walking (Herzog et al., 1989). For every continuous walking segment, the median of the gait cycle series (*Cycles*) was calculated to reflect the gait cadence of this minute from the time domain, which was denoted as *median cycle*. To assess the gait cadence from the frequency domain, the power spectral density (PSD) of walking was obtained by applying the fast Fourier transformation to the filtered magnitude ($\bar{r}_i$) of the acceleration signals of every continuous walking segment. The peak frequency (Sun et al., 2010) of the 0.5-3–Hz band (reflecting walking) (Weiss et al., 2013) of the PSD was used to reflect the main rhythm of steps from the frequency domain, which was denoted as *peak frequency*. For gait force, we calculated the median of the *Force* series (*median force*) to represent the average power of all steps in the minute.

Table 6.1. Short-term and long-term gait features extracted and their short descriptions.

Gait feature	Description
Short-term gait features	
Median cycle (seconds)	Median of gait cycles in the 1-minute walking segment
Peak frequency (Hz)	Peak frequency in the PSD[a] of the magnitude of 1-minute acceleration signals
Median force (m/s^2)	Median of gait force in the 1-minute walking segment
Long-term gait features	
25th percentile of median cycle	25th percentile of median gait cycle values of all walking segments[b]
50th percentile of median cycle	Median of median gait cycle values of all walking segments
75th percentile of median cycle	75th percentile of median gait cycle values of all walking segments
SD of median cycle	Standard deviation of median gait cycle values of all walking segments
25th percentile of peak frequency	25th percentile of peak frequency values of all walking segments
50th percentile of peak frequency	Median of peak frequency values of all walking segments
75th percentile of peak frequency	75th percentile of peak frequency values of all walking segments
SD of peak frequency	Standard deviation of peak frequency values of all walking segments
25th percentile of median force	25th percentile of median gait force values of all walking segments
50th percentile of median force	Median of median gait force values of all walking segments
75th percentile of median force	75th percentile of median gait force values of all walking segments
SD of median force	Standard deviation of median gait force values of all walking segments

[a] PSD: power spectral density (from 0.5 Hz to 3 Hz).

[b] All detected continuous walking segments (defined in the Methods section) in a feature window (3 days for the Long Term Movement Monitoring data set and 14 days for the Remote Assessment of Disease and Relapse–Major Depressive Disorder data set).

c) Long-Term Gait Features

For each of the short-term gait features (*median cycle, peak frequency,* and *median force*), we calculated four statistical second-order features (25th percentile, median, 75th percentile, and SD) from all detected continuous walking segments during a feature window.

Previous studies suggested that the extreme values of gait characteristics over the long term could reflect the optimal or worst walking performance of the participant, which could in turn reflect physical or mental conditions better than the median value (Rispens et al., 2015). Therefore, we used *25th percentile*, *median*, and *75th percentile* second-order statistics to represent three levels of walking performance (low, medium, and high) during a feature window. For example, faster walking during a feature window could represent *high-performance walking*, which may not be affected by other factors such as fatigue and the crowded environment. *High-performance walking* could be represented by the *75th percentile of peak frequency* and the *25th percentile of median cycle* in a feature window, which is expected to be closely associated with depression status. The variance of daily-life walking in a feature window was measured by the SD.

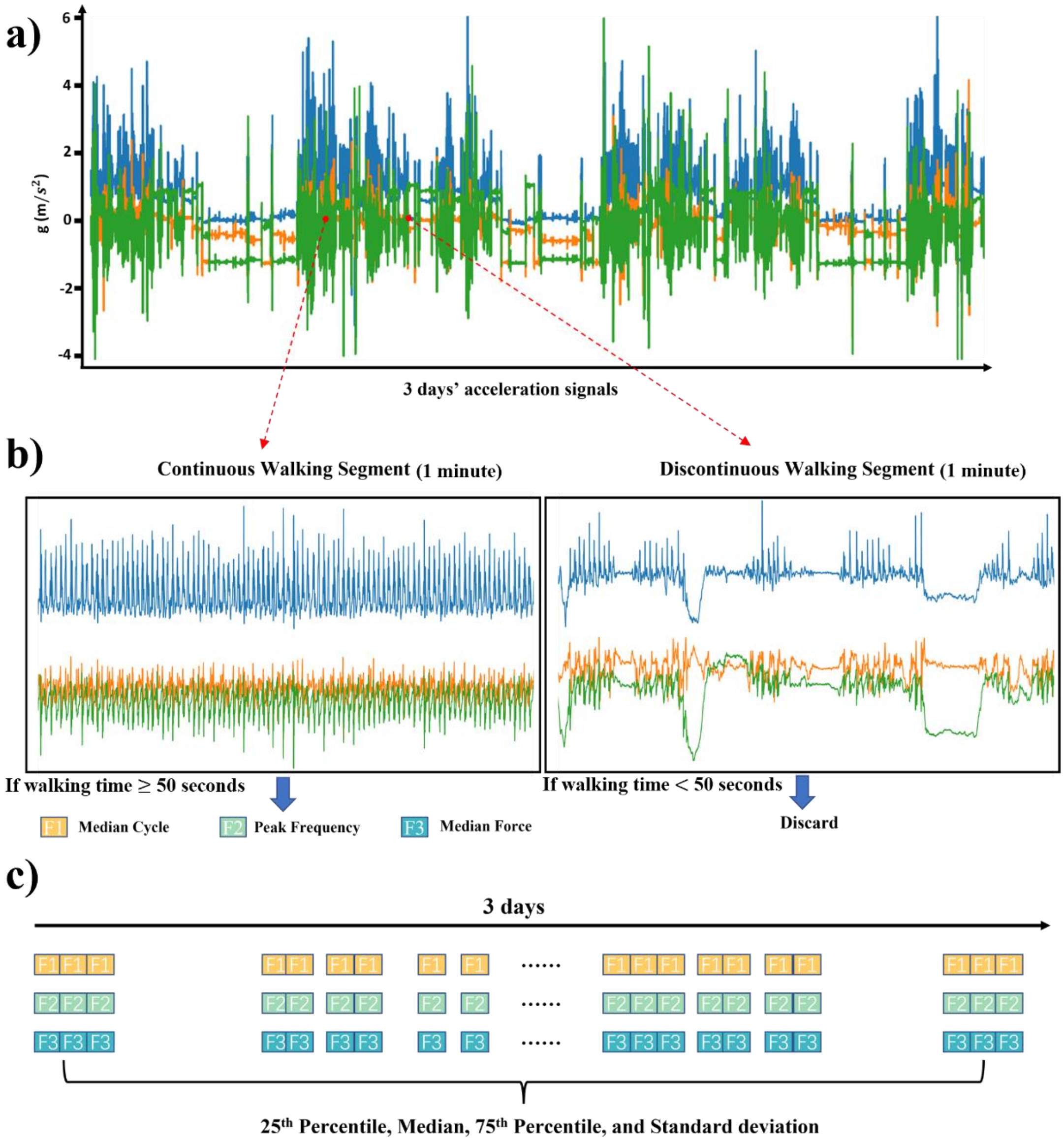
d) Laboratory Gait Features Extracted from Laboratory Walking Tests in the LTMM Data Set

We also extracted *median cycle, peak frequency,* and *median force* from the 1-minute acceleration signals of laboratory walking tests in the LTMM data set. For reading convenience, we denote these as laboratory gait features.

Inclusive Criteria for Data Missingness in the RADAR-MDD-KCL Data Set

The raw acceleration signals were remotely collected by mobile phones in the RADAR-MDD-KCL study. Possibly due to the high battery consumption and network traffic for uploading the raw signal, the missing rate of acceleration signals was relatively high. To reduce the impact of missingness, a PHQ-8 period (14 days) included in this study should have at least 3 days (aligned with the LTMM data set) with more than 50% acceleration signals (Saeb et al., 2016; Zhang et al., 2021).

Figure 6.2. A schematic diagram of long-term gait feature extraction for the Long-Term Movement Monitoring dataset. a) 3-axis acceleration signals of 3 consecutive days; b) examples of continuous and discontinuous walking segments and three short-term gait features (definitions in Table 6.1) were extracted from each continuous walking segment; c) long-term gait feature extraction: 25th percentile, median, 75th percentile, and standard deviation of short-term gait feature values of all continuous walking segments over 3 days for each participant.


a)
6
4
2
0
-2
-4
g (m/s²)
3 days' acceleration signals
b)
Continuous Walking Segment (1 minute)
Discontinuous Walking Segment (1 minute)
If walking time ≥ 50 seconds
If walking time < 50 seconds
F1 Median Cycle
F2 Peak Frequency
F3 Median Force
Discard
c)
3 days
F1
F2
F3
......
25th Percentile, Median, 75th Percentile, and Standard deviation

6.2.4 Statistical Analyses

For the LTMM data set, Spearman coefficients (Spearman, 1987) were calculated to assess associations between the GDS-15 score and gait features (3 laboratory gait features and 12 long-term gait features). As the data in the RADAR-MDD-KCL data set are longitudinal (repeated PHQ-8 measurements for each participant), a series of pairwise linear mixed-effects regression models (Laird & Ware, 1982) with random participant intercepts were performed to explore the association between the PHQ-8 score and each of the 12 long-term gait features (no laboratory tests were included in the RADAR-MDD-KCL data set). Age, gender, and the number of comorbidities (see Supplementary Table 1, Appendix B) were considered as covariates. The Benjamini-Hochberg method was used for multiple-comparison corrections in both data sets (Benjamini & Hochberg, 1995).

To test whether long-term gait features could explain additional data variance in depression scores relative to laboratory gait features, we built two nested multivariate linear regression models without and with long-term gait features for the GDS-15 score (denoted as Model A and Model B; Equations 4 and 5) in the LTMM data set. Specifically, predictors of Model A are age, gender, and the 3 laboratory gait features, while predictors of Model B are age, gender, the 3 laboratory gait features, and the 12 long-term gait features. The coefficient of determination (R^2) was calculated for both models to estimate how much data variance was explained by predictors. Then, the likelihood ratio test (Glover & Dixon, 2004) was used to test whether Model B fit the GDS-15 score better than Model A. Since the laboratory walking test was not included in the RADAR-MDD-KCL data set, the likelihood ratio test was only performed in the LTMM data set.

Model A: GDS-15=Age+Gender+3 laboratory gait features **(4)**

Model B: GDS-15=Age+Gender+3 laboratory gait features+12 long-term gait features **(5)**

6.3 Results

6.3.1 Data Summary

The 71 participants in the LTMM data set have a mean age of 78.36 (SD 4.71) years with 18 (25%) participants having potential depressive disorders (GDS-15≥5) and 69.82 (SD 9.65) hours of acceleration signals per participant. The RADAR-MDD-KCL data set, according to the data inclusion criteria, contains 659 PHQ-8 records collected from 215 participants and corresponding 99,445 hours (average 463 hours per participant). The cohort in the RADAR-MDD-KCL data set has a mean age of 43.36 (SD 15.12) years with the majority being women (75%), and half of the PHQ-8 records indicated potential depression symptoms (PHQ-8≥10). The average missing rate of acceleration signals collected by phones in the RADAR-MDD-KCL data set (70.60%) was significantly higher than that of the acceleration signals collected by the wearable device in the LTMM data set (3.03%). A summary of the demographics, and distributions of depression scores and available acceleration signals for participants in the LTMM and the RADAR-MDD-KCL data sets is shown in Table 6.2. The heatmaps of correlations between the 12 long-term gait features of the LTMM and RADAR-MDD-KCL data sets are presented in Figure 6.3.

Table 6.2. Demographics and distributions of depression scores and available acceleration signals of participants in the two data sets.

Characteristic	LTMM[a] (N=71)	RADAR-MDD-KCL[b] (N=215)
Age (years), mean (SD)	78.36 (4.71)	43.36 (15.12)
Female, n (%)	46 (65%)	162 (75%)
Depression score, mean (SD)	GDS-15[c]: 3.18 (2.81)	PHQ-8[d]: 9.67 (5.84)
Potential depressive episode (GDS-15≥5) and PHQ-8≥10), n (%)[e]	18 (25%)	330 (50%)
Number of completed depression questionnaires[f]	71	659
Number of completed depression questionnaires per participant, mean (SD)	1 (0)	3.09 (2.76)
Length of total available acceleration signals (hours)	4817	99,445
Length of available acceleration signals (hours) for each GDS-15/PHQ-8 record[g], mean (SD)	69.82 (9.65)	98.77 (105.20)
Average missing rate of acceleration signals (%)	3.03	70.60
Number of continuous walking segments[h] detected from each GDS-15/PHQ-8 record, mean (SD)	73.48 (66.98)	113.24 (170.48)

[a] LTMM: Long Term Movement Monitoring.

[b] RADAR-MDD-KCL: subset of the Remote Assessment of Disease and Relapse–Major Depressive Disorder data set collected from King's College London, United Kingdom.

[c] GDS-15: 15-item Geriatric Depression Scale.

[d] PHQ-8: 8-item Patient Health Questionnaire.

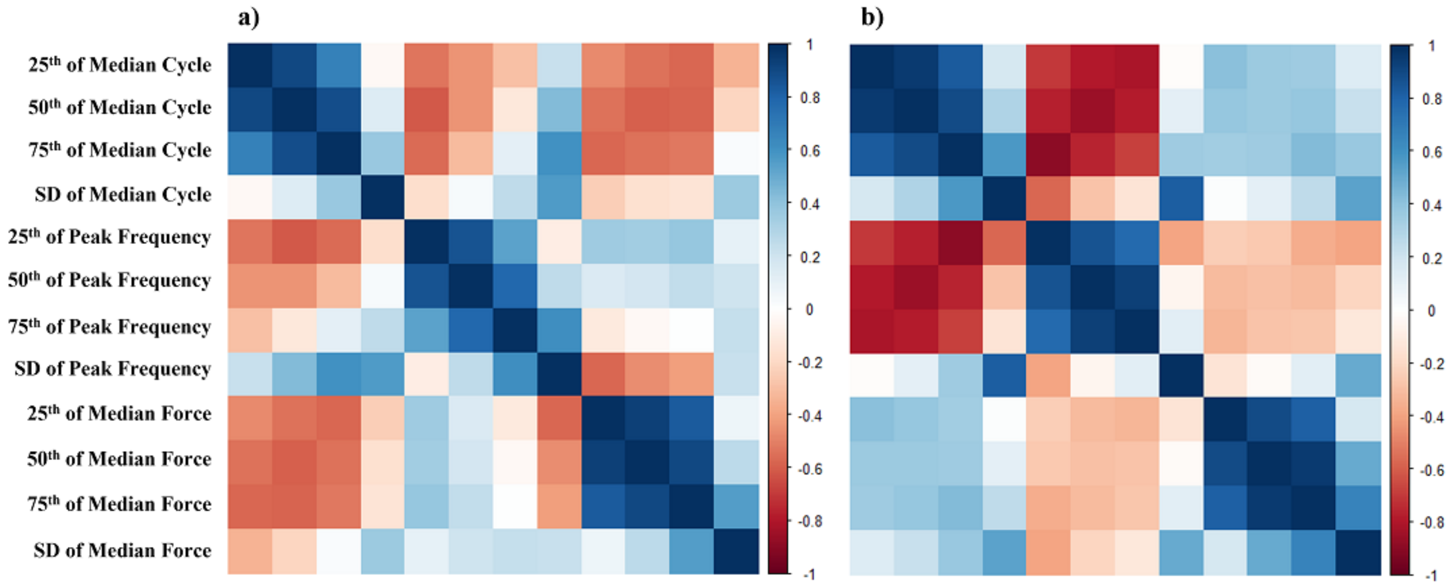
[e] Based on the total number of completed questionnaires.

[f] The RADAR-MDD-KCL data set has multiple PHQ-8 records for each participant, which was conducted every 2 weeks.

[g] We regarded acceleration signals in the 14 days before a PHQ-8 record. For the GDS-15 record, we considered acceleration signals of all 3-day activities after enrollment.

[h] Continuous walking segment was defined as 1-minute acceleration signals with at least 50 seconds of walking (see Methods section).

Figure 6.3. Heatmaps of correlations between 12 long-term gait features of the Long-term Movement Monitoring dataset (a) and RADAR-MDD-KCL dataset (b).



6.3.2 Associations Between Gait Features and the GDS-15 Score in the LTMM Data Set

The Spearman correlations between the GDS-15 score and gait features (both laboratory and long-term gait features) in the LTMM data set are shown in Table 6.3. We found that a higher GDS-15 score was significantly correlated with a larger median of gait cycles, lower peak frequency, and smaller median gait force in the 1-minute laboratory walking test. For the long-term period, a higher GDS-15 score was significantly correlated with lower variance of gait force and slower cadence of high-performance walking and 75th percentile of peak frequency during 3-day activities.

Table 6.3. Spearman correlations between the 15-item Geriatric Depression Scale score and gait features, including laboratory and long-term gait features, in the Long-Term Movement Monitoring data set.

Feature[a]	ρ	*P* value[b]
Laboratory gait features extracted from the 1-minute laboratory walking test		
Median cycle	0.39	.001
Peak frequency	–0.32	.01
Median force	–0.25	.04
Long-term gait feature extracted from 3-day activities		
25th percentile of median cycle	0.31	.01
50th percentile of median cycle	0.13	.29
75th percentile of median cycle	0.02	.86
SD of median cycle	–0.24	.06
25th percentile of peak frequency	–0.02	.85
50th percentile of peak frequency	–0.09	.45
75th percentile of peak frequency	–0.27	.03
SD of peak frequency	–0.12	.33
25th percentile of median force	0.02	.85
50th percentile of median force	–0.01	.98
75th percentile of median force	–0.10	.41
SD of median force	–0.30	.02

[a]Definitions of gait features in this table are provided in Table 6.1 and the Methods section.

[b]*P* values were adjusted by the Benjamini-Hochberg method for correction of multiple comparisons.

6.3.3 Associations Between Long-Term Gait Features and the PHQ-8 Score in the RADAR-MDD-KCL Data Set

The pairwise linear mixed-effects models performed in the RADAR-MDD-KCL data set revealed a significant and negative link between the PHQ-8 score and the gait cadence of *high-performance walking* during the 14 days before submitting PHQ-8 records. Specifically, the *25th percentile of median cycle* was positively associated with the PHQ-8 score; that is, for every increase of 0.1 seconds in the median gait cycle of *high-performance walking*, the PHQ-8 score increased by 0.606 points. Likewise, the

75th percentile of peak frequency was negatively associated with the PHQ-8 score, indicating that a reduction of 0.1 Hz in the peak frequency of *high-performance walking* was associated with an increase of 0.26 PHQ-8 points. Other long-term gait features were not found to be significantly associated with the PHQ-8 score in the RADAR-MDD-KCL data set. A summary of all 12 linear mixed-effects regression models is provided in Table 6.4.

Table 6.4. Twelve pairwise linear mixed-effects models for exploring associations between long-term gait features and depression symptom severity (8-item Patient Health Questionnaire) in the RADAR-MDD-KCL data set[a].

Long-term gait feature[b]	Estimate	SE	*df*	*t* value	*P* value[c]
25th percentile of median cycle	6.06	2.72	648.75	2.23	.03
50th percentile of median cycle	3.98	2.51	639.41	1.59	.11
75th percentile of median cycle	2.49	2.08	653.72	1.20	.23
SD of median cycle	2.87	4.41	631.11	0.65	.52
25th percentile of peak frequency	–1.50	1.02	656.44	–1.46	.15
50th percentile of peak frequency	–1.93	1.05	650.76	–1.83	.07
75th percentile of peak frequency	–2.62	1.01	634.70	–2.60	.01
SD of peak frequency	0.21	1.86	600.50	0.12	.91
25th percentile of median force	–0.57	2.24	637.46	–0.25	.80
50th percentile of median force	0.88	1.79	655.77	0.49	.62
75th percentile of median force	0.44	1.66	656.37	0.26	.79
SD of median force	2.05	3.78	602.90	0.54	.59

[a] RADAR-MDD-KCL: Subset of Remote Assessment of Disease and Relapse–Major Depressive Disorder collected from King's College London.

[b] Definitions of daily-life gait features are provided in Table 6.1 and the Methods section.

[c] *P* values were adjusted by the Benjamini-Hochberg method for correction of multiple comparisons.

6.3.4 Results of the Likelihood Ratio Test in the LTMM Data Set

The regression model with long-term gait features (Model B) achieved better performance (R^2=0.30) than the model without long-term gait features (Model A)

(R^2=0.06). We found that the 12 long-term gait features extracted from 3-day activities could explain an extra 24% data variance (an increase of 0.24 in R^2) of GDS-15 scores relative to the laboratory gait features and participants' demographics. The likelihood ratio test showed that Model B fitted GDS-15 scores significantly better than Model A (χ^2=32.91>$\chi^2_{0.05}$(12), *P*=.001). The detailed results of the two nested regression models are shown in Supplementary Table 2 Appendix B.

6.4 Discussion

6.4.1 Principal Findings

This study retrospectively used two ambulatory data sets for exploring the associations between depression symptom severity and daily-life gait characteristics. We extracted 12 long-term gait features to describe the distribution and variance of gait cadence and force over a long-term period and link daily-life gait patterns with a self-reported depression score. The main findings of this study are (1) higher depression symptom severity is significantly associated with lower gait cadence of *high-performance walking* (faster walking in all continuous walking segments) over a long-term period; (2) long-term daily-life walking has the potential to provide additional information for predicting depression symptom severity relative to laboratory gait characteristics and demographics; and (3) wearable devices and mobile phones both have potential to capture the associations between daily gait and depression.

The results of Spearman correlations between laboratory gait features and the GDS-15 score in the LTMM data set are consistent with previous studies (Brandler et al., 2012;

Lemke et al., 2000; Michalak et al., 2009, 2011; Pieruccini-Faria et al., 2018; Radovanović et al., 2014; J. B. Sanders et al., 2016; Sloman et al., 1987; van Iersel et al., 2005); that is, the participants with more severe depression symptoms were more likely to have slower gait cadence (longer median of gait cycles and lower gait frequency) and smaller gait force in laboratory walking tests.

For daily-life walking, this study used the faster walking (*75th percentile of peak frequency* and *25th percentile of median cycle*) in all detected continuous walking segments to represent *high-performance walking* during a feature window (3 days for LTMM and 14 days for RADAR-MDD-KCL). Only gait cadence of *high-performance walking* was found to be significantly and negatively associated with depression symptom severity, whereas gait patterns under *medium/low-performance walking* were not significantly associated with the depression score. This finding was consistent in both the LTMM and RADAR-MDD-KCL data sets. A potential reason is that the walking performance in real-world scenarios may be affected by multiple factors (such as walking during the day or at night, walking under fatigue or walking after rest, and walking to a destination or navigating a crowded supermarket) (Rispens et al., 2015); therefore, the lower walking performance may not fully reflect the participant's physical or mental conditions. Therefore, from the main finding of this study, we speculated that faster steps over a long-term period could represent the high performance of participants' walking, which could be closely associated with their depression status.

In the LTMM data set, we found that the variance of gait force (*SD of median force*) in

3-day activities was significantly and negatively associated with the depression symptom severity, indicating that participants with higher depression symptom severity were likely to have relatively monotonous walking over 3 days. However, the feature was not significantly associated with the PHQ-8 score in the RADAR-MDD-KCL data set. One reason is that the magnitude (r_i) (explained in the Step Detection Algorithm section) of the acceleration signals depends on the location of the accelerometers attached to the body (Derawi & Bours, 2013). As acceleration signals in the RADAR-MDD-KCL data set were collected by mobile phones, the variable locations of phones when attached to participants' bodies (such as in the hand, handbag, and pocket) affected the magnitude of acceleration signals. Therefore, the magnitude of phone-collected acceleration signals cannot fully reflect the gait force.

Results of regression models and the likelihood test in the LTMM data set illustrated the importance of monitoring daily-life gait in real-world settings. Laboratory gait features and demographics in LTMM data only explained a small proportion of data variance of the GDS-15 score (R^2=0.06), whereas long-term gait features extracted from 3-day activities could explain an extra 24% of data variance (R^2=0.30). This finding supported that long-term daily-life walking has the potential to provide additional information for predicting depression symptom severity relative to laboratory gait characteristics and demographics. Further, this finding also indicated that the laboratory walking test may be affected by several factors such as subjective psychological factors and laboratory-controlled conditions, which may not fully reflect the condition of a participant's mental health (Notthoff et al., 2018; Rispens et al., 2015).

Since there were no laboratory tests in the RADAR-MDD-KCL data set, the comparison between laboratory gait features and long-term daily-life gait features was not performed in the RADAR-MDD-KCL. We will consider adding laboratory tests at enrollment in future digital depression studies.

6.4.2 Limitations

Although we found that wearables and mobile phones have the potential to capture the associations between depression and daily-life gait patterns, both devices have some limitations. Wearables could collect relatively complete walking data; however, wearing sensors may not be suitable for long-term monitoring. Mobile phones could be used for long-term monitoring without user burden, but the missing rate of mobile phone acceleration signals is relatively high. The findings of this study support that the links between gait and depression could still be revealed from the limited and sparse daily-life walking acceleration signals. Missingness is a common challenge in remote digital studies (Onnela, 2021), which may be caused by high battery consumption, network traffic for uploading the raw acceleration signals, and the Android operating system moderation of resources. According to the findings of this study, a possible solution to reduce missingness is uploading gait cycles instead of uploading raw acceleration signals in future long-term monitoring research. This is not difficult to implement, as most current smartphones have real-time step detection functions or apps (Silva et al., 2020; Stavropoulos et al., 2020). Furthermore, the self-reported PHQ-8 data may be subject to recall bias. We may consider implementing ecological momentary assessments with passive gait data collection in future research.

The hyperparameters in step detection and feature extraction need further investigation. We considered using a 1-minute window size for step detection and 50 seconds for continuous walking segment selection based on previous studies (Ihlen et al., 2015; Weiss et al., 2013) and our experience. The feature window sizes for the two data sets are different due to the different study designs. However, the optimal hyperparameters are still unclear and will be investigated in future research.

Gait features extracted in this study were simple and statistically based, which were used to illustrate the importance of daily walking in our initial analysis. More features such as nonlinear features will be considered in future research.

Gait characteristics could be affected by some physical diseases, neurological disorders, and age (Del Din et al., 2016; Helbostad et al., 2007; Rodgers et al., 1999). Although none of the participants had any cognitive or gait/balance disorders in the LTMM data set and the number of comorbidities and age were considered as covariates in the RADAR-MDD-KCL data set, physical comorbidities and other comorbidities may have different impacts on the gait characteristics. We will consider a wider range of comorbidities and investigate them further in future research.

6.4.3 Conclusion

In summary, the findings of this study showed that significant links between depression symptom severity and daily-life gait characteristics could be captured in different data sets and by different accelerometer devices. Long-term daily-life walking patterns could provide additional value for understanding depression manifestations relative to gait patterns in laboratory walking tests, which illustrated the importance of long-term gait monitoring. The gait cadence of high-performance walking in daily life has the potential to be an indicator for monitoring depression severity, which may contribute to developing clinical tools to remotely monitor mental health in real-world settings.

Acknowledgments

The Remote Assessment of Disease and Relapse–Central Nervous System (RADAR-CNS) project has received funding from the Innovative Medicines Initiative (IMI) 2 Joint Undertaking under grant agreement number 115902. This Joint Undertaking receives support from the European Union’s Horizon 2020 Research and Innovation Program and the European Federation of Pharmaceutical Industries and Associations (EFPIA). This communication reflects the views of the RADAR-CNS consortium and neither IMI nor the European Union and EFPIA are liable for any use that may be made of the information contained herein. The funding bodies have not been involved in the design of the study, the collection or analysis of data, or the interpretation of data. This study represents independent research partly funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London, and Maudsley NHS Foundation Trust and King’s College London. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health and Social Care. We thank all the members of the RADAR-CNS patient advisory board for their contribution to the device selection procedures, and their invaluable advice throughout the study protocol design. This research was reviewed by a team with experience of mental health problems and their careers, who have been specially trained to advise on research proposals and documentation through Feasibility and Acceptability Support Team for Researchers (FAST-R), a free, confidential service in England provided by the NIHR Maudsley Biomedical Research Centre via King’s College London and South London and Maudsley NHS Foundation Trust. We thank all GLAD Study volunteers for their participation, and gratefully acknowledge the NIHR BioResource, NIHR BioResource centers, NHS Trusts and staff for their contribution. We also acknowledge NIHR BRC, King’s College London, South London and Maudsley NHS Trust and King’s Health Partners. We thank the NIHR, NHS Blood and Transplant, and Health Data Research UK as part of the Digital Innovation Hub Program.

CO is supported by the UK Medical Research Council (MR/N013700/1) and King’s College London member of the MRC Doctoral Training Partnership in Biomedical Sciences.

RJBD is supported by the following: (1) NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King’s College London, United Kingdom; (2) Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health

Agency (Northern Ireland), British Heart Foundation, and Wellcome Trust; (3) The BigData@Heart consortium, funded by the IMI-2 Joint Undertaking under grant agreement number 116074. This Joint Undertaking receives support from the European Union's Horizon 2020 Research and Innovation Program and EFPIA; it is chaired by DE Grobbee and SD Anker, partnering with 20 academic and industry partners and ESC; (4) the NIHR University College London Hospitals Biomedical Research Centre; (5) the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London; (6) the UK Research and Innovation London Medical Imaging and Artificial Intelligence Centre for Value Based Healthcare; and (7) the NIHR Applied Research Collaboration South London (NIHR ARC South London) at King's College Hospital NHS Foundation Trust.

Authors' Contributions

YZ extracted and integrated the questionnaire and raw accelerometer data for the analysis, planned and performed the analysis, and drafted the manuscript. MH and VAN gained funding and co-led the Remote Assessment of Disease and Relapse–Central Nervous System program. MH is the principal investigator for the Remote Assessment of Disease and Relapse–Major Depressive Disorder study. RJBD, AAF, YR, ZR, PC, HS, and CS have contributed to the development of the RADAR-base platform used for data collection and management across sites, data protection, security, and storage. YZ, AAF, S Sun, NC, SV, PL, LQ, DCM, MH, and RJBD contributed to the design of the study. FM, KMW, CO, AI, FL, S Siddi, S Simblett, JMH, BWJHP, MH contributed to data collection. AAF, IMG, AR, VAN, TW, PA, MH, and RJBD contributed to the administrative, technical, and clinical support of the study. All authors were involved in reviewing the manuscript, had access to the study data, and provided direction and comments on the manuscript.

Conflicts of Interest

SV and VAN are employees of Janssen Research and Development LLC. PA is employed by the pharmaceutical company H Lundbeck A/S. DCM has accepted honoraria and consulting fees from Otsuka Pharmaceuticals, Optum Behavioral Health, Centerstone Research Institute, and the One Mind Foundation; has received royalties from Oxford Press; and has an ownership interest in Adaptive Health, Inc. MH is the principal investigator of RADAR-CNS, a private public precompetitive consortium that receives funding from Janssen, UCB, Lundbeck, MSD, and Biogen. The other authors have no conflicts of interest to declare.

References

Abedi, P., Nikkhah, P., & Najar, S. (2015). Effect of pedometer-based walking on depression, anxiety and insomnia among postmenopausal women. *Climacteric: The Journal of the International Menopause Society*, *18*(6), 841–845.

Adolph, D., Tschacher, W., Niemeyer, H., & Michalak, J. (2021). Gait Patterns and Mood in Everyday Life: A Comparison Between Depressed Patients and Non-depressed Controls. *Cognitive Therapy and Research*, *45*(6), 1128–1140.

Althubaiti, A. (2016). Information bias in health research: Definition, pitfalls, and adjustment methods. *Journal of Multidisciplinary Healthcare*, 211.

Atrsaei, A., Corrà, M. F., Dadashi, F., Vila-Chã, N., Maia, L., Mariani, B., Maetzler, W., & Aminian, K. (2021). Gait speed in clinical and daily living assessments in Parkinson's disease patients: Performance versus capacity. *NPJ Parkinson's Disease*, *7*(1), 24.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A

Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 289–300.

Brandler, T. C., Wang, C., Oh-Park, M., Holtzer, R., & Verghese, J. (2012). Depressive Symptoms and Gait Dysfunction in the Elderly. *The American Journal of Geriatric Psychiatry*, *20*(5), 425–432.

D'Ath, P., Katona, P., Mullan, E., Evans, S., & Katona, C. (1994). Screening, detection and management of depression in elderly primary care attenders. I: The acceptability and performance of the 15 item Geriatric Depression Scale (GDS15) and the development of short versions. *Family Practice*, *11*(3), 260–266.

Del Din, S., Godfrey, A., Galna, B., Lord, S., & Rochester, L. (2016). Free-living gait characteristics in ageing and Parkinson's disease: Impact of environment and ambulatory bout length. *Journal of Neuroengineering and Rehabilitation*, *13*(1), 46.

Deligianni, F., Guo, Y., & Yang, G.-Z. (2019). From Emotions to Mood Disorders: A Survey on Gait Analysis Methodology. *IEEE Journal of Biomedical and Health Informatics*, *23*(6), 2302–2316.

Derawi, M., & Bours, P. (2013). Gait and activity recognition using commercial phones. *Computers & Security*, *39*, 137–144.

Devaux, M., & Sassi, F. (2016). Social disparities in hazardous alcohol use: Self-report bias may lead to incorrect estimates. *European Journal of Public Health*, *26*(1), 129–134.

Friedrich, M. J. (2017). Depression Is the Leading Cause of Disability Around the World. *JAMA*, *317*(15), 1517.

Glover, S., & Dixon, P. (2004). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, *11*(5), 791–806.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C. K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), E215-220.

Große, J., Petzold, M. B., Brand, R., & Ströhle, A. (2021). Step Away from Depression—Study protocol for a multicenter randomized clinical trial for a pedometer intervention during and after in-patient treatment of depression. *International Journal of Methods in Psychiatric Research*, *30*(1).

Harman, J. S., Edlund, M. J., & Fortney, J. C. (2004). Disparities in the Adequacy of Depression Treatment in the United States. *Psychiatric Services*, *55*(12), 1379–1385.

Hawton, K., Casañas i Comabella, C., Haw, C., & Saunders, K. (2013). Risk factors for suicide in individuals with depression: A systematic review. *Journal of Affective Disorders*, *147*(1–3), 17–28.

Helbostad, J. L., Leirfall, S., Moe-Nilssen, R., & Sletvold, O. (2007). Physical fatigue affects gait characteristics in older persons. *The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences*, *62*(9), 1010–1015.

Herzog, W., Nigg, B. M., Read, L. J., & Olsson, E. (1989). Asymmetries in ground reaction force patterns in normal human gait: *Medicine & Science in Sports & Exercise*, *21*(1), 110–114.

Ihlen, E. A. F., Weiss, A., Helbostad, J. L., & Hausdorff, J. M. (2015). The Discriminant Value of Phase-Dependent Local Dynamic Stability of Daily Life Walking in Older Adult Community-Dwelling Fallers and Nonfallers. *BioMed Research International*, *2015*, 402596.

Kamphuis, M. H., Stegenga, B. T., Zuithoff, N. P. A., King, M., Nazareth, I., de Wit, N. J., & Geerlings, M. I. (2012). Does recognition of depression in primary care affect outcome? The PREDICT-NL study. *Family Practice*, *29*(1), 16–23.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B. W., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of Affective Disorders*, *114*(1–3), 163–173.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–974.

Lemke, M. R., Wendorff, T., Mieth, B., Buhl, K., & Linnemann, M. (2000). Spatiotemporal gait patterns during over ground locomotion in major depression compared with healthy controls. *Journal of Psychiatric Research*, *34*(4–5), 277–283.

Lenox-Smith, A., Macdonald, M. T. B., Reed, C., Tylee, A., Peveler, R., Quail, D., & Wildgust, H. J. (2013). Quality of Life in Depressed Patients in UK Primary Care: The FINDER Study. *Neurology and Therapy*, *2*(1–2), 25–42.

Lerner, D., Adler, D. A., Chang, H., Berndt, E. R., Irish, J. T., Lapitsky, L., Hood, M. Y., Reed, J., & Rogers, W. H. (2004). The Clinical and Occupational Correlates of Work Productivity Loss Among Employed Patients With Depression. *Journal of Occupational & Environmental Medicine*, *46*(6), S46–S55.

Levine, D., Richards, J., Whittle, M., & Whittle, M. (Eds.). (2012). *Whittle's gait analysis* (5th ed). Churchill Livingstone/Elsevier.

Lewinsohn, P. M., Solomon, A., Seeley, J. R., & Zeiss, A. (2000). Clinical implications of "subthreshold" depressive symptoms. *Journal of Abnormal Psychology*, *109*(2), 345–351.

Marron, J. J., Labrador, M. A., Valle, A. M., Lanvin, D. F., & Rodriguez, M. G. (2016). Multi sensor system for pedestrian tracking and activity recognition in indoor environments. *International Journal of Ad Hoc and Ubiquitous Computing*, *23*(1/2), 3.

Matcham, F., Barattieri Di San Pietro, C., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., Folarin, A. A., Haro, J. M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V., Mohr, D. C., Myin-Germeys, I., Narayan, V., Bwjh, P., Ranjan, Y., Rashid, Z., Rintala, A., … Meyer, N. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry*, *19*(1), 1–11.

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K. M., Annas, P., de Girolamo, G., Difrancesco, S., Haro, J. M., Horsfall, M., Ivan, A., Lavelle, G., Li, Q., Lombardini, F., Mohr, D. C., Narayan, V. A., Oetzmann, C., Penninx, B. W. J. H., Bruce, S., … Hotopf, M. (2022). Remote Assessment of Disease and

Relapse in Major Depressive Disorder (RADAR-MDD): Recruitment, retention, and data availability in a longitudinal remote measurement study. *BMC Psychiatry*, *22*(1), 1–19.

McKercher, C. M., Schmidt, M. D., Sanderson, K. A., Patton, G. C., Dwyer, T., & Venn, A. J. (2009). Physical Activity and Depression in Young Adults. *American Journal of Preventive Medicine*, *36*(2), 161–164.

Michalak, J., Troje, N. F., Fischer, J., Vollmar, P., Heidenreich, T., & Schulte, D. (2009). Embodiment of sadness and depression—Gait patterns associated with dysphoric mood. *Psychosomatic Medicine*, *71*(5), 580–587.

Michalak, J., Troje, N., & Heidenreich, T. (2011). The effects of mindfulness-based cognitive therapy on depressive gait patterns. *Journal of Cognitive and Behavioral Psychotherapies*, *11*(1), 13–27.

Notthoff, N., Drewelies, J., Kazanecka, P., Steinhagen-Thiessen, E., Norman, K., Düzel, S., Daumer, M., Lindenberger, U., Demuth, I., & Gerstorf, D. (2018). Feeling older, walking slower—But only if someone's watching. Subjective age is associated with walking speed in the laboratory, but not in real life. *European Journal of Ageing*, *15*(4), 425–433.

Onnela, J.-P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology*, *46*(1), 45–54.

Pieruccini-Faria, F., Muir-Hunter, S. W., & Montero-Odasso, M. (2018). Do depressive symptoms affect balance in older adults with mild cognitive impairment? Results from the "gait and brain study." *Experimental Gerontology*, *108*, 106–111.

Radovanović, S., Jovičić, M., Marić, N. P., & Kostić, V. (2014). Gait characteristics in patients with major depression performing cognitive and motor tasks while walking. *Psychiatry Research*, *217*(1–2), 39–46.

Ranjan, Y., Rashid, Z., Stewart, C., Conde, P., Begale, M., Verbeeck, D., Boettcher, S., Hyve, Dobson, R., Folarin, A., & RADAR-CNS Consortium. (2019). RADAR-Base: Open Source Mobile Health Platform for Collecting, Monitoring, and Analyzing Data Using Sensors, Wearables, and Mobile Devices. *JMIR MHealth and UHealth*, *7*(8), e11734.

Rispens, S. M., van Schooten, K. S., Pijnappels, M., Daffertshofer, A., Beek, P. J., & van Dieën, J. H. (2015). Do Extreme Values of Daily-Life Gait Characteristics Provide More Information About Fall Risk Than Median Values? *JMIR Research Protocols*, *4*(1), e4.

Rodgers, M. M., Mulcare, J. A., King, D. L., Mathews, T., Gupta, S. C., & Glaser, R. M. (1999). Gait characteristics of individuals with multiple sclerosis before and after a 6-month aerobic training program. *Journal of Rehabilitation Research and Development*, *36*(3), 183–188.

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018). Correlations Between Objective Behavioral Features Collected From Mobile and Wearable Devices and Depressive Mood Symptoms in Patients With Affective Disorders: Systematic Review. *JMIR MHealth and UHealth*, *6*(8), e165.

Saeb, S., Lattie, E. G., Schueller, S. M., Kording, K. P., & Mohr, D. C. (2016). The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ*, *2016*(9), 1–15.

Sanders, J. B., Bremmer, M. A., Comijs, H. C., Deeg, D. J. H., & Beekman, A. T. F. (2016). Gait Speed and the Natural Course of Depressive Symptoms in Late Life; An Independent Association With Chronicity? *Journal of the American Medical Directors Association*, *17*(4), 331–335.

Sanders, R. D., & Gillig, P. M. (n.d.). Gait and its assessment in psychiatry. *Psychiatry (Edgmont)*, *7*(7), 38.

Schrijvers, D., Hulstijn, W., & Sabbe, B. G. C. (2008). Psychomotor symptoms in depression: A diagnostic, pathophysiological and therapeutic tool. *Journal of Affective Disorders*, *109*(1–2), 1–20.

Silva, A. G., Simões, P., Queirós, A., Rodrigues, M., & Rocha, N. P. (2020). Mobile Apps to Quantify Aspects of Physical Activity: A Systematic Review on its Reliability and Validity. *Journal of Medical Systems*, *44*(2), 51.

Sloman, L., Pierrynowski, M., Berridge, M., Tupling, S., & Flowers, J. (1987). Mood, depressive illness and gait patterns. *Canadian Journal of Psychiatry. Revue Canadienne De Psychiatrie*, *32*(3), 190–193.

Sobin, C., & Sackeim, H. A. (1997). Psychomotor symptoms of depression. *The American Journal of Psychiatry*, *154*(1), 4–17.

Spearman, C. (1987). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, *100*(3/4), 441.

Stavropoulos, T. G., Andreadis, S., Mpaltadoros, L., Nikolopoulos, S., & Kompatsiaris, I. (2020). Wearable Sensors and Smartphone Apps as Pedometers in eHealth: A Comparative Accuracy, Reliability and User Evaluation. *2020 IEEE International Conference on Human-Machine Systems (ICHMS)*, 1–6.

Sun, L.-P., Zheng, X.-D., Shou, H., Li, J.-S., & Li, Y.-D. (2010). Quantitative prediction of channel sand bodies based on seismic peak attributes in the frequency domain and its application. *Applied Geophysics*, *7*(1), 10–17.

van Iersel, M. B., Haitsma, A., Olde Rikkert, M. G. M., & Benraad, C. E. M. (2005). Quantitative gait analysis to detect gait disorders in geriatric patients with depression. *Journal of the American Geriatrics Society*, *53*(8), 1441–1442.

Wang, Y., Wang, J., Liu, X., & Zhu, T. (2021). Detecting Depression Through Gait Data: Examining the Contribution of Gait Features in Recognizing Depression. *Frontiers in Psychiatry*, *12*, 661213.

Weiss, A., Brozgol, M., Dorfman, M., Herman, T., Shema, S., Giladi, N., & Hausdorff, J. M. (2013). Does the Evaluation of Gait Quality During Daily Life Provide Insight Into Fall Risk? A Novel Approach Using 3-Day Accelerometer Recordings. *Neurorehabilitation and Neural Repair*, *27*(8), 742–752.

Williams, E. I., & Wallace, P. (1993). Health checks for people aged 75 and over. *Occasional Paper (Royal College of General Practitioners)*, *59*, 1–30.

Yesavage, J. A., & Sheikh, J. I. (1986). 9/Geriatric Depression Scale (GDS): Recent Evidence and Development of a Shorter Version. *Clinical Gerontologist*, *5*(1–2), 165–173.

Yogev-Seligmann, G., Hausdorff, J. M., & Giladi, N. (2008). The role of executive function and attention in gait: EF and Gait. *Movement Disorders*, *23*(3), 329–342.

Young, A. S., Klap, R., Sherbourne, C. D., & Wells, K. B. (2001). The quality of care for depressive and anxiety disorders in the United States. *Archives of General Psychiatry*, *58*(1), 55–61.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., Laiou, P., Matcham, F., Oetzmann, C., Lamers, F., Siddi, S., Simblett, S., Rintala, A., Mohr, D. C., Myin-Germeys, I., Wykes, T., Haro, J. M., … Dobson, R. J. B. (2021). Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: Preliminary longitudinal study. *JMIR MHealth and UHealth*, *9*(7), 1–19.

Chapter 7

Associations Between Depression Symptom Severity and the Circadian Rhythm Patterns Extracted from Passive Wearable Data

Background: Depression is known to be closely associated with the circadian rhythm of individuals. Estimating the circadian rhythm can help to monitor the evolution of depression. However, the traditional assessments of the circadian rhythm are laboratory-based and unsuitable for larger cohorts and long-term monitoring in real-world settings.

Objectives: This chapter aimed to approximate the circadian rhythm using passive wearable data and explore the associations between depression symptom severity and circadian rhythm patterns in a large, multicenter, longitudinal data set.

Methods: Participants' depression symptom severity was self-reported by the 8-item Patient Health Questionnaire (PHQ-8) via mobile phones every 2 weeks. Fitbit's heart rate, steps, and sleep data in the preceding 2 weeks of each PHQ-8 record were analyzed respectively using the Cosinor models. Then, the parameters of these Cosinor models were extracted as features to represent the average level, variance, peak hour, and strength of circadian rhythmicity of participants' behaviors. Linear mixed-effect models were used to explore associations between the PHQ-8 score and circadian rhythm features. Likewise, the seasonal impact on circadian rhythm features was also investigated by linear mixed-effect models.

Results: This study included 8090 PHQ-8 records with corresponding Fitbit data from 489 participants (median [IQR] age of 48 [31.0, 58.0] years; 378 [77.30%] females). Higher depression symptom severity was found to be significantly associated with lower circadian rhythmicity, lower level and variance of activity, lower heart rate variance, later peak hour of heart rate, and later sleep time. In agreement with the literature, the season was found to have a significant impact on the extracted circadian rhythm features.

Conclusion: This chapter indicated that the circadian rhythm patterns derived from wearable data have the potential to be indicators of depression. Findings in this chapter may provide a basis for the development of clinical applications of remote mental health monitoring in real-world settings.

7.1 Introduction

The circadian rhythm is an internal clock related to endogenous oscillations of an approximately 24-hour period, which affects and regulates the timing of nearly all human behavioral and physiological activities and has extensive associations with individuals' physical and mental health (Partch, Green, & Takahashi, 2014). Mental disorders, such as depression, are associated with disturbances in the circadian rhythm and manifest as abnormal behaviors (Walker, Walton, DeVries, & Nelson, 2020). For example, depressed people have been reported to have a larger variance in sleep and more irregular behaviors than healthy controls (Alvaro, Roberts, & Harris, 2013; Weyerer & Kupfer, 1994). Therefore, measuring individuals' circadian rhythm could be useful for tracking the progression of depression. The gold standard for estimating the circadian rhythm of individuals is tracking melatonin in blood, urine, or saliva samples in a constant light environment (Duffy & Dijk, 2002; Keijzer, Smits, Duffy, & Curfs, 2014). However, these methods are unsuitable for large cohort studies and long-term monitoring in real-world settings (Bowman et al., 2021). Thus, there is a need for easy-to-use approaches for estimating the circadian rhythm of participants in long-term cohort studies.

With the development of sensor technologies, wearable devices can passively and cost-efficiently capture individuals' daily behaviors in real-world settings (Lee, Kim, Park, & Jeon, 2021). Several past studies have attempted to approximate the circadian rhythm using some behavioral rhythms (such as Sleep-Wake rhythm, Rest-Activity rhythm, and circadian rhythm in heart rate [CRHR]) measured by passive data from wearable devices (Carr et al., 2018; Moraes et al., 2013; Robillard et al., 2015; Slyepchenko et al., 2019; Smagula et al., 2018; White, Rumble, & Benca, 2017). Several significant associations between depression severity and wearable-derived circadian rhythm patterns were found in these studies (Carr et al., 2018; Moraes et al., 2013; Robillard et al., 2015; Slyepchenko et al., 2019; Smagula et al., 2018; White et al., 2017). It was shown that disruptions in sleep-wake cycles, such as later sleep onset time, insomnia, hypersomnia, and worse sleep quality, were associated with higher depression severity (Robillard et al., 2015; Slyepchenko et al., 2019; White et al., 2017). For the Rest-Activity rhythm, higher depression severity was correlated with a later activity peak, lower activity level, and weaker circadian rhythmicity (Moraes et al., 2013; Smagula et al., 2018; White et al., 2017). Further, Carr et al found that depressed people tend to have desynchronized CRHR with Sleep-Wake and Rest-Activity rhythms compared to healthy controls (Carr et al., 2018).

However, because the majority of previous studies were cross-sectional, associations between wearable-measured circadian rhythm patterns and depression at both individual and cohort levels have not been fully explored. Further, the seasonal effects (Adamsson, Laike, & Morita, 2017) on circadian rhythm were not considered in previous studies, perhaps due to their short study durations. To address these limitations, this chapter explored associations between depression symptom severity and wearable-derived circadian rhythm patterns using a large, multicenter, longitudinal digital depression dataset. Sleep-Wake rhythm, Rest-Active rhythm, and CRHR were estimated from Fitbit's sleep, step, and heart rate data, respectively. Multilevel analysis was used to investigate the patterns of data at both between-participant and within-participant levels, with the season and demographics as confounding variables.

7.2 Methods

7.2.1 Data Set

Participants and Settings

The data used in this chapter came from an EU digital depression study, Remote Assessment of Disease and Relapse-Major Depressive Disorder (RADAR-MDD), which remotely monitored over 600 participants' daily activities for up to 2 years (Matcham et al., 2019). More details about participant retention and data availability of the RADAR-MDD data set are reported in (Matcham et al., 2022).

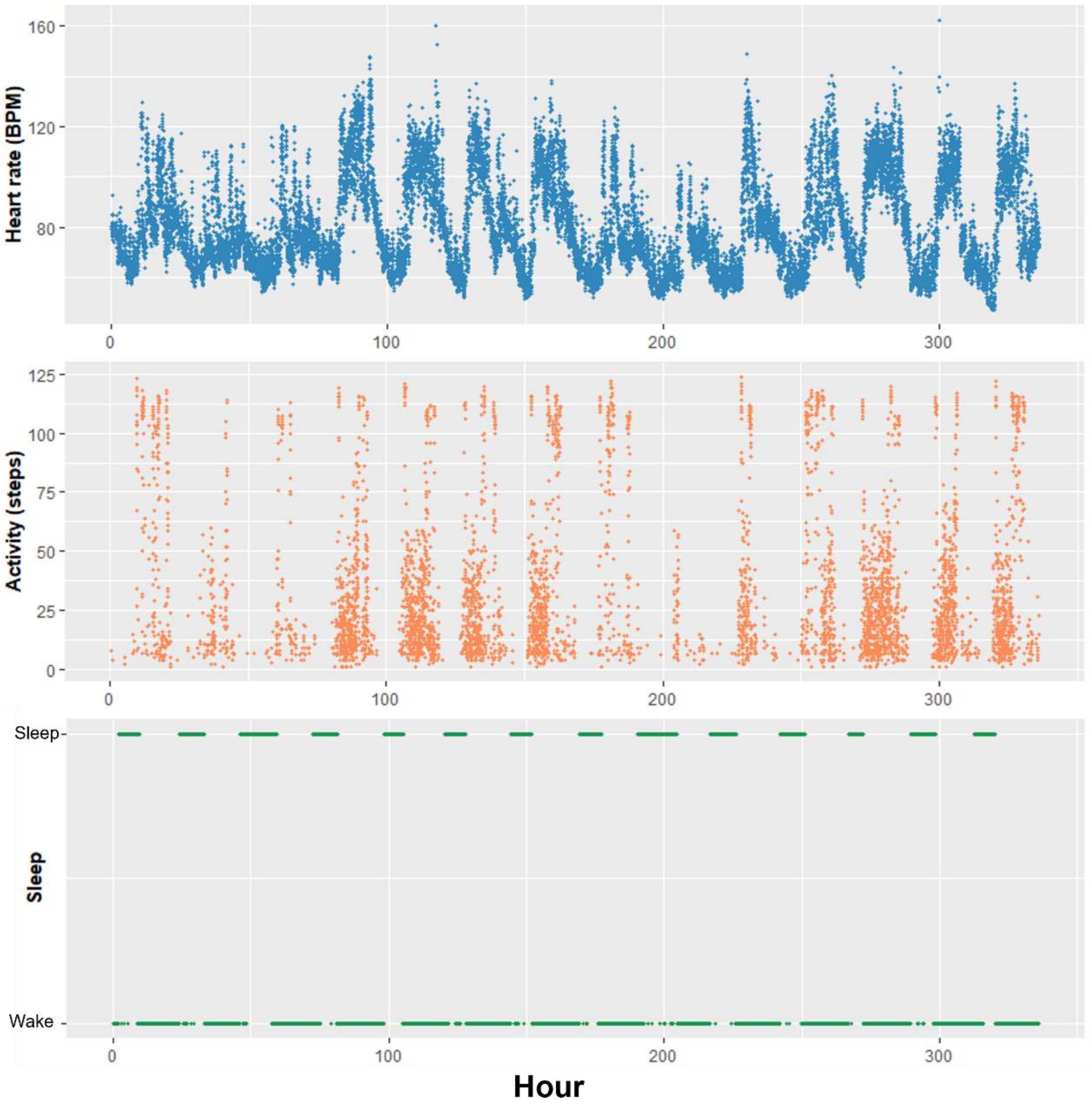
Depression Symptom Severity

The participant's depression symptom severity was self-reported by the 8-item Patient Health Questionnaire (PHQ-8) (Kroenke et al., 2009) every two weeks via mobile phone. The total score of the PHQ-8 ranges from 0 to 24, with increasing severity of depression symptom severity (Kroenke et al., 2009).

Fitbit Data

Participants' sleep, steps, and heart rates (HR) were continuously monitored by Fitbit devices. **HR data:** Fitbit provides an estimate of HR every 5 seconds based on the embedded photoplethysmography sensor. However, due to signal quality and other technical factors, some sample points lack heart rate data. Therefore, to obtain robust heart rate trends, the average heart rate value for every minute was calculated. **Step data:** Fitbit continuously measures the number of steps taken in one minute. **Sleep data:** Fitbit provides a sleep label (“awake”, “light sleep”, “deep sleep”, and “rapid eye movement [REM]”) every 30 seconds based on the integrated algorithm (Bian et al., 2017). Although Fitbit has limited accuracy in discriminating different sleep phases (light, deep, and REM sleep), Fitbit shows promise in identifying sleep-wake status (de Zambotti, Goldstone, Claudatos, Colrain, & Baker, 2018; Liang & Chapa-Martell, 2019). Therefore, a binary time series was used to represent sleep-wake status, where 0 indicates awake and 1 represents sleep (light, deep, and REM sleep). For the time period without sleep labels but having HR data, it was labeled as 0 (awake); otherwise, it is regarded as missing data. Figure 7.1 shows an example of a participant’s processed Fitbit data during the preceding 14 days of a PHQ-8 record.

Figure 7.1. An example of a participant's processed HR, step, and sleep Fitbit data during the preceding 14 days of a PHQ-8 record.



7.2.2 Data Inclusion Criteria

The time interval of 14 days before each of the PHQ-8 records was regarded as a "PHQ-8 interval". To reduce the impact of missing data, PHQ-8 intervals included in the present analysis should have at least 70% of HR and step data and 10 days of sleep recordings.

7.2.3 Circadian Rhythm Feature Extraction

There were several mathematic methods were developed to evaluate the circadian rhythm from passive behavioral signals, such as spectrum approaches and Cosinor-based models (Refinetti, Cornélissen, & Halberg, 2007). Compared with spectrum techniques (e.g., Fourier method (Duhamel & Vetterli, 1990)), the Cosinor-based method is more practical for estimating characteristics (such as mesor, amplitude, and acrophase) of circadian rhythms (Refinetti et al., 2007) and applicable to missing data (Cornelissen, 2014). Since the aim of this chapter is to explain the association between the circadian rhythm and depression severity, it is important to extract more explanatory characteristics of the circadian rhythm. Therefore, the Cosinor model is adopted for the subsequent analyses.

The Cosinor model makes the following assumptions: 1) the individuals' behavioral data follow a Cosinor function with a period of 24 hours; and 2) the parameters do not change during the estimated time interval (Cornelissen, 2014). In this chapter, three Cosinor models were fitted with the processed HR, step, and sleep data of each PHQ-8 interval, respectively, using the "cosinor" R package.

The parameters of the Cosinor models were regarded as circadian rhythm features to describe the participant's behavioral patterns in a PHQ-8 interval. These parameters include 1) mesor (the average value of the fitted behavioral curve), 2) amplitude (the difference between the maximum value and the mean value of the cosine wave, providing a measure of the variance of rhythm), 3) acrophase (a time index indicating the time when the modeled behavioral rhythm reaches its peak), and 4) R^2 (a goodness-of-fit measure of the Cosinor model, providing a measure of the strength of the circadian rhythmicity). These four parameters were extracted from both HR and step data and were denoted as *HR_Mesor, HR_Amplitude, HR_Acrophase, HR_R2, Step_Mesor, Step_Amplitude, Step_Acrophase,* and *Step_R2,* respectively. Since sleep status was represented by a binary variable, only the acrophase and R^2 of the fitted Cosinor model were extracted and denoted as *Sleep_Acrophase* and *Sleep_R2,* respectively. In total, 10 circadian rhythm features were extracted from each PHQ-8 interval.

7.2.4 Statistical Analysis

The linear mixed-effects regression model was used (Singer, Willett, & Willett, 2003) to explore associations between depression symptom severity and circadian rhythm features at both between-participant and within-participant levels (Singer et al., 2003). A two-level linear mixed-effects regression model with a random participant-specific intercept was built to regress the PHQ-8 score with each circadian rhythm feature. At the between-participant level, age, gender, and season (summer and winter) were considered as covariates to adjust the individual differences. Note that in this chapter, summer (April to October) and winter were distinguished based on daylight saving time in the EU.

To further investigate the seasonal impact on the circadian rhythm, I also constructed a series of linear mixed-effect regression models considering each of the circadian rhythm features as the outcome variable and the season as the predictor variable. Likewise, age and gender were also regarded as confounding variables at the between-participant level.

The Benjamini-Hochberg approach was applied to all models for multiple comparison corrections (Benjamini & Hochberg, 1995). Data preprocessing and analyses were performed in R software, and linear mixed-effect models were implemented using the "lmerTest" R package. The significant level was set to adjusted *P* value < .05.

7.3 Results

7.3.1 Data Summary

According to our data inclusion criteria, 8090 PHQ-8 intervals from 489 participants were selected in this study. The median (IQR) age of the cohort is 48 (31.0, 58.0) years, and the majority (77.30%; N=378) are females. The median score (IQR) of all selected PHQ-8 records is 9.0 (5.0, 14.0) and the median (IQR) number of PHQ-8 intervals per participant is 15.0 (5.0, 25.0).

7.3.2 Associations Between Circadian Rhythm Features and Depressive Symptom Severity

Table 7.1 summarizes the results of 10 linear mixed-effects models, each of which explored the association between the PHQ-8 score and one of the circadian rhythm features. Except for *HR_Mesor* and *Step_Acrophase*, the remaining 8 circadian rhythm features were significantly associated with the PHQ-8 score.

For both *Step_Mesor* and *Step_Amplitude*, strong negative associations with the PHQ-8 score were observed. For example, an increase of 10 steps in step mesor was associated with a reduction of 3.8 points in the PHQ-8 score (Beta = -0.38 t =-14.09, *P* <.001). Similarly, there was a decrease of 2.4 PHQ-8 points for every increase of 10 steps in step amplitude (Beta = -0.24, t =-11.24, *P* <.001). The amplitude of the Cosinor model of HR data (*HR_Amplitude*) was also negatively associated with depression symptom severity (Beta = -0.20, t =-8.83, *P* <.001), that is, with a decrease of 2.0 points in the PHQ-8 score for every 10 BPM (beats per minute) rise in the HR amplitude.

The R^2 coefficients of three Cosinor models (HR, step, and sleep) were all significantly and negatively associated with the PHQ-8 score. Among them, the R^2 of the step Cosinor model had the greatest impact on the severity of depressive symptoms, with a decrease of 1.57 PHQ-8 scores for every 0.1 increase in R^2 (Beta = -15.72, t =-8.14, *P* <.001). The R^2 of the sleep and HR models had relatively small impacts on PHQ-8 scores, for every 0.1 increase in R^2, the PHQ-8 score increased by around 0.27 points (Sleep: Beta = -2.70, t =-5.57, *P* <.001; HR: Beta = -2.73, t =-4.71, *P* <.001).

Further, positive but relatively weak associations between the PHQ-8 score and the acrophase of sleep and HR models (Sleep: Beta = 0.16, t =3.86, *P* <.001; HR: Beta = 0.1, t =2.91, *P* =.004) were found. For every hour late in the *Sleep_Acrophase* and *HR_Acrophase*, the PHQ-8 score increased by 0.1 and 0.16 points, respectively.

Table 7.1. A summary of results of 10 linear mixed-effect models, each of which explored the association between the PHQ-8 score and one of circadian rhythm features considering age, gender, and season as covariates at the between-participant level.

Feature[a]	Coeff	SE	t value	*P* value[b]
HR_Mesor	-0.01	0.01	-1.00	.32
HR_Amplitude	-0.20	0.02	-8.83	<.001
HR_Acrophase	0.10	0.03	2.91	.004
HR_R2	-2.73	0.58	-4.71	<.001
Step_Mesor	-0.38	0.03	-14.09	<.001
Step_Amplitude	-0.24	0.02	-11.24	<.001
Step_Acrophase	0.02	0.04	0.44	.66
Step_R2	-15.72	1.93	-8.14	<.001
Sleep_Acrophase	0.16	0.04	3.86	<.001
Sleep_R2	-2.70	0.49	-5.57	<.001

[a] Definitions of circadian rhythm features are explained in the Method section.

[b] All *P* values were adjusted by the Benjamini-Hochberg method for multiple comparisons.

7.3.3 The Seasonal Impact on Circadian Rhythm Features

The results of linear mixed-effect models for exploring the seasonal impact on circadian rhythm features are summarized in Table 7.2. A considerable seasonal impact on the acrophase of Cosinor models was observed. Specifically, compared with winter, summer has approximately 1 hour later acrophase of the HR and sleep models (HR: Beta = 1.05, *P* < .001; Sleep: Beta = 0.96, *P* < .001), and about 20 minutes later acrophase of the step model (Beta = 0.37, *P* <.001). The boxplots of the acrophase of HR, step, and sleep models for every month and season (winter and summer) are respectively shown in Figure 7.2. Further, the seasonal impacts on several other circadian rhythm features were significant but relatively small. Notably, gender also has significant effects on the acrophase of all three Cosinor models. Compared with females, males have 44.4, 28.8, and 19.2 minutes later acrophase of HR, step, and sleep models, respectively.

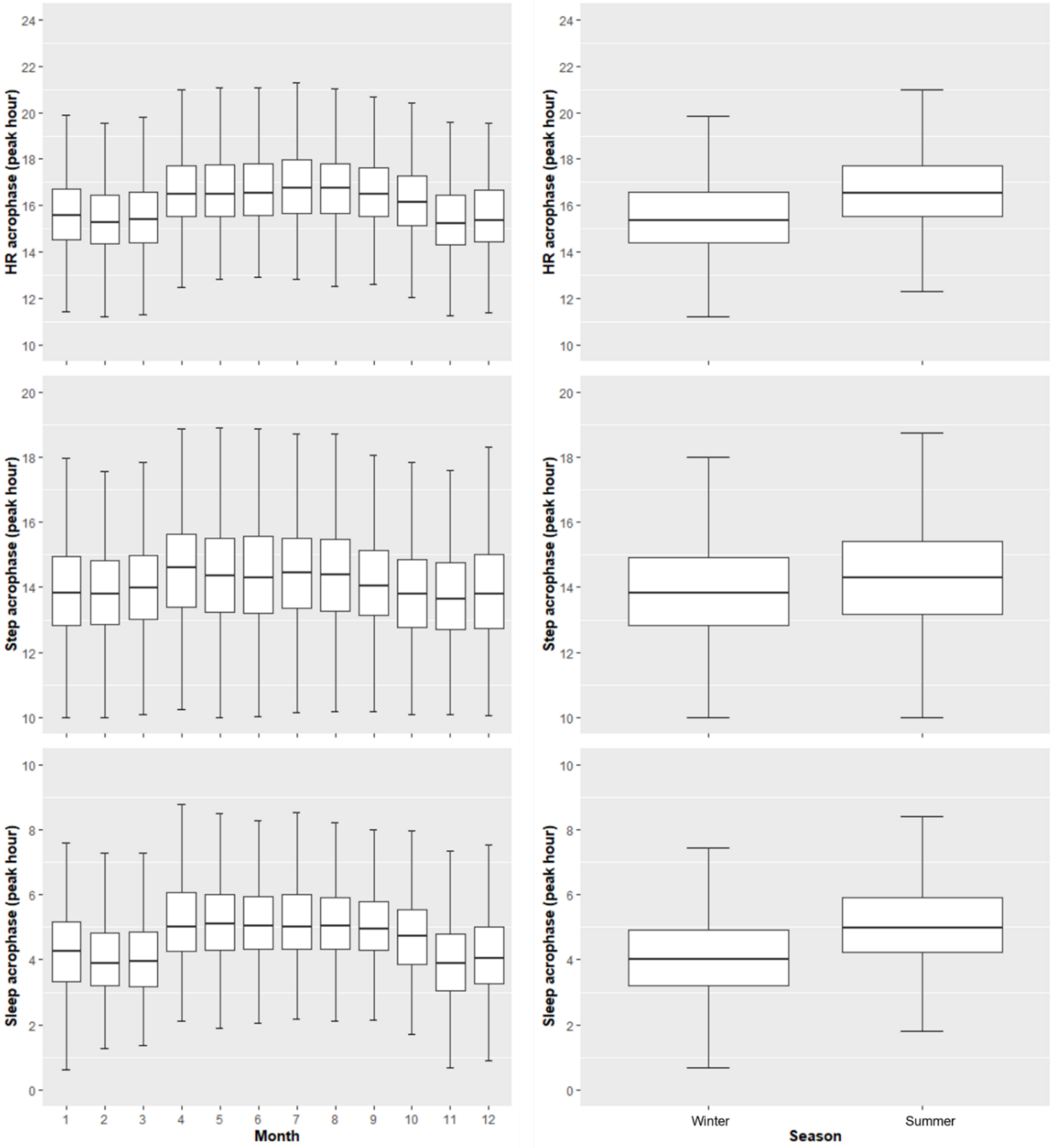
Table 7.2. A summary of results for 10 linear mixed-effect models, each of which explored the seasonal impact on one of the circadian rhythm features with considering age and gender as covariates.

Feature[a]	Summer	P value[b]	Age	P value	Male	*P* value
HR_Mesor	-0.17	.01	-0.05	.05	-2.21	.01
HR_Amplitude	0.54	<.001	-0.03	<.001	0.81	.01
HR_Acrophase	1.05	<.001	-0.03	<.001	0.74	<.001
HR_R2	0.03	<.001	0.001	.06	0.01	.46
Step_Mesor	0.16	<.001	-0.003	.67	0.43	.10
Step_Amplitude	-0.01	.89	0.01	.36	0.01	.98
Step_Acrophase	0.37	<.001	-0.04	<.001	0.48	<.001
Step_R2	0.001	.19	0.001	<.001	-0.01	.05
Sleep_Acrophase	0.96	<.001	-0.01	<.001	0.32	.03
Sleep_R2	0.003	.11	0.001	.92	-0.04	<.001

[a] Definitions of circadian rhythm features are explained in the Method.

[b] All *P* values were adjusted by the Benjamini-Hochberg method for multiple comparisons.

Figure 7.2. The boxplots of the acrophase of HR, step, and sleep Cosinor models for every month and season (winter and summer).



7.4 Discussion

7.4.1 Principal Findings

This study approximated individuals' circadian rhythm using passive Fitbit data and explored associations between depression symptom severity and circadian rhythm patterns. In our large, multicenter, longitudinal data set, I found that higher depression symptom severity was associated with the following two-week behavioral patterns: 1) weaker circadian rhythmicity, 2) later sleep time, 3) later peak hour of HR, 4) lower HR variation, and 5) lower average level and variance in movement. Furthermore, in agreement with the literature, the season was found to have a significant impact on the extracted circadian rhythm features.

The R^2 of the Cosinor model was used to measure the circadian rhythmicity of participants' behaviors. The greater the goodness-of-fit (R^2), the better the Cosinor model can explain the passive Fitbit data, and the more regular the 24-hour periodicity of the participants' behaviors. The negative link between circadian rhythmicity and depression symptom severity found in this study may be explained by the fact that participants with high PHQ-8 scores may have some depressive symptoms, such as insomnia, hypersomnia, low motivation, and decreased organization, which may result in irregular behaviors. Saeb et al also found a similar link between 24-hour movement and depression severity using GPS data (Saeb et al., 2015). In our previous studies, we found that the circadian rhythmicity in Bluetooth and location data was significantly and negatively associated with depression symptom severity using a frequency-domain measure (Zhang et al., 2021b; Zhang et al., 2022).

The mesor and amplitude of Cosinor models were used to represent the average level and variance of behaviors (HR and steps). The average level and variance of movements are both found to have substantial and negative associations with depression symptom severity, which is expected and consistent with previous research (McKercher et al., 2009; Saeb et al., 2015; Weyerer & Kupfer, 1994). In previous survey-based research, depressed people were found to be more sedentary than healthy controls (Weyerer & Kupfer, 1994). This finding has also been demonstrated in several digital depression studies using passive mobile data, such as steps and GPS data (McKercher et al., 2009; Saeb et al., 2015).

Heart rate can be influenced by many individual and environmental factors, such as age (Antelmi et al., 2004), cardiovascular disease (Fox et al., 2007), autonomic nervous system (Agelink, Boz, Ullrich, & Andrich, 2002), and environmental temperature (Schnell et al., 2013). Therefore, this may explain why the average heart rate level was not significantly correlated with depression symptom severity in our diverse cohort. I found that the amplitude of the HR model is significantly and negatively correlated with depression symptom severity. The range of heart rate fluctuations may be affected by activity level (Mølgaard, Sørensen, & Bjerregaard, 1991), and sleep (Viola et al., 2002). A larger range in heart rate fluctuations indicates a higher level of activity and more deep sleep time (Mølgaard et al., 1991; Viola et al., 2002), both of which are correlated with reduced depression severity.

The peak hour of the highest level of daily behavior was represented using the acrophase of the Cosinor model. I found that later sleep time was associated with higher depression symptom severity, which is consistent with previous studies (Robillard et al., 2015; White et al., 2017; Zhang et al., 2021a). However, past research revealed that depressed people have a later peak hour of movement than healthy controls (Smagula et al., 2018; White et al., 2017), which is inconsistent with the results in this chapter. A potential reason is that the seasonal impacts on circadian rhythms are large in high-latitude regions (Adamsson et al., 2017). In our EU data set, I found that the peak hours of individuals' behaviors during summer are much later than those during winter (Table 7.2). In places of high latitude, extreme differences in the amount of daily light exposure between summer and winter may influence the variation of individuals' melatonin and cortisol concentrations (Adamsson et al., 2017).

7.4.2 Limitations

The causal relationships between depression and wearable-derived circadian rhythm were not investigated in this initial study. Previous longitudinal studies have reported some bidirectional relationships between depression and the circadian rhythm over time (Maglione et al., 2014; Smagula et al., 2015). In future research, we will investigate if these causal relationships could be captured by wearable devices in our data set. Also, as nearly half of our data were gathered during the COVID-19 pandemic, COVID-19 and some of its restrictions (such as national lockdown, keeping social distance, and the encouragement of working from home) may have influenced participants' behaviors (particularly activities) (Sun et al., 2020). Future data collection will be required to investigate the generalizability of the findings in this chapter. Furthermore, the division of seasons in this study is based on daylight saving time (summertime) in the EU. However, other factors, such as temperature and light exposure, can also affect the circadian rhythm (Adamsson et al., 2017). Therefore, the seasonal impact on circadian rhythm needs further investigation in future research. Circadian rhythm features extracted in this chapter are only based on the Cosinor model. I will also use other approaches, such as Fourier methods, to measure the circadian rhythms and compare the results in future research.

7.4.3 Conclusions

This study found significant and negative links between depression symptom severity and circadian rhythmicity derived from passive wearable data. The higher depression symptom severity was also found to be associated with some other extracted circadian rhythm patterns, including lower movement level and variability, lower HR variance, and later sleep time. Furthermore, I found the season has a considerable impact on the patterns of circadian rhythm, which needs further investigation. This chapter indicated that wearable-derived circadian rhythm patterns have the potential to be indicators of depression, which may help the development of clinical applications for remote mental health monitoring in the future.

References

Adamsson, M., Laike, T., & Morita, T. (2017). Annual variation in daily light exposure and circadian change of melatonin and cortisol concentrations at a northern latitude with large seasonal differences in photoperiod length. *Journal of physiological anthropology, 36*(1), 1-15.

Agelink, M. W., Boz, C., Ullrich, H., & Andrich, J. (2002). Relationship between major depression and heart rate variability.: Clinical consequences and implications for antidepressive treatment. *Psychiatry Research, 113*(1-2), 139-149.

Alvaro, P. K., Roberts, R. M., & Harris, J. K. (2013). A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. *Sleep, 36*(7), 1059-1068.

Antelmi, I., De Paula, R. S., Shinzato, A. R., Peres, C. A., Mansur, A. J., & Grupi, C. J. (2004). Influence of age, gender, body mass index, and functional capacity on heart rate variability in a cohort of subjects without heart disease. *The American journal of cardiology, 93*(3), 381-385.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological), 57*(1), 289-300.

Bian, J., Guo, Y., Xie, M., Parish, A. E., Wardlaw, I., Brown, R., . . . Perry, T. T. (2017). Exploring the association between self-reported asthma impact and Fitbit-derived sleep quality and physical activity measures in adolescents. *JMIR mHealth and uHealth, 5*(7), e7346.

Bowman, C., Huang, Y., Walch, O. J., Fang, Y., Frank, E., Tyler, J., . . . Sen, S. (2021). A method for characterizing daily physiology from widely used wearables. *Cell reports methods, 1*(4), 100058.

Carr, O., Saunders, K. E., Bilderbeck, A. C., Tsanas, A., Palmius, N., Geddes, J. R., . . . Goodwin, G. M. (2018). Desynchronization of diurnal rhythms in bipolar disorder and borderline personality disorder. *Translational psychiatry, 8*(1), 1-9.

Cornelissen, G. (2014). Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling, 11*(1), 1-24.

de Zambotti, M., Goldstone, A., Claudatos, S., Colrain, I. M., & Baker, F. C. (2018). A validation study of Fitbit Charge 2™ compared with polysomnography in adults. *Chronobiology international, 35*(4), 465-476.

Duffy, J. F., & Dijk, D.-J. (2002). Getting through to circadian oscillators: why use constant routines? *Journal of biological rhythms, 17*(1), 4-13.

Duhamel, P., & Vetterli, M. (1990). Fast Fourier transforms: a tutorial review and a state of the art. *Signal processing, 19*(4), 259-299.

Fox, K., Borer, J. S., Camm, A. J., Danchin, N., Ferrari, R., Lopez Sendon, J. L., . . . Tendera, M. (2007). Resting heart rate in cardiovascular disease. *Journal of the American College of Cardiology, 50*(9), 823-830.

Keijzer, H., Smits, M. G., Duffy, J. F., & Curfs, L. M. (2014). Why the dim light melatonin onset (DLMO) should be measured before treatment of patients with circadian rhythm sleep disorders. *Sleep medicine reviews, 18*(4), 333-339.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., & Mokdad, A. H. (2009). The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders, 114*(1-3), 163-173.

Lee, S., Kim, H., Park, M. J., & Jeon, H. J. (2021). Current advances in wearable devices and their sensors in patients with depression. *Frontiers in Psychiatry, 12*, 672347.

Liang, Z., & Chapa-Martell, M. A. (2019). Accuracy of Fitbit wristbands in measuring sleep stage transitions and the effect of user-specific factors. *JMIR mHealth and uHealth, 7*(6), e13384.

Maglione, J. E., Ancoli-Israel, S., Peters, K. W., Paudel, M. L., Yaffe, K., Ensrud, K. E., & Stone, K. L. (2014). Subjective and objective sleep disturbance and longitudinal risk of depression in a cohort of older women. *Sleep, 37*(7), 1-9.

Matcham, F., Barattieri di San Pietro, C., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., . . . Lamers, F. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): a multi-centre prospective cohort study protocol. *BMC psychiatry, 19*(1), 1-11.

Matcham, F., Leightley, D., Siddi, S., Lamers, F., White, K. M., Annas, P., . . . Horsfall, M. (2022). Remote Assessment of Disease and Relapse in Major Depressive Disorder (RADAR-MDD): recruitment, retention, and data availability in a longitudinal remote measurement study. *BMC psychiatry, 22*(1), 1-19.

McKercher, C. M., Schmidt, M. D., Sanderson, K. A., Patton, G. C., Dwyer, T., & Venn, A. J. (2009). Physical activity and depression in young adults. *American journal of preventive medicine, 36*(2), 161-164.

Mølgaard, H., Sørensen, K. E., & Bjerregaard, P. (1991). Circadian variation and influence of risk factors on heart rate variability in healthy subjects. *The American journal of cardiology, 68*(8), 777-784.

Moraes, C. Á., Cambras, T., Diez-Noguera, A., Schimitt, R., Dantas, G., Levandovski, R., & Hidalgo, M. P. (2013). A new chronobiological approach to discriminate between acute and chronic depression using peripheral temperature, rest-activity, and light exposure parameters. *BMC psychiatry, 13*(1), 1-10.

Partch, C. L., Green, C. B., & Takahashi, J. S. (2014). Molecular architecture of the mammalian circadian clock. *Trends in cell biology, 24*(2), 90-99.

Refinetti, R., Cornélissen, G., & Halberg, F. (2007). Procedures for numerical analysis of circadian rhythms. *Biological rhythm research, 38*(4), 275-325.

Robillard, R., Hermens, D. F., Naismith, S. L., White, D., Rogers, N. L., Ip, T. K., . . . Smith, K. L. (2015). Ambulatory sleep-wake patterns and variability in young people with emerging mental disorders. *Journal of Psychiatry and Neuroscience, 40*(1), 28-37.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research, 17*(7), e4273.

Schnell, I., Potchter, O., Epstein, Y., Yaakov, Y., Hermesh, H., Brenner, S., & Tirosh, E. (2013). The effects of exposure to environmental factors on Heart Rate Variability: An ecological perspective. *Environmental Pollution, 183*, 7-13.

Singer, J. D., Willett, J. B., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*: Oxford university press.

Slyepchenko, A., Allega, O. R., Leng, X., Minuzzi, L., Eltayebani, M. M., Skelly, M., . . . Frey, B. N. (2019). Association of functioning and quality of life with objective and subjective measures of sleep and biological rhythms in major depressive and bipolar disorder. *Australian & New Zealand Journal of Psychiatry, 53*(7), 683-696.

Smagula, S. F., Ancoli-Israel, S., Blackwell, T., Boudreau, R., Stefanick, M. L., Paudel, M. L., . . . Group, O. F. i. M. R. (2015). Circadian rest–activity rhythms predict future increases in depressive symptoms among community-dwelling older men. *The American Journal of Geriatric Psychiatry, 23*(5), 495-505.

Smagula, S. F., Krafty, R. T., Thayer, J. F., Buysse, D. J., & Hall, M. H. (2018). Rest-activity rhythm profiles associated with manic-hypomanic and depressive symptoms. *Journal of psychiatric research, 102*, 238-244.

Sun, S., Folarin, A. A., Ranjan, Y., Rashid, Z., Conde, P., Stewart, C., . . . Simblett, S.

(2020). Using smartphones and wearable devices to monitor behavioral changes during COVID-19. *Journal of medical Internet research, 22*(9), e19992.

Viola, A. U., Simon, C., Ehrhart, J., Geny, B., Piquard, F., Muzet, A., & Brandenberger, G. (2002). Sleep processes exert a predominant influence on the 24-h profile of heart rate variability. *Journal of biological rhythms, 17*(6), 539-547.

Walker, W. H., Walton, J. C., DeVries, A. C., & Nelson, R. J. (2020). Circadian rhythm disruption and mental health. *Translational psychiatry, 10*(1), 1-13.

Weyerer, S., & Kupfer, B. (1994). Physical exercise and psychological health. *Sports Medicine, 17*(2), 108-116.

White, K. H., Rumble, M. E., & Benca, R. M. (2017). Sex differences in the relationship between depressive symptoms and actigraphic assessments of sleep and rest-activity rhythms in a population-based sample. *Psychosomatic medicine, 79*(4), 479.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Bendayan, R., Ranjan, Y., . . . Dobson, R. J. B. (2021a). Relationship between major depression symptom severity and sleep collected using a wristband wearable device: multicenter longitudinal observational study. *JMIR mHealth and uHealth, 9*(4), e24604.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Ranjan, Y., Rashid, Z., . . . Dobson, R. J. B. (2021b). Predicting depressive symptom severity through individuals' nearby bluetooth device count data collected by mobile phones: preliminary longitudinal study. *JMIR mHealth and uHealth, 9*(7), e29840.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Vairavan, S., Bendayan, R., . . . Dobson, R. J. B. (2022). Longitudinal Relationships Between Depressive Symptom Severity and Phone-Measured Mobility: Dynamic Structural Equation Modeling Study. *JMIR mental health, 9*(3), e34898.

Chapter 8

Discussion and Future Work

8.1 Summary of Key Findings and Contributions

Chapter 2 performed a novel investigation into long-term participant retention and engagement from a large European multinational remote digital study for depression, the RADAR-MDD study. A significantly higher participant retention rate was found in the RADAR-MDD study than in previous remote digital health studies. Several real-world factors, including sociodemographics, usage of study apps, and depression severity, were shown to be associated with participant retention and engagement patterns in the remote digital health study. Our key findings were as follows: 1) Participants with higher depression severity at the time of enrollment contributed fewer data both actively and passively. 2) Participants with shorter survey response/completion times tend to complete more surveys and keep wearing the Fitbit device for a significantly longer period. 3) We found that older participants contributed more data and had a lower risk of disengaging from the study app. 4) Passive data gathered from wearables without additional participant burden showed greater data contiguity and participant retention than active survey data over the long term. We found a considerable proportion of the participants who completed fewer surveys continued to share passive Fitbit data for significantly longer. Together, these data-driven findings could help improve the design of future remote digital health research studies to enable equitable and balanced health data collection from diverse target populations.

Using the RADAR-MDD data set, **Chapters 3–7** of this thesis sequentially extracted features of sleep, sociability, mobility, daily gait, and circadian rhythm from passive data streams gathered from mobile phones and Fitbit devices and analyzed relationships between these behavioral features and depression symptom severity. The main findings and contributions of these chapters are summarized below in terms of feature engineering and associations with depression symptom severity.

Feature engineering. Several novel behavioral features were proposed in this thesis. Based on sleep-related knowledge, **Chapter 3** proposed several Fitbit features to characterize participants' sleep in the following 5 aspects: sleep architecture, sleep stability, sleep quality, insomnia, and hypersomnia. This thesis also attempted to adapt several widely used features from other research fields (e.g., signal processing) to digital depression research. Specifically, **Chapter 4** utilized multiscale entropy and frequency-domain features to measure the complexity and periodicity of sociability via Bluetooth data. Likewise, **Chapter 5** leveraged frequency-domain features to estimate the periodicity of movement through location data. Finally, this thesis also adjusted several existing features to make them more suitable for long-term monitoring. For instance, previous digital studies with short study periods only estimated one home location for calculating the Homestay feature, whereas **Chapter 5** estimated multiple home locations for the Homestay feature for several complex conditions (e.g., moving house, traveling, and business trips) in the long-term follow-up. These novel behavioral features can aid in a more accurate description of individuals’ behaviors in real-world settings and improve the performance of depression prediction models.

Associations with depression symptom severity. A number of significant associations between behavioral features and depression symptom severity were observed in these chapters. Specifically, **Chapter 3** found that higher depression symptom severity was associated with more awakenings during sleep, higher sleep variation, worse sleep efficiency, and more days of insomnia or hypersomnia. In **Chapter 4**, we observed that as participants' depression symptoms worsen, their Bluetooth device count data becomes lower, more monotonous, and less regular, indicating abnormalities in their social activities. **Chapter 6** revealed that higher depression severity was associated with slower gait cadence during high-performance walking. **Chapter 7** demonstrated that higher depression severity was associated with weaker circadian rhythmicity approximated from passive behavioral data of Fitbit devices. Notably, **in Chapter 5,** we found not only the negative linkage between mobility and depression (association), but also that several mobility features (Homestay, Location Entropy, and Residential Location Count) might influence the subsequent changes in depression symptom severity (longitudinal association or cross-lagged effects).

The contributions of these findings are outlined as follows. First, some previously discovered associations between depression and specific behaviors (sleep, circadian rhythms, and mobility) were examined and reaffirmed in a large multicenter digital dataset, which indicated the generalizability and robustness of these associations. Second, we also found some significant associations of depression with additional data streams that had not been explored in previous work (Bluetooth and daily walking data), illustrating the need to gather these data streams. Third, the analysis of the cross-lagged effects indicated that some behavioral features have the potential to predict the severity of depression in advance, which may aid in preventing depression relapse or deterioration. Together, these findings highlighted the importance of passive long-term monitoring of individuals' behaviors in naturalistic settings for depression research and provided a basis for future clinical applications in remote digital health research.

8.2 Future Work

8.2.1 Suggestions for Data Collection in Future Digital Health Studies

Participant Engagement

We found several real-world characteristics, especially depression severity, had substantial effects on participant participation in a digital depression study. Non-uniform participant engagement in study apps may introduce bias into the real-world data collection, impacting the generalizability and robustness of findings. There is an urgent need for further research to understand the underlying causes that influence participant engagement. Several techniques and strategies, including co-developing research apps, reducing the user burden, collecting passive data effectively, and dynamic motivation based on real-time engagement analysis, may be utilized in future digital health studies to increase participant engagement.

Operating System and Brand of Mobile Devices

The RADAR-MDD study is based on the Android operating system only. The impact of the types of smartphone operating systems on passive data collection is unclear. Also, we observed in Chapter 2 that the phone brand has a considerable influence on passive data collection. Therefore, additional research is needed to investigate intra-device/brand differences within and across Android and iOS phones to enable the robust and equitable collection of passive data.

Private Passive Data Collection

The strategies for private passive data collection need to be improved in future digital research. Several passive data streams related to private information were obfuscated or not recorded in the RADAR-MDD study. For example, the MAC address and types of Bluetooth devices were not recorded in the study, and the raw locations were obfuscated by adding a unique and random reference location. Several contextual details, such as the types of locations and owners of Bluetooth devices, were lost. Future research may leverage other techniques to overcome these limitations, such as using the hashed MAC address for Bluetooth data and the Places API of Google Maps (*Google Maps Platform*, 2022) for collecting location context.

Passive Data Resolution

The resolution of some passive data streams may need to be adjusted. Bluetooth count data was gathered on an hourly basis in the RADAR-MDD study. However, previous research suggests scanning nearby Bluetooth devices every five minutes to acquire sufficient temporal resolution for capturing dynamic changes. Therefore, I recommend increasing the resolution of Bluetooth data collection. Regarding the gait data, we found a high missing rate of phone accelerometry data, which may be the result of the high battery/network consumption of uploading the raw acceleration signals (50Hz). Based on our findings in Chapter 6, one of the potential options for future gait monitoring is to record gait features (e.g., gait cycles) provided by phone functions or passive apps instead of gathering raw signals.

8.2.2 Analysis Plans for Future Research

Feature Engineering

Depression is characterized by a variety of symptoms and manifestations (American Psychiatric Association, 2013; Lewinsohn, 1975). The extraction of features from a range of passive behavioral data streams can better characterize the behaviors of individuals and understand the associations of these behaviors with depression. This thesis did not conduct an in-depth investigation into several data streams, such as phone usage, app usage, battery consumption, and weather information. Some patterns in these data streams were reported to be associated with depression severity by previous research, but they were not examined using large data sets (David, Roberts, & Christenson, 2018; Doryab, Min, Wiese, Zimmerman, & Hong, 2014; Rohani, Faurholt-Jepsen, Kessing, & Bardram, 2018; Saeb et al., 2015). Therefore, we planned to design and extract features from these data streams in the RADAR-MDD data set in future research. In addition, we will combine several data streams and extract additional features. For example, features of heart rate during walking or sleeping can be computed using multiple passive data streams from Fitbit. Additionally, Bluetooth and location data streams can be combined to produce a more accurate approximation of sociability.

Some features extracted in this thesis and corresponding data streams need more future investigations. For example, regarding the behaviors in answering remote surveys, there is a need to extract more features (such as time spent on each subitem) and explore their relationships with depression severity in future research. Additionally, although the Fitbit data streams of heart rate and steps were used to estimate the circadian rhythms in Chapter 7, additional features (e.g., heart rate variability and step count) can be extracted in the future to correlate heart rate/activity with depression directly.

Furthermore, the selection of the window size for feature extraction needs further discussion. In this thesis, one of the strategies for describing individuals' behaviors over time is to first extract behavioral features for each day and then extract second-order statistics (e.g., mean and SD) for each of these daily features over a period (e.g., 14 days) preceding a depression assessment (e.g., PHQ-8). However, the optimal feature window size is still unclear. In future studies, we will attempt smaller slots (e.g., morning, afternoon, evening, and night) for each day and different lengths of time windows before depression assessments in future research.

Missing Data, Data Quality, and Data Imputation

Missing data has always been a challenge in mobile health research (Onnela, 2021). This thesis applied the threshold method for data inclusion criteria. Specifically, if the integrity of passive data of a certain day falls below a specified threshold (50% used in this thesis), the behavioral features for that day will not be calculated. However, the optimal threshold is still unclear, and different missing rates may have different impacts on feature values. For example, according to a threshold of 50%, data completeness of 60% and 90% both exceed the threshold; however, their effects on the feature values may vary. Therefore, further sensitivity analysis for threshold selection is required in future research.

For machine and deep learning, missing data must be imputed before being inputted into models. There are several data imputation techniques, including linear interpolation, spline interpolation, the Gaussian Process, and machine learning-based methods (Hasan et al., 2021). Finding adequate imputation algorithms for missing data in longitudinal smartphone and wearable data gathered from the real world is a valuable and challenging topic in future research.

Subgroup Analysis

Depression manifests differently in various people. Similarly, behaviors are also affected by various characteristics of people, such as age, gender, physical condition, and employment status, in real-world settings. Consequently, several previous research demonstrated that depression had varying correlations with behaviors across genders and age groups (Doryab et al., 2014; Pratap et al., 2020). As the RADAR-MDD study is an observational cohort study with an open enrollment strategy, performing association analysis directly on the entire diverse population may obscure or distort some connections. Therefore, it may be helpful to identify subgroups of participants with similar patterns in behavioral trajectories and then perform subgroup analysis to understand various depressive manifestations and their correlations with behaviors. Common techniques for trajectory clustering include K-means, hierarchical clustering, and the hidden Markov model (Casolla, Cuomo, Di Cola, & Piccialli, 2019; Morris & Trivedi, 2011). Using K-means clustering models, we have identified three long-term participant engagement patterns in Chapter 2 based on the completeness of the gathered data. We will perform appropriate clustering methods on passive behavioral data for future subgroup analysis.

Synthetic and external control groups for the RADAR-MDD study

The RADAR-MDD study is an observational cohort study, that is, all participants who met the eligibility criteria were recruited without stratifying or randomizing participants' sociodemographic characteristics (Matcham et al., 2019). With this recruitment strategy, the RADAR-MDD study successfully recruited a large cohort of participants. However, the absence of a control group limits some data analysis. If we divide the cohort into depressed and asymptomatic groups based on scores of depression questionnaires, the sociodemographic characteristics of these two groups will be imbalanced. The group comparison would be biased by confounding variables. Fortunately, some statistical methods, such as propensity matching scores (Caliendo & Kopeinig, 2008), can create a synthetic control group (Ko et al., 2021; Thorlund, Dron, Park, & Mills, 2020). In future research, we will apply this methodology to the RADAR-MDD cohort to generate the depressed and asymptomatic groups with similar distributions of sociodemographics.

Furthermore, we also have access to some other large observational cohort data sets, e.g., Covid Collab (Stewart et al., 2021) and GLAD (Davies et al., 2019) data sets, with similar data collection settings (Fitbit data and depression questionnaires). We are planning to use the propensity matching scores to create external synthetic healthy control groups for the RADAR-MDD cohort.

Free Speech Analysis

In the RADAR-MDD study, a subset of participants was asked to do some speech tasks every two weeks (Dineley et al., 2021; Matcham et al., 2019). One of these tasks was a free-speech activity in which participants were asked to talk about their expectations and plans for the next week (Dineley et al., 2021). We plan to apply an open-source automatic speech recognition system, Whisper (Radford et al., 2022), to transfer these unscripted speech records to text and then leverage some NLP approaches (e.g., sentiment analysis and topic models (Stappen et al., 2021)) to explore the sentiment and topics reflected in the free speech task and their links with depression.

Mixed-effects Machine Learning Models

The linear mixed-effects regression model is the most utilized statistical model in this thesis. Since the depression assessments and corresponding passive data used in this thesis are longitudinal (i.e., repeated measurements for each participant), the linear mixed-effects regression model is an appropriate method to estimate data patterns at both individual and cohort levels. However, linear models can only explore linear connections, which are insufficient for real-world circumstances. Therefore, some past studies leveraged machine learning models to predict the severity of depression, considering nonlinear correlations (De Angel et al., 2022). However, conventional machine learning models assume training data are independently and identically distributed (i.i.d. assumption). Nevertheless, this assumption is violated in longitudinal studies where a high degree of correlation is exhibited at the individual level. Consequently, ignoring the data's underlying correlations may lead to mediocre model performance and misleading findings (Hajjem, Bellavance, & Larocque, 2010; Sela & Simonoff, 2012). To address this limitation, GPBoost, a mixed-effects machine learning model, combining boosting tree models with mixed effects, was developed recently (Sigrist, 2020). This innovative algorithm performed better than conventional machine learning techniques (e.g., Random Forest and XGBoost) in some recent applications of longitudinal research, including driver fatigue prediction (Zhou et al., 2022), COVID-19 disease severity prediction (Sokhansanj & Rosen, 2022), and colon cancer analysis (Levy et al., 2021). We plan to leverage this algorithm to predict depression symptom severity in our longitudinal data set in the future.

Deep Learning Models

Shallow machine learning methods (e.g., Support Vector Machine, Random Forest, and XGBoost) significantly rely on time-consuming and domain-specific manual feature engineering (Janiesch, Zschech, & Heinrich, 2021). Although these hand-crafted features can assist in explaining model results, some data patterns may not be fully explored due to human bias (Janiesch et al., 2021). Deep learning models with deeply nested network architectures have the capability to automatically extract discriminative features with minimal human effort (Janiesch et al., 2021). In particular, there are many excellent time-series deep learning structures, including LSTM (Graves, 2012), DeepGlo (Sen et al., 2019), and LSTNet (Lai et al., 2018), that can learn the temporal patterns in time series data. Several past studies (Espino-Salinas et al., 2022; Jacobson & Bhattacharya, 2022) have leveraged deep learning methods for assessing mental health status using passive monitoring data and achieved satisfactory results. However, due to the limited sample size and study length, the generalizability and robustness of their models need to be validated in large longitudinal data sets. In future research, we will develop suitable deep learning frameworks for evaluating depression severity using smartphone and wearable data in the RADAR-MDD data set.

8.3 Final Remarks

This thesis demonstrated that the associations between depression and the behaviors of individuals can be captured by mobile phones and wearable devices in real-world settings. The behavioral characteristics derived from passive data streams have the potential to predict the severity of depressive symptoms. The findings of this thesis provide the basis for developing future clinical tools for remote monitoring of mental health status and trajectory with minimal user burden. Future research will require the collaborative efforts of participants, clinicians, software engineers, and data scientists to overcome challenges including data quality, participant retention, and individual differences as discussed above.

References

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys, 22*(1), 31-72.

Casolla, G., Cuomo, S., Di Cola, V. S., & Piccialli, F. (2019). Exploring unsupervised learning techniques for the Internet of Things. *IEEE Transactions on Industrial Informatics, 16*(4), 2621-2628.

David, M. E., Roberts, J. A., & Christenson, B. (2018). Too much of a good thing: Investigating the association between actual smartphone use and individual well-being. *International Journal of Human–Computer Interaction, 34*(3), 265-275.

Davies, M. R., Kalsi, G., Armour, C., Jones, I. R., McIntosh, A. M., Smith, D. J., Walters, J. T. R., Bradley, J. R., Kingston, N., Ashford, S., Beange, I., Brailean, A., Cleare, A. J., Coleman, J. R. I., Curtis, C. J., Curzons, S. C. B., Davis, K. A. S., Dowey, L. R. C., Gault, V. A., … Breen, G. (2019). The Genetic Links to Anxiety and Depression (GLAD) Study: Online recruitment into the largest recontactable study of depression and anxiety. *Behaviour Research and Therapy, 123*, 103503.

De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., . . . Mohr, D. C. (2022). Digital health tools for the passive monitoring of depression: a systematic review of methods. *NPJ Digital Medicine, 5*(1), 1-14.

Dineley, J., Lavelle, G., Leightley, D., Matcham, F., Siddi, S., Peñarrubia-María, M. T., White, K. M., Ivan, A., Oetzmann, C., Simblett, S., Dawe-Lane, E., Bruce, S., Stahl, D., Ranjan, Y., Rashid, Z., Conde, P., Folarin, A. A., Haro, J. M., Wykes, T., … The RADAR-CNS Consortium, -. (2021). Remote Smartphone-Based Speech Collection: Acceptance and Barriers in Individuals with Major Depressive Disorder. *Interspeech 2021*, 631–635.

Doryab, A., Min, J. K., Wiese, J., Zimmerman, J., & Hong, J. (2014). *Detection of behavior change in people with depression.* Paper presented at the Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence.

Espino-Salinas, C. H., Galván-Tejada, C. E., Luna-García, H., Gamboa-Rosales, H., Celaya-Padilla, J. M., Zanella-Calzada, L. A., & Tejada, J. I. G. (2022). Two-Dimensional Convolutional Neural Network for Depression Episodes Detection in Real Time Using Motor Activity Time Series of Depresjon Dataset. *Bioengineering, 9*(9), 458.

Google Maps Platform. (2022). https://developers.google.com/maps/documentation/places/web-service

Graves, A. (2012). Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, 37-45.

Hajjem, A., Bellavance, F., & Larocque, D. (2010). Generalized mixed effects regression trees. *Mixed Effects Trees and Forests for Clustered Data, 34*.

Hasan, M. K., Alam, M. A., Roy, S., Dutta, A., Jawad, M. T., & Das, S. (2021). Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked, 27*, 100799.

Jacobson, N. C., & Bhattacharya, S. (2022). Digital biomarkers of anxiety disorder symptom changes: Personalized deep learning models using smartphone sensors accurately predict anxiety symptoms from ecological momentary assessments. *Behaviour research and therapy, 149*, 104013.

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets, 31*(3), 685-695.

Ko, Y.-A., Chen, Z., Liu, C., Hu, Y., Quyyumi, A. A., Waller, L. A., . . . Martin, G. S. (2021). Developing a synthetic control group using electronic health records: Application to a single-arm lifestyle intervention study. *Preventive medicine reports, 24*, 101572.

Lai, G., Chang, W.-C., Yang, Y., & Liu, H. (2018). Modeling long-and short-term temporal patterns with deep neural networks. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 95–104.

Levy, J. J., Bobak, C. A., Nasir-Moin, M., Veziroglu, E. M., Palisoul, S. M., Barney, R. E., . . . Vaickus, L. J. (2021). *Mixed Effects Machine Learning Models for Colon Cancer Metastasis Prediction using Spatially Localized Immuno-Oncology Markers.* Paper presented at the PACIFIC SYMPOSIUM ON BIOCOMPUTING 2022.

Lewinsohn, P. M. (1975). The behavioral study and treatment of depression *Progress in behavior modification* (Vol. 1, pp. 19-64): Elsevier.

Matcham, F., Barattieri Di San Pietro, C., Bulgari, V., De Girolamo, G., Dobson, R., Eriksson, H., Folarin, A. A., Haro, J. M., Kerz, M., Lamers, F., Li, Q., Manyakov, N. V., Mohr, D. C., Myin-Germeys, I., Narayan, V., Bwjh, P., Ranjan, Y., Rashid, Z., Rintala, A., … Meyer, N. (2019). Remote assessment of disease and relapse in major depressive disorder (RADAR-MDD): A multi-centre prospective cohort study protocol. *BMC Psychiatry, 19*(1), 1–11.

Medsker, L. R., & Jain, L. (2001). Recurrent neural networks. *Design and Applications, 5*, 64-67.

Morris, B. T., & Trivedi, M. M. (2011). Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE transactions on pattern analysis and machine intelligence, 33*(11), 2287-2301.

Onnela, J.-P. (2021). Opportunities and challenges in the collection and analysis of digital phenotyping data. *Neuropsychopharmacology, 46*(1), 45-54.

Pratap, A., Neto, E. C., Snyder, P., Stepnowsky, C., Elhadad, N., Grant, D., . . . Wilbanks, J. (2020). Indicators of retention in remote digital health studies: a cross-study evaluation of 100,000 participants. *NPJ digital medicine, 3*(1), 1-10.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision. *Technical report, OpenAI, 2022*.

Rohani, D. A., Faurholt-Jepsen, M., Kessing, L. V., & Bardram, J. E. (2018). Correlations between objective behavioral features collected from mobile and wearable devices and depressive mood symptoms in patients with affective disorders: systematic review. *JMIR mHealth and uHealth, 6*(8), e9691.

Saeb, S., Zhang, M., Karr, C. J., Schueller, S. M., Corden, M. E., Kording, K. P., & Mohr, D. C. (2015). Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research, 17*(7), e4273.

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: a data mining approach for longitudinal and clustered data. *Machine learning, 86*(2), 169-207.

Sen, R., Yu, H.-F., & Dhillon, I. S. (2019). Think globally, act locally: A deep neural network approach to high-dimensional time series forecasting. *Advances in Neural Information Processing Systems, 32*.

Sigrist, F. (2020). Gaussian process boosting. *arXiv preprint arXiv:2004.02653*.

Sokhansanj, B. A., & Rosen, G. L. (2022). Predicting COVID-19 disease severity from SARS-CoV-2 spike protein sequence by mixed effects machine learning. *Computers in biology and medicine*, 105969.

Stappen, L., Baird, A., Cambria, E., & Schuller, B. W. (2021). Sentiment Analysis and Topic Recognition in Video Transcriptions. *IEEE Intelligent Systems, 36*(2), 88–95.

Stewart, C., Ranjan, Y., Conde, P., Rashid, Z., Sankesara, H., Bai, X., . . . Folarin, A. A. (2021). Investigating the Use of Digital Health Technology to Monitor COVID-19 and Its Effects: Protocol for an Observational Study (Covid Collab Study). *JMIR research protocols, 10*(12), e32587.

Thorlund, K., Dron, L., Park, J. J., & Mills, E. J. (2020). Synthetic and external controls in clinical trials–a primer for researchers. *Clinical epidemiology, 12*, 457.

Zhou, F., Alsaid, A., Blommer, M., Curry, R., Swaminathan, R., Kochhar, D., . . . Tijerina, L. (2022). Predicting Driver Fatigue in Monotonous Automated Driving with Explanation using GPBoost and SHAP. *International Journal of Human–Computer Interaction, 38*(8), 719-729.

Appendix A

Supplementary Material to Chapter 2

Supplementary Table 1. A summary of characteristics of 313 participants with a longer observation period (94 weeks) in the RADAR-MDD study, with comparisons across the three study sites.

Characteristics	Total	KCL	CIBER	VUmc	P value
Number of participants, n	313	206	91	16	
Age (median [IQR])	51.00 [37.00, 59.00]	47.00 [32.00, 58.00]	54.00 [47.50, 60.00]	39.50 [33.50, 52.25]	<.001
Female, n (%)	235 (75.1)	155 (75.2)	67 (73.6)	13 (81.3)	.81
Marital status, n (%)					.06
Single/separated/divorced/widowed	155 (49.5)	109 (52.9)	36 (39.6)	10 (62.5)	
Married/cohabiting/LTR	158 (50.5)	97 (47.1)	55 (60.4)	6 (37.5)	
Ethnicity, n (%)					<.001
White	188 (84.7)	174 (84.5)	-	14 (87.5)	
Black	9 (4.1)	8 (3.9)	-	1 (6.2)	
Asian	9 (4.1)	9 (4.4)	-	0 (0)	
Other	16 (7.2)	15 (7.3)	-	1 (6.2)	
Employed, n (%)	135 (43.1)	106 (51.5)	19 (20.9)	10 (62.5)	<.001
Having children, n (%)	173 (55.3)	97 (47.1)	70 (76.9)	6 (37.5)	<.001
Years in education (median [IQR])	15.00 [12.00, 18.00]	17.00 [14.00, 19.00]	11.00 [9.00, 14.00]	17.00 [15.00, 21.50]	<.001
Annual income, n (%)					.003
<15,000 (£/€)	86 (27.5)	53 (25.7)	32 (35.2)	1 (6.3)	
15,000-55,000 (£/€)	181 (57.8)	116 (56.3)	53 (58.2)	12 (75.0)	
>55000 (£/€)	44 (14.1)	36 (17.5)	6 (6.6)	2 (12.5)	
Accommodation, n (%)					.04
Own outright/with mortgage	177 (56.6)	105 (51.0)	63 (69.2)	9 (56.3)	
Renting	110 (35.1)	82 (39.8)	21 (23.1)	7 (43.8)	
Living rent-free	23 (7.4)	18 (8.7)	5 (5.5)	0 (0)	
Baseline PHQ-8 score (median [IQR])	11.00 [7.00, 16.00]	8.00 [5.25, 13.00]	15.00 [10.00, 18.00]	9.00 [7.00, 10.00]	<.001
Having comorbidities, n (%)	176 (56.2)	113 (54.9)	60 (65.9)	3 (18.8)	.002
Taking depression medication, n (%)	208 (66.5)	116 (56.3)	84 (92.3)	8 (50.0)	<.001
Number of contact logs (median [IQR])	9.00 [6.00, 14.00]	12.00 [8.00, 16.00]	5.00 [3.00, 8.00]	3.50 [2.75, 5.25]	<.001
Smartphone brand, n (%)					<.001
Motorola	136 (43.7)	109 (53.2)	22 (24.4)	5 (31.2)	
Samsung	85 (27.3)	54 (26.3)	23 (25.6)	8 (50.0)	
Other	90 (28.9)	42 (20.5)	45 (50.0)	3 (18.8)	

Supplementary Table 2. Proportional hazards assumption tests (using the scaled Schoenfeld residuals) for 3 Cox Proportional-Hazards models of Phone-Active, Phone-Passive, and Fitbit-Passive data streams during the first 43 weeks of the RADAR-MDD study, respectively. All 3 Cox models passed the global proportional hazards assumption tests.

Predictor	Phone-Active		Phone-Passive		Fitbit-Passive	
	χ^2	P value	χ^2	P value	χ^2	P value
Age	0.39	0.98	5.34	0.25	6.23	0.18
Gender	0.63	0.43	5.37	0.02	0.65	0.42
Marital status	1.82	0.18	2.19	0.14	0.78	0.38
Employment	1.86	0.17	1.08	0.30	1.72	0.19
Having children	0.11	0.74	0.13	0.72	1.34	0.25
Years in education	2.69	0.10	0.15	0.70	0.86	0.35
Annual income	1.43	0.49	1.40	0.50	0.66	0.72
Accommodation	2.03	0.36	2.76	0.25	2.37	0.31
Baseline PHQ-8 score	0.23	0.63	0.76	0.38	0.36	0.55
Having comorbidities	1.74	0.19	0.04	0.85	0.04	0.84
Taking depression medication	0.72	0.40	0.68	0.41	0.10	0.75
Study Site	0.50	0.78	3.23	0.20	4.42	0.11
Brand of smartphone	0.46	0.79	4.16	0.13	0.61	0.74
GLOBAL	12.80	0.89	30.11	0.07	21.90	0.35

Supplementary Table 3. Proportional hazards assumption tests (using the scaled Schoenfeld residuals) for 3 Cox Proportional-Hazards models of Phone-Active, Phone-Passive, and Fitbit-Passive data streams during the first 94 weeks of the RADAR-MDD study, respectively. All 3 Cox models passed the global proportional hazards assumption tests.

Predictor	Phone-Active		Phone-Passive		Fitbit-Passive	
	χ^2	P value	χ^2	P value	χ^2	P value
Age	9.90	0.04	1.93	0.75	6.69	0.15
Gender	1.58	0.21	2.26	0.13	2.86	0.09
Marital status	0.07	0.79	0.01	0.91	0.05	0.82
Employment	1.39	0.24	0.02	0.89	3.69	0.06
Having children	2.25	0.13	1.26	0.26	5.14	0.02
Years in education	0.39	0.54	0.75	0.39	1.58	0.21
Annual income	0.13	0.94	0.32	0.85	2.15	0.34
Accommodation	0.88	0.65	1.55	0.46	2.45	0.29
Baseline PHQ-8 score	0.62	0.43	1.28	0.26	0.33	0.56
Having comorbidities	2.38	0.12	0.10	0.76	1.48	0.22
Taking depression medication	2.83	0.09	0.67	0.41	0.02	0.88
Study Site	1.64	0.44	0.88	0.65	0.04	0.98
Brand of smartphone	0.73	0.69	0.01	0.99	2.90	0.23
GLOBAL	25.18	0.20	13.70	0.85	25.52	0.18

Supplementary Table 4. Summary participants' characteristics across three distinct engagement subgroups of Phone-Active data for the first 43 weeks of the RADAR-MDD study. The median and interquartile range (IQR) of continuous variables and the count number and percentage of category variables are reported in the table.

Characteristics	C1	C2	C3	P value
Number of participants, n	231	179	204	
Age	53.00 [34.00, 61.50]	45.00 [31.00, 55.50]	48.00 [32.00, 57.25]	.003
Male, n (%)	52 (22.5)	42 (23.5)	55 (27.0)	.53
Number of biweekly surveys	20.00 [18.00, 21.00]	13.00 [11.00, 15.00]	4.00 [1.00, 6.00]	<.001
PHQ-8 response time (minutes)	73.68 [31.31, 215.77]	148.08 [54.05, 322.27]	302.36 [122.30, 527.10]	<.001
PHQ-8 completion time (seconds)	50.29 [37.92, 68.96]	49.42 [40.01, 66.95]	61.56 [46.12, 83.00]	<.001
RSES response time (minutes)	67.94 [24.11, 198.42]	134.96 [44.55, 304.56]	274.09 [108.18, 518.50]	<.001
RSES completion time (seconds)	54.54 [43.78, 72.39]	55.75 [43.99, 74.28]	69.24 [50.34, 100.52]	<.001
Site, n (%)				<.001
CIBER	47(20.3)	48 (26.8)	51 (25.0)	
KCL	119 (51.5)	102 (57.0)	129 (63.2)	
VUmc	65 (28.1)	29 (16.2)	24 (11.8)	
Married Status, n (%)				.24
Single/separated/divorced/widowed	114 (49.4)	97 (54.2)	117 (57.4)	
Married/cohabiting/LTR	117 (50.6)	82 (45.8)	87 (42.6)	
Years in education	16.00 [12.00, 19.00]	16.00 [13.00, 19.00]	15.00 [12.00, 19.00]	.90
Having children, n (%)	116 (50.2)	88 (49.2)	100 (49.0)	.78
Employed, n (%)	90 (39.0)	80 (44.7)	88 (43.1)	.06
Annual income, n (%)				.73
<15,000 (£/€)	48 (20.8)	48 (26.8)	56 (27.5)	
15,000-55,000 (£/€)	135 (58.4)	102 (57.0)	111 (54.4)	
more than 55000 (£/€)	40 (17.3)	25 (14.0)	33 (16.2)	
Accommodation, n (%)				.47
Own outright/with mortgage	131 (56.7)	90 (50.3)	102 (50.0)	
Renting	82 (35.5)	73 (40.8)	81 (39.7)	
Living rent-free	13 (5.6)	15 (8.4)	18 (8.8)	
Baseline PHQ-8 score	9.00 [6.00, 15.00]	10.00 [8.00, 15.00]	13.00 [7.00, 17.00]	.003
Having comorbidities, n (%)	109 (47.2)	86 (48.0)	116 (56.9)	.09
Taking depression medication, n (%)	145 (62.8)	126 (70.4)	129 (63.2)	.22
Number of contact logs	3.00 [2.00, 5.00]	5.00 [3.00, 7.00]	5.00 [2.00, 9.00]	<.001
Brand of smartphone, n (%)				.56
Motorola	99 (43.4)	66 (37.1)	75 (37.7)	
Samsung	72 (31.6)	56 (31.5)	66 (33.2)	
Other	57 (25.0)	56 (31.5)	58 (29.1)	

Supplementary Table 5. Summary participants' characteristics across three distinct engagement subgroups of Phone-Passive data for the first 43 weeks of the RADAR-MDD study. The median and interquartile range (IQR) of continuous variables and the count number and percentage of category variables are reported in the table.

Characteristics	C1	C2	C3	P value
Number of participants, n	259	148	207	
Age	52.00 [36.50, 61.00]	46.50 [30.75, 56.25]	46.00 [30.50, 57.50]	.01
Male, n (%)	66 (25.5)	36 (24.3)	47 (22.7)	.79
Days with phone passive data	283.00 [257.00, 298.00]	167.00 [142.25, 205.25]	32.00 [4.00, 67.50]	<.001
PHQ-8 response time (minutes)	121.70 [45.13, 327.68]	99.76 [40.89, 289.39]	150.94 [52.23, 340.70]	.21
PHQ-8 completion time (seconds)	53.63 [41.84, 72.91]	48.08 [36.46, 66.11]	55.29 [39.75, 74.47]	.01
RSES response time (minutes)	116.13 [36.36, 347.98]	81.50 [31.10, 254.39]	160.42 [39.65, 321.28]	.14
RSES completion time (seconds)	61.08 [47.33, 83.46]	51.87 [41.31, 70.32]	63.55 [46.68, 82.90]	.001
Site, n (%)				<.001
CIBER	60 (23.2)	24 (16.2)	62 (30.0)	
KCL	165 (63.7)	82 (55.4)	103 (49.8)	
VUmc	34 (13.1)	42 (28.4)	42 (20.3)	
Married Status, n (%)				.26
Single/separated/divorced/widowed	131 (50.6)	77 (52.0)	120 (58.0)	
Married/cohabiting/LTR	128 (49.4)	71 (48.0)	87 (42.0)	
Years in education	15.00 [12.00, 18.00]	17.00 [14.00, 20.00]	16.00 [12.00, 19.00]	.005
Having children, n (%)	142 (54.8)	66 (44.6)	96 (46.4)	.15
Employed, n (%)	106 (40.9)	64 (43.2)	88 (42.5)	.87
Annual income, n (%)				.80
<15,000 (£/€)	65 (25.1)	31 (20.9)	56 (27.1)	
15,000-55,000 (£/€)	151 (58.3)	83 (56.1)	114 (55.1)	
more than 55000 (£/€)	37 (14.3)	29 (19.6)	32 (15.5)	
Accommodation, n (%)				.08
Own outright/with mortgage	150 (57.9)	71 (48.0)	102 (49.3)	
Renting	90 (34.7)	67 (45.3)	79 (38.2)	
Living rent-free	17 (6.6)	7 (4.7)	22 (10.6)	
Baseline PHQ-8 score	9.00 [6.00, 15.00]	10.00 [7.00, 14.00]	12.00 [8.00, 17.00]	.001
Having comorbidities, n (%)	135 (52.1)	75 (50.7)	101 (48.8)	.78
Taking depression medication, n (%)	161 (62.2)	95 (64.2)	144 (69.6)	.24
Number of contact logs	5.00 [2.00, 7.00]	4.00 [2.00, 7.00]	4.00 [2.00, 7.00]	.30
Brand of smartphone, n (%)				<.001
Motorola	147 (57.0)	63 (42.9)	30 (15.0)	
Samsung	71 (27.5)	55 (37.4)	68 (34.0)	
Other	40 (15.5)	29 (19.7)	102 (51.0)	

Supplementary Table 6. Summary participants' characteristics across three distinct engagement subgroups of Fitbit-Passive data for the first 43 weeks of the RADAR-MDD study. The median and interquartile range (IQR) of continuous variables and the count number and percentage of category variables are reported in the table.

Characteristics	C1	C2	C3	P value
Number of participants, n	407	99	108	
Age	48.00 [32.00, 58.50]	45.00 [31.50, 54.00]	51.50 [36.00, 61.00]	.06
Male, n (%)	99 (24.3)	26 (26.3)	24 (22.2)	.79
Days with Fitbit passive data	294.00 [274.00, 301.00]	156.00 [132.00, 190.00]	18.00 [0.00, 67.00]	<.001
PHQ-8 response time (minutes)	113.51 [38.68, 288.84]	170.78 [53.70, 470.75]	161.81 [65.91, 410.24]	.007
PHQ-8 completion time (seconds)	50.28 [39.19, 66.99]	52.80 [37.71, 81.08]	64.20 [50.71, 84.31]	<.001
RSES response time (minutes)	97.68 [29.95, 288.16]	143.47 [56.43, 423.82]	177.56 [65.53, 381.02]	.009
RSES completion time (seconds)	56.40 [44.29, 73.28]	60.74 [44.95, 87.61]	69.43 [53.61, 96.26]	<.001
Site, n (%)				<.001
CIBER	77 (18.9)	29 (29.3)	40 (37.0)	
KCL	237 (58.2)	54 (54.5)	59 (54.6)	
VUmc	93 (22.9)	16 (16.2)	9 (8.3)	
Married Status, n (%)				.06
Single/separated/divorced/widowed	204 (50.1)	61 (61.6)	63 (58.3)	
Married/cohabiting/LTR	203 (49.9)	38 (38.4)	45 (41.7)	
Years in education	16.00 [13.00, 19.00]	15.00 [13.00, 20.00]	14.50 [11.00, 18.00]	.004
Having children, n (%)	191 (46.9)	48 (48.5)	65 (60.2)	.03
Employed, n (%)	179 (44.0)	37 (37.4)	42 (38.9)	.47
Annual income, n (%)				.08
<15,000 (£/€)	90 (22.1)	35 (35.4)	27 (25.0)	
15,000-55,000 (£/€)	232 (57.0)	53 (53.5)	63 (58.3)	
more than 55000 (£/€)	75 (18.4)	8 (8.1)	15 (13.9)	
Accommodation, n (%)				.30
Own outright/with mortgage	215 (52.8)	44 (44.4)	64 (59.3)	
Renting	157 (38.6)	46 (46.5)	33 (30.6)	
Living rent-free	30 (7.4)	8 (8.1)	8 (7.4)	
Baseline PHQ-8 score	9.00 [6.00, 15.00]	11.00 [7.00, 16.00]	13.00 [9.00, 17.50]	<.001
Having comorbidities, n (%)	197 (48.4)	49 (49.5)	65 (60.2)	.09
Taking depression medication, n (%)	264 (64.9)	63 (63.6)	73 (67.6)	.82
Number of contact logs	4.00 [2.00, 7.00]	5.00 [3.00, 8.00]	4.00 [2.00, 6.50]	.09
Brand of smartphone, n (%)				.047
Motorola	168 (41.6)	39 (39.4)	33 (32.4)	
Samsung	137 (33.9)	27 (27.3)	30 (29.4)	
Other	99 (24.5)	33 (33.3)	39 (38.2)	

Supplementary Table 7. Summary participants' characteristics across four distinct engagement subgroups of Phone-Active data for the first 94 weeks of the RADAR-MDD study. The median and interquartile range (IQR) of continuous variables and the count number and percentage of category variables are reported in the table.

Characteristics	C1	C2	C3	C4	P value
Number of participants, n	82	63	59	109	
Age	56.00 [46.25, 63.00]	49.00 [32.50, 59.00]	48.00 [30.00, 57.50]	47.00 [35.00, 54.00]	<.001
Male, n (%)	24 (29.3)	15 (23.8)	13 (22.0)	26 (23.9)	.75
Number of biweekly surveys	41.00 [37.25, 44.00]	25.00 [22.00, 30.00]	18.00 [14.50, 24.00]	5.00 [1.00, 8.00]	<.001
PHQ-8 response time (minutes)	84.20 [31.44, 204.10]	114.91 [71.77, 275.07]	158.96 [48.80, 338.91]	185.82 [71.58, 469.91]	.01
PHQ-8 completion time (seconds)	50.78 [37.80, 66.44]	47.26 [38.60, 57.76]	49.67 [39.45, 64.32]	55.58 [41.74, 81.56]	.04
RSES response time (minutes)	62.88 [25.62, 195.36]	104.12 [41.94, 359.53]	176.02 [40.42, 351.18]	194.83 [53.30, 457.22]	.008
RSES completion time (seconds)	52.23 [42.13, 71.35]	52.09 [44.01, 67.51]	54.67 [42.18, 69.99]	66.35 [49.14, 98.78]	<.001
Site, n (%)					.001
CIBER	27 (32.9)	14 (22.2)	19 (32.2)	31 (28.4)	
KCL	45 (54.9)	49 (77.8)	35 (59.3)	77 (70.6)	
VUmc	10 (12.2)	0 (0.0)	5 (8.5)	1 (0.9)	
Married Status, n (%)					.43
Single/separated/divorced/widowed	35 (42.7)	33 (52.4)	28 (47.5)	59 (54.1)	
Married/cohabiting/LTR	47 (57.3)	30 (47.6)	31 (52.5)	50 (45.9)	
Years in education	15.50 [12.00, 18.00]	16.00 [12.00, 19.00]	15.00 [12.00, 18.00]	15.00 [12.00, 19.00]	.87
Having children, n (%)	32 (39.0)	25 (39.7)	28 (47.5)	54 (49.5)	.47
Employed, n (%)	28 (34.1)	34 (54.0)	25 (42.4)	48 (44.0)	.13
Annual income, n (%)					.84
<15,000 (£/€)	20 (24.4)	18 (28.6)	18 (30.5)	30 (27.5)	
15,000-55,000 (£/€)	50 (61.0)	38 (60.3)	34 (57.6)	59 (54.1)	
more than 55000 (£/€)	12 (14.6)	7 (11.1)	6 (10.2)	19 (17.4)	
Accommodation, n (%)					.08
Own outright/with mortgage	58 (70.7)	33 (52.4)	31 (52.5)	55 (50.5)	
Renting	20 (24.4)	28 (44.4)	22 (37.3)	40 (36.7)	
Living rent-free	3 (3.7)	2 (3.2)	5 (8.5)	13 (11.9)	
Baseline PHQ-8 score	9.00 [5.75, 13.00]	10.00 [7.00, 14.00]	13.00 [8.50, 17.50]	13.00 [7.00, 17.00]	.005
Having comorbidities, n (%)	39 (47.6)	39 (61.9)	32 (54.2)	66 (60.6)	.24
Taking depression medication, n (%)	57 (69.5)	38 (60.3)	41 (69.5)	72 (66.1)	.65
Number of contact logs	7.00 [5.00, 10.00]	11.00 [7.00, 17.00]	9.00 [5.00, 13.50]	10.00 [5.00, 14.00]	.002
Brand of smartphone, n (%)					.13
Motorola	45 (55.6)	23 (36.5)	22 (37.3)	46 (42.6)	
Samsung	17 (21.0)	18 (28.6)	15 (25.4)	35 (32.4)	
Other	19 (23.5)	22 (34.9)	22 (37.3)	27 (25.0)	

Supplementary Table 8. Summary participants' characteristics across four distinct engagement subgroups of Phone-Passive data for the first 94 weeks of the RADAR-MDD study. The median and interquartile range (IQR) of continuous variables and the count number and percentage of category variables are reported in the table.

Characteristics	C1	C2	C3	C4	P value
Number of participants, n	122	53	61	77	
Age	54.00 [41.00, 62.75]	46.00 [35.00, 53.00]	48.00 [37.00, 55.00]	50.00 [36.00, 59.00]	.009
Male, n (%)	33 (27.0)	9 (17.0)	16 (26.2)	20 (26.0)	.54
Days with phone passive data	607.00 [538.25, 639.00]	417.00 [362.00, 465.00]	230.00 [179.00, 286.00]	31.00 [3.00, 86.00]	<.001
PHQ-8 response time (minutes)	107.42 [40.78, 320.76]	123.17 [35.15, 246.15]	107.89 [37.38, 329.76]	174.82 [63.87, 261.17]	.81
PHQ-8 completion time (seconds)	51.78 [41.09, 66.86]	47.00 [37.99, 61.73]	46.49 [37.06, 61.44]	55.58 [41.17, 75.53]	.16
RSES response time (minutes)	105.32 [33.88, 336.49]	104.53 [21.69, 290.55]	83.92 [33.37, 280.32]	178.30 [64.61, 336.32]	.55
RSES completion time (seconds)	57.42 [45.97, 74.53]	53.36 [43.56, 68.84]	50.79 [41.02, 72.28]	64.11 [48.09, 97.14]	.06
Site, n (%)					.007
CIBER	26 (21.3)	15 (28.3)	17 (27.9)	33 (42.9)	
KCL	92 (75.4)	35 (66.0)	37 (60.7)	42 (54.5)	
VUmc	4 (3.3)	3 (5.7)	7 (11.5)	2 (2.6)	
Married Status, n (%)					.85
Single/separated/divorced/widowed	60 (49.2)	24 (45.3)	30 (49.2)	41 (53.2)	
Married/cohabiting/LTR	62 (50.8)	29 (54.7)	31 (50.8)	36 (46.8)	
Years in education	14.50 [12.00, 18.00]	17.00 [13.00, 18.00]	16.00 [13.00, 19.00]	15.00 [12.00, 19.00]	.55
Having children, n (%)	69 (56.6)	26 (49.1)	35 (57.4)	43 (55.8)	.85
Employed, n (%)	52 (42.6)	24 (45.3)	29 (47.5)	30 (39.0)	.89
Annual income, n (%)					.15
<15,000 (£/€)	32 (26.2)	14 (26.4)	15 (24.6)	25 (32.5)	
15,000-55,000 (£/€)	69 (56.6)	29 (54.7)	39 (63.9)	44 (57.1)	
more than 55000 (£/€)	21 (17.2)	8 (15.1)	7 (11.5)	8 (10.4)	
Accommodation, n (%)					.82
Own outright/with mortgage	71 (58.2)	32 (60.4)	33 (54.1)	41 (53.2)	
Renting	41 (33.6)	18 (34.0)	23 (37.7)	28 (36.4)	
Living rent-free	10 (8.2)	3 (5.7)	4 (6.6)	6 (7.8)	
Baseline PHQ-8 score	8.50 [5.25, 14.00]	11.00 [6.00, 17.00]	10.50 [7.25, 15.75]	13.00 [9.00, 17.00]	.02
Having comorbidities, n (%)	74 (60.7)	26 (49.1)	33 (54.1)	43 (55.8)	.53
Taking depression medication, n (%)	73 (59.8)	36 (67.9)	44 (72.1)	55 (71.4)	.24
Number of contact logs	10.50 [7.00, 17.00]	9.00 [6.00, 14.00]	10.00 [6.00, 15.00]	6.00 [3.00, 10.00]	<.001
Brand of smartphone, n (%)					<.001
Motorola	81 (66.9)	22 (41.5)	19 (31.1)	14 (18.4)	
Samsung	17 (14.0)	11 (20.8)	19 (31.1)	43 (56.6)	
Other	23 (19.0)	20 (37.7)	23 (37.7)	19 (25.0)	

Supplementary Table 9. Summary participants' s characteristics across four distinct engagement subgroups of Fitbit-Passive data for the first 94 weeks of the RADAR-MDD study. The median and interquartile range (IQR) of continuous variables and the count number and percentage of category variables are reported in the table.

Characteristics	C1	C2	C3	C4	P value
Number of participants, n	153	58	52	50	
Age	53.00 [40.00, 60.00]	45.00 [32.00, 54.75]	43.00 [30.50, 53.00]	54.50 [40.75, 63.00]	<.001
Male, n (%)	38 (24.8)	18 (31.0)	10 (19.2)	12 (24.0)	.56
Days with Fitbit passive data	634.00 [586.00, 655.00]	426.50 [358.25, 480.00]	218.00 [162.75, 264.00]	61.50 [2.50, 100.50]	<.001
PHQ-8 response time (minutes)	106.57 [38.25, 252.87]	166.12 [59.78, 305.03]	198.13 [39.73, 378.89]	160.82 [63.15, 259.01]	.56
PHQ-8 completion time (seconds)	47.46 [39.41, 58.66]	48.08 [39.29, 63.49]	57.93 [41.96, 77.13]	67.27 [48.78, 92.50]	.001
RSES response time (minutes)	100.70 [27.77, 234.81]	143.48 [38.42, 316.74]	180.24 [18.92, 425.51]	149.08 [50.24, 331.34]	.56
RSES completion time (seconds)	52.44 [43.61, 66.07]	53.18 [42.18, 73.52]	67.99 [48.09, 94.11]	71.15 [54.14, 106.58]	.001
Site, n (%)					.11
CIBER	35 (22.9)	15 (25.9)	20 (38.5)	21 (42.0)	
KCL	109 (71.2)	39 (67.2)	30 (57.7)	28 (56.0)	
VUmc	9 (5.9)	4 (6.9)	2 (3.8)	1 (2.0)	
Married Status, n (%)					.39
Single/separated/divorced/widowed	70 (45.8)	34 (58.6)	27 (51.9)	24 (48.0)	
Married/cohabiting/LTR	83(54.2)	24 (41.4)	25 (48.1)	26 (52.0)	
Years in education	16.00 [12.00, 19.00]	15.00 [13.00, 18.00]	15.00 [12.00, 20.00]	14.00 [11.00, 17.00]	.19
Having children, n (%)	84 (54.9)	27 (46.6)	26 (50.0)	36 (72.0)	.17
Employed, n (%)	74 (48.4)	24 (41.4)	18 (34.6)	19 (38.0)	.27
Annual income, n (%)					.32
<15,000 (£/€)	37 (24.2)	17 (29.3)	17 (32.7)	15 (30.0)	
15,000-55,000 (£/€)	87 (56.9)	34 (58.6)	31 (59.6)	29 (58.0)	
more than 55000 (£/€)	29 (19.0)	6 (10.3)	4 (7.7)	5 (10.0)	
Accommodation, n (%)					.08
Own outright/with mortgage	94 (61.4)	26 (44.8)	23 (44.2)	34 (68.0)	
Renting	50 (32.7)	23 (39.7)	24 (46.2)	13 (26.0)	
Living rent-free	8 (5.2)	8 (13.8)	5 (9.6)	2 (4.0)	
Baseline PHQ-8 score	9.00 [7.00, 13.75]	12.00 [6.00, 18.00]	13.00 [7.00, 17.00]	13.00 [8.75, 17.00]	.11
Having comorbidities, n (%)	81 (52.9)	31 (53.4)	27 (51.9)	37 (74.0)	.05
Taking depression medication, n (%)	102 (66.7)	37 (63.8)	37 (71.2)	32 (64.0)	.84
Number of contact logs	10.00 [7.00, 16.00]	8.00 [5.00, 14.00]	9.00 [5.75, 12.25]	6.00 [3.00, 9.75]	<.001
Brand of smartphone, n (%)					.45
Motorola	72 (47.4)	27 (46.6)	20 (38.5)	17 (34.7)	
Samsung	39 (25.7)	17 (29.3)	17 (32.7)	12 (24.5)	
Other	41 (27.0)	14 (24.1)	15 (28.8)	20 (40.8)	

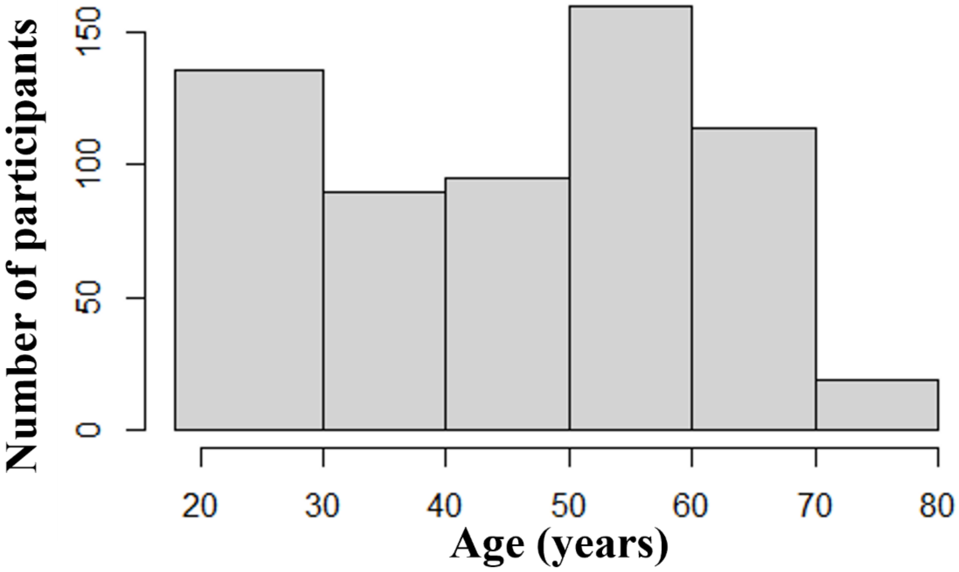
Supplementary Table 10. Summary of ethnicity difference across three distinct engagement subgroups of Phone-Active, Phone-Passive, and Fitbit-Passive data streams for the first 43 weeks of the RADAR-MDD study (ethnicity data was available for KCL and VUmc sites).

Data stream	C1	C2	C3	P value
Phone-Active				<.001
White	175 (95.1%)	110 (84.0%)	119 (77.8%)	
Black	3 (1.6%)	3 (2.3%)	8 (5.2%)	
Asian	2 (1.1%)	2 (1.5%)	12 (7.8%)	
Other	4 (2.2%)	16 (12.2%)	14 (9.2%)	
Phone-Passive				.001
White	174 (87.4%)	115 (92.7%)	115 (79.3%)	
Black	10 (5.0%)	1 (0.8%)	3 (2.1%)	
Asian	2 (1.0%)	4 (3.2%)	10 (6.9%)	
Other	13 (6.5%)	4 (3.2%)	17 (11.7%)	
Fitbit-Passive				.003
White	296 (89.7%)	57 (81.4%)	51 (75.0%)	
Black	9 (2.7%)	4 (5.7%)	1 (1.5%)	
Asian	6 (1.8%)	4 (5.7%)	6 (8.8%)	
Other	19 (5.8%)	5 (7.1%)	10 (14.7%)	

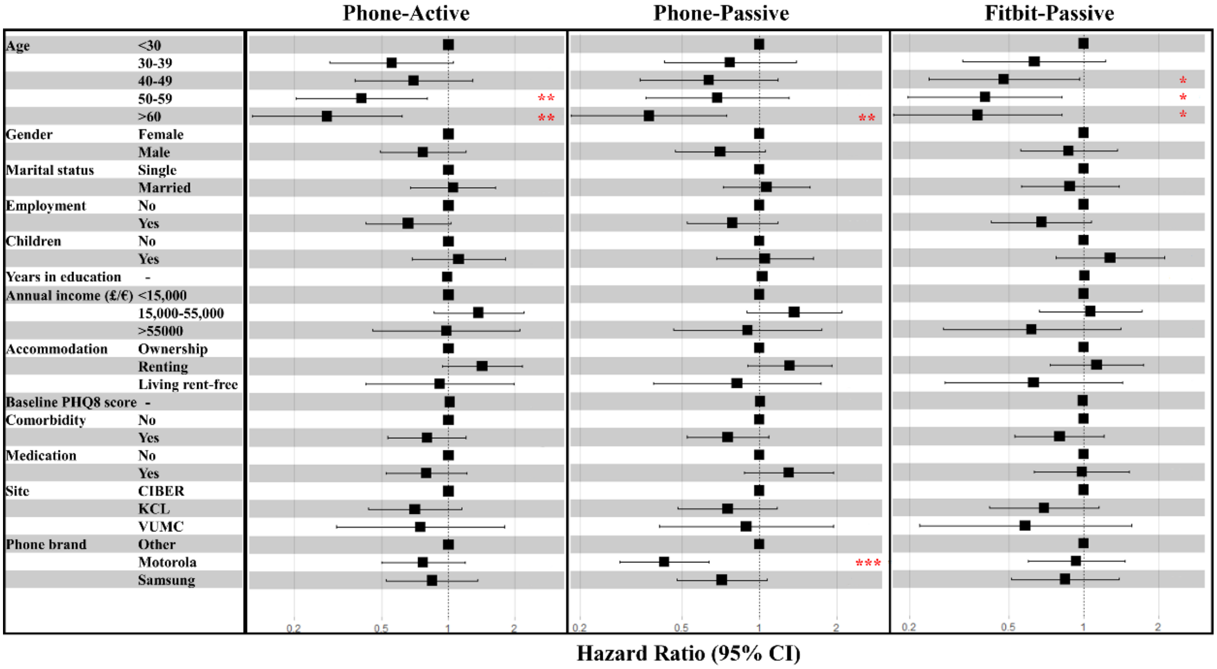
Supplementary Table 11. A list of 19 comorbidities that recorded at the enrollment of the RADAR-MDD study.

Number	Comorbidity
1	Asthma
2	Chronic bronchitis
3	Other chest trouble
4	Diabetes
5	Stomach or other digestive disorder
6	Liver trouble
7	Kidney trouble
8	Rheumatoid arthritis
9	Osteoarthritis
10	Heart trouble
11	Cancer
12	High blood pressure
13	Multiple Sclerosis
14	Epilepsy/fits
15	Stroke
16	Other neurological trouble
17	Migraine
18	Back trouble
19	Other

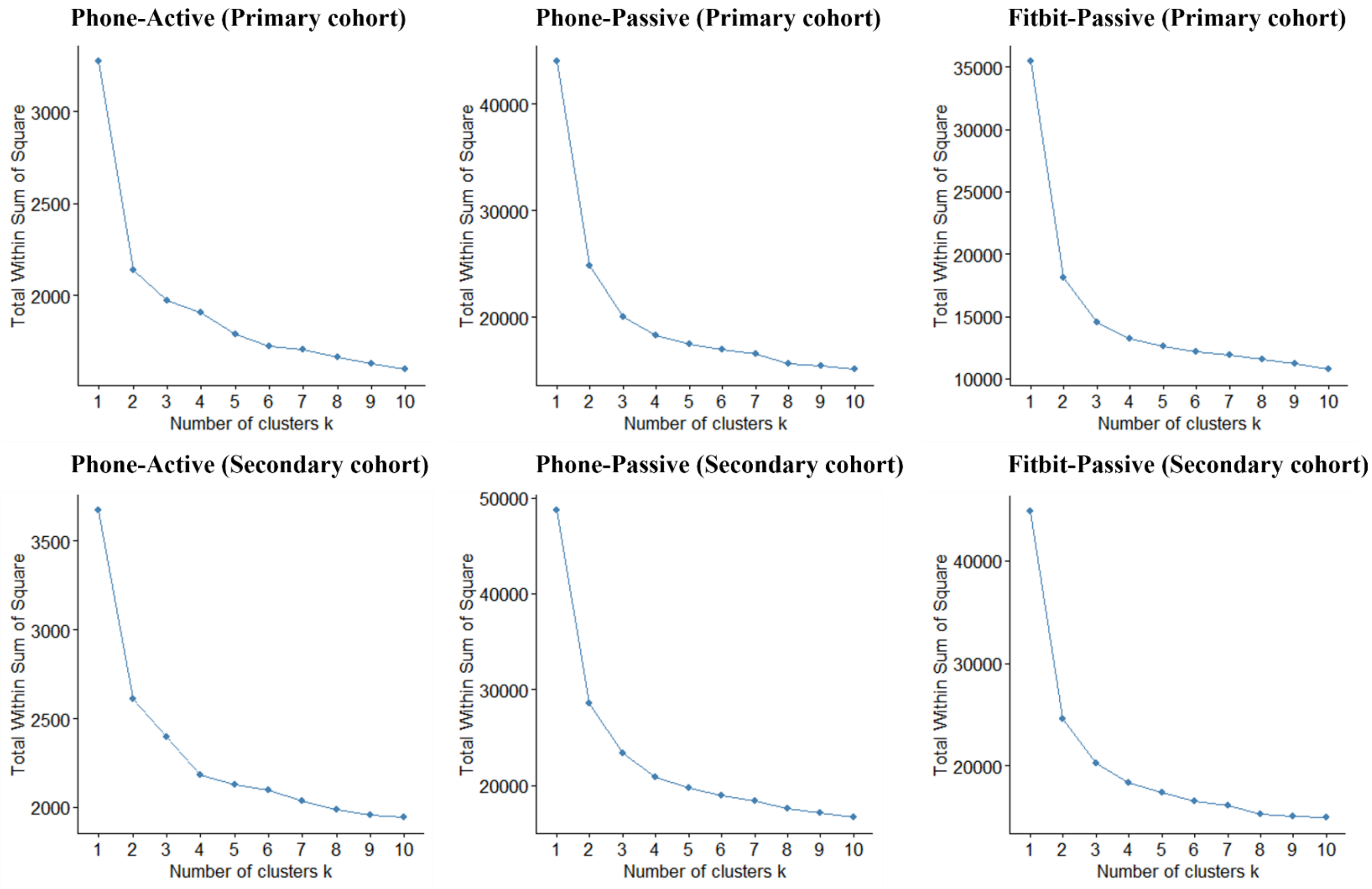
Supplementary Figure 1. The histogram of age distribution for 614 participants in the RADAR-MDD study.



Supplementary Figure 2. The hazard ratio plots of Cox Proportional-Hazards models for assessing the impact of multiple variables of interest on the participant retention time in the study of the secondary cohort (94-week observation period) for the Phone-Active, Phone-Passive, and Fitbit-Passive data streams, respectively. Significance levels: p < .05 *, p < .01 **, and p < .001 ***.



Supplementary Figure 3. Comparison of within-cluster variations across data streams using different cluster sizes (N=1-10) for K-means clustering. The optimal numbers of clusters for primary (43-week observation period) and secondary (94-week observation period) cohorts are 3 and 4, respectively.



Phone-Active (Primary cohort)

Phone-Passive (Primary cohort)

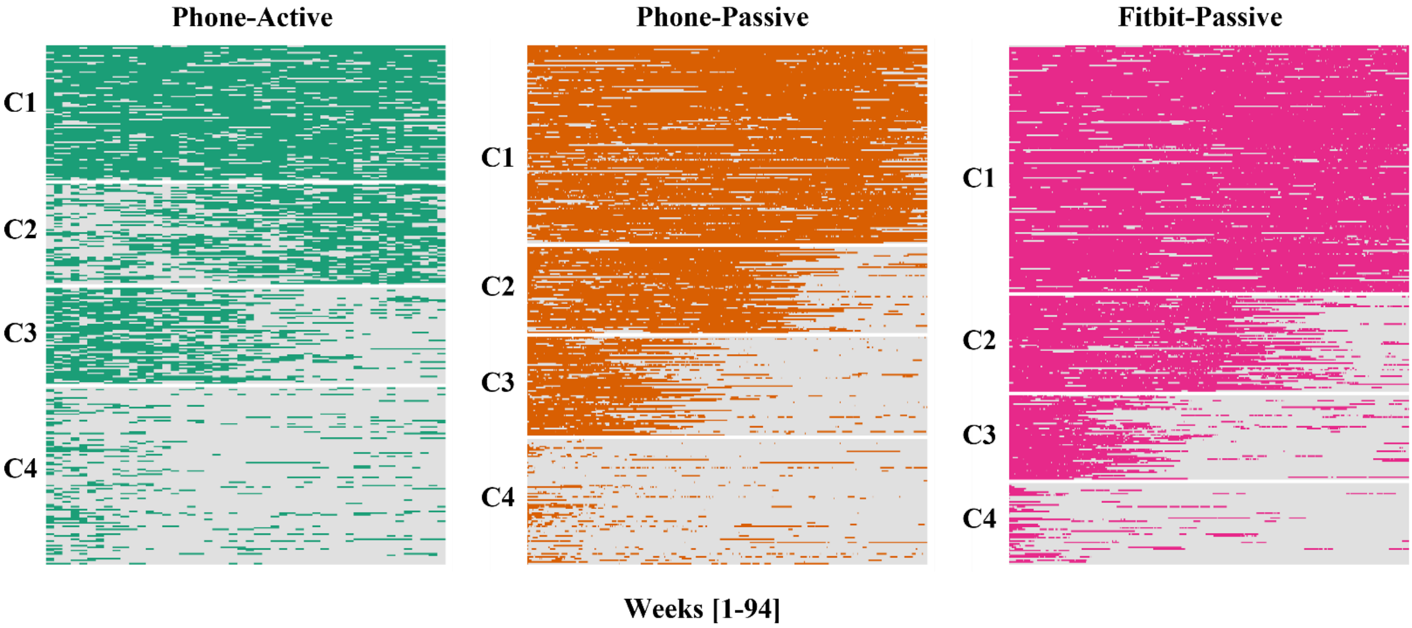
Fitbit-Passive (Primary cohort)

Phone-Active (Secondary cohort)

Phone-Passive (Secondary cohort)

Fitbit-Passive (Secondary cohort)

Supplementary Figure 4. Heatmaps of participant longitudinal engagement patterns for the three data streams in the longer observation period (94 weeks), clustered using K-means clustering. In each heatmap, each row represents a data-availability vector of one participant (described in Methods), and subgroups were arranged from the most engaged cluster to the least engaged cluster (C1-C4).



Appendix B

Supplementary Material to Chapter 6

Supplementary Table 1. The list of comorbidities that recorded at the enrollment session of the RADAR-MDD-KCL dataset[a].

Number	Comorbidity
1	Asthma
2	Chronic bronchitis
3	Other chest trouble
4	Diabetes
5	Depression
6	Stomach or other digestive disorder
7	Liver trouble
8	Kidney trouble
9	Rheumatoid arthritis
10	Osteoarthritis
11	Heart trouble
12	Cancer
13	High blood pressure
14	Multiple Sclerosis
15	Epilepsy/fits
16	Stroke
17	Other neurological trouble
18	Migraine
19	Back trouble
20	Other

[a] RADAR-MDD-KCL: A subset of the Remote Assessment of Disease and Relapse – Major Depressive Disorder data set, which was collected from King's College London, United Kingdom.

Supplementary Table 2. Results and performance of two nested linear regression models with (Model B) and without (Model A) long-term gait features in the Long-Term Movement Monitoring dataset.

Feature[a]	Model A			Model B		
	Estimate	SE[b]	P value	Estimate	SE	P value
(Intercept)	9.49	11.57	0.42	17.36	22.98	0.45
Age	-0.08	0.08	0.35	-0.07	0.09	0.46
Gender	-0.06	0.85	0.95	-0.90	1.03	0.39
Median Cycle	0.02	0.08	0.77	0.05	0.13	0.67
Peak Frequency	0.38	1.92	0.85	-0.43	2.10	0.84
Median Force	-2.43	2.22	0.28	-1.81	2.82	0.52
25th percentile of Median Cycle	—[c]	—	—	0.19	0.24	0.42
50th percentile of Median Cycle	—	—	—	-0.54	0.39	0.17
75th percentile of Median Cycle	—	—	—	0.29	0.16	0.08
SD of Median Cycle	—	—	—	-0.11	0.09	0.25
25th percentile of Peak Frequency	—	—	—	8.54	7.07	0.23
50th percentile of Peak Frequency	—	—	—	-4.25	8.70	0.63
75th percentile of Peak Frequency	—	—	—	-8.33	4.30	0.06
SD of Peak Frequency	—	—	—	12.45	8.12	0.13
25th percentile of Median Force	—	—	—	2.75	21.69	0.90
50th percentile of Median Force	—	—	—	1.23	22.37	0.96
75th percentile of Median Force	—	—	—	5.38	28.31	0.85
SD of Median Force	—	—	—	-27.53	53.83	0.61
R^2		0.06			0.30	
LR test[d]: χ^2			32.91			
LR test: P value			.001			

[a] Definitions of gait features in this table are shown in Table 1.

[b] SE: standard error.

[c] Not applicable.

[d] The critical value of the likelihood ratio statistic: $\chi^2_{0.05}(12) = 21.03$.

Appendix C

Supplementary Material to Chapter 2

Participant Retention and Response Rates

Supplementary Table 1 provides an overview of participant retention and response rates for the three data streams in the RADAR-MDD study. The participant retention rates were 54.6% (N = 335), 47.7% (N = 293), and 67.6% (N = 415) after 43 weeks (primary cohort defined in Chapter 2) for Phone-Active, Phone-Passive, and Fitbit-Passive data streams, respectively. Participants shared 12.00 [5.00, 18.00] bi-weekly surveys, 182.50 [54.25, 272.00] days of phone passive data, and 267.50 [135.25, 299.00] days of Fitbit passive data. The median passive data availability of a smartphone is 12.05 [2.34, 18.82] hours per day whereas the median Fitbit wear time is 19.79 [12.46, 22.38] hours per day. Supplementary Table 1 also provides the participant retention rates and response rates for a longer observation (94 weeks, as stated in Chapter 2).

Monetary Incentive

It is recognized that compensation for participant time and monetary incentives increase participation (Bentley & Thacker, 2004; Simblett et al., 2018). Although participants were not compensated for completing surveys, sharing phone passive data, and wearing Fitbit devices, they did receive £15/€20 for enrollment, £5/€10 for clinical assessments (every 3 months), and £10/€10 for each additional qualitative interview (e.g., 1-year interview) and were permitted to keep the Fitbit device after the study was complete. The existing monetary incentives could increase participants' willingness to remain engaged in the study, which may be one of the reasons for the high participant retention of the RADAR-MDD study. However, the monetary incentives may affect the generalizability of the findings in real-world settings (i.e., no incentives).

Supplementary Table 1. A summary of participant retention rates and response rates for the three data streams (phone active surveys, phone passive data, and Fitbit data) in the RADAR-MDD study. The primary cohort (43 weeks) and Secondary cohort (94 weeks) are defined in Chapter 2. Note, for data density and daily data, medians and interquartile ranges (IQR) are reported.

	Primary Cohort (43 weeks)	Secondary Cohort (94 weeks)
Number of participants, n	614	313
Participant retention		
Phone-Active, n (%)	335 (54.6)	151 (48.2)
Phone-Passive, n (%)	293 (47.7)	123 (39.3)
Fitbit-Passive, n (%)	415 (67.6)	169 (54.0)
Data Density		
Phone-Active (Bi-weeks)	12.00 [5.00, 18.00]	21.00 [7.00, 33.00]
Phone-Passive (Days)	182.50 [54.25, 272.00]	369.00 [147.00, 584.00]
Fitbit-Passive (Days)	267.50 [135.25, 299.00]	480.00 [221.00, 630.00]
Daily Data		
Phone-Passive (Hours)	12.05 [2.34, 18.82]	12.99 [4.44, 18.33]
Fitbit-Passive (Hours)	19.79 [12.46, 22.38]	18.25 [11.99, 21.61]

Supplementary Material to Chapters 3-7

Behavioral Characteristics and Demographics

The baseline behavioral characteristics (sleep, sociability, mobility, steps, and circadian rhythm) stratified by participant demographics are summarized in Supplementary Table 2. Older participants had shorter sleep duration, lower sociability, lower mobility, and fewer daily steps than younger participants. Compared with males, female participants slept longer. Participants who have children had shorter sleep, lower sociability, and fewer daily steps than those without children. Employed participants had higher sociability and mobility, and more daily steps than participants without jobs. Among income levels, participants of the highest income level had the highest levels of sociability, mobility, and daily steps.

Supplementary Table 2. A summary of baseline behavioral characteristics stratified by participant demographics. Note, median values and interquartile ranges (IQR) are displayed in this table.

	Sleep[a]	Sociability[b]	Mobility[c]	Step[d]	Circadian Rhythm[e]
Age					
<30	8.06 [7.58, 8.49]	20.47 [13.16, 30.13]	-8.79 [-11.75, -7.56]	8376 [6142, 10582]	0.24 [0.19, 0.31]
30-39	8.08 [7.39, 8.58]	17.75 [11.65, 27.27]	-10.05 [-11.85, -8.15]	8724 [6405, 11123]	0.26 [0.20, 0.32]
40-49	7.73 [7.11, 8.38]	19.15 [11.57, 25.92]	-9.32 [-10.96, -8.12]	8503 [5535, 10854]	0.22 [0.15, 0.27]
50-59	7.67 [6.99, 8.13]	14.02 [10.29, 22.43]	-10.52 [-12.90, -8.73]	7908 [5659, 10683]	0.25 [0.19, 0.32]
>60	7.48 [6.84, 8.07]	12.82 [9.36, 19.21]	-11.25 [-14.31, -9.64]	7929 [5600, 10628]	0.26 [0.20, 0.34]
Gender					
Female	7.84 [7.25, 8.38]	16.01 [10.68, 25.02]	-10.30 [-12.77, -8.21]	8216 [5675, 10628]	0.25 [0.19, 0.32]
Male	7.48 [6.78, 8.28]	14.69 [10.43, 23.28]	-10.19 [-12.35, -8.41]	8402 [6117, 11182]	0.24 [0.19, 0.31]
Marital Status					
Married	7.83 [7.24, 8.38]	15.09 [10.86, 22.80]	-10.15 [-12.49, -8.32]	8395 [5878, 10811]	0.25 [0.20, 0.32]
Single	7.76 [7.04, 8.31]	16.79 [10.31, 26.71]	-10.43 [-12.81, -8.20]	8063 [5593, 10737]	0.25 [0.17, 0.31]
Children					
No	8.00 [7.42, 8.56]	18.93 [12.25, 29.55]	-9.57 [-11.86, -7.72]	8487 [6153, 10866]	0.24 [0.18, 0.31]
Yes	7.51 [6.83, 8.13]	13.91 [9.73, 20.59]	-10.82 [-13.29, -8.95]	7954 [5348, 10573]	0.25 [0.19, 0.32]
Employment					
No	7.77 [7.12, 8.37]	13.67 [9.50, 21.26]	-11.01 [-13.65, -9.02]	7833 [5460, 10663]	0.25 [0.19, 0.33]
Yes	7.82 [7.09, 8.37]	19.86 [12.50, 29.62]	-9.14 [-11.04, -7.41]	8609 [6582, 10804]	0.24 [0.19, 0.30]
Income					
below minimum	7.82 [7.13, 8.56]	13.62 [9.20, 25.73]	-11.67 [-13.80, -9.07]	7774 [5278, 10128]	0.25 [0.18, 0.32]
15,000-55,000	7.75 [7.07, 8.24]	15.32 [10.65, 22.91]	-10.19 [-12.10, -8.30]	8235 [5845, 10860]	0.25 [0.19, 0.31]
more than 55000	7.84 [7.25, 8.42]	20.05 [13.25, 32.25]	-9.02 [-11.30, -7.38]	8988 [7414, 11182]	0.24 [0.20, 0.31]

[a] Sleep is measured by total sleep time (hours) defined in Chapter 3.

[b] Sociability is measured by the mean Bluetooth count defined in Chapter 4.

[c] Mobility is measured by location variance defined in Chapter 5 (larger location variance higher mobility)

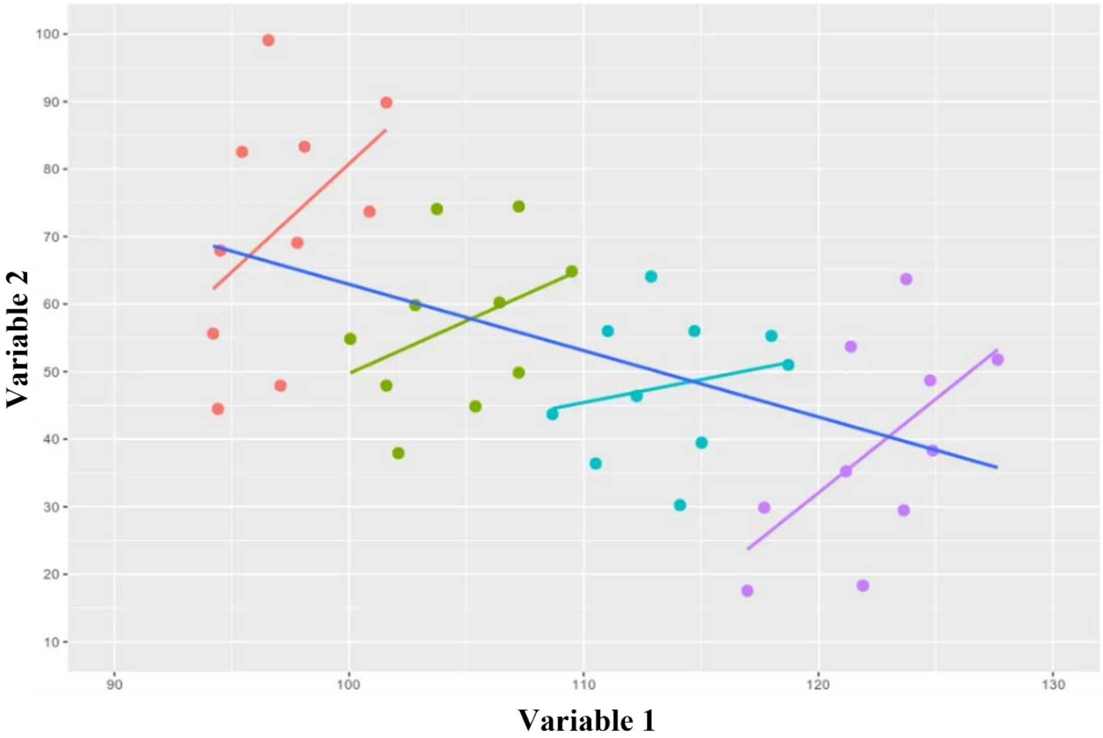
[d] Step is the daily step count.

[e] Circadian rhythm is the strength of rhythmicity measured by the R^2 of the fitted Cosinor function using the step and HR signals (defined in Chapter 7).

The Mixed-Effect Linear Regression Model

Since the dataset utilized in this thesis (RADAR-MDD) is longitudinal (each participant had repeated measurements), the independence assumption of some widely used association methods (e.g., Pearson correlations and simple linear regression) is violated. Since the life habits and behavioral patterns of participants differed across distinct demographics, ignoring underlying correlations at the participant level may obscure or distort the associations between behavioral characteristics and depression severity (Hajjem, Bellavance, & Larocque, 2010). Supplementary Figure 1 illustrates this issue: Variable 1 has a positive effect on Variable 2 at the individual level (represented by colors), but a negative (incorrect) association (blue) is found by the pooled data directly (ignore the data dependence).

Supplementary Figure 1. A schematic plot for the issue of data dependence in longitudinal data (adapted from Oskolkov, 2020).



The mixed-effect linear regression model is an appropriate method for accounting both within and between participant patterns across time in longitudinal data (Laird & Ware, 1982). Specifically, the linear mixed model incorporates fixed and random effects (Laird & Ware, 1982). A fixed effect is a parameter that does not change (represents the general association in the cohort), whereas the random effect might vary among participants (e.g., average activity level for each participant) (Laird & Ware, 1982). In addition, the linear relationships are easy to be used to explain the effect magnitude and direction (positive or negative) of behavioral features on depression severity. Hence, the mixed-effect linear regression models were leveraged in this thesis.

Normality Test

The linear mixed model assumes the errors (residuals) are normally distributed. As mentioned in Chapter 3, this assumption was checked by the histograms of residuals and the Shapiro-Wilk test (Yap & Sim, 2011). If the residuals are not normally distributed, the Box-Cox transformation was performed (Box & Cox, 1964).

Limitation of Linear Mixed-Effect Model

However, linear models can only explore linear connections, which are insufficient for real-world circumstances. To address this limitation, I plan to leverage GPBoost (a mixed-effects machine learning model) (Sigrist, 2020) to explore the non-linear relationships in future research (mentioned in Chapter 8).

Supplementary Material to Chapter 4

Multiscale Entropy Feature

Some nonlinear characteristics (e.g., complexity and periodicity) of individuals' behaviors are associated with the severity of depression. For instance, depression may lead to a misalignment of the circadian rhythm and make people's life rhythms more irregular, which could be reflected in the behavioral signals that are more complicated and chaotic (Walker, Walton, DeVries, & Nelson, 2020). However, it is difficult to measure these nonlinear characteristics using statistical features. Multiscale entropy (MSE) analysis has been widely used in the signal processing field, which can measure the complexity of the signal at different time scales (Costa, Goldberger, & Peng, 2005). Therefore, I selected to apply MSE to measure the complexity of the Bluetooth sequence from 1-hour timescale to 1-day (24 hours) timescale in this chapter.

Leave-All-Out and Leave-One-Out Training Strategies

Individual variation in behavioral patterns (e.g., average activity/social level) due to diverse life habits and demographics is a challenge that can affect the accuracy of forecasting depression severity. The aim of this chapter is to utilize a hierarchical Bayesian linear regression model to conduct individualized (personalized) depression prediction. Specifically, for each participant, the individual parameters (intercept and slope) are trained based on this participant's historical data (Bluetooth features and PHQ-8 score) while these individual parameters are drawn from the cohort parameter distribution (Busk et al., 2020). This training strategy guaranteed that both individual and cohort patterns are considered in the hierarchical Bayesian linear regression model. This method can also be implemented in real-world settings, such as requiring a new user to remotely complete PHQ-8 questionnaires during the first few weeks and then predicting the user's depression status in the subsequent weeks.

Based on this strategy, we examined two potential application-specific scenarios:

1. Leave all out (LAO): For all participants, the individual parameters are trained on all historical data, and only PHQ-8 scores at the next time point will be predicted. The advantage of LAO is that all participants have more and more historical data to train the individual parameters over time, which may allow the model to learn the individual patterns more effectively. The disadvantage of LAO is that users are required to continually complete questionnaires, which causes a high user burden.
2. Leave one out (LOO): For one participant, the individual parameters (intercept and slope) were trained only based on the first two PHQ-8 questionnaires (4 weeks), while the cohort parameters are trained using all data of other participants. The advantage of LOO is that a new participant just only needs to complete 2 PHQ-8 questionnaires (low user burden). However, the disadvantage is that the individual patterns may not be trained sufficiently (lack of data).

These two training strategies are described in detail in Chapter 4 and a prior study (Busk et al., 2020).

Limitations

One of the limitations of Chapter 4 is that the types and the MAC addresses of Bluetooth devices were not recorded in order to protect the privacy of participants and passers. This makes it difficult to estimate the number of people in the vicinity of the participant. Due to this limitation, this chapter did not explain in depth the actual meaning behind the changes in the Nearby Bluetooth device count data. It is indeed a limitation of the software implementation for this data modality and stems from the design of an earlier version of the RADAR-base platform. Our research team noted that MAC address hash and device types should be collected in future research and implementation of these improvements is currently under review in the RADAR-base platform and will likely be available in the next version release.

The time-series cross-validation method utilized in this chapter is another limitation. As mentioned above, the scenario considered in this chapter is to predict participants' subsequent depressive severity based on some of their history data (behavioral features and PHQ8 scores). However, this strategy is not suitable for "new" users who lack past data. To address this limitation, I will hold out some participants as test data and train a prediction model on the rest of the participants in future research.

Supplementary Material to Chapter 6

Gait Patterns and Step Count

Gait patterns and cumulative step count are two distinct measures of a person's walking. Gait reflects walking characteristics, such as cadence and force, whereas step count shows an individual's mobility or activity level. Since changes in gait and mobility are both essential manifestations of depression (Sobin & Sackeim, 1997; Weyerer & Kupfer, 1994), it is valuable to extract these two measures of daily walking for depression monitoring. Several prior studies have revealed a negative link between the step count and depression severity via wearables or mobile phones (Abedi, Nikkhah, & Najar, 2015; McKercher et al., 2009). For gait patterns, although laboratory gait parameters are found to be significantly correlated with depression, the gait characteristics of daily life walking in real-world settings have yet to be fully investigated (Zhang et al., 2022). Since the step count data provided by Fitbit devices (Charge 2 and Charge 3) cannot evaluate the cadence and force of steps, Chapter 6 aimed to extract gait patterns of daily-life walking and investigate their associations with depression using raw acceleration signals gathered by smartphones and wearables. In future research, I will include gait and step features as well as other behavioral features (e.g., sleep, Bluetooth, and GPS) in multimodal models for predicting depression.

Supplementary Material to Chapter 7

Measurements for Circadian Rhythm

Many mathematic techniques were developed to evaluate the circadian rhythm using passive behavioral data (Refinetti, Cornélissen, & Halberg, 2007). Spectrum approaches estimate the strength of circadian rhythm by calculating the frequency power in the circadian range of the signal spectrum obtained via some specific time-frequency transform techniques (such as Fourier analysis, Enright method, and Lomb – Scargle periodogram). Since the circadian rhythms can be thought of as smooth rhythms with added noise, the Cosinor-based methods estimate the circadian rhythm by fitting behavioral data to a Cosinor function (Cornelissen, 2014). With the aim of Chapter 7 to explain the association between the circadian rhythm and depression severity, it is required to extract more explanatory characteristics of the circadian rhythm. Generally speaking, ideal rhythmic processes can be fully characterized by four parameters: MESOR (mean level), period, amplitude, and phase (Refinetti et al., 2007). Based on the assumption of a known circadian period (24h), the Cosinor-based method is a reliable and practical tool for the computation of MESOR, amplitude, and phase (acrophase) of circadian rhythms (Cornelissen, 2014), whereas the spectrum method cannot extract these temporal characteristics (Duhamel & Vetterli, 1990). Moreover, Cosinor-based models are suitable for non-equidistant data (e.g., missing data) which is prevalent in wearable data (Cornelissen, 2014). Therefore, the Cosinor-based model is utilized in this chapter.

References

Abedi, P., Nikkhah, P., & Najar, S. (2015). Effect of pedometer-based walking on depression, anxiety and insomnia among postmenopausal women. *Climacteric, 18*(6), 841-845.

Bentley, J. P., & Thacker, P. G. (2004). The influence of risk and monetary payment on the research participation decision making process. *Journal of medical ethics, 30*(3), 293-298.

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Methodological), 26*(2), 211-243.

Busk, J., Faurholt-Jepsen, M., Frost, M., Bardram, J. E., Vedel Kessing, L., & Winther, O. (2020). Forecasting mood in bipolar disorder from smartphone self-assessments: hierarchical bayesian approach. *JMIR mHealth and uHealth, 8*(4), e15028.

Cornelissen, G. (2014). Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling, 11*(1), 1-24.

Costa, M., Goldberger, A. L., & Peng, C.-K. (2005). Multiscale entropy analysis of biological signals. *Physical review E, 71*(2), 021906.

Duhamel, P., & Vetterli, M. (1990). Fast Fourier transforms: a tutorial review and a state of the art. *Signal processing, 19*(4), 259-299.

Hajjem, A., Bellavance, F., & Larocque, D. (2010). Generalized mixed effects regression trees. *Mixed Effects Trees and Forests for Clustered Data, 34*.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, 963-974.

McKercher, C. M., Schmidt, M. D., Sanderson, K. A., Patton, G. C., Dwyer, T., & Venn,

A. J. (2009). Physical activity and depression in young adults. *American journal of preventive medicine, 36*(2), 161-164.

Refinetti, R., Cornélissen, G., & Halberg, F. (2007). Procedures for numerical analysis of circadian rhythms. *Biological rhythm research, 38*(4), 275-325.

Sigrist, F. (2020). Gaussian process boosting. *arXiv preprint arXiv:2004.02653*.

Simblett, S., Greer, B., Matcham, F., Curtis, H., Polhemus, A., Ferrão, J., . . . Wykes, T. (2018). Barriers to and facilitators of engagement with remote measurement technology for managing health: systematic review and content analysis of findings. *Journal of medical Internet research, 20*(7), e10480.

Sobin, C., & Sackeim, H. A. (1997). Psychomotor symptoms of depression. *American Journal of Psychiatry, 154*(1), 4-17.

Oskolkov, N. (2020). How Linear Mixed Model Works. *Medium.* https://towardsdatascience.com/how-linear-mixed-model-works-350950a82911

Walker, W. H., Walton, J. C., DeVries, A. C., & Nelson, R. J. (2020). Circadian rhythm disruption and mental health. *Translational psychiatry, 10*(1), 28.

Weyerer, S., & Kupfer, B. (1994). Physical exercise and psychological health. *Sports Medicine, 17*, 108-116.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation, 81*(12), 2141-2155.

Zhang, Y., Folarin, A. A., Sun, S., Cummins, N., Vairavan, S., Qian, L., . . . Stewart, C. (2022). Associations Between Depression Symptom Severity and Daily-Life Gait Characteristics Derived From Long-Term Acceleration Signals in Real-World Settings: Retrospective Analysis. *JMIR mHealth and uHealth, 10*(10), e40667.