Multimodal Behavioral Sensing for Precision Mental Health Care

Prerna Chikersal

CMU-HCII-23-107
August 2023

Human-Computer Interaction Institute
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA - 15213, USA

Thesis Committee:
Anind Dey (co-chair; UW)
Mayank Goel (co-chair; CMU)
Geoff Kaufman (CMU)
Andrew Campbell (Dartmouth)
Mary Czerwinski (Microsoft Research)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

This work was supported partly by the United States Department of Defense (CDMRP MS190178), Microsoft Research, Carnegie Mellon University's (CMU) Provost's Office, CMU's Center for Machine Learning and Health Fellowship, the Carnegie Bosch Initiative, and CMU's Human-Computer Interaction Institute.

Keywords: machine learning, feature selection, Mobile sensing, mobile health, depression, multiple sclerosis, disability, mental health, ecological momentary assessments.

Abstract

Mental health disorders are increasing in occurrence. They are the largest cause of disability worldwide and the strongest predictor of suicide. Despite their prevalence, the majority of affected people either never seek support, or receive limited to no support from under resourced health systems. Further, finding the right treatment for a specific person is a time consuming and inefficient process, as most interventions are based on studies that find the best treatment for a "typical" patient, rather than tailoring interventions to the patient's genes, environment and lifestyle. Hence, to increase access and efficiency of mental health care, there is a need to develop digital tools that make medicine more precise by using data-driven insights and predictions to aid diagnosis, monitoring, prevention, and treatment of mental health disorders.

This PhD thesis focuses on developing computational methods and models that use user-generated data from multiple data sources including passively sensed smartphone and wearable sensor data, text messages exchanged between users, and the users' interaction logs with web or mobile apps, to analyze or predict mental health outcomes with the goal of making the diagnosis and treatment of mental health disorders more efficient and precise. The biggest problem in precision mental health care is the curse of dimensionality in terms of the feature space, outcomes, and patients. That is, to tailor diagnosis, prevention and treatment to each individual, we need to collect and analyze enormous amounts of data associated with the person's behaviors, environment, and other in-situ features, the person's outcomes (*e.g.* multiple morbidities, outcomes related to mental and physical health), and contextual, occupational, demographic, and other confounding variables that can affect the patient's health. The curse of dimensionality when working with such multimodal and multidimensional data lowers the reliability of analysis and modeling, and decreases the interpretability of the findings.

During my PhD, I addressed the curse of dimensionality challenge through 5 studies. My thesis makes the following contributions: (1) Presents a machine learning based feature selection method that mitigates the curse of dimensionality in the feature space by decomposing and iteratively reducing the feature space during feature selection. (2) Demonstrates the generalizability of the approach in detecting depression, change in depression, and loneliness, as well as forecasting these outcomes several weeks in advance. (3) Demonstrates that behavioral changes resulting from the stay-at-home mandates during the pandemic are predictive of health outcomes during the stay-at-home period for patients with multiple sclerosis. (4) Demonstrates how we can categorize supporters or identify patient phenotypes based on multiple co-morbid outcomes, thereby mitigating the curse of dimensionality with respect to multimorbidities. (5) Presents a method that visualizes and identifies support strategies that work best in an online mental health intervention for patients in a specific context or situation. (6) Demonstrates that accounting for the patient's history or behav-

ioral context improves model performance for the longitudinal monitoring of some health outcomes such as fatigue.

This work has the potential to minimize suffering, by enabling early diagnosis and frequent monitoring of health outcomes using passively sensed longitudinal behavioral data. My work also had implications for more effective treatments through personalization, and improving the patients' awareness about their own health and treatment.

Acknowledgements

Completing my PhD would've been impossible without the tremendous support, guidance, and wisdom I received from three exceptional groups of individuals. I am immensely grateful for these wonderful people, and would like to take this opportunity to thank them.

Superheroes that enabled and contributed to my research: This dissertation would not exist without the wisdom, hard work, and support of a global community of researchers from whom I had the honor of learning.

My PhD Advisors – To Anind Dey and Mayank Goel, thank you for being my sources of inspiration and wisdom. Thank you for selflessly sharing your knowledge and expertise with me, and for providing invaluable feedback and guidance throughout my PhD. For the difficult times when things moved slowly, thank you for encouraging me and never losing faith in me, and for pushing me when I needed pushing.

My PhD Committee – To Andrew Campbell, Mary Czerwinski, and Geoff Kaufman, I am grateful for your constructive feedback and the insightful questions, which helped me scope, plan and execute my research agenda and complete my dissertation.

My Research Collaborators – My PhD work was enriched by many fruitful collaborations with several researchers from various disciplines in academia and industry. To Zongqi Xia at UPMC, thank you for all your hard work in designing and executing our research related to patients with multiple sclerosis. I'm grateful for your invaluable insights and prompt feedback, which fundamentally shaped my thesis. To Anja Thieme, Danielle Belgrave, and others at Microsoft Research and SilverCloud Health, thank you for giving me the opportunity to apply machine learning research to real-world products and patients. Thank you for creating a vibrant and inclusive work environment, and for your insightful and passionate guidance. To Brian Smith, Shree Nayar, Karl Bayer, and others I met at Snapchat, thank you for pushing me outside my comfort zone and giving me the opportunity to work on a project very different from my PhD research, and thank you for your guidance and feedback throughout that project. I am in awe of your passion and creativity for developing user friendly products that change how people communicate and interact.

The village that raised me: I was able to embark and persist in this journey, entirely due to the unwavering and consistent support and encouragement, nurturing love and boundless patience of the village that raised me.

My Family and Chosen Family – To my sweet Dad and departed grandparents, thank you for your unconditional love and for always supporting me in everything I have ever set out to do. I can never thank you enough for the many hours you toiled for my upbringing

and education. Dad, thank you for inspiring me to pursue computer science and ultimately a PhD. To Juvina, Laxmi, Shreya, Nishtha, Rinsha, and Maria, thank you for being my best friends and chosen family, and for always being there for me in every way possible whenever I have needed you.

My Therapist – To Perry Henschke, words fall short in expressing the immense positive impact you've had on my life. Your unwavering belief in me, gentle encouragement, and constant support have been instrumental in helping me complete my PhD. Thank you for nurturing the parts of me that needed healing, for helping me understand myself better, for imparting the wisdom to detach my self-worth from my work, and for empowering me to create a fulfilling life.

My Beloved Dog – The best decision I've ever made in my life, was to adopt my little cavapoo puppy. In the two years I've had him, Kulfi has brought immense joy, playfulness, and companionship into my life, and for that I am beyond grateful.

Supporters who helped me through: Several others played a key role in my development as a researcher, getting me into the PhD program, and supporting me through it.

My Former Advisors and Collaborators – To Erik Cambria, Laura Dabbish, Anitha Woolley, Maria Tomprou, and Young Ji Kim, thank you for advising and collaborating with me on my first few research projects. Thank you supporting my PhD applications and paving the path to this journey.

My CMU Friends and Colleagues – To Sai Ganesh Swaminathan, Vikram Kamath, Ishani Chatterjee, Samantha Reig, Stephanie Valencia, Fannie Lie, and my many other peers and colleagues, thank you for advising and supporting me in so many personal and professional steps throughout my PhD journey.

HCII Staff – To Queenie Kravitz and all the HCII staff, thank you for everything you do to help facilitate the work that students and researchers like me to!

As I begin my career in the industry, I cannot wait to pay forward all the support, encouragement, and wisdom I have received from all of you! Thank you!

Contents

1 Introduction **1**
- 1.1 The Curse of Dimensionality Challenge . . . 2
  - 1.1.1 The curse of dimensionality in the feature space (C1) . . . 3
  - 1.1.2 The curse of dimensionality with respect to multiple co-morbid outcomes (C2) . . . 3
  - 1.1.3 The curse of dimensionality presented by the diversity in patient characteristics (C3) . . . 4
- 1.2 Thesis Outline and Contributions . . . 5
  - 1.2.1 Study Framework . . . 5
  - 1.2.2 Addressing the Curse of Dimensionality Challenge . . . 8
  - 1.2.3 Contributions Across All Studies . . . 11

2 Related Work **13**
- 2.1 Exploring the relationship between behavioral features and mental health . . . 13
  - 2.1.1 Depression . . . 13
  - 2.1.2 Loneliness . . . 14
- 2.2 Detecting mental health outcomes . . . 15
  - 2.2.1 Depression . . . 15
  - 2.2.2 Loneliness . . . 18
- 2.3 Detecting health outcomes in patients in multiple sclerosis . . . 18
- 2.4 Understanding large-scale mental health behavioral data . . . 19
  - 2.4.1 Clustering and Text Mining in (Large-Scale) Mental Health Data . . . 19
  - 2.4.2 Using Association Rules for Behavioral Pattern Mining . . . 20

3 S1: Detecting Depression and Loneliness In College Students **21**
- 3.1 Introduction . . . 21
- 3.2 Data Collection . . . 23
  - 3.2.1 Participants and Recruitment . . . 23
  - 3.2.2 Participant-reported Depression and Loneliness Measures (Ground Truth) 24
  - 3.2.3 Passive Data Collection . . . 25
- 3.3 Data Processing and Analysis . . . 26
  - 3.3.1 Feature Extraction . . . 27
  - 3.3.2 Handling Missing Features . . . 35
  - 3.3.3 Modeling . . . 36
- 3.4 Results for Detecting Depression . . . 38
  - 3.4.1 Descriptive Statistics . . . 39
  - 3.4.2 Detecting Post-semester Depression . . . 40

3.4.3 Detecting Change in Depression . . . 42
3.5 Results for Detecting Loneliness . . . 42
3.5.1 Descriptive Statistics . . . 42
3.5.2 Detecting Post-semester Loneliness and Change in Loneliness . . . 43
3.6 Discussion . . . 45
3.6.1 Observations about Selected Features . . . 45
3.6.2 Comparison with Other Machine Learning Approaches . . . 46
3.6.3 Implications for Longitudinal Studies and Opportunities to Improve Model Performance . . . 49
3.6.4 Implications for Privacy and Technical Limitations . . . 50
3.6.5 Generalizability of our approach to detecting post-semester and change in loneliness . . . 51
3.6.6 Extending to Other Health Outcomes and Opportunities for Combining with Verbal and Non-Verbal Behaviors, and Genomic Data . . . 52
3.7 Conclusion . . . 52
3.8 Addressing the Curse of Dimensionality . . . 53
3.8.1 W.r.t. the feature space (C1) . . . 53
3.8.2 W.r.t. multiple co-morbidities (C2) . . . 54

4 S2: Forecasting End of Semester Depression In College Students 55
4.1 Introduction . . . 55
4.2 Prediction Models for Predicting Future Depressive Symptoms . . . 55
4.3 Results for Early Prediction of Future Depressive Episodes . . . 56
4.4 Implications for Interventions . . . 58
4.5 Conclusion . . . 60
4.6 Addressing the Curse of Dimensionality . . . 60
4.6.1 W.r.t. the feature space (C1) . . . 60

5 S3: Predicting the Mental Health of People with Multiple Sclerosis during the COVID-19 Stay-at-Home Period 62
5.1 Introduction . . . 62
5.2 Methods . . . 63
5.2.1 Participants . . . 63
5.2.2 Study Design . . . 63
5.2.3 Survey Response and Patient-Reported Outcomes . . . 64
5.2.4 Sensor Data Collection . . . 65
5.2.5 Mediation Analysis . . . 65

5.2.6 Data Processing and Machine Learning Analysis 65
5.3 Results 71
5.3.1 Participant Characteristics 71
5.3.2 Interrelated Outcomes 72
5.3.3 Predicting Outcomes during the Stay-at-Home Period 74
5.4 Discussion 75
5.5 Conclusion 77
5.6 Addressing the Curse of Dimensionality 78
5.6.1 W.r.t. the feature space (C1) 78
5.6.2 W.r.t. multiple co-morbidities (C2) 78

6 S4: Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention 79
6.1 Introduction 79
6.2 Background for Human Support in Online Mental Health Therapy 81
6.2.1 Modalities & Benefit of Human Support in iCBT 81
6.2.2 Human Support Behaviors & their Effectiveness in iCBT 81
6.3 The iCBT Intervention 82
6.3.1 Frequency & Format of Supporter Interactions 83
6.3.2 Dataset Description 83
6.4 Identifying Clusters of Successful Support 85
6.4.1 Change & Improvement Rates as Clinical Outcomes 85
6.4.2 Clustering Supporters Based on Support Outcomes 86
6.5 Identifying Successful Support Strategies 87
6.5.1 Methodology: <Strategy> Across ¡Context¿ Bins 89
6.5.2 Results: Strategies Used in Successful Messages 91
6.6 Context-specific Patterns of Support Success 94
6.6.1 Extracting Multidimensional Context → Strategy Patterns 95
6.6.2 Results: Salient Context-Specific Support Strategies 97
6.7 Discussion 97
6.7.1 Identifying Effective Context-Specific Support Strategies 98
6.7.2 Data-Enabled, Personalized Mental Health Support 98
6.7.3 Understanding (Big) Data in Digital Health Interventions 99
6.8 Conclusion 100
6.9 Addressing the Curse of Dimensionality 101
6.9.1 W.r.t. multiple co-morbidities (C2) 101
6.9.2 W.r.t. diversity in patient characteristics (C3) 101

7 S5: Predicting Periodically Assessed Multiple Sclerosis Outcomes Using Passively Sensed Behaviors and Ecological Momentary Assessments 102
7.1 Introduction . . . 102
7.2 Methods . . . 103
7.2.1 Participants . . . 104
7.2.2 Study Design . . . 104
7.2.3 Survey Response and Patient-Reported Outcomes . . . 104
7.2.4 Sensor and EMA Data Collection . . . 105
7.2.5 Data Processing and Machine Learning . . . 106
7.3 Results . . . 112
7.4 Discussion . . . 117
7.5 Conclusion . . . 119
7.6 Addressing the Curse of Dimensionality . . . 120
7.6.1 W.r.t. multiple co-morbidities (C2) . . . 120
7.6.2 W.r.t. diversity in patient characteristics (C3) . . . 120

8 Thesis Conclusion and Future Work 121
8.1 Key Takeaways from the Thesis . . . 121
8.2 Future Work . . . 123
8.2.1 Increasing sample sizes . . . 123
8.2.2 Deploying predictive models and studying their acceptability . . . 123
8.2.3 Explore the feasibility and utility phenotyping patients based on multiple co-morbidities . . . 123
8.2.4 ML-driven personalized interventions and user-driven self-experimentation 124

References 126

Chapter 1
Introduction

The global prevalence and burden of mental illnesses is huge and is consistently increasing [251]. Nearly one in five U.S. adults live with a mental illness at any given time [156]. Further, mental illnesses disproportionately affect certain populations such as college students (one in three prevalence) [122] and people with chronic medical conditions like multiple sclerosis [187, 35, 240]. The coronavirus disease 2019 (COVID-19) pandemic and the ensuing response (*e.g.,* lockdown and social distancing) have also been reported to negatively impact mental health, leading to an increase in the prevalence of mental illnesses [54, 184, 78, 258, 132, 179, 137]. Mental health disorders are linked with lower productivity, performance, and participation in schools, universities, and workplaces [104, 115, 75, 63]. They are also the leading cause of disability [87] and suicide [126, 127, 174]. As a result, the prevention, early diagnosis, and treatment of mental health disorders has become a public health priority worldwide [269].

Several treatments for mental health disorders exist (*e.g.* psychotherapy, medication), and are effective. However, the vast majority of people with mental health disorders never seek help due to stigma or lack of awareness [9, 47, 74, 155]. Some of those who seek help receive inadequate care due to inequitable access to care and under-resourced healthcare systems [223, 29, 49]. Further, finding the right treatment for a specific person is typically a time consuming and inefficient process, as most interventions are based on studies that find the best treatment for a "typical" patient, rather than tailoring interventions to the patient's genes, environment and lifestyle. This causes many patients to prematurely drop out of treatment [180, 61]. Hence, there is a need to develop digital tools that increase access to mental health care while making the treatment of mental health disorders more precise by using data-driven insights and predictions to aid diagnosis, monitoring, prevention, and treatment of mental health disorders.

Precision health is an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle. In the case of mental health, factors such as a person's symptoms, behaviors, environment, and social context can often play a larger role than genetics [22, 214], and can be captured by technological devices and our interactions with them. Over the last decade, mobile phones and wearable devices have become ubiquitous. Built-in sensors on mobile phones and wearable fitness trackers allow us to passively and unobtrusively collect information such as location, communication, environment, phone usage, physical activity, and sleep. Previous work has found such data to be related to mental health disorders like depression [65, 124, 216, 266, 217, 20]. There has also been some limited work on predicting mental health outcomes using such data [216, 32, 262, 80, 267]. Further, there have been recent developments in mobile apps [215, 90, 150],

computerized psycho-educational and psycho-therapeutic interventions [10, 53, 209, 271], or chat-based programs [84, 134, 227] to complement, and expand access to, psychotherapy.

Despite these recent technological developments, there are many challenges in precision medicine that have yet not been fully addressed. Previous research has estimated that the patient's behavior and lifestyle is responsible for 40% of their health, genetics is responsible for 30% of their health, environment and social factors are responsible for 20% of their health, and access to healthcare is responsible for 10% of their health [151]. Hence, in order to be precise, precision medicine models must generate, link, and learn from a variety of data sources from different kinds of patients who may have one or more medical conditions. A patient's health state can be characterized by a multitude of signals including medical imaging, clinical variables, genome sequencing, conversations between clinicians and patients, and continuous signals from personal devices such as smartphones and wearables, among others [21]. While such high dimensional data creates new opportunities for higher-precision diagnosis, prognosis, and tracking, it also introduces one of the biggest challenges in precision medicine *i.e. the curse of dimensionality*, which makes it harder to develop robust algorithms that generalize to real-world scenarios. This thesis focuses on developing and presenting novel methods that address the curse of dimensionality challenge with respect to feature space, multiple outcomes stemming from co-morbid medical conditions, and diversity in patient contexts and characteristics.

1.1 The Curse of Dimensionality Challenge

The curse of dimensionality, first introduced by Bellman [19], indicates that the number of samples needed to estimate an arbitrary function with a given level of accuracy grows exponentially with respect to the number of input variables or features (*i.e.*, dimensionality) of the function. When the number of features increases greatly without increasing the sample size, the dataset is left with several "blind spots" which are contiguous regions of feature space without any observations. These blind spots can result in highly variable models (*e.g.*, vastly different selected features) and highly variable estimates of true model performance, across different subsamples of the same dataset. Further, datasets containing a very large number of features generally have multiple correlated features, which lead to inaccurate model estimates and cause certain important features to be incorrectly excluded during feature selection (the multicollinearity problem) [254]. Hence, algorithms that are developed for diagnosis, monitoring, and treatment of diseases, must take steps to mitigate the curse of dimensionality problem.

To address the curse of dimensionality problem in my thesis, I divided the problem into three sub-problems, and then carried out studies that addressed one or more of these sub-problems. In this section, I explain these sub-problems and give a very brief overview of how I addressed them in my studies.

1.1.1 The curse of dimensionality in the feature space (C1)

Precision health involves combining billions of datapoints per individual (covering genomic, environmental, medical and lifestyle data), which can lead to models with low reliability and stability, and make insights harder to interpret [70, 25, 111, 272]. Similar to genomic data, user-generated behavioral data often have many dimensions *i.e.*, the number of features is typically much larger than the sample size, which depends on the number of participants and the number of outcomes collected per participant. This is because mental health outcomes are usually collected using well-validated self-reported questionnaires, and so, collecting more data is often infeasible due to the increased survey burden on the user. Further, previous work has shown that behavioral features aggregated over different time slices should be included in our analysis as they can encode very different information. For example, Chow et al. [46] found no relationship between depression and time spent at home during 4-hour time windows, but they found that people who are more depressed tend to spend more time at home between 10:00 AM and 6:00 PM. Hence, including behavioral features from different time slices is important. However, including behavioral features from multiple time slices can greatly increase the size of the feature space. Hence, during my thesis work, I used a combination of feature space decomposition during feature selection and late sensor fusion to integrate data from multiple modalities while reducing the dimensionality of the feature space. This approach outperformed off-the-shelf machine learning approaches that don't decompose the feature space by 18%.

1.1.2 The curse of dimensionality with respect to multiple co-morbid outcomes (C2)

Multimorbidity is defined as the co-occurrence of two or more chronic conditions [168]. Mental health conditions are frequently chronic and co-morbid with other mental health and medical conditions. In fact, the presence of one or more medical conditions in people with a mental health disorder is reported to be close to 70% [94]. For example, over 85% of people with depression also have anxiety, and vice-versa [95]. Hence, the presence of multimorbidities in patients is the rule, rather than the exception. When diagnosing or treating a condition, it is important to consider other co-morbid conditions, as the co-morbid

conditions could be the most important factor during diagnosis or treatment of the primary condition or could be totally unrelated. Despite the prevalence and importance of multimorbidities, multimorbidities are frequently ignored in clinical trials and research studies. In fact, in many studies, people with co-morbid conditions are explicitly excluded [189]. Precision medicine studies also similarly neglect multimorbities [181, 230]. In this thesis, I have incorporated multimorbities into my studies in two ways: (i) evaluating whether my computational approaches generalize to outcomes that are frequently co-morbid with the primary outcome, (ii) leveraging co-morbid outcomes to compute one final outcome that can be used to generate predictions or data-driven insights.

1.1.3 The curse of dimensionality presented by the diversity in patient characteristics (C3)

Multimodal mental health sensing studies typically capture and analyze data passively collected in the wild *i.e.*, while the participants are engaged in their daily routine or regular interaction with apps on their devices. Hence, these studies are essentially uncontrolled natural experiments in which we cannot control for various confounding variables such as behaviors resulting from a person's history, occupation, demographics, environment, and other contextual variables, that may effect the relationship between the in-the-moment behavioral features and mental health outcomes. These patient characteristics can be static such as certain demographics or dynamic such as behavioral history which contextualizes current behaviors. For example, in a study with college students, the time spent at home was negatively correlated with well-being [27, 26], whereas, in a study with older cancer patients, the time spent at home was positively correlated with well-being [31]. Boukhechba et al. [25] talks about "how this inverted effect may be the result of the dramatic differences between college students and breast cancer patients (with an average age of 60 years) undergoing active treatment. For example, while college students are generally expected to frequently engage in social interactions that take place outside their home environment, recently diagnosed breast cancer patients may relish the opportunity to spend time with family members and friends at home, rather than receiving medical care at a hospital". In my thesis, I explored and developed statistical and machine learning approaches that allow us to analyze data and build models while accounting for important confounding contextual variables. I did this by – (i) accounting for inter-participant differences in situations where we had multiple samples from the same participant, (ii) clustering patients into sub-groups that share similar characteristics and accounting for their membership in a subgroup during the modeling stage. Accounting for important confounding contextual variables in this

way enabled the generation of data-driven insights revealing effective support strategies or interventions for patients in different contexts, and (iii) using behavioral features from the previous time period (context features) and behavioral features from the current time period (action features) to predict health outcomes at frequent intervals over time.

1.2 Thesis Outline and Contributions

So far, I've completed the following 4 studies:

- Study 1: Detecting Depression and Loneliness In College Students (chapter 3)
- Study 2: Forecasting End of Semester Depression In College Students (chapter 4)
- Study 3: Predicting the Mental Health of People with Multiple Sclerosis during the COVID-19 Stay-at-Home Period (chapter 5)
- Study 4: Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention (chapter 6)
- Study 5: Predicting Periodically Assessed Multiple Sclerosis Outcomes Using Passively Sensed Behaviors and Ecological Momentary Assessments (chapter 7)

Most multimodal behavioral sensing studies are designed based on a general study framework. Below, I describe this study framework, and discuss how the above studies fit into that framework. Understanding the study framework is important as study design choices often inform the curse of dimensionality problems we will face and thus need to address. In this section, I also discuss the contributions through the 5 completed studies. These contributions are primarily related to how different aspects of the curse of dimensionality problem were addressed in these studies.

1.2.1 Study Framework

Figure 1.1 shows the study framework for multimodal behavioral sensing studies for precision health, and how the 5 studies above fit into it.

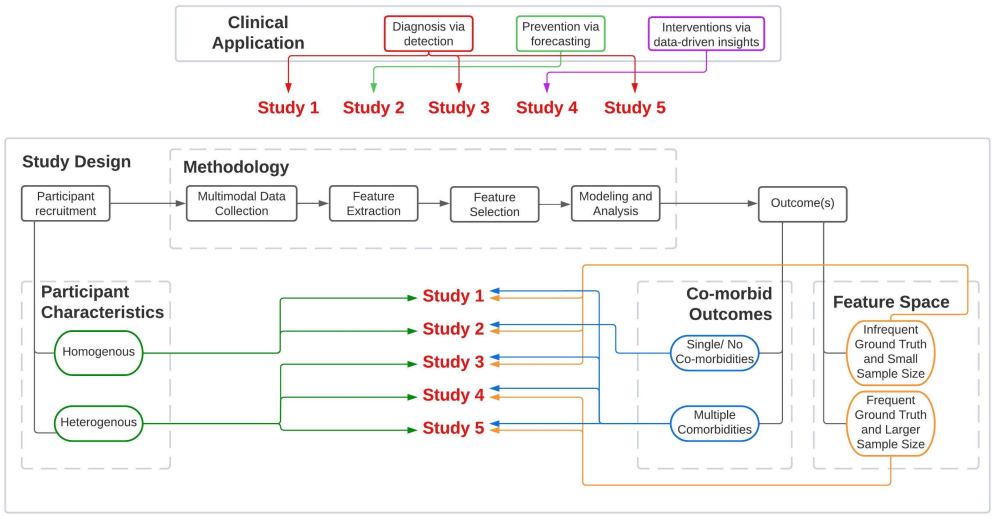


Figure 1.1: Multimodal Behavioral Sensing for Precision Mental Health Care – Study Framework

Clinical Applications

Multimodal behavioral sensing studies can focus on a wide variety of clinical applications. In my work, I focus on 3 clinical applications, which are as follows:

- **Diagnosis or monitoring via detection:** Using multimodal behavioral data to predict health outcomes directly associated with that data. For example, predicting depression at the end of two weeks using data from that two week period.
- **Prevention via forecasting:** Using multimodal behavioral data to predict health outcomes days or weeks into the future using data from the past. For example, predicting depression at the end of the month using data from the first week of the month.
- **Interventions via data-driven insights:** Analyzing existing digital health interventions to derive data-driven insights that can then be used to improve the intervention with the goal of improving outcomes.

Other clinical applications such as personalizing interventions to the user's sensed context with the goal of improving outcomes are outside the scope of this thesis. The clinical application impacts all aspects of the study design process and methodology.

Patient characteristics

While deciding on a study design, we typically start out by identifying the target population we'd like to recruit from. In my work, the target populations can be loosely divided into two groups – groups of patients with homogenous routines and behaviors (*e.g.* college students living on-campus in the same year of study and semester), and heterogenous (*e.g.* patients with multiple sclerosis with different occupations, lifestyles, and ages, recruited over time on a rolling basis). Patient characteristics greatly influences feature extraction. For example, for a homogenous population of college students or employees of the same company, we can extract behavioral features related to different locations on their university campus or workplace with reasonable certainty.

Outcomes

The chosen outcomes for analysis can be single outcomes representing one health condition, or multiple outcomes representing multiple co-morbid conditions that frequently co-occur in patients. Predicting multiple co-morbid outcomes can give a more holistic view of the patient's health. An objective assessment of different aspects of the patient's health can help clinicians better diagnose and treat conditions. For example, the symptoms of depression and anxiety frequently co-occur and effect each other. Clinicians may treat someone with high depression and high anxiety differently from someone with high depression and low anxiety. While both these groups will benefit from medication and/or therapeutic interventions for depression, the former group may not show significant improvement in depression until their anxiety is also treated. Further, predicting multiple outcomes using the same analysis pipeline allows us to assess the generalizability of our approach to other outcomes. There is also value in combining multiple outcomes, for example, to identify one target outcome to optimize for in digital interventions.

Feature Space

The frequency of the ground truth collected affects the dimensionality of the feature space. When the ground truth collected is infrequent, we end up with smaller sample sizes and higher dimensional spaces as we have too many behavioral data for a few patient observations. When the ground truth collected is frequent, we end up with larger sample sizes and possibly lower dimensional spaces as the ratio of features to patient observations is smaller than the former case. The dimensionality of the feature space greatly impacts the methodology, as high dimensional spaces often require more complex feature selection approaches in order to be create robust and stable models.

1.2.2 Addressing the Curse of Dimensionality Challenge

Below, I discuss how the 5 studies I completed during my PhD addressed the curse of dimensionality challenge w.r.t. the feature space (C1), multiple co-morbid outcomes (C2), and diversity in patient characteristics (C3).

S1: Detecting Depression and Loneliness In College Students (chapter 3)

In study 1, we present a machine learning approach that uses data from smartphones and fitness trackers of 138 college students to identify students that experienced depressive symptoms at the end of the semester and students whose depressive symptoms worsened over the semester. We also applied the same pipeline to detecting loneliness in college students. Our approach allowed us to detect the presence of post-semester depressive symptoms with an accuracy of 85.7%, and the change in depressive symptom severity with an accuracy of 85.4% [41]. For loneliness, our approach achieved an accuracy of 80.2% in detecting the binary level of loneliness and an accuracy of 88.4% in detecting change in the loneliness level [66].

Challenges it addresses and how:

Our novel approach is a feature extraction technique that allows us to select meaningful features indicative of depressive symptoms from longitudinal data by integrating data from multiple time slices and sensors while decomposing and reducing the dimensionality of the feature space (addresses C1). It allowed us to detect the presence of post-semester depressive symptoms with an accuracy of 85.7%, whereas simpler approaches that don't reduce the dimensionality of the feature space achieve an accuracy of only 64.6%. Depression and loneliness are frequently co-morbid. By detecting both depression and loneliness with high accuracy, we are able to show that our approach can be used to identify multiple co-morbid conditions (addresses C2).

S2: Forecasting End of Semester Depression In College Students (chapter 4)

In study 2, we used the same pipeline to forecast or predict post-semester and change in depression before the end of the semester. We were able to predict these outcomes with an accuracy of >80%, 11-15 weeks before the end of the semester, allowing ample time for preemptive interventions [41].

Challenges it addresses and how:

To predict depression at the end of a specific week x of the semester, we had to train and test multiple models containing features from week 0 to week x. Models containing data from fewer weeks had a smaller feature space, whereas models containing data from more weeks had a larger feature space, as the features from different weeks were concatenated before being input into the model. By being able to achieve reasonable model performance across multiple weeks and by achieving a high accuracy 11-15 weeks before the end of the semester, we show that our approach is able to handle feature spaces of different sizes (addresses C1).

S3: Predicting the Mental Health of People with Multiple Sclerosis during the COVID-19 Stay-at-Home Period (chapter 5)

In study 3, we present a machine learning approach leveraging passive sensor data from smartphones and fitness trackers of 56 multiple sclerosis (MS) patients to predict their health outcomes in a natural experiment during a state-mandated "stay-at-home" period due to a global pandemic. The algorithm detects depression with an accuracy of 82.5% (65% improvement over baseline; f1-score: 0.84), high global MS symptom burden with an accuracy of 90% (39% improvement over baseline; f1-score: 0.93), severe fatigue with an accuracy of 75.5% (22% improvement over baseline; f1-score: 0.80), and poor sleep quality with an accuracy of 84% (28% improvement over baseline; f1-score: 0.84) [42].

Challenges it addresses and how:

From a methodological standpoint, we first extract features that capture behavioral changes due to the "stay-at-home" order. Then, we adapt and apply the algorithm used in studies 1 and 2 to these behavioral change features to predict 4 health problems frequently co-morbid in patients with MS – the presence of depression, high global MS symptom burden, severe fatigue, and poor sleep quality during the "stay-at-home" period. By predicting multiple outcomes with a high accuracy, we show that our approach generalizes to multiple co-morbid outcomes in patients with MS, thus addressing C2. Further, we achieve these results by decomposing and reducing the dimensionality of the feature space during feature selection, thus addressing C1.

S4: Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention (chapter 6)

In study 4, we present an analysis of 234,735 messages from supporters to clients on an online mental health intervention platform to discover how different support strategies correlate with clinical outcomes over time. We describe our machine learning methods for:

(i) clustering supporters based on client outcomes; (ii) extracting and analyzing linguistic features from supporter messages; and (iii) identifying context-specific patterns of support. Our findings indicate that concrete, positive and supportive feedback from supporters that reference social behaviors are strongly associated with better outcomes; and show how their importance varies dependent on different client situations [43].

Challenges it addresses and how:

For clustering supporters, we present a novel method that takes into account clients' severity of depression and anxiety, which are often co-morbid. In doing so, we are able to assess the effectiveness of the intervention based on two co-morbid conditions combined together as one outcome, thus addressing C2. After identifying support strategies that are consistently associated with better clinical outcomes, we present a method for identifying context-specific patterns of support. That is, we explore how considering a combination of multiple client context variables might change the relationship between support strategies and client outcomes, thus addressing C3.

S5: Predicting Periodically Assessed Multiple Sclerosis Outcomes Using Passively Sensed Behaviors and Ecological Momentary Assessments (chapter 7)

In study 5, I predicted periodically assessed multiple sclerosis (MS) related outcomes and associated mental health outcomes in patients with multiple sclerosis (pwMS) using passively sensed behaviors and ecological momentary assessments (EMAs). By predicting periodically assessed outcomes, this study enables passive and low cost longitudinal monitoring of clinically relevant health outcomes for pwMS.

For each participant, the algorithm detects depression every two weeks with an accuracy of 80.6% (35.5% improvement over baseline; f1-score: 0.76). It also detects the following outcomes every 4 weeks – high global MS symptom burden with an accuracy of 77.3% (51.3% improvement over baseline; f1-score: 0.77), severe fatigue with an accuracy of 73.8% (45.0% improvement over baseline; f1-score: 0.74), and poor sleep quality with an accuracy of 72.0% (28.1% improvement over baseline; f1-score: 0.70).

Challenges it addresses and how:

In study 4, I addressed the curse of dimensionality challenge presented by diversity in patient characteristics (C3), but I did not do so while predicting mental health outcomes from passively smartphone and wearable sensor data. Hence, in study 5, in order to predict health outcomes at the end of each time period, I focused on leveraging behavioral features from the previous time period (context features) in addition to behavioral features from the

current time period (action features) to help the model better capture the patients' in-situ behaviors in the context of the behavioral characteristics they expressed in their history or contextual behaviors (C3). Further, by predicting multiple outcomes at short periodic intervals with a high accuracy, we show that our approach generalizes to multiple co-morbid outcomes in patients with MS, thus addressing C2.

1.2.3 Contributions Across All Studies

My PhD thesis has made the following contributions:

1. Presented a feature selection method that mitigates the curse of dimensionality in the feature space by decomposing and iterative reducing the feature space during feature selection. In the future, leveraging my method or other approaches based on feature space decomposition and/or late fusion can help researchers obtain good model performance while preserving valuable information by including features from many different temporal slices. – Completed during studies 1, 2, and 3.

2. Demonstrated the generalizability of this approach in detecting depression, change in depression, and loneliness, as well as forecasting these outcomes several weeks in advance. This shows that the feature space decomposition and late fusion techniques used in my work perform well for multiple outcomes, bolstering our confidence in the validity of our approach. – Completed during studies 1 and 2.

3. Demonstrated that behavioral changes resulting from the stay-at-home mandates during the pandemic are predictive of health outcomes during the stay-at-home period for patients with multiple sclerosis. This shows that behavioral changes occurring upon the onset of a natural disaster can be used to predict a patient's mental health during the disaster. In the future, researchers should consider behavioral change instead of depending solely on current behaviors to predict mental health outcomes. Further, this also shows that it is possible to use sensor data to predict mental health outcomes in patients with an underlying neurological condition. Also, demonstrated the generalizability of this approach to predict multiple comorbid health outcomes. – Completed during study 3.

4. Demonstrated how we can categorize supporters or identify patient phenotypes based on multiple co-morbid outcomes, thereby mitigating the curse of dimensionality with respect to multimorbidities. Categorize supporters or identify patient phenotypes helps us combine multiple outcomes into one target outcome. In the future, researchers can

use this to simplify analysis when deriving data-driven insights to assess the effectiveness of an intervention or to understand the detection model. – Partially completed during study 4. Potential contribution for study 6.

5. Presented a method that visualizes and identifies support strategies that work best in an online mental health intervention for patients in a specific context or situation. The results showed that looking at the relationship between current behaviors and outcomes alone isn't sufficient, as patient contexts can change the relationship between behaviors and outcomes. For example, when analyzing the relationship between supporter behaviors and client outcomes, we found shorter messages to correlate with better outcomes. However, when we also factored in patient contexts into our analysis, we found that while shorter messages correlate with better outcomes for engaged clients, longer messages correlate with better outcomes when the client has very low engagement. Hence, while analyzing the relationship between behaviors and outcomes can reveal interesting general insights, researchers should also factor in patient contexts to derive more precise or personalized insights, and they may apply our method to their work to do so. Further, the insights derived from this work can be used to improve human supporter training, and can be incorporated into a system that helps human supporters by recommending support strategies for a participant in a specific context. – Completed during study 4.

6. Demonstrated that using features capturing the patient's past behaviors (context features) in addition to their recent behaviors (action features) improves model performance for the longitudinal monitoring of some health outcomes such as fatigue. This highlights the importance of accounting for diversity in patient histories, characteristics, or context when predicting health outcomes. Also, demonstrated the generalizability of this approach to enable periodic monitoring of multiple comorbid health outcomes. – Completed during study 5.

7. Demonstrated that adding ecological momentary assessments to models using passively sensed behaviors to predict health outcomes does not significantly improve performance. While adding pre-survey EMA did improve the prediction of biweekly depression, the improvement was small and only requires two days of EMA from a 14 day period. – Completed during study 5.

Chapter 2
Related Work

The related work can be divided into 4 sections: (i) Exploring the relationship between behavioral features and mental health, (ii) Detecting mental health outcomes, (iii) Detecting health outcomes in patients in multiple sclerosis, and (iv) Understanding large-scale mental health behavioral data.

2.1 Exploring the relationship between behavioral features and mental health

In this section, we describe previous research that has explored the statistical relationships between behavioral features from smartphones and wearable devices, and the two mental health outcomes that are the focus of this thesis – depression and loneliness.

2.1.1 Depression

The Diagnostic and Statistical Manual of mental disorders (DSM-5) [7] describes several depressive disorders, most prevalent of which are Major Depressive Disorder (MDD) and Persistent Depressive Disorder (PDD). People with these disorders experience similar symptoms over different periods of time (*e.g.,* at least 2 weeks for MDD and at least 2 years for PDD). These symptoms include *depressed mood, diminished interest or pleasure in almost all activities, sleep disturbances* such as insomnia or hypersomnia, *psychomotor agitation or retardation, fatigue or loss of energy, feelings of worthlessness or guilt, diminished concentration,* and *recurrent thoughts of suicide.* Many of these symptoms manifest as verbal, non-verbal or daily behaviors [64] and can be passively sensed with limited user involvement.

Automated techniques for identifying depressive symptoms can be grouped based on the type of behavioral symptoms they sense – verbal [56, 212], non-verbal [48, 243, 5, 4, 121, 226, 225, 3, 229], or daily. Our approach focuses on daily behaviors that can be sensed using smartphones and fitness trackers, which allows for depression detection and longitudinal symptom monitoring.

Daily behaviors are related to communication, movement patterns, smartphone use, sleep, and physical activity, which can be sensed using sensors embedded in smartphones and fitness trackers. Features indicative of daily behaviors can be extracted from sensor data to capture behavioral symptoms of depression.

Doryab *et al.* [65] explored detection of behavior change in people with major depression from smartphone data. Their pilot study of three participants (2 female and 1 male) over 4

months found an inverse relationship between the number of outgoing calls and depression scores over time with the male patient, and a direct relationship between the number and duration of outgoing calls and depression scores over time with the female patients. A study with 216 college students [124] demonstrated a direct relationship between Internet use and depression, *i.e.*, students with depressive symptoms used the Internet significantly more than non-depressed students. They also switched more frequently between email, chat rooms, social media, video watching, and games. Saeb *et al.* [216] explored the relationship between depression severity score and mobile data including location traces and phone usage in 28 adults over a two-week period and found significant correlations between participants' depression scores (from a standardized assessment) and Location features such as location variation, regularity in movement over days ("circadian movement") and evenness in time spent across locations ("location entropy"). They also found significant correlations between phone usage features such as usage duration and frequency. They replicated the same results using Location features on another dataset [266] containing data from 48 college students over a 10-week period [217]. This dataset was originally collected as part of the StudentLife study at Dartmouth [266] which revealed significant correlations between depression scores and sleep duration, conversation duration, as well as frequency and number of collocations. Further analysis of the dataset showed significant relationships between change in depression scores and features such as sleep duration, speech duration, and geospatial activity (from locations and WiFi scans) [20].

2.1.2 Loneliness

A few studies have explored the relationships between single behavioral features, such as level of physical activity, mobility, social interactions, and loneliness [20, 266, 88]. Wang et al [266] analyzed smartphone data collected from 40 students over a spring semester and found negative correlations between loneliness and activity duration for day and evening times, traveled distance, and indoor mobility during the day. A related study from the same group found statistically significant correlations (P¡.01) between kinesthetic activities and change in loneliness but no relationship between loneliness and sleep duration, geospatial activity, or speech duration [20]. Gao et al [88] found that people with higher levels of loneliness made or received fewer phone calls and used certain types of apps, such as health and fitness, social media, and Web browsing, more frequently than those with low levels of loneliness. Our data mining approach, in addition to providing similar behavioral features to those reported by Wang et al [266], presents an innovative method for extracting the combined behavioral patterns in our participant population. For example, we can observe that compared with students with a low level of loneliness, students with a high level of

loneliness unlock their phones in different time segments during weekends, spend less time off-campus during evening hours on weekends, and socialize less during evening hours on weekdays.

2.2 Detecting mental health outcomes

In this section, we describe previous research that has applied machine learning to behavioral features from smartphones and wearable devices in order to predict the two mental health outcomes that are the focus of this thesis – depression and loneliness.

2.2.1 Depression

The statistical relationships described in section 2.1.1 suggests that machine learning models could be used to detect depression. As summarized in Table 2.1, existing work has made important strides in this domain. Saeb *et al.* [216] were able to achieve a leave-one-participant-out accuracy of 86.5% for distinguishing between participants with depressive symptoms and those without depressive symptoms. However, they collected data from 28 adults over a short two-week period and used only one feature from the Location sensor in their machine learning model. Further, cross-validation was not used for feature selection, thus reducing the generalizability of their model. Canzian and Musolesi [32] trained personal models for each of their 28 users using features related to mobility patterns from location data, to detect periods in which users experience an unusual depressed mood. Their models achieved high sensitivity and specificity values, which means that for most of the users, they were able to detect periods of depressed mood (related to sensitivity) while generating few false alarms (related to specificity). They also extended their approach to predicting depressive symptom severity 1-14 days in advance. Wahle *et al.* [262] detect biweekly depression in 36 participants over 2-10 weeks using a very limited set of features from location, physical activity, phone usage, calls, texts, and WiFi scans, and achieve an accuracy of 61.5%. Farhan *et al.* [80] detect biweekly depression in 79 college-age participants over 7-8 months using location data as input to their model and clinical evaluations as their ground truth, and achieve a F1 score of 0.82. Wang *et al.* [267] detect depression on a week-by-week basis using features from smartphone and wearable data as input and weekly subjective assessments as ground truth from 68 undergraduates over two 9-week terms, and achieved 81.5% recall and 69.1% precision. In addition to some of the above features, they used campus-specific features such as time spent in dorm and time spent at study places. Mehrotra and

Musolesi [153] used autoencoders for automatically extracting features from the raw GPS data, and achieved better results than "hand-crafted" location features.

All of this earlier work has heavily relied on frequent assessment of depression (weekly or biweekly). In real world situations, the mental health status of individual people is often unknown which makes the above mentioned approaches less usable and realistic. In this thesis (study 1 described in chapter 3), we address this specific issue through developing a machine learning pipeline capable of detecting depression without frequent ground truth data. Further, while subjective measures for depression are evaluated for their "sensitivity to change" [165, 18, 130], the same has not been done for depression models based on passive sensing. That is, we do not know if existing ML methods for depression detection work well because they capture transient depressive symptoms or latent characteristics known to increase the risk of depression (*e.g.* early major life events [164], thought patterns [249]). In addition to detecting post-semester depression, we also detect change in depression, thereby resulting in ML models that capture transient depressive symptoms.

Finally, predicting depression in advance is a very useful task as it can allow us to intervene *before* the onset or worsening of symptoms. Subjective measures for depression are designed to measure symptom severity at a particular time by directly asking the participant about their symptoms. Passive sensing models, however, have the potential to do more than that, as they may be able to capture early behavioral signs of depression that even the participant may not be aware of. Other than the study in [32] which attempted to predict depression 0-2 weeks in advance, we are unaware of any research in early prediction of depression. With our approach, in study 2 (described in chapter 4), we can predict the post-semester depression with an accuracy of $> 80\%$ as early as week 5 of the 16 weeks-long semester, giving clinicians a larger window of time for interventions.

Table 2.1: Related Work for Depression Detection. For study 1 (last row), note that "all" results are obtained using all features, while "best" results are obtained via a feature ablation study (see chapter 3)

Reference	**Part.**	**Duration**	**Sensors**	**Outcome**	**Accuracy**	**Other Metrics**
[216]	28 adults	2 wks	Location (only 1 feat.)	Dep. at end of 2 wks	86.5%	
[32]	28 adults	avg. 10 wks/user	Location	Detecting dep. over different periods of time, and predicting dep. 1-14 days in advance.		Sensitivity= 0.71/Specificity= 0.87
[262]	36 people	Variable	Smartphone sensors	Dep. biweekly	61.5%	F1=0.62
[80]	79 col. age	7-8 mos	Location	Clinical dep. biweekly		F1=0.82
[267]	68 col. studs	18 wks	Smartphone sensors (light, GPS, accelerometer, microphone, screen status) & heart rate sensor	Dep. weekly		F1=0.75
[153]	28 adults	avg. 10 wks/user	Location	Detecting dep. over different periods of time. No early prediction.		Sensitivity= 0.77/Specificity= 0.91
Study 1 (see: chapter 3)	**138 col. studs**	**16 wks**	**Smartphone sensors (bluetooth, calls, GPS, microphone, screen status) & wear. fitness tracker (steps, sleep**	Post-semester dep.	85.7% (best); 82.3% (all)	F1=0.82 (best); 0.78 (all)
				Change in dep.	85.4% (best); 75.9% (all)	F1=0.80 (best); 0.67 (all)
				Explored predicting the above 2 outcomes 1-15 weeks in advance. Results: > 80% accuracy 11-15 weeks ahead of the end of semester, for both prediction problems.		

.

2.2.2 Loneliness

Pulekar et al [198] analyzed data logs of social interactions, communication, and smartphone activity collected from 9 college students over to 2 weeks to detect loneliness and its relationship with personality traits. The study reports 90% accuracy in classifying loneliness using the smartphone features that were mostly correlated with the loneliness score. However, the small sample size, the short duration of the data collection phase, and missing details in the machine learning approach, especially the classification evaluation, make the results difficult to generalize and build on. Sanchez et al [220] used machine learning to infer the level of loneliness in 12 older adults who used a mobile app for one week. Call logs and global positioning system (GPS) coordinates were collected from the phones. A total of 4 models for family loneliness, spousal loneliness, social loneliness, and existential crisis were built with a reported accuracy of 91.6%, 83.3%, 66.6%, and 83.3%, respectively. However, similar to the results of the study by Pulekar et al, these results may fail to generalize because of the small sample and short duration of data collection.

2.3 Detecting health outcomes in patients in multiple sclerosis

Previous work on using using passive sensing in pwMS is very sparse. Most prior work has focused on the feasibility of using smartphones and wearable devices to sense behaviors in this population. However, a few studies have found preliminary correlations between passively sensed behaviors and MS outcomes. For example, Newland *et al.* explored real-time depth sensors at home to identify gait disturbance and falls in 21 MS patients [169]. Other studies reported correlations between passively sensed physical activity and disability worsening in pwMS [231, 23, 245]. Chitnis *et al.* examined the gait, mobility, and sleep of 25 pwMS over 8 weeks using sensors mounted on their wrist, ankle and sternum, and reported correlations among gait-related features (*e.g.* turn angle, maximum angular velocity), sleep and activity, and disability outcomes [45].

Previous work on *predicting health outcomes for pwMS using passively sensed behaviors* also very limited. Tong *et al.* used passively sensed sleep and activity data collected from 198 pwMS over 6 months to predict their fatigue severity and overall health scores, and achieved good performance in line with acceptable instrument errors [255]. To our knowledge, this thesis (study 3 described in 5) is the first to use passively sensed behavior changes to predict multiple inter-related clinically relevant health outcomes in MS, including depression, disability, fatigue, and sleep quality. While several studies used passively sensed data from

the general population to report behavioral changes during the COVID-19 pandemic [246, 182, 190, 113], our work provides the first real-world evidence of potential clinical utility of passively sensed behavioral changes to predict health outcomes during the unique stay-at-home period in a population with a chronic neurological disorder and complex health needs.

Further, to our knowledge, this thesis (study 5 described in 7) is also the first to use passively sensed behavior features to predict multiple inter-related clinically relevant health outcomes in MS, including depression, disability, fatigue, and sleep quality, repeatedly at periodic intervals with the goal of enabling health monitoring over time.

2.4 Understanding large-scale mental health behavioral data

Recent years have seen a growth in research exploring ML for mental health behavior data as captured by wearable health trackers [90]; mobile phones [33, 57, 67, 266, 274, 66]; social media [36, 125, 171, 186, 218]; or electronic healthcare records (EHR) [2, 257]. For the wealth of data that can be collected by these technologies, the fields of ML and data mining provide computational methods that can help improve our understanding of human behaviors and predicting or optimizing clinical outcomes [120]. Frequently applied methods that are particularly relevant to the approach taken in this paper are: clustering, text-mining and association rule mining (ARM).

2.4.1 Clustering and Text Mining in (Large-Scale) Mental Health Data

Clustering is an unsupervised ML technique to identify behavioral patterns in data through commonalities in each data piece; it is often used to identify features in unstructured data (*e.g.* [2, 38, 186]). For study 4 (described in chapter 6), two types of work are particularly relevant. Firstly, to better understand the behaviors of therapists engaged in text-based (SMS) counseling, Althoff et al. [6] clustered therapists, based on client outcomes, into groups of 'more' and 'less' successful counselors; and then compared how their communications differed using linguistic analysis. We followed a similar approach. Secondly, to identify support behaviors in thousands of anonymous supporter messages, we employ text-mining, which uses natural language processing (NLP) techniques to extract linguistic features or topics from large-scale text. In mental health, text-mining has been used to better understand

discussion topics in online mental health communities [171, 186]; to study mental expressions [218], the receipt of social support [232], or cognitive change [197]. Few works seek to specifically aid moderators of online support communities in their work practices, *e.g.* by identifying helpful and unhelpful comments [36, 125]. Outside of social media, text mining is used to predict suicide risks from SMS messages [173], suicide notes [192], and EHRs [2] to aid care provision. By extending this body of work in study 4, we seek to better understand mental health support through a linguistic analysis of supporter messages as part of an iCBT intervention.

2.4.2 Using Association Rules for Behavioral Pattern Mining

Similar to clustering, association rule mining (ARM) is a common data mining technique for extracting behavioral patterns in data (*e.g.* [37, 66, 183, 274]). Here, the focus is on discovering interesting relations between variables in large data sets such as how patterns of certain data characteristics (*e.g.* client opinions, types of symptoms, demographics) relate to desirable outcomes (*e.g.* help-seeking behaviors, clinical score) [92, 183, 278]. In study 4, we show how we adapted an ARM algorithm to extract patterns of context-specific best practices of support.

Chapter 3

S1: Detecting Depression and Loneliness In College Students

3.1 Introduction

Depression is a common and serious mental health disorder that is especially prevalent among college students. In 2013, the percentage of college students in the United States that reported having difficulty functioning in the last 12 months due to depression was over 33%. Depression has been found to affect academic participation, productivity, and performance [104, 115], and may double the likelihood of dropping out from college [75]. Further, depression is the most common disorder among people with suicidal behaviors [127, 126, 174]. It is estimated that approximately 11.2% of undergraduates seriously considered suicide and 2.1% attempted suicide in 2015-2016.

Although treatment for depression is effective and includes a variety of methods, such as psychotherapy and medication, a large number of affected students do not seek treatment [74, 96]. Commonly reported barriers to seeking treatment include the belief that stress is a normal part of student life and treatment is not needed. Furthermore, students may not be aware that they are experiencing not only stress, but also depression [55]. Tools used to monitor the severity of depressive symptoms rely on periodic self-reports that are subjective and if administered too often may reduce compliance. Hence, there is a need to develop more efficient methods to monitor and identify changes in depressive symptoms in college students, and predict future depressive episodes.

Built-in sensors on mobile phones and wearable fitness trackers allow us to passively and unobtrusively collect information such as location, communication, environment, phone usage, physical activity, and sleep. Previous work has shown that such information is linked to depressive symptoms, such as social isolation and sleep disturbances [7]. Measuring the severity of depressive symptoms using such sensors could enable continuous depression detection, prediction before onset, and longitudinal symptom monitoring in-the-wild. Ultimately, it creates the potential for technology-mediated real-time interventions that support the diagnosis, treatment, and prevention of depression. As a result, over the past few years, researchers have conducted several studies that use statistics to understand the relationship between sensor data from phones and wearables, and depression [65, 124, 216, 266, 217, 20]. A growing body of research also focuses on using machine learning to detect depression using sensor data [216, 32, 262, 80, 267], and there has been some initial work on predicting depression in advance as well [32].

---

http://www.apa.org/monitor/2014/09/cover-pressure.aspx
http://www.acha-ncha.org/reports_ACHA-NCHAIIc.html

Depression, however, is a long-term health problem that needs to be continuously monitored and managed. Although mobile and wearable technology make the long-term monitoring of depression possible, some issues remain. Machine learning (ML) methods used for detecting and predicting depression rely on subjective ground truth acquired through psychological questionnaires such as BDI-II [237]. ML models are trained to detect these scores and their output is compared with these scores to measure their accuracy of prediction. Obtaining ground truth from users with depression or any mental health problem frequently over a long period of time is not sustainable as frequent requests to complete questionnaires will over time become an extra burden especially when the user is experiencing severe symptoms. Nevertheless, so far, all existing research in detecting and inferring depression has relied on frequent measurement of depression status (*e.g.*, every week). Further, while existing research has evaluated ML methods for detecting the presence of depressive symptoms, whether or not these methods can capture changes in depressive symptoms is unexplored.

In this study, we present a machine learning approach that uses data from mobile and wearable sensors to detect and monitor depression and change in depression at any time point, with limited ground truth data. Although our approach can be generalized to any chronic and longitudinal health problem, we evaluate it in the context of depression. We use data from smartphones and wearable fitness trackers from 138 students at an R-1 Carnegie-classified US University to identify students who experienced depressive symptoms or whose depressive symptoms worsened by the end of a semester.

To test the generalizability of our approach to other outcomes, we use our pipeline to predict loneliness in addition to depression. Loneliness has been associated with higher risk for developing depression, and is often co-morbid with depression [109].

Our machine learning approach advances the research in mobile health and analysis as follows:

1. To build machine learning models that can make accurate predictions from long-term data without frequent ground truth acquisition (in our case only two measurements at the beginning and end of semester), data needs to be processed and aggregated without losing key behavioral information during different time periods that may be useful in detecting and predicting depression. Therefore, we extract fine-grained features to capture behavioral markers in different time windows with varying granularity during the day, week, and semester. Although this step results in a number of features (>60,000) that is significantly larger than the number of samples (138 students), the hierarchical and incremental modeling component and stable feature selection in the pipeline are capable of identifying the most significant features, *i.e.*, features that are commonly chosen in most validation runs. We evaluate our approach by identifying students that

have post-semester depressive symptoms using data collected over one semester (16 weeks) from the smartphones and fitness trackers of 138 college students, and achieve an accuracy of 85.7%. We demonstrate that our method outperforms off-the-shelf ML methods such as Lasso and K-Nearest Neighbors.

2. We also evaluate our approach on its ability to detect change in depressive symptoms. To the best of our knowledge, our work is the first to detect change in depressive symptom severity without any knowledge of the students' initial or previous depression severity. We detect whether students' depressive symptom severity changed with an accuracy of 85.4%.

3. Our machine learning pipeline achieved an accuracy of 80.2% in detecting the binary level of loneliness and an 88.4% accuracy in detecting change in the loneliness level. By being able to detect both depression and loneliness in the student sample, we demonstrate that our pipeline can successfully identify multiple co-morbid outcomes, thus giving us a more holistic view of the participants' health. To the best of our knowledge, we are the first to detect loneliness in college students over the course of a semester.

3.2 Data Collection

In this section, we describe the participant recruitment and the data collection process (including participant-reported depression and loneliness measures and passively sensed data from smartphones and fitness trackers).

3.2.1 Participants and Recruitment

Participants in the study were from a pool of first-year undergraduate students at a Carnegie-classified R-1 University in the United States. Students were eligible to participate in the study if they were enrolled as a full-time student on campus for the semester and owned a data plan-enabled smartphone running iOS or Android. The research team advertised the study *via* emails and posts to student mailing lists and Facebook groups. Students were invited to our lab to be screened for eligibility, provide informed consent, download a mobile application to track sensor data from their smartphones and receive a Fitbit Flex 2 to track steps and sleep. After enrollment, the students completed initial depression and loneliness questionnaires online. They also gave us the phone numbers of their top-10 family members, friends on-campus, and friends off-campus, which were used to compute certain

"calls"-related features (see section 3.3.1). Data was collected from smartphone and Fitbit sensors as described in Section 3.2.3 and was continuously recorded over the duration of the study: one semester (16 weeks).

Out of the 188 first-year college students initially recruited, 138 completed the study and the depression and loneliness questionnaires at the beginning and the end of the study. The questionnaires were delivered via email and administered using Qualtrics – an online survey platform [201]. For their participation, the participants were allowed to keep the Fitbit Flex 2 and were compensated up to USD $205 spread over different points in time – $10 after the baseline appointment, $20 after the pre-semester depression and loneliness questionnaires, $25 after week 1, $40 after week 7, $60 after week 15, $25 after post-semester depression and loneliness questionnaires, and $25 bonus for compliance.

3.2.2 Participant-reported Depression and Loneliness Measures (Ground Truth)

The Beck Depression Inventory-II (BDI-II) [18, 68] is a widely used psychometric test for *measuring the severity of depressive symptoms*, and has been validated for college students [237, 68]. It contains 21 questions, with each answer being scored on a scale of 0-3. *Higher scores indicate more severe depressive symptoms.* For college students, the cut-offs on this scale are 0-13 (no or minimal depression), 14-19 (mild depression), 20-28 (moderate depression) and 29-63 (severe depression) [68].

The revised University of California, Los Angeles (UCLA) loneliness scale is a well-validated and commonly used measure of general feelings of loneliness [213]. It contains 20 questions, with each answer being scored on a scale of 1-4. The total loneliness scores ranged from 20 to 80 with *higher scores indicating higher levels of loneliness.* There are no standard cut-offs for this questionnaire and each study creates their own arbitrary categorizations informed by previous work.

The semester spanned over 16 weeks towards the end of which exams start and continue into the 17^{th} week. Since we expected compliance for answering the post-semester depression and loneliness questionnaires during exams to be low, we concluded the study at the end of week 16. Participants answered questions from BDI-II and the UCLA Loneliness scale at the beginning (week 1) and at the end (week 16) of the semester, which gave us their pre-semester and post-semester depression scores indicating the severity of depressive symptoms, and pre-semester and post-semester loneliness scores indicating the severity of loneliness. From these scores, we calculated ground truth for four outcomes, as follows:

1. *Post-semester Depression (Binary):* All participants with no or minimal depression

(post-semester BDI-II score < 14) at the end of the semester were classified as *"not having depression"*. While all participants with mild, moderate, or severe depression (post-semester BDI-II score $>= 14$) at the end of the semester were classified as *"having depression"*.

2. *Change in Depression (Binary):* We compare the pre-semester depression severity levels to the post-semester depression severity levels to obtain the *change in depression severity levels*. Using the standardized thresholds listed above, we assessed both pre-semester and post-semester BDI-II scores as being in one of four levels: no or minimal, mild, moderate, or severe. The depression severity levels did not improve for any participant. If there was no change in depression severity levels for a participant, the participant's *"Change in Depression"* was classified as *"did not worsen"*, otherwise it was classified as *"worsens"*.

3. *Post-semester Loneliness (Binary):* Participants with a UCLA loneliness score of 40 and below, were classified as having *"low loneliness"*. Whereas, participants with a UCLA loneliness score of over 40, were classified as having *"high loneliness"*. Doryab et al. [66] describes how these cut-offs were determined.

4. *Change in Loneliness (3-class):* We calculate change in loneliness as the difference between the pre-semester loneliness scores and the post-semester loneliness scores to obtain 3 categories: *increased loneliness*, *decreased loneliness*, and *loneliness levels remain the same*.

3.2.3 Passive Data Collection

We installed the AWARE framework [82] – a data collection mobile application with supporting backend and network infrastructure to collect sensor data unobtrusively from students' smartphones. This enabled us to record nearby bluetooth addresses, location, phone usage (*i.e.*, when the screen status changed to *on* or *off* and *locked* or *unlocked*), and call logs for incoming, outgoing and missed calls. In order to assess calls to close contacts, we asked participants to provide phone numbers of family members, friends on-campus, and friends off-campus that they most frequently contact. We also used a conversation plugin for AWARE (same as the one used by [266]) which makes audio inferences such as silence, voice, noise, or unknown. Further, we equipped participants with a Fitbit Flex 2 which records the number of steps and sleep status (asleep, awake, restless, or unknown). Calls, conversation, and phone usage are event-based sensor streams, whereas Bluetooth, location, sleep, and steps are sampled time series. These time series data streams were sampled at

different rates, due to the capabilities of the hardware being used. Bluetooth and Location coordinates are sampled at 1 sample per 10 minutes, sleep at 1 sample per minute, and steps at 1 sample per 5 minutes.

Data from AWARE was deidentified and automatically transferred over WiFi to our backend server on a regular basis, and data from the wearable Fitbit was retrieved using the Fitbit API at the end of the study. Participants were asked to keep their phone and Fitbit charged and carry/wear them at all times.

To maintain the participants' privacy and confidentiality, we stored all identifiable information (*e.g.* names, contact information) separate from their deidentified survey and sensor data. Only a few authorized members of the research team had access to the participants' identifiable information. All data sources – identifiable or not were password protected for security. At the University where this research was conducted, the Institutional Review Board (IRB) reviewed, oversaw and approved all procedures.

3.3 Data Processing and Analysis

This section describes the data processing and analysis pipeline, that consisted of 4 main steps:

1. Feature extraction to acquire sets of behavioral and behavioral change features from different sensors over different time slices (Section 3.3.1).

2. Handling missing features (Section 3.3.2).

3. Machine learning to detect Post-semester Depression and Change in Depression (Section 3.3.3), which involved:

    a) Detecting depression outcomes using 1-feature set models (*i.e.*, models containing features from one sensor).

    b) Combining detection probabilities given by these 1-feature set models to obtain a final detection label for our two outcomes.

4. Further, we slightly modified step (3) for different outcomes and different sensor combinations (also in Section 3.3.3).

This pipeline is illustrated in Figure 3.1 and explained in the subsections below.

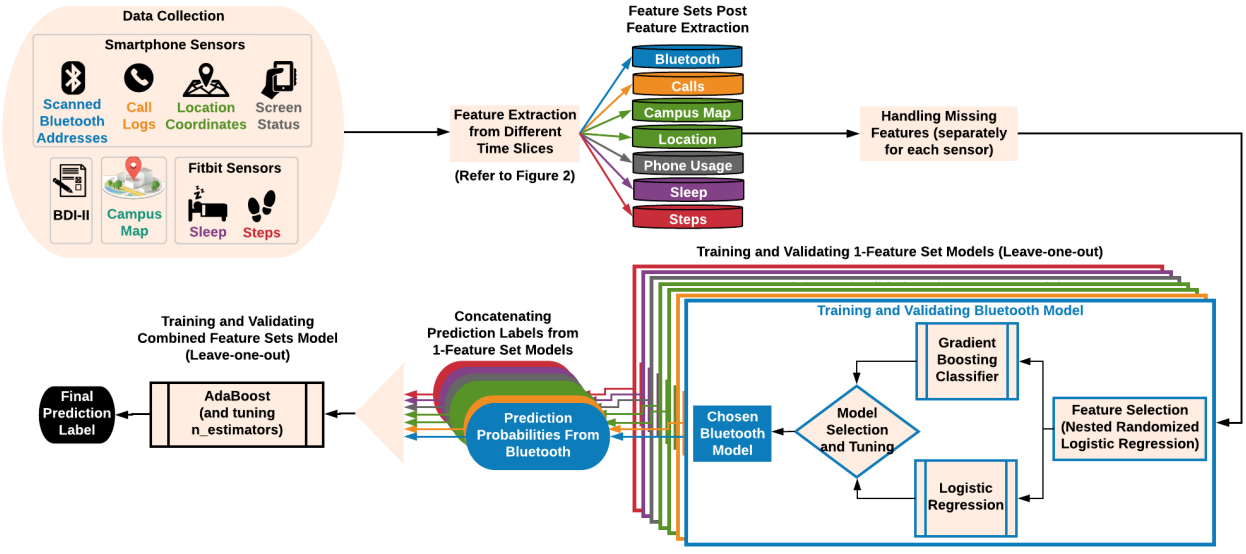


Figure 3.1: Pipeline for the data processing and analysis.

3.3.1 Feature Extraction

We computed *seven* feature sets from the collected data: Bluetooth, Calls, Location, Campus Map, Phone Usage, Steps, and Sleep. These feature sets were chosen because they have the potential to capture depressive symptoms described in the DSM-5 [7]. Location and Campus Map features capture users' mobility patterns; Calls features capture communication patterns; Bluetooth features can reflect both mobility and communication patterns; and Steps capture physical activity. Together they can be strong indicators of social withdrawal and diminished interest or pleasure in almost all activities, especially social and occupational activities. Further, fatigue or loss of energy can cause users to take longer to perform certain tasks, which may also be represented by these features. Sleep disturbances such as insomnia or hypersomnia are often present in people with depression [176]. Depression also causes diminished concentration which can affect phone usage [58, 133].

Due to some technical issues with AWARE's conversations plugin, many conversation inferences were missing. Hence, we used available conversation inferences to inform a Campus Map feature called social duration (explained later), but did not extract a Conversations feature set.

The feature extraction approach for each of the seven feature sets is described in section 3.3.1. These features were extracted over different temporal slices (see section 3.3.1). We

also extracted behavioral change counterparts for every feature (see section 3.3.1). As a result, we obtained 14 feature matrices for the seven feature sets and their behavioral change counterparts as explained in section 3.3.1. *We did not include pre-semester depression/ loneliness scores or labels as features in any of our models.*

Bluetooth Feature Set

Bluetooth features were calculated from the scanned bluetooth addresses recorded by the Bluetooth sensor in the smartphone, and can be used to sense the user's social context [71, 39, 275, 172, 40]. While a relationship between a bluetooth feature and depression has been found [266], bluetooth features have not been used to *detect* depression or loneliness.

Scanned bluetooth addresses can be clustered into the participant's own devices ("self" – scanned most often), family/ roommate/ office mate's devices ("related" – scanned less often than "self" but more often than "others"), and other people's devices ("others" – scanned least often) to help us estimate how many different people the participant meets, thereby capturing social activity and collocated communication. Since a participant may or may not be living with family or a roommate or be sharing an office, we clustered scanned addresses twice. First, the addresses were clustered into two categories for "self" *vs.* "others" ($K = 2$ clusters), then into three clusters - "self" *vs.* "related" *vs.* "others" ($K = 3$ clusters), and then chose the model which fit the data best out of the two sets. This process is described below.

1. We calculated the number of days each unique bluetooth address was scanned at least once.
   That is, $number\_of\_days_{bti}$.

2. We calculated the average frequency of each unique bluetooth address.
   That is, $average\_frequency_{bti} = \frac{total\_count_{bti}}{number\_of\_days_{bti}}$.

3. We Z-normalized the $number\_of\_days_{bti}$ and $average\_frequency_{bti}$ in order to give equal weight to both while optimizing score in step 4.

4. For each bluetooth address, we computed $Score = number\_of\_days_{bti} + average\_frequency_{bti}$.

5. We used K-means clustering to cluster $Score$ from step 4 for all bluetooth addresses using K=2 and K=3.

6. We chose the model with K=2 if sum of squared distances between clustered points and cluster centers was smaller than what we get with K=3. Otherwise we chose model with K=3.

7. If model with K=2 was chosen, the cluster with higher scores contained the participant's own devices ("self"), while the other cluster contained other people's devices ("others"). If the model with K=3 was chosen, the cluster with highest scores contained the participant's own devices ("self"), the cluster with lowest scores contained other people's devices ("others"), and the remaining cluster contained devices of the participant's partners, roommates, or officemates ("related").

Once the bluetooth addresses scanned were clustered into "self" *vs.* "others" or "self" *vs.* "related" *vs.* "others", we extracted the *number of unique devices*, *number of scans of most and least frequent device*, and *sum, average, and standard deviation of the number of scans of all devices* from all devices (*i.e.*, ignoring clusters), "self" devices, "related" devices, or "others" devices.

It is important to note that we do not have the bluetooth addresses of devices belonging to the user or people related to the user. We are using the frequency of occurrence of the devices scanned to heuristically 'guess' these clusters/ categories. Wang *et. al.* [266] used the number of collocated bluetooth devices to estimate the user's social context, however these devices may or may not belong to other people. However, these devices would also include the user's own devices, and hence may not accurately represent the user's social context. By using the frequency of occurrence of these devices to obtain 3 clusters, we build on previous work by attempting to separate the devices that are more likely to (1) belong to the user ("self"), (2) belong to people the user meets/ sees regularly ("related"), and (3) belong to other people ("others"). If the user does not meet many people regularly, then K=2 may fit the data better than K=3, thus giving us devices that are more likely to (1) belong to the user ("self") and (2) belong to other people ("others").

Calls Feature Set

Calls features were calculated using the call logs from the smartphone. We extracted the following features: *Number and duration of incoming, outgoing, and missed calls* and the *number of correspondents in total*.

Location Feature Set

Location features are derived from the Location 'virtual' sensor of the smartphone which uses proprietary algorithms to come up with the best estimate of location based on available GPS, WiFi and Celltower signals. We extracted the following Location features:

Location variance (sum of the variance in latitude and longitude coordinates), *log of location variance*, *total distance traveled*, *average speed*, and *variance in speed*. *Circadian*

movement [216] was calculated using the Lomb-Scargle method [195]. It encodes the extent to which a person's location patterns follow a 24-hour circadian cycle. Then, we carried out the following processing steps:

(a) Speed of the person was calculated from the distance covered and time elapsed between two samples. Samples with speed > 1 km/h were labeled as "moving", else "static" [216, 217].

(b) Samples labeled as "static" were clustered using DBSCAN [77] to find significant places visited by the participant. When we clustered all data and extracted each feature per week or per half-semester, we used global clusters. When we first split the data into weeks or half-semesters and then extracted features from each temporal slice, we used local clusters. Temporal slicing is discussed in section 3.3.1.

These steps allowed us to extract: *number of significant places*, *number of transitions between places*, *radius of gyration* [32], *time spent at top-3 (most frequented) local and global clusters*, *percentage of time spent moving*, and *percentage of time spent in insignificant or rarely visited locations* (labeled as -1 by DBSCAN). We also calculated *statistics related to length of stay at clusters* such as maximum, minimum, average, and standard deviation of length of stay at local and global clusters. *Location entropy* and *normalized location entropy* across local and global clusters were also calculated (implemented using the method in [216]). Location entropy will be higher when time is spent evenly across significant places. Calculating features for both local and global clusters allowed us to capture different behaviors related to the user's overall location patterns (global) and the user's location patterns within a time slice (local). For example, time spent at top-3 global and local clusters captures the time spent at places of overall significance to the user and places significant to the user in a particular time slice (*e.g.,* mornings on weekends).

We assume the place most visited by the participant at night to be their home location. To approximate the home location, we performed steps (a) and (b) above on the location coordinates from all nights (12am to 6am) and assumed the center of the most frequented cluster to be the participant's home location center. Since we don't know the radius of the home, we calculated two home-related features *time spent at home assuming home to be within 10 meters of the home location center*, and *time spent at home assuming home to be within 100 meters of the home location center*.

The 100m threshold is the default geofencing radius used by automation systems like HomeKit and https://www.home-assistant.io/, while the 10m threshold corresponds to the accuracy of GPS in an urban environment [157].

Campus Map Feature Set

We also analyze the user's location patterns in relation to their college campus. First, we obtained a campus map of the participants' University. Then, we marked out the campus boundary and different types of buildings on campus by creating polygons on Google Maps using GmapGIS. We annotated six types of buildings and spaces – Greek houses that hold the most social events, all Greek houses, student apartments, residential halls, athletic facilities, and green spaces. As academic buildings in this University are often collocated with other spaces, we assume any on-campus space not belonging to these six categories to be an academic building. For every location sample, we assigned one of eight *location type labels* (6 building/space types, academic, off-campus). Then, the following features were extracted for each type of space: *time spent at each location type in minutes*, *percentage time spent at each location type*, *number of transitions between different spaces*, *number of bouts (or continuous periods of time) at space*, *number of bouts during which participant spends 10, 20, or 30 minutes at the same space*, and *minimum, maximum, average, and standard deviation of length of bouts at each space*.

Campus map features also include two multimodal features – *study duration* and *social duration*, as implemented by Wang *et al.* [265, 267]. These features fuse data from Location, Phone Usage, Conversation, and Steps sensors. Study duration was calculated by fusing location type labels with data from the phone usage and steps sensors. A participant was assumed to be studying if they spent 30 minutes or more in an academic building while being sedentary (fewer than 10 steps) and having no interaction with their phone. Social duration was calculated by fusing location type labels with data from the conversation sensor. A participant was assumed to be social if they spent 20 minutes or more in any of the residential buildings or green spaces and the conversation sensor inferred human voice or noise for 80% or more of that time. Study duration was only calculated in academic buildings, while social duration was only calculated in residential buildings or green spaces.

Phone Usage Feature Set

Phone Usage features were calculated using the screen status sensor in the smartphone, which recorded screen status (on, off, lock, unlock) over time. We extracted the following phone usage features:

Number of unlocks per minute, *total time spent interacting with the phone*, *total time the screen was unlocked*, *the hour of the days the screen was first unlocked or first turned on*, *the hour of the days the screen was last unlocked, locked, and turned on*, and *the maximum,*

`http://www.gmapgis.com/`

minimum, average, and standard deviation of length of bouts (or continuous periods of time) during which the participant is interacting with the phone and when the screen is unlocked. A participant is said to be "interacting" with the phone between when the screen status is "unlock" and when the screen status is "off" or "lock".

Sleep Feature Set

Sleep features were calculated from the sleep inferences (asleep, restless, awake, unknown) over time returned by the Fitbit API. The following features were calculated:

Number of asleep samples, *number of restless samples*, *number of awake samples*, *number of unknown samples* (still detected as sleep), *weak sleep efficiency* (sum of number of asleep and restless samples divided by sum of number of asleep, restless, and awake samples), *strong sleep efficiency* (sum of number of asleep samples divided by sum of number of asleep, restless, and awake samples), *count, sum, average, maximum, and minimum length of bouts during which the participant was asleep, restless, or awake* as well as the *start and end time of longest and shortest bouts during which the participant was asleep, restless, or awake.* We include 3 summary statistics – count, sum, and average length of asleep/ restless/ awake bouts as individual features, despite them being dependent on each other, because we want to consider the "interaction" between these features. For example, say larger average length per asleep bout and smaller number of asleep bouts correlate with better mental health outcomes, the relationship between average length per asleep bout and mental health may still be dependent on the number of asleep bouts. Very high number of asleep bouts could indicate disturbed sleep or polyphasic hypersomnia, such that even with high average length per asleep bout, the mental health outcomes could be poor.

Steps Feature Set

Steps features were calculated from the step counts over time returned by the Fitbit API. The following features were calculated:

Total number of steps and *maximum number of steps taken in any 5 minute period* were extracted as features. Other features were extracted from "bouts", where a "bout" is a continuous period of time during which a certain characteristic is exhibited. Examples of such features include *total number of active or sedentary bouts* [11], and *maximum, minimum,*

Sleep captured by Fitbit is accurate +/- 45min [147, 50, 279].

Interaction models are commonly used in statistics (see: http://www.medicine.mcgill.ca/epidemiology/Joseph/courses/EPIB-621/interaction.pdf). For example, let x_1 = mean, x_2 = count. Then, the interaction term is $x_1 x_2$ = sum.

and average length of active or sedentary bouts. We also calculated *minimum, maximum, and average number of steps over all active bouts.* A bout is said to be sedentary if the user takes less than 10 steps during each 5 minute interval within the bout. As soon as the user takes more than 10 steps in any 5 minute interval, they switch to an active bout.

Temporal Slicing

Our temporal slicing approach helps us extract behavioral features from different time slices. Past work has shown that this approach can better elicit the relationship between a feature and depression. For example, Chow *et al.* [46] found no relationship between depression and time spent at home during 4-hour time windows, but they found that people who are more depressed tend to spend more time at home between 10:00 AM and 6:00 PM. Similarly, Saeb *et al.* [217] found that the same behavioral feature calculated over weekdays and weekends can have a very different effect on depression.

Each feature described in section 3.3.1 was extracted from 45 temporal slices or time segments as illustrated in figure 3.2. First, we fetched all available data (spanning over multiple days of the study) from a certain epoch or time of the day (all day, night *i.e.*, 12am-6am, morning *i.e.*, 6am-12pm, afternoon *i.e.*, 12pm-6pm, evening *i.e.*, 6pm-12am) and for certain days-of-the-week (all days of the week, weekdays only *i.e.*, Monday-Friday, weekends only *i.e.*, Saturday-Sunday). Then, we calculated features from this data aggregated over different levels of granularity (whole semester, two halves of the semester, weekly). Since there are 5 epochs, 3 days-of-the-week segmentations, and 3 levels of granularity, we get $5 \times 3 \times 3 = 45$ time slices. Each location feature is calculated over these 45 time slices. Note that the two halves of the semester are not perfect halves. For simplicity, we refer to weeks 1-6 as the first half (before midterms) and weeks 7-16 (midterms and after midterms) as the second half. We also investigated the effect of removing the spring break weeks (week 8 and 9 as the spring break was mid-week 8 to mid-week 9) on detecting the two outcomes, and while our findings were inconclusive, this may be worthy of future study.

Behavioral Change Features

Behavioral change features capture changes in behaviors over 16 weeks. These features can be abstractly characterized as the change in slope for each behavioral feature over the semester. For this purpose, we only use features computed weekly (*i.e.*, using granularity "weeks"). This gives us 15 time slices (for 5 epochs $\times$ 3 days-of-the-week options) for which

A threshold of 10 steps is often used to ignore 'false steps' [224, 247]. Previous work has also used 10 steps as a threshold to detect sedentary behavior [105, 248].

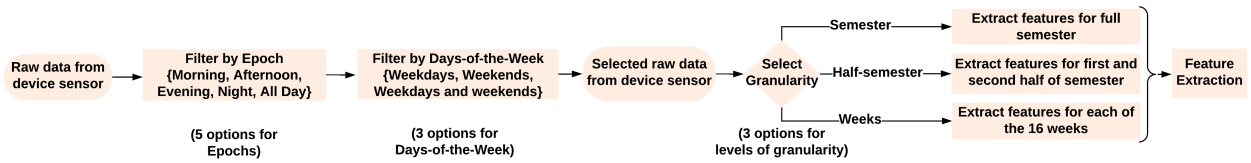


Figure 3.2: For each sensor, each feature was extracted from 45 time slices. First, raw data from the device sensor was preprocessed and then filtered by an epoch and a days-of-the-week option. Features (Let NS be number of features derived from each sensor) were then extracted from the selected raw data according to 3 levels of granularity – per semester (NS features), per half-semester ($2 * NS$ features), and per week ($16 * NS$ features).

we have weekly values of every behavioral feature described in section 3.3.1. We compute the behavioral change feature for each behavioral feature using their weekly values over 16 weeks. We follow the same method employed by [265], to test whether their approach works on our dataset:

- *Slope:* We fit a linear regression model to the values of the feature over 16 weeks. "Slope" is the slope of this linear regression line.
- *Slope first half and second half:* We fit two separate linear regression models to the values of the feature over weeks 1-6 (*i.e.*, before midterms) and weeks 7-16 (*i.e.*, midterms and after midterms) of the semester. "Slope first half" and "Slope second half" are the slopes of these linear regression lines.
- *Breakpoint:* Each student's breakpoint is the week after which the student's behavior (represented by the feature value) begins to change. This is calculated by fitting a piecewise linear regression model with two segments with each of the 16 weeks as a breakpoint. "Breakpoint" is the week that when used as a breakpoint gives the best model as determined by Bayesian Information Criterion (BIC).
- *Slope before and after Breakpoint:* A piecewise linear regression model with two segments is fit to the feature values over 16 weeks with the final "Breakpoint". The slope of the first line segment is "Slope before Breakpoint" and the slope of the second line segment is the "Slope after Breakpoint".

Defining the Feature Matrix

After feature extraction we obtain a feature matrix for each of the seven feature sets derived from different sensors, as well as their behavioral change counterparts (*i.e.*, 14 feature matrices in total). In each of these feature matrices, each sample or record contains features extracted from 1 student. We aggregate our features over different timeslices (see section 3.3.1) – over different weeks, in the two 'halves' of the semester, and across the whole semester. The features from all these time slices are concatenated to form the feature vector for each student. By investigating features from individual weeks, we aim to capture the variability in a person's behavior in different time periods. For example, the midterm week may have a greater impact on depression/ loneliness than the Spring break week.

3.3.2 Handling Missing Features

Missing features are the result of missing data. While we occasionally miss data from all sensors due to non-semantic reasons (*i.e.*, technical issues such as the phone/ app stopped working, the data was not transferred on time, or the server was down, and compliance-related issues such as the user withdrew permissions for the app), we often miss data due to semantic reasons. For example, if the user does not sleep at all during a time period, we will get no sleep data. If the user does not make any calls during a time period, we will get no calls data. Hence, instead of completely ignoring missing data, since we do not know if it was not collected or whether it did not exist to be collected, we have tried to encode it in our features.

A feature (*i.e.*, feature value during a temporal slice) being missing for a large number of people can indicate non-semantic issues such as server-side problems. Hence, we excluded all features that were missing for more than 30 participants. Further, a participant missing a very large number of features can indicate non-semantic issues such as the phone/ app not working, or that they withdrew permissions. Hence if a participant was missing more than 20% of all features from a feature set, we removed that participant. The "30 participants" and "20% features" thresholds were determined empirically by plotting the number of participants and features remaining for different threshold values and observing where the curve falls off. All the remaining missing features were imputed as "-1" as their "missingness" may be due to semantic reasons and can be useful information for the classifier. The same features calculated over different time slices were viewed independently, such that if a feature was missing for a week for over 30 people, we only removed that feature from that week. In the end, we were left with roughly 79-110 participants and thousands of features for every feature set. The exact numbers were different across feature sets as missing features in each

feature set were handled separately. That is, a participant was excluded from a feature set only if they were missing 20% or more of the total number of features in that feature set.

3.3.3 Modeling

We use machine learning to build detection models for depression and loneliness. Our modeling approach consists of the steps below, *each using leave-one-out cross-validation to minimize over-fitting*. That is, we train a separate model to detect an outcome for each participant, and that model does not include the participant in question, during feature selection or training. It is important to remember that each sample contains features from 1 participant only (recall section 3.3.1), such that leave-one-out or leave-one-sample-out is actually leave-one-person-out.

Our model generation process uses the following steps:

1. *Stable Feature Selection* using Randomized Logistic Regression while leveraging the semantic structure of the temporal slices (section 3.3.3).

2. *Training and Validating 1-Feature Set Models* for each of the seven feature sets: Bluetooth, Calls, Campus Map, Location, Phone Usage, Sleep, and Steps (section 3.3.3).

3. *Obtaining the Final Label for the Outcome* by combining detection probabilities from 1-feature set models (section 3.3.3).

4. *Classifying Different Outcomes* by slightly modifying the pipeline to detect post-semester depression, change in depression, post-semester loneliness, and change in loneliness (section 3.3.3).

We describe these steps in the following sections.

Feature Selection

After handling missing data, we have 79-110 people (depending on the sensor used) and thousands of features for each feature set. So, the sample size is very small in comparison to the number of features. Hence, feature selection is a crucial step of the pipeline. Moreover, it is essential to select stable features, that is the set of selected features should remain stable when we remove or replace a small number of people. For this purpose, we tried a number

of feature selection methods but all of them selected unstable features. That is, the features selected greatly varied across cross-validation folds.

Randomized Logistic Regression [154] is a method that creates several random subsamples of the training dataset (200 in our case), computes a logistic regression on each subsample, and selects features by optimizing their importance across all subsamples. That is, a feature is selected if the average of its logistic regression coefficients across all subsamples is above a specified selection threshold, which is treated as a model parameter and tuned during cross-validation. This usually results in a stable set of selected features. However, in our case, since the number of features in each feature set is significantly larger than the sample size, randomized logistic regression also did not work.

To address this problem, we decomposed our feature space for each feature set (*e.g.*, for bluetooth) by grouping features from the same time slices, and performed randomized logistic regression on each of these groups. The selected features from all groups (*i.e.*, all time slices) were then concatenated to give a *new and much smaller* set of features. Then, randomized logistic regression was performed again, this time on this *new* set of features to extract the final selected features for the feature set, thereby *nesting* the process. We call this method Nested Randomized Logistic Regression, and used it to extract selected features for each of the seven 1-feature set models.

This method was performed in a *leave-one-out manner* such that the model used to detect an outcome for a person did not include that person during the feature selection process.

Training and Validating 1-Feature Set Models (Model Selection and Tuning)

For each feature set, we built a model of the selected features from that feature set to detect an outcome. We used leave-one-out cross-validation (same as leave-one-person-out – see section 3.3.1) to choose the model and parameters for that model. We tried two types of learning algorithms – Logistic Regression and Gradient Boosting Classifier. Logistic Regression was tried because our feature selection approach was based on Logistic Regression, while Gradient Boosting was tried because it can perform well on a noisy dataset, learn complex non-linear decision boundaries via boosting and has been effectively used to detect

We selected features using recursive feature elimination or that give k highest scores from the model, p-values below alpha based on a FPR test, p-values below alpha based on ANOVA test, and p-values below alpha based on Pearson's correlation.

$Best(F_s) = sel(concatenate[sel(F_{s1}), sel(F_{s2}), ..., sel(F_{sT})])$ where F_{si} = features from feature set s and time slice i (e.g. calls features from the mornings on weekdays calculated weekly), T = total number of time slices, and $sel(...)$ is the Randomized Logistic Regression Function. $T = 45$ for regular feature sets and $T = 15$ for behavioral change feature sets. $Best(F_s)$ are the final features selected from feature set s and are given as input to the 1-feature set model for feature set s.

similar outcomes in previous work [264]. We chose the model and model parameters using accuracy as a metric for post-semester and change in depression and loneliness. The chosen 1-feature set model gave us detection probabilities for each outcome label.

Combining Detection Probabilities from 1-Feature Set Models to Obtain Combined Models

The detection probabilities from all seven 1-feature set models were concatenated into a single feature vector and given as input to an ensemble classifier, *i.e.*, AdaBoost with Gradient Boosting Classifier as a base estimator, which then outputted the final label for the outcome. For post-semester and change in depression, and post-semester loneliness only the detection probabilities of class label "1" were concatenated. Whereas, for change in loneliness, detection probabilities of all class labels were concatenated. The "n_estimators" parameter was tuned during leave-one-out cross-validation to get the best combined model.

We also carried out a *feature ablation study* to analyze the effect that different feature sets have on the performance of the models, thereby understanding their salience. For this purpose, we concatenated detection probabilities from specific 1-feature set models instead of all seven 1-feature set models. We do this for *all possible combinations of 1-feature set models*, in order to analyze the usefulness of each feature set. There are seven 1-feature set models and 120 combinations of feature sets, as total combinations = combinations with 2 feature sets + ... + combinations with 7 feature sets = $\sum_{r=1}^{7} \binom{7}{r} = 120$.

Classifying Different Outcomes

The pipeline described in the sections above was used to detect two outcomes – post-semester depression, change in depression, post-semester loneliness, and change in loneliness. For all outcomes, we used the pipeline *as described above* without excluding any students and using *accuracy as the metric for model selection and tuning.*

3.4 Results for Detecting Depression

In this section, we present our results for detecting depression. First, we report descriptive statistics about the prevalence of depression in our sample of college students. Then, we report the results obtained for detecting post-semester and change in depression. *It is im-*

The maximum number of estimators at which boosting is terminated.

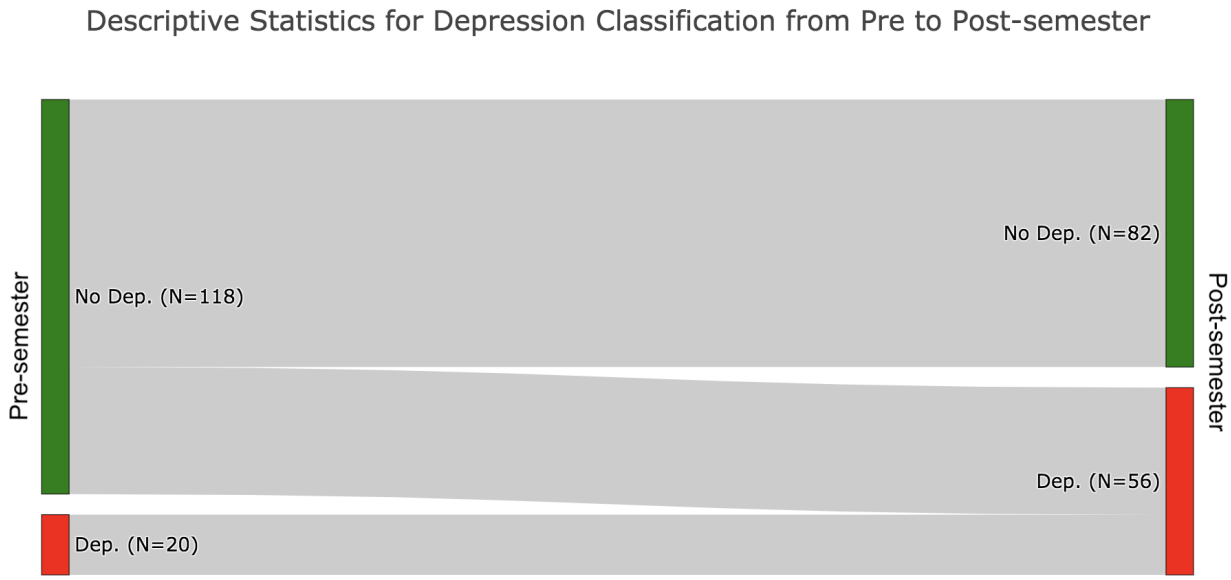


Figure 3.3: Shows how depression status ("no dep." vs "dep.") changed from pre to post-semester.

portant to note that none of our models contained pre-semester depression scores or labels as features.

3.4.1 Descriptive Statistics

As mentioned in section 3.2.2, the four severity levels of depression specified by BDI-II are symptoms reflecting no or minimal depression (score 0-13), mild depression (score 14-19), moderate depression (score 20-28), and severe depression (score 29-63). At the beginning of the semester, 14.5% *i.e.*, 20 out of the 138 participants who completed the study were categorized as having mild (13 participants), moderate (5 participants), or severe (2 participants) depression. At the end of the semester, this number significantly increased to 40.6% *i.e.*, 56 out of the 138 participants were categorized as having mild (25 participants), moderate (19 participants), or severe (12 participants) depression (see figure 3.3). While the number of students with depression almost tripled by the end of the semester, the post-semester depression rate is comparable to the 33% estimated by the American Psychological Association

http://www.apa.org/monitor/2014/09/cover-pressure.aspx

for US universities. So, depression statistics at the study University are not surprising or unusual.

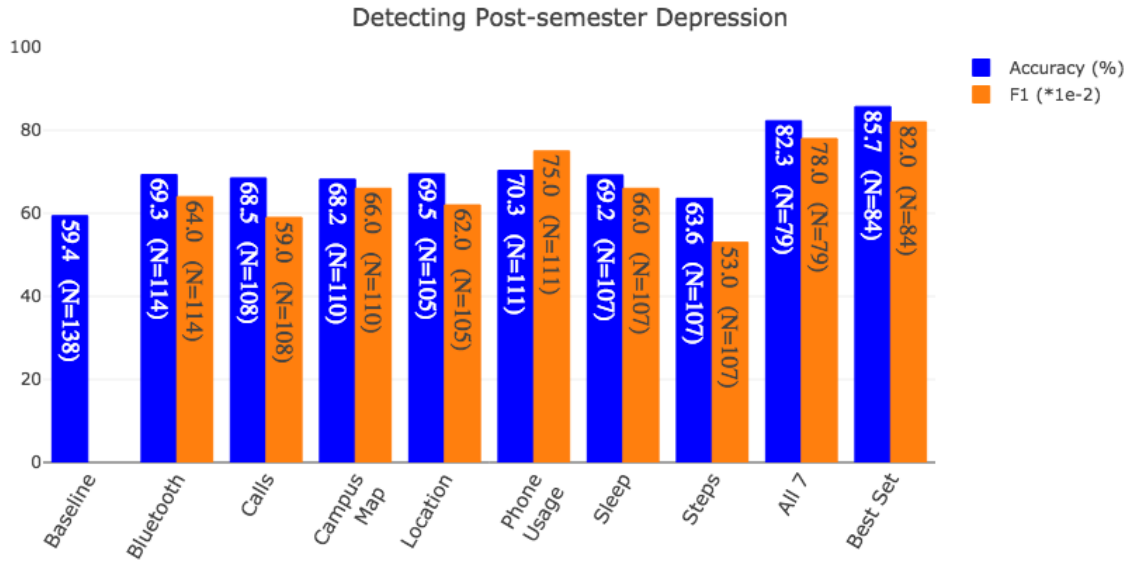
On comparing BDI-II scores from the beginning and end of the semester, we found that the scores of 23 people improved by an average of 2.8, the scores of 99 people got worse by an average of 8.7, and the scores of 16 participants did not change at all. However, on comparing depression severity levels (thresholded scores) from the beginning and the end of the semester, we found that none of the 23 people showed improvement significant enough to improve their depression severity levels. So, the depressive severity levels of none of the participants got better. In fact, depression severity levels did not worsen for 65.9% *i.e.*, 91 out of 138 participants, while they worsened for 34.1% *i.e.*, 47 participants.

3.4.2 Detecting Post-semester Depression

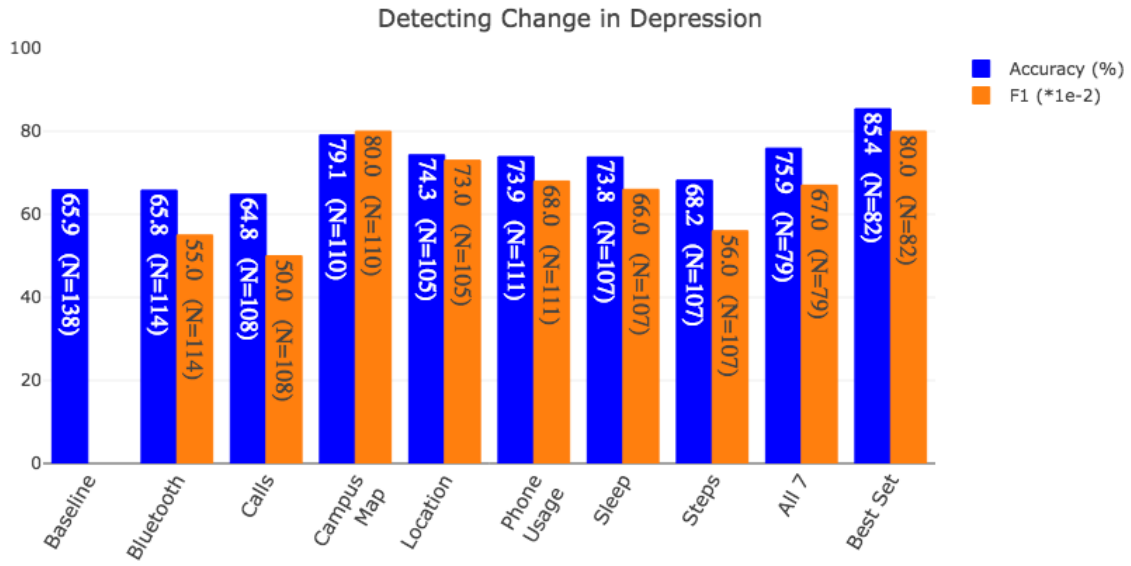
Figure 3.4a shows accuracies obtained by the seven 1-feature set models, the 7-feature sets model, and the best set model for detecting *post-semester depression* (*i.e.*, "depression" vs. "no depression"). The *7-feature sets model* is obtained by combining all seven feature sets, while the *best set model* is the model that gives us the best accuracy out of the 120 different combinations of feature sets tried during the feature ablation study. The number of people (*i.e.*, sample size N) may be different for models containing different feature sets since handling missing features for different feature sets will remove a different number of people from the analysis.

If we detect all students as having "no depression" (majority class), we obtain an accuracy of 59.4% (baseline) for detecting post-semester depression. The 7-feature sets model was significantly better than this baseline (using McNemar Test [256, 60], $X^2 = 10.1$ and $p < 0.01$) and obtained accuracy of 82.3% (N = 79). The best accuracy obtained using a 1-feature set model was 70.3% (N = 111) using "Phone Usage". The best set accuracy was 85.7% (N=84) obtained using a model containing 4 feature sets: "Bluetooth", "Calls", "Phone Usage", "Steps" from the feature ablation study. The best set model significantly outperformed the baseline (using McNemar Test [256, 60], $X^2 = 13.4$ and $p < 0.01$), but its performance was not significantly better than the 7-feature sets model (using McNemar Test [256, 60], $X^2 = 0.5$ and $p = 0.48$).

We found that behavioral features are *better* than behavioral change features for detecting post-semester depression. Hence, we only use behavioral features to detect post-semester depression. The behavioral change features were calculated using the method employed by [265] that assumed that the weekly features have a linear relationship. It is possible that these features don't work well on our dataset because the linearity assumption is false.



(a) Detecting Post-semester Depression. Best 1-feature set model contains {Phone Usage}. Best set model contains {Bluetooth, Calls, Phone Usage, Steps}.



(b) Detecting Change in Depression. Best 1-feature set model contains {Campus Map}. Best set model contains {Bluetooth, Campus Map, Phone Usage, Sleep}.

Figure 3.4: Shows accuracies and F1 scores obtained for detecting (a) Post-semester Depression, and (b) Change in Depression. Accuracies and F1 scores are reported for 1-feature set models, the 7-feature set model *i.e.*, model combining detections from all feature sets ("All 7"), and the best set model *i.e.*, the model that gives us the best accuracy during the feature ablation study and thus contains the best set of feature sets ("Best set"). F1 score for (a) is the F1 score of the "depression" class, and F1 score for (b) is the F1 score of the "worsens" class.

Therefore, future work should investigate other methods that do not assume linearity for calculating behavioral change features.

3.4.3 Detecting Change in Depression

Figure 3.4b shows accuracies obtained by the seven 1-feature set models, the 7-feature sets model, and the best set model for detecting *change in depression* (*i.e.*, "did not worsen" vs. "worsens").

If we detect all students as "did not worsen" (majority class), we obtain an accuracy of 65.9% (baseline) for detecting change in depression. The 7-feature sets model was marginally significantly better than this baseline (using McNemar Test [256, 60], $X^2 = 3.6$ and $p = 0.06$) and obtained an accuracy of 75.9% (N = 79). The best accuracy obtained using a 1-feature set model was 79.1% (N = 110) using "Campus Map". The best set accuracy was 85.4% (N = 82) obtained using a model containing 4 feature sets: "Bluetooth", "Campus Map", "Phone Usage", and "Sleep" from the feature ablation study. The best set model significantly outperformed the baseline (using McNemar Test [256, 60], $X^2 = 12.4$ and $p < 0.01$) and the 7-feature sets model (using McNemar Test [256, 60], $X^2 = 4.5$ and $p < 0.05$).

We found that behavioral features are *better* than behavioral change features for detecting post-semester depression. Hence, we only use behavioral features to detect change in depression.

3.5 Results for Detecting Loneliness

In this section, we present our results for detecting loneliness. First, we report descriptive statistics about the prevalence of loneliness in our sample of college students. Then, we report the results obtained for detecting post-semester and change in loneliness. *It is important to note that none of our models contained pre-semester loneliness scores or labels as features.*

3.5.1 Descriptive Statistics

For loneliness, we analyzed the UCLA loneliness scores from both pre-semester and post-semester questionnaires, and found that 63.8% of the participants fell into the "high loneliness" category pre-semester, while 58.8% of participants fell into the "high loneliness" category post-semester. When comparing pre-semester and post-semester loneliness scores, we found that 47% of participants showed increased levels of loneliness towards the end of

the semester, 47% of participants showed decreased levels of loneliness towards the end of the semester, and loneliness levels remained the same for only 6% of participants.

3.5.2 Detecting Post-semester Loneliness and Change in Loneliness

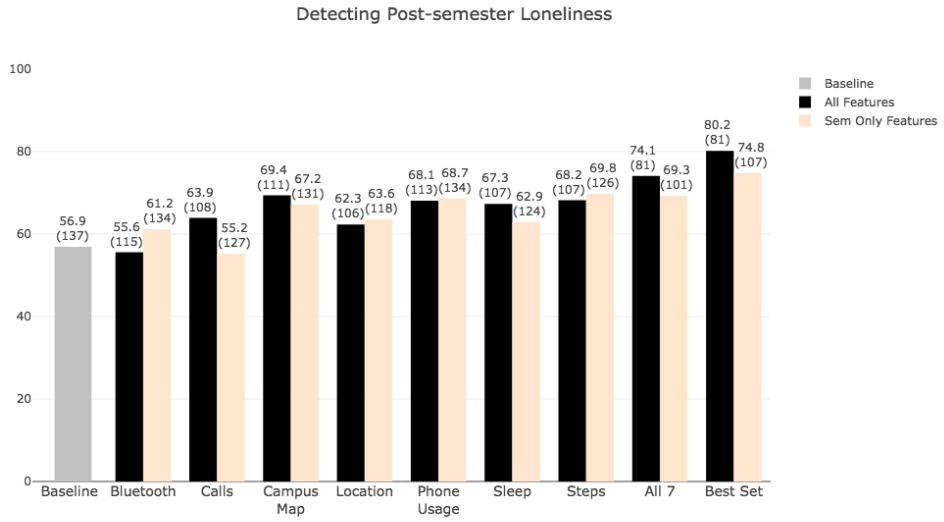
We ran our machine learning pipeline to infer 2 outcomes: post-semester loneliness level (low or high) and change in loneliness level (increased, decreased, and unchanged). For both outcomes, we used the set of all-epochs features extracted from all time slices and time slices as described in the processing section, as well as semester-aggregated (semester-level) features. Our goal was to identify a minimal set of features capable of accurately inferring loneliness level. Whereas the all-epochs features provided the opportunity to analyze behavior on a more fine-grained level, the semester-level features provided a reduced set that described the overall behavior of each participant during the semester. Figures 3.5a and 3.5b show the accuracy results for both outcomes and their comparison with the baseline. The graphs show the accuracy obtained from sensor-specific features (1-feature set), all feature sets combined, and the set that provides the best overall accuracy.

Post-semester Loneliness

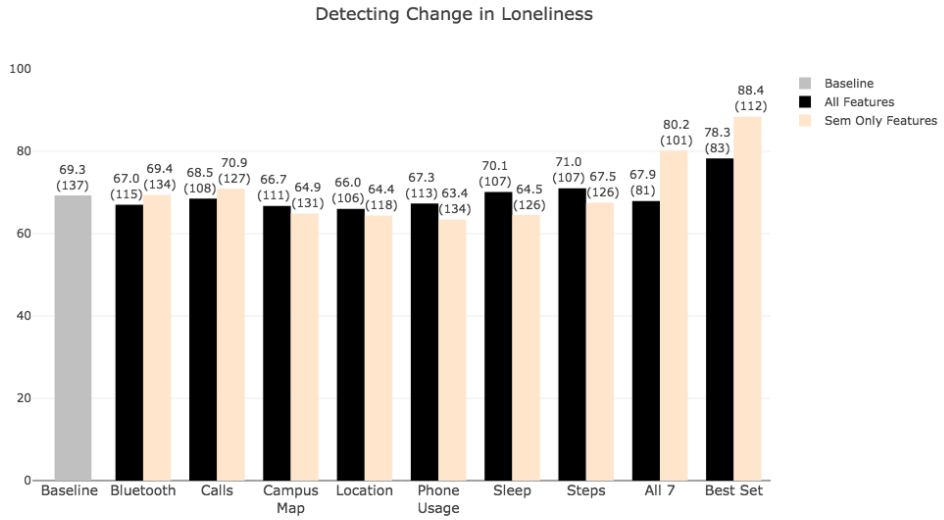
For detection of post-semester loneliness, our machine learning pipeline achieved the highest accuracy of 80.2%, using all-epochs features in the best feature set that included call logs, location, location map, screen, sleep, and steps. This accuracy was 6.1% higher than the accuracy obtained from using all 7 feature sets (74.1%) and indicated that including Bluetooth features contributed to performance reduction. The all-epochs-Bluetooth-only features provided 55.6% accuracy, confirming their low prediction power in detecting post-loneliness level. Except for Bluetooth, all other feature sets and their combinations achieved a higher (by at least 5.4%) accuracy than the baseline measure. The semester-level features, including the combination of Bluetooth, location, screen, and steps, provided the best set accuracy of 74.8%, which was 5.5% higher than the all features set (69.3%). In general, the analysis with all-epochs features included provided better results than the analysis with semester-level features for detecting loneliness level (5.4% higher accuracy).

Change in Loneliness

Detection of change in loneliness provided slightly different results. Using all-epochs features, the best feature set including calls and screen state achieved 78.3% accuracy, whereas the



(a) Detecting Post-semester Loneliness (high or low loneliness).



(b) Detecting Change in Depression (increased, decreased, or unchanged loneliness).

Figure 3.5: Detection of post-semester and change in loneliness using all-epochs features and semester-level features. Each bar shows the accuracy followed by the number of samples used in the analysis in parentheses; the gray bar represents the baseline accuracy as measured by the percentage of samples belonging to the majority class.

best set obtained with semester-level features—which included Bluetooth, calls, location, and location map—achieved 88.4% accuracy. In contrast to the post-semester loneliness detection model, where the analysis with all-epochs features provided better results, in these models for detecting change, the semester-level features contributed to higher accuracy using the best set (88.4% with semester only vs 78.3% with all-epochs features) and all 7 sets (80.2% with semester only vs 67.9% with all-epochs features).

3.6 Discussion

In this section, we discuss our observations about the selected features, compare our approach with existing ML approaches, and discuss the implications of longitudinal studies, interventions, privacy and technical limitations, and combining daily behaviors with verbal and non-verbal behaviors and genomic data.

3.6.1 Observations about Selected Features

We made some interesting observations when we informally analyzed the features selected by the best 1-feature set models for detecting post-semester and change in depression. For this purpose, we took features that were selected in at least one fold, and made different graphs to visualize how many of them came from each feature type, week, epoch, and days-of-the-week. For both outcomes, there is a significant negative correlation between week number and the number of features selected from each week (post-semester depression: r = -0.94 and p ¡ 0.0001, and change in depression: r = -0.73 and p ¡ 0.0020). That is, more features are selected from earlier weeks and fewer features are selected from later weeks. For all feature sets except sleep and steps, features from nights are selected less often. We interpret this to mean that the participants' social context at night captured using bluetooth, calls, campus map, and location, and their phone usage at night are not predictive of depression. For post-semester depression and change in depression, the most frequently selected features for Bluetooth, calls, and campus map come from afternoons and evenings, and the most frequently selected features for phone usage come from the "all day" epoch. We interpret this to mean that the participants' social context in the afternoons and evenings is the most predictive of depression and change in depression, while their phone usage throughout the day is more predictive of depression and change in depression than

Bluetooth, calls, and campus map features from the afternoons and evenings likely reflect social context, as the subjective 'number of social interactions' (in-person and otherwise) reported by participants during weeks 1, 7, and 15, are significantly more in the afternoons and evenings than in the mornings.

their phone usage during specific times of the day. For both outcomes, selected Bluetooth features are related to the devices of "others" and features related to devices of "self" are rarely selected. This shows that the Bluetooth features we calculate to encode proximity to "others" are able to successfully capture depression. Our study is also the first to use Bluetooth features to detect depression. Features such as maximum length of sedentary bouts from steps, maximum length of awake bouts during sleep, time spent in green spaces were most frequently selected for post-semester depression and change in depression. This shows that long periods of time with no exercise, periods of disturbed sleep at night, as well as time spent outdoors are some of the features that are most predictive of depression and change in depression.

The significant negative correlation between weeks and features selected from each week explains why we are able to achieve an accuracy of >80% very early in the semester using our prediction models, thus enabling depression prevention early in the semester (see chapter 4).

Above, we lightly reflected on the features selected, because very little is known about the relationship between behavior *in the wild* and depressive symptoms. Most previous work in this space has only looked at behavior in the therapist's office or behavior self-reported by the participants in retrospect. Hence, validating these features will require qualitatively analyzing participant and clinician experience, which is beyond the scope of this project. That said, the above findings should help explain some of the selected features. To do more would be speculative.

3.6.2 Comparison with Other Machine Learning Approaches

We compared the results obtained by our novel ML pipeline for detection post-semester depression with two baseline methods – K-Nearest Neighbors and Lasso. For both these methods, we detected Post-semester Depression using models trained on each feature set as well as a model trained on all feature sets. Table 3.1 shows that our *method outperformed both these methods* for almost every 1-feature set model and for the "all" feature sets model. Comparing the average number of feature selected across all folds, reveals that *our method selected a smaller number of features than the other two methods*. That is, our feature selection approach is more stringent and selects more meaningful features from a large set of features that may often be correlated, as compared to traditional approaches. Our method outperforms traditional approaches for the following reasons:

Lasso performs regression. We apply use threshold of 0.5 on the score returned by Lasso to achieve binary outcomes.

1. *Selecting Stable Features by Using Randomized Logistic Regression:* Randomized Logistic Regression selects features by performing Logistic Regression on several subsamples of the training data and selecting features that perform the best across most subsamples. This method leads to a more stable and useful set of selected features as it reduces overfitting by diversifying the training samples. This method and its adaptations hence work well for highly dimensional feature spaces [281, 268].

2. *Reducing Correlation Between Features During Training by Decomposing the Feature Space Using Data Sources and Temporal Slices:* Some existing ensemble classification methods partition the feature space into smaller subsets using various techniques, learn separate models for each subset, and combine their predictions to get the final prediction. Partitioning the feature space can reduce correlation between features and further diversify the training data that each model is trained on, thereby improving performance [193, 69, 28, 159, 210]. Leveraging the same idea, we decompose the feature space and learn separate 1-feature set models for each data source (*e.g.*, Bluetooth, Location) because we expect different data sources to contain overlapping and correlated behavioral information. For example, step counts (and features derived from step counts) will usually be low when location variance is low. Further, for each 1-feature set model, our novel feature selection approach applies randomized logistic regression on subsets of features from different temporal slices (see section 3.3.3). We do this because features from the same data source can correlate across different temporal slices. For example, a person with low physical activity may have low step count related features in several temporal slices.

Hence, our ML pipeline outperforms other approaches by jointly tackling three challenges of working with behavioral data – multiple modalities (*i.e.* collected from various data sources), high dimensionality with correlated features, and small sample sizes (resulting from logistical and privacy-related limitations during data collection).

Features are said to be 'stable' when they don't vary greatly across folds or with minor perturbations of the training data. There is no definite method of quantifying stability.

Table 3.1: Comparing our method for detecting Post-semester Depression with 2 Baselines – K-Nearest Neighbors and Lasso. Our method performs better than the two baselines for all feature sets by employing a more stringent and robust feature selection strategy that consistently selects fewer but useful features.
KEY – "N": Sample Size, "F1": F1 score of the "depression" class.

Feature Set	**N**	**Total Features**	**Method**	**Model Parameters**	**Accuracy**	**F1**	**No. of Features Selected (avg. across folds)**
Bluetooth	114	3202	KNN	K=2	53.5	.18	N/A
			Lasso	Alpha=0.7	60.5	.52	229
			Our Method	NRL (C=0.5, scaling=0.5, sample_fraction=0.80, selection_threshold=0.20) Model = GBC	**69.3**	**.64**	**73**
Calls	108	606	KNN	K=1	58.3	.46	N/A
			Lasso	Alpha=1.0	55.6	.33	57
			Our Method	NRL (C=0.5, scaling=0.7, sample_fraction=0.80, selection_threshold=0.80) Model = LogR (C=0.5 - same as NRL)	**68.5**	**.59**	**7**
Campus Map	110	23873	KNN	K=8	61.8	.46	N/A
			Lasso	Alpha=0.9	57.3	.53	140
			Our Method	NRL (C=0.3, scaling=0.5, sample_fraction=0.80, selection_threshold=0.20) Model = GBC	**68.2**	**.66**	**63**
Location	105	10238	KNN	K=7	61.9	.56	N/A
			Lasso	Alpha=0.3	50.5	.46	513
			Our Method	NRL (C=0.35, scaling=0.5, sample_fraction=0.85, selection_threshold=0.60) Model = LogR (C=0.35 - same as NRL)	**69.5**	**.62**	**10**
Phone Usage	111	15447	KNN	K=8	60.4	.35	N/A
			Lasso	Alpha=0.3	47.7	.37	261
			Our Method	NRL (C=0.6, scaling=0.5, sample_fraction=0.80, selection_threshold=0.50) Model = GBC	**70.3**	**.75**	**3**
Sleep	107	5890	KNN	K=10	49.5	.41	N/A
			Lasso	Alpha=1.0	44.9	.34	282
			Our Method	NRL (C=0.6, scaling=0.45, sample_fraction=0.80, selection_threshold=0.20) Model = GBC	**69.2**	**.66**	**74**
Steps	107	3055	KNN	K=9	**66.4**	.50	N/A
			Lasso	Alpha=0.3	62.6	.57	305
			Our Method	NRL (C=0.75, scaling=0.6, sample_fraction=0.80, selection_threshold=0.20) Model = LogR (C=0.75 - same as NRL)	63.6	**.53**	**80**
All	79	62311	KNN	K=3	64.6	.58	N/A
			Lasso	Alpha=0.7	59.5	.53	480
		Predictions from the 7 feature sets	Our Method	Combined predictions from the all seven 1-feature set models using AdaBoost (n_estimators=100)	**82.3**	**.78**	**310**

3.6.3 Implications for Longitudinal Studies and Opportunities to Improve Model Performance

Chapter 2 (especially table 2.1 in chapter 2) shows that our depression detection results are either better than or comparable to the current state-of-the-art. Further, our change of depression detection and depression prediction extend the current state-of-the-art. However, depression detection and prediction using mobile and wearable sensing is a fairly novel area of research, and there are significant opportunities to improve the accuracy of our models in future work. To this end, we have identified two possible kinds of sources of errors: (1) Errors that occur due to modeling, and (2) Errors that occur due to poor quantity or quality of data collected. Opportunities to mitigate these errors are described below.

The small sample size of our dataset contributes greatly to the errors that occur due to modeling. Increasing the sample size for training by collecting data from more people will increase the robustness and generalizability of our models and reduce error due to variance (*i.e.*, error due to small fluctuations in the training data), thereby improving accuracy. For this study, we started out with 188 participants but were left with 138 participants by the end of the study. 50 participants either dropped out, failed to answer depression questionnaires, or were missing much of their passively collected data due to technical issues. Hence, in order to increase the sample size, researchers will have to take a multi-pronged approach by (1) recruiting more participants, (2) encouraging compliance and reducing drop-out rates by offering additional or more engaging incentives (*e.g.*, interventions to improve their well-being), and (3) improving quantity or quality of data collected. Further, some participants may exhibit behavioral symptoms that are different from the rest of the population. Hence, in the future, researchers should investigate building personal models for each participant, such that each personal model contains weekly samples from 1 participant only, in order to predict the weekly depression labels for that participant. This kind of study will be challenging though, since self-report data will have to be collected over a much longer period of time.

We are currently repeating this study with a new cohort of first year undergraduate students from the same University and a subset of the now second year undergraduate students whose data was used in the analysis presented in this study. This will allow us to compare behaviors from the same participants 1 year apart and their effect on depression, as well as build more stable models by training on a larger sample size and reporting test accuracies. This study is also being repeated at another University which will allow us to compare behavioral symptoms of depression and test the validity of our models across universities. We have also significantly improved our system and protocol for monitoring data collection daily throughout the study, which should greatly improve the quality of data

collected.

To improve the quantity or quality of passively collected data, researchers need to monitor data collection daily and promptly address technical issues that cause noisy or missing data, as they arise. To this end, we have implemented a dashboard that shows us the amount of data received by the server from each participant daily. This allows us to reach out to participants and resolve data collection or data transmission issues that are causing missing data. In addition, efforts to encourage compliance and reduce drop-out rates will help improve the quantity or quality of data collected through questionnaires.

3.6.4 Implications for Privacy and Technical Limitations

The results of our feature ablation study show that we do *not* need data from all the sensor streams we recorded. In fact, combining features from fewer sensor streams often leads to better performance. For example, for detecting post-semester depression, a model containing features from all 7 sensors give us an accuracy of 82.3% while a model containing data from 4 sensors gives us an accuracy of 85.7%. This demonstrates an opportunity for algorithms that minimize data collection burden (*e.g.,* privacy, data transfer rate) while maximizing value (*i.e.*, model performance metrics like accuracy) for detecting mental health outcomes. As an example, consider our detection results. In both our detection outcomes for depression (post-semester and change in depression), the best set model did not include Location, while for one outcome, it did include Campus Map. This means that from a privacy perspective and from a battery usage perspective, detailed and granular Location data is not needed, and instead human-understandable location (*i.e.*, Campus Map) labels are sufficient for well-performing models. As stated earlier, the human cost of obtaining Campus Map and Calls features is higher than for the other feature sets. Anyone implementing a detection system like the ones we have proposed in this study, has to trade off this burden against the loss in accuracy that they might induce (*e.g.*, 3.2% loss in post-semester detection, and 8.9% loss in detecting change). Further, any features or feature sets that do not contribute to our best models means a reduction in the amount of data transferred from the phone to a back-end server. This also reduces battery usage, and potential financial costs to the participant depending on the data plan they have paid for.

To optimize for these types of burdens, Early *et al.* [72] present a method that dynamically chooses sensors and switches between them during data collection, thereby reducing data collection costs while achieving equivalent or better model performance. This method can be extended to our work in detecting mental health outcomes in college students.

3.6.5 Generalizability of our approach to detecting post-semester and change in loneliness

We used our approach to predict loneliness and estimated overall levels of loneliness and change in loneliness with a high accuracy of 80.2% and 88.4%, respectively. For predicting post-semester loneliness, the average accuracy obtained across all feature sets using LASSO was 56.7% and using our approach was 65%. Table 3.2 shows a comparison between our novel feature selection approach *i.e.* Nested Randomized Logistic Regression and LASSO-based feature selection. Compared with LASSO, the average number of selected stable features (features selected in all cross-validation folds) is 3 times smaller in our approach. Hence, our approach outperforms Lasso by substantially reducing the size of the feature vector. These findings for loneliness are aligned with what we discussed for depression in section 3.6.2. Hence, our method generalizes well to detecting loneliness and is able to detect depression and loneliness related outcomes with high accuracy by mitigating the effects of the curse of dimensionality in the feature space. This gives us strong reasons to believe that our approach will also generalize well to other outcomes that are frequently co-morbid with depression such as anxiety, chronic medical conditions, and sleep issues.

			Number of features selected during cross-validation process			
Feature set	**Number of features**	**Number of samples**	**LASSO**		**NRLR**	
			In all folds	**In at least one fold**	**In all folds**	**In at least one fold**
Bluetooth	3201	115	203	1026	278	1864
Calls	605	108	30	134	34	142
Campus Map	16,381	111	66	455	12	161
Location	10,237	106	345	784	14	124
Screen	15,446	113	96	467	8	52
Sleep	5889	107	87	534	23	266
Steps	3055	107	270	485	0	8
Average	7831	110	157	555	53	374

LASSO: least absolute shrinkage and selection operator
NRLR: Nested Randomized Logistic Regression – Our novel feature selection approach

Table 3.2: The list of feature sets with the number of features and data samples used in the machine learning pipeline after handling missing values and the number of features selected by our approach (Nested Randomized Logistic Regression) and LASSO during the cross-validation process for detecting post-semester loneliness

Other than the study by Pulekar et al. [198] that analyzed 2 weeks of data from 9 students using a small set of features from smartphones only and the study by Sanchez et al. [220] that inferred different types of loneliness in 12 older adults using one week of mobile data,

we are unaware of any existing study to detect loneliness from longitudinal passive sensing data using machine learning.

3.6.6 Extending to Other Health Outcomes and Opportunities for Combining with Verbal and Non-Verbal Behaviors, and Genomic Data

We evaluated our ML pipeline in the context of depression and then assessed its generalizability to detecting loneliness, but it can be generalized to any chronic and longitudinal health problem. Further, depression has temperamental (cognitive), environmental (*e.g.*, childhood experiences, lifestyle), and genetic and physiological prognostic and risk factors [7, 17]. While we are able to detect depression by sensing daily behaviors, incorporating verbal and non-verbal behaviors and genomic data into our model will lead to a more holistic and unified model of depression [17]. This can help us predict depression before its onset more accurately, estimate prognosis after onset, and develop a better understanding of depression and its causes, thereby enabling more effective treatments and interventions for depression. We can do this by capturing cognitive (*e.g.,* negative beliefs [16]) and environmental (*e.g.*, abuse) factors using verbal behaviors from ecological momentary assessments [234] and social media posts [56], physiological (*e.g.*, response to stress) factors using wearable physiological sensors (*e.g.*, heart rate sensors) and hormonal testing (*e.g.*, saliva testing for stress hormones), and genetic factors using genomic sequencing. Large initiatives such as the UCLA Depression Grand Challenge and the Precision Medicine Initiative are already working on combining these different sources of data to detect and understand depression and other health-related outcomes. We plan to contribute to these initiatives by open sourcing our feature extraction library which will allow researchers to extract tens of thousands of behavioral and behavioral change features from a wide variety of sensor streams.

3.7 Conclusion

In this study, we present a new feature selection approach that allows us to select meaningful features even when the number of features is significantly larger than the sample size. This approach enables models that detect depression at specific time points while considering a large set of features computed over the previous several weeks. We evaluate our approach by

https://grandchallenges.ucla.edu/depression/
https://ghr.nlm.nih.gov/primer/precisionmedicine/initiative

identifying students that have post-semester depressive symptoms using data collected over one semester (16 weeks) from the smartphones and fitness trackers of 138 college students, and achieve an accuracy of 85.7%. Further, we detect whether students' depressive symptom severity changed with an accuracy of 85.4%, and the levels of change with an accuracy of 82.9%. Models that detect change in depression are novel, and will likely be better at evaluating interventions than diagnostic models. Hence, our work has significant implications for depression detection and monitoring, and longitudinal symptom monitoring in-the-wild. Ultimately, it creates the potential for technology-mediated interventions that support the diagnosis, treatment, and prevention of depression. For example, a system built on data from these sensors can provide real-time feedback and alert the user before a depressive episode occurs. Such interventions could help increase awareness and motivate students to seek treatment and affect behavior change.

We also assess the generalizability of our approach in detecting loneliness which is frequently co-morbid with depression, and achieve a high accuracy of 80.2% and 88.4% for detecting post-semester and change in depression, respectively. We discuss how our novel feature selection approach outperforms existing approaches by more efficiently reducing the feature space, thereby mitigating the curse of dimensionality in the feature space.

In the future, features related to daily behaviors from our work can be combined with features related to verbal and non-verbal behaviors, and genomic data to develop a better understanding of depression, loneliness, and other co-morbid conditionms, and their causes, predict mental health conditions before their onset and prognosis after onset, thereby enabling more effective and personalized treatments and interventions for mental health.

3.8 Addressing the Curse of Dimensionality

This section explains how this study addressed the curse of dimensionality challenge first introduced in chapter 1.

3.8.1 W.r.t. the feature space (C1)

In this study, we developed a novel feature selection technique called Nested Randomized Logistic Regression (NRLR) that allowed us to select meaningful features indicative of depression and loneliness from longitudinal data by integrating data from multiple time slices and sensors while decomposing and reducing the dimensionality of the feature space.

In section 3.6.2 we discuss how NRLR outperforms and selects a much smaller number of features when detecting post-semester depression than two off-the-shelf approaches – KNN

and Lasso. Similarly, in section 3.6.5, we discuss how NRLR outperforms and selects a much smaller number of features when detecting post-semester loneliness than Lasso. This shows that our approach increases model performance by effectively reducing high dimensional feature spaces, thereby mitigating the effects of the curse of dimensionality in the feature space.

While the results show that our approach mitigates the curse of dimensionality in large feature spaces, does it also work for smaller feature spaces? That is, say we collect data from fewer weeks, can we continue to use this approach, or would we need to build a different pipeline from scratch? To answer this question, I leveraged this approach to predict depression several weeks in advance in study 2 (chapter 4). Further, in study 3 (chapter 5), I explore the generalizability of this approach to health outcomes in a sample of participants from all walks of life (as opposed to the homogeneous sample of first year college students in this study).

3.8.2 W.r.t. multiple co-morbidities (C2)

Our approach is able to detect with high accuracy 4 outcomes related to depression and loneliness which are often co-morbid health conditions. This shows that our approach is able to generalize well to identify multiple comorbid conditions. Sections 3.6.2 and 3.6.5 also show that method produces good results for both depression and loneliness using the same "mechanism", that is, by reducing the size of the feature space. These consistent findings give us strong reasons to believe that our approach will generalize well to outcomes beyond depression and loneliness too. By being predicting multiple co-morbid outcomes to give clinicians a more holistic picture of the patient's health, our approach addresses the curse of dimensionality w.r.t. multiple co-morbid outcomes.

Given the complex multidimensional nature of mental health, I have had to address the challenge of the curse of dimensionality w.r.t. multiple co-morbidities in almost every study I ran. In studies 1 (chapter 3) and 3 (chapter 5), I address this challenge by demonstrating our ability to predict multiple co-morbid outcomes. Whereas, in study 4 (chapter 6), I address this challenge by combining multiple co-morbid outcomes into one outcome to be used as the target outcome to optimize via intervention.

Chapter 4

S2: Forecasting End of Semester Depression In College Students

4.1 Introduction

In study 1 (see chapter 3), we presented a machine learning approach that uses data from mobile and wearable sensors to detect and monitor depression and change in depression at any time point, with limited ground truth data. In this study, we use the *same dataset and pipeline* to predict depression several weeks in advance. Previous work on prediction has only looked into predicting depression 0-2 weeks in advance and it may not leave enough time for interventions [32]. Our work is the first to demonstrate that it is possible to predict depression several weeks in advance. We are able to identify students who will have depressive symptoms by the end of the semester with an accuracy of 81.3%, 11 weeks before the semester ends.

For information about details and methodology, refer to chapter 3.

4.2 Prediction Models for Predicting Future Depressive Symptoms

Being able to predict post-semester depression and change in depression, using data from a limited number of weeks from the beginning of the semester can help us identify students at-risk for depression and get them treatment early. For each week, we trained 1-feature set models on features from the beginning of the semester to the end of that week, and combined all available 1-feature set models to obtain the final outcome label for that week.

To understand this clearly, it is important to recall (from section 3.3.1 in chapter 3) that we only have 1 sample per person and the sample or feature vector for each person contains features averaged over different levels of granularity – each week, each half-semester, and the full semester. So when we exclude a week from our analysis, we exclude all features averaged over that week as well as features averaged over the full semester and the half-semester that that week belongs to. For example, in week 1, the feature vector for each person will only contain features averaged over week 1. Whereas, for week 15, the feature vector for each person will contain features averaged over each week from week 1 to 15, as well as features averaged over the first half of the semester. Model parameters were tuned at each time step for all these models.

Canzian and Musolesi [32] investigated the possibility of predicting depression 1-14 days in advance using location features, and achieved acceptable results 13-14 days in advance. In fact, they obtained very similar results at different time points in their analysis. For

example, results obtained 13 days in advance were as good as the results obtained 0 days in advance (see Figure 9 of [32]). Based on their results, we hypothesize that we do not need data from 16 weeks to predict depression, and we do not expect the prediction accuracy to monotonically increase as we add features from subsequent weeks. Even though our detection model contains all the features from the previous weeks' prediction models, we hypothesize that it is possible for some prediction models to outperform the detection model since feature selection in machine learning is rarely optimal. Features from certain weeks can add "noise" to the model and reduce the accuracy obtained after those weeks. For example, students may deviate from their regular behavior during weeks 6-9 which include preparing for midterm exams, and spring break, and weeks 15-16 which include submitting final projects and preparing for final exams.

4.3 Results for Early Prediction of Future Depressive Episodes

This section describes initial results obtained for predicting future depressive episodes using data from the beginning of the semester up to a certain number of weeks until the prediction point. It addresses the question "How early can we predict the two outcomes and with what accuracy?"

Figure 4.1 contains 2 sub-figures, corresponding to our two outcomes. In each graph **on the left side**, the horizontal axis indicates the week up to which features are included in a model and the vertical axis indicates the accuracy and F1 score that the model obtains. For example, "7" on the horizontal axis means we include features from the start of week 1 to the end of week 7, and the corresponding value on the vertical axis indicates the accuracy a model trained on features from weeks 1 to 7. The best 5 models (with highest accuracies) are labeled. We combine all seven 1-feature set models at each time step, and tune model parameters for them. As mentioned in 3.3.1, we concatenate features from different weeks in order to capture the variability in behaviors across weeks. In the graphs **on the right side**, at each time step, we take the predictions for every participant made by all models up to that time step (as shown in the graph on the left side) and use majority voting to determine the final prediction for every participant. For example, if at least 50% of the models at weeks "1", "2", and "3" predict a participant p as "may have depression", only then will participant p be labeled as "may have depression" in week 3. The graphs on the right side show the final performance obtained when majority voting is applied to the predictions of the models whose performance is shown in the graphs on the left side.

As explained in section 4.2, **for the graphs on the left side**, we do not see the prediction accuracy monotonically increase as we add features from subsequent weeks. This is expected and also aligned with previous work [32]. In fact, these prediction models (trained on features from fewer weeks) sometimes outperform the corresponding detection model (trained on features from all weeks) because feature selection in machine learning is rarely optimal. Further, these weeks also have semantic meaning, such that adding data from certain weeks can increase predictive power or introduce noise, thereby affecting accuracy. For example, students have midterms from the beginning of week 7 and 1-2 days into week 8, and spring break during the remainder of week 8 and most of week 9. They typically return to school towards the end of week 9, and weeks 10 and 11 are their first two weeks of regular schoolwork after spring break. While we know what happens in these weeks and our prediction accuracy in the following sections peaks and drops for specific weeks, we cannot associate causality to these results since we do not have any ground truth to support such findings. For example, while most students should have midterms in week 7 or the first 1-2 days of week 8, we don't know the specific days they had their midterms and there may be students who had no midterms at all.

The instability of model performance across weeks makes it harder for the university staff carrying out interventions to trust the output of the model in any one week. Hence, we propose that university staff should look at the predictions from all models previously trained before each time step, and contact participants that are repeatedly labeled as at-risk. Mathematically, this can be achieved using majority voting. In figure 4.1, **the graphs on the right side** show that after majority voting, performance of the models greatly stabilizes across the 16 weeks. Hence, instead of trusting the output of the prediction model from a specific week, we recommend that the university staff contact at-risk participants every week as long as they have been predicted as at-risk by at least 50% of the models trained until that week.

Predicting Post-semester Depression

The baseline for predicting post-semester depression is 59.4% (see section 3.4a). Out of the five best prediction models, the model which allows for the earliest prediction needs data from weeks 1 to 5 and achieves an accuracy of 81.3% (N = 80), as shown in figure 4.1a (left). Hence, we are able to predict post-semester depression with an accuracy significantly better than the baseline as early as the end of week 5. In figure 4.1a (right), we see that the performance of the prediction models increases quite steadily across the 16 weeks when using majority voting. Therefore, contacting at-risk participants that were labeled as "may

have depression" by at least 50% of the models trained until the end of each week, is more reliable and can be repeated every week.

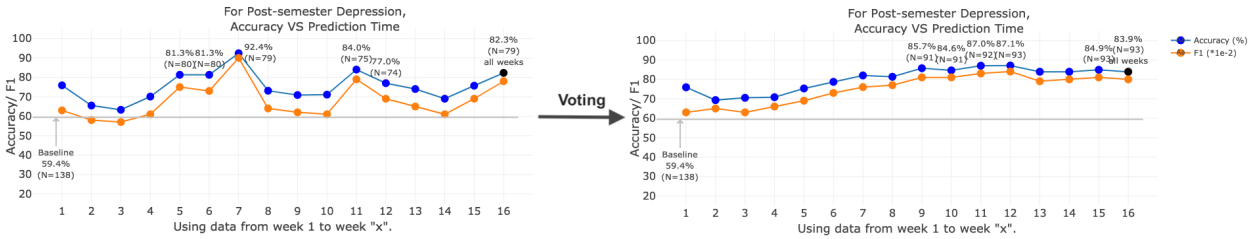
Predicting Change in Depression

The baseline for predicting change in depression is 65.9% (see section 3.4.3). Out of the five best prediction models, the model which allows for earliest prediction needs data only from weeks 1 to 2 and achieves an accuracy of 88.1% (N = 84), as shown in figure 4.1b (left). Hence, we are able to predict change in depression with an accuracy significantly better than the baseline as early as the end of week 2. In figure 4.1b (right), we see that when using majority voting, the performance of the prediction models increases quite steadily across the 16 weeks, with weeks 7 and 9 being the only exceptions. Therefore, contacting at-risk participants that were labeled as "depression may worsen" by at least 50% of the models trained until the end of each week, is more reliable and can be repeated every week.

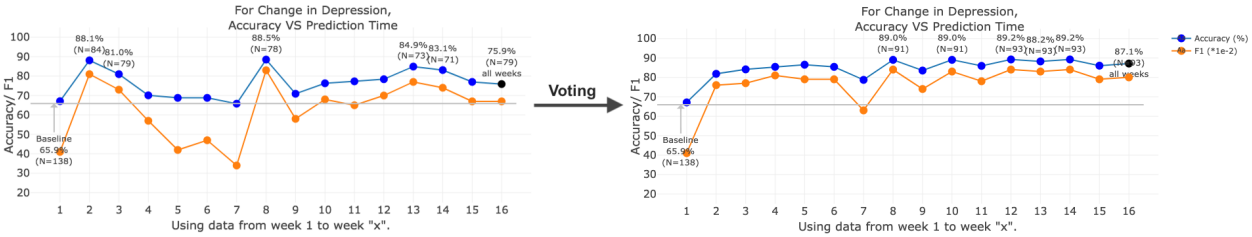
4.4 Implications for Interventions

Our machine learning approach enables building behavior models for early detection and prediction of change in depression without frequent ground truth data. This provides opportunities for timely interventions and treatments. We have discussed the implications of this work with mental health experts. They seem very excited about this research as they believe that this system can help them screen students for depression more efficiently. They also want to help us take this research forward by identifying modified behaviors that can be targeted during behavioral change interventions to improve depressive symptoms. In our sample of 138 participants, 40.6% students were found to have depressive symptoms post-semester, however only 17.4% students self-reported seeking counseling and psychological services. In subsequent studies with in-built interventions, we plan to use our system to identify students with depression and reach out to them through the student counseling center. Detecting post-semester depression allows us to identify students who may have a depressive disorder at the end of the semester. Detecting change in depression allows us to identify students who have worsened, such that we can intervene urgently and more aggressively if needed. Models for detecting change in depression may also be more sensitive to changes resulting from interventions, and hence, better at evaluating their effectiveness.

The drop in performance in weeks 7 and 9 is probably due to atypical behavior during midterms and spring break.



(a) Accuracy and F1 over time for predicting Post-semester Depression using data from limited number of weeks starting at the beginning of the semester.



(b) Accuracy and F1 over time for predicting Change in Depression using data from limited number of weeks starting at the beginning of the semester.

Figure 4.1: Accuracies over time obtained when predicting (a) Post-semester Depression, and (b) Change in Depression using data from a limited number of weeks starting at the beginning of the semester (week 1 to week x). In the graphs on the left side, for each time step, detections from all available feature sets are combined to get the corresponding accuracy. In the graphs on the right side, at time step x, majority voting is used on predictions from models at weeks 1 to x (shown on the left) to obtain more stable performance across the 16 weeks.

Further, since our models are understandable, that is, they are built using meaningful behavioral features, they can be used to inform therapists treating students about the relationship between the students' behaviors and depression. As a result, therapists will be able to make more informed choices about which interventions would be most effective for each student. Students will also be able to participate in technological self-help interventions. For example, students can be shown visualizations of their sensed behaviors (features from our model) and their relationship with depression, thereby enabling guided self-reflection and planning for behavioral change.

The prediction models that predict post-semester depressive state and change in depression weekly, enable us to reach out to students who may be at-risk for depression as early

as 1 to 5 weeks into a semester, in order to execute interventions to preempt depressive symptoms. However, we also find that the performance of these models trained at the end of each week varies over the 16 week period, instead of monotonically increasing. While this is expected behavior (and seen in previous work [32]) due to the weeks having semantic meaning, it makes it harder for university staff carrying out interventions to trust the output of the model at the end of any one week. Hence, to address this problem, we carried out additional analysis (*i.e.*, majority voting) and accordingly suggest an intervention strategy that utilizes our models. That is, we recommend that instead of trusting the output of one model at the end of a specific week, university staff should contact students predicted to be at-risk at the end of each week by a majority of *all* the models trained until that time point. We show that using this strategy would result in more stable accuracy and F1 values across the 16 weeks of the semester, and can thus be trusted more.

Detecting and monitoring depression in a large sample of students can also help inform policy changes at the university level, such as increasing outreach for psychological services, hiring more mental health professionals, and deciding drop deadlines for courses.

4.5 Conclusion

Our work is the first to demonstrate that it is possible to predict depression several weeks in advance with an accuracy of >80% (*e.g.*, 81.3%, 11 weeks before the end of the semester). Being able to predict depression several weeks in advance can allow time for preemptive interventions that can prevent depressive symptoms before their onset.

4.6 Addressing the Curse of Dimensionality

This section explains how this study addressed the curse of dimensionality challenge first introduced in chapter 1.

4.6.1 W.r.t. the feature space (C1)

In this study, we applied a novel feature selection technique (first developed in study 1) that allowed us to select meaningful features indicative of depression from longitudinal data by integrating data from multiple time slices and sensors while decomposing and reducing the dimensionality of the feature space. We were able to use this technique to predict post-semester depression using data from week 1 to week x of the semester (where $x = 2$

to 16 weeks of the semester). Since the features from different weeks are concatenated to form the feature vector for each participant, this showed that our feature selection approach can handle feature spaces of different sizes by successfully reducing them in a way that mitigates the curse of dimensionality in the feature space. In study 3 (chapter 5), I explore the generalizability of this approach to health outcomes in a sample of participants from all walks of life (as opposed to the homogeneous sample of first year college students in this study).

Chapter 5

S3: Predicting the Mental Health of People with Multiple Sclerosis during the COVID-19 Stay-at-Home Period

5.1 Introduction

The coronavirus disease 2019 (COVID-19) pandemic and the ensuing response (*e.g.,* lockdown and social distancing) have broad negative impacts on physical and mental health worldwide [54, 184, 78, 258, 132, 179, 137]. The effect is more pronounced for people with chronic neurological diseases such as multiple sclerosis (MS) [166, 280, 277]. People with MS (pwMS) have a significantly higher burden of mental health comorbidities than the general population. In particular, pwMS have a 50% lifetime prevalence of depression, 2-3 times higher than the general population [187, 35, 240]. Given its association with higher disability and mortality, depression is a major comorbidity that lowers the quality of life [235, 81, 187, 282, 59, 86, 12, 244]. Further, pwMS have greater COVID-19 risk due to certain immune disease-modifying therapies as well as their physical disability, and many have experienced drastic change in their neurological care due to the pandemic [140]. Concerns for COVID-19, coupled with decreased social support and healthcare access during the pandemic, have contributed to even higher stress and depression in pwMS [277, 260, 145, 141].

During the pandemic, digital technologies have become invaluable for supporting social interaction, healthcare access, and health monitoring. Digital health tools can also measure an individual's mental health profile based on passive (non-invasive) tracking. Given the complexity and heterogeneity of real-world behaviors, models that leverage different aspects of an individual's daily behaviors are necessary to accurately predict mental health status. Relevant to depression in pwMS, clinicians could use this digital passive sensing approach to potentially identify patients who require urgent health interventions.

Past research has leveraged passively generated data from personal digital devices (*e.g.,* smartphones and fitness trackers) to capture human behavior and predict health outcomes. This moment-by-moment, in situ quantification of the individual-level human phenotype using data from personal digital devices is known as digital phenotyping [114]. Previous works using passively sensed smartphone and wearable data to predict physical disability and fatigue in pwMS have been exploratory in assessing the feasibility of data collection and the preliminary association between sensed behaviors and outcomes [169, 231, 45]. However, the clinical applicability of digital phenotyping to inform clinical outcomes in pwMS in the real world has not yet been established.

Here, we present a machine learning approach leveraging data from the smartphones and fitness trackers of pwMS to predict their health outcomes during a mandatory "stay-at-home" period of the pandemic. Building on an existing analytical pipeline [41], we

quantified behavioral changes during the "stay-at-home" period when compared to the preceding period, and used the changes to predict the presence of patient-reported outcomes of depression, neurological disability, fatigue, and poor sleep quality during the "stay-at-home" period. This study differentiates from prior studies by examining the clinical utility of digital phenotyping with passive sensors for predicting health outcomes during the early wave of the COVID-19 pandemic in a unique natural experiment. The study has relevance for predicting the health outcomes of patients with chronic and complex conditions beyond MS during major stressful scenarios (*e.g.,* pandemics, natural disasters) that could considerably alter behaviors.

5.2 Methods

To briefly summarize our approach, we used data from 3 sensors in participants' smartphones (calls, location, screen activity) and 3 sensors in participants' fitness trackers (heart rate, sleep, steps) to predict patient-reported outcomes of depression, global MS symptom burden, fatigue, and sleep quality. We computed behavioral features from these 6 sensors before and during the stay-at-home period, and took the difference as a measure of behavioral change resulting from the stay-at-home mandate. We then used changes in behavioral features to predict the outcomes.

5.2.1 Participants

The study included adults 18 years or older with a neurologist-confirmed MS diagnosis who owned a smartphone (Android or iOS) and enrolled in the Prospective Investigation of Multiple Sclerosis in the Three Rivers Region (PROMOTE) study, a clinic-based natural history study at the University of Pittsburgh Medical Center (UPMC) [140, 139, 146]. The institutional review boards of University of Pittsburgh and Carnegie Mellon University approved the study. All participants provided written informed consent.

5.2.2 Study Design

Participants downloaded a mobile application to capture sensor data from their own smartphones and additionally received a Fitbit Inspire HR to track steps, heart rate, and sleep. Data were continuously collected from smartphone and Fitbit sensors of 56 participants during the study period (16 November 2019 to 15 May 2020, including the local stay-at-home period).

All the participants completed data collection for a pre-defined period of 12 weeks while 39 agreed to extend data collection for an additional 12 weeks (for a total of 24 weeks). Six participants who did not have sufficient data during the period before the stay-at-home mandate were excluded from the machine learning analysis.

5.2.3 Survey Response and Patient-Reported Outcomes

All participants completed a baseline questionnaire, which queried their demographics and baseline health outcomes, on the Saturday following enrollment. During the study, participants completed additional questionnaires as described below at intervals according to each questionnaire. All questionnaires were administered online using the secure, web-based Research Electronic Data Capture (REDCap) system [101, 100].

Depression: We used the Patient Health Questionnaire (PHQ-9) to measure the severity of depression symptoms *once every two weeks* [131]. PHQ-9 contained 9 questions, with each answer being scored on a scale of 0-3. Higher scores indicated more severe depressive symptoms.

Global MS symptom burden: We used the Multiple Sclerosis Rating Scale - Revised (MSRS-R) to measure global MS symptom burden and neurological disability *once every four weeks* [270]. MSRS-R assessed eight neurological domains (walking, upper limb function, vision, speech, swallowing, cognition, sensory, bladder and bowel function; each domain scored as 0 to 4, with 0 indicating the absence of symptom and 4 indicating higher symptom burden and more severe disability).

Fatigue: We used the 5-item version of the Modified Fatigue Impact Scale (MFIS-5) to measure the impact of fatigue on cognitive, physical and psychosocial function *once every four weeks* [152]. Each item in MFIS-5 was scored on a five-point Likert scale from 0 (never) to four (almost always). Higher scores indicated more severe fatigue.

Sleep quality: We used the Pittsburgh Sleep Quality Index (PSQI) to measure sleep disturbances *once every four weeks* [30]. PSQI comprised 19 individual items, with seven component scores (each on a 0-3 scale) and one composite score (0 to 21, where higher scores indicating a poorer sleep quality).

For each outcome, we averaged the measures collected during the stay-at-home-period and then dichotomized the resulting outcomes using thresholds. For "Depression", PHQ-9 scores were dichtomized as ">=5: presence of depression" and "<5: absence of depression". For "Global MS symptom burden", MSRS-R scores were dichotomized as ">=6.4: higher burden" and "<6.4: lower burden". For "Fatigue", MSIF-5 scores were dichotomized as ">=8: high fatigue" and "<8: low fatigue". For "Sleep quality", PSQI scores were dichtomized as ">=9: poorer sleep quality" and "<9: better sleep quality". The thresholds

for depression and sleep quality were based on previous works [131, 83]. Given the lack of consensus from the literature, we calculated the median scores of the global MS symptom burden and fatigue in a larger dataset of 104 pwMS, of which the 56 pwMS in this study represented a subgroup (with data collection encompassing the stay-at-home period), and used the median scores as the thresholds.

5.2.4 Sensor Data Collection

Each participant installed a mobile application based on the AWARE framework [82] which provided backend and network infrastructure that unobtrusively collected from smartphones the location, screen usage (*i.e.*, when the screen status changed to *on* or *off* and *locked* or *unlocked*), and call logs (for incoming, outgoing and missed calls). Further, participants wore a Fitbit Inspire HR that captured the number of steps, sleep status (asleep, awake, restless, or unknown), and heart rate. Calls and screen usage were event-based sensor streams, whereas location, steps, sleep, and heart rate were time series sensor streams. We sampled location coordinates at 1 sample per 10 minutes, and steps, sleep, and heart rate at 1 sample per minute.

Data from AWARE were deidentified and automatically transferred over WiFi to a study server at regular intervals. Data from the Fitbit were retrieved using the Fitbit API at the end of the data collection. Participants were asked to keep their devices charged and carry their phone and wear Fitbit at all times.

To protect confidentiality, we removed identifiable information (*e.g.* names, contact information) from survey and sensor data prior to analysis. We followed the standard practice for sensor data security.

5.2.5 Mediation Analysis

Mediation analysis was performed using the non-dichotomized outcomes *i.e.* the average of the patient-reported outcomes collected during the stay-at-home-period. Process Macro in SPSS was used for mediation analysis [102].

5.2.6 Data Processing and Machine Learning Analysis

The data processing and analysis pipeline (Figure 5.1a) built on our prior work [41] and involved several steps:

1. Feature extraction from sensors over time slices to identify behavior changes.

2. Handling missing features.
3. Machine learning to predict patient-reported health outcomes during the stay-at-home period:
    a) Using 1-sensor models (*i.e.*, models containing features from one sensor).
    b) Combining 1-sensor models to obtain the best model for each outcome.

Feature Extraction

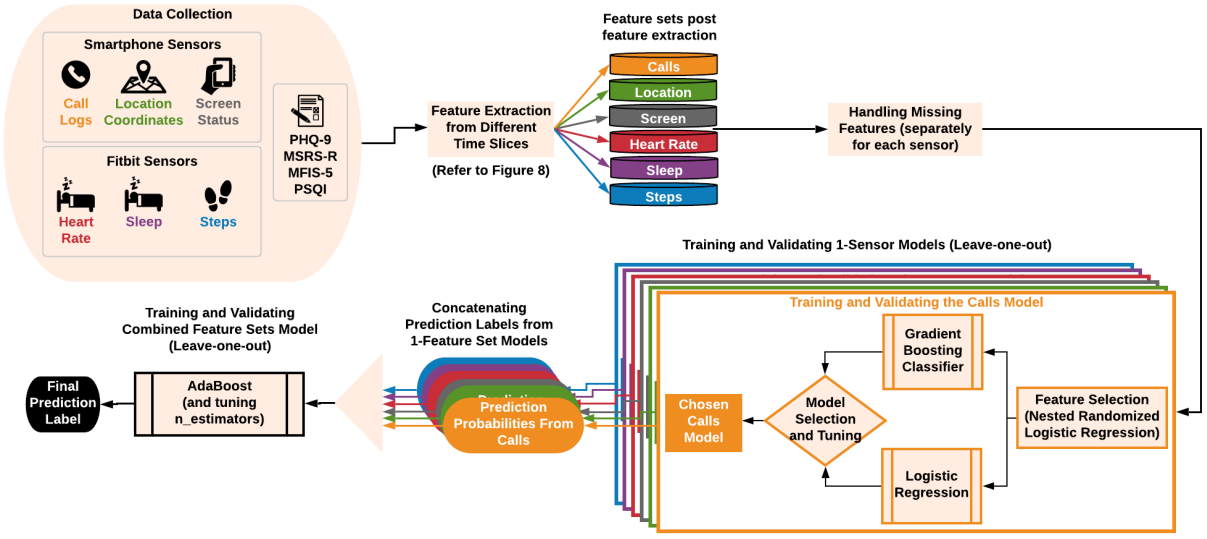
We computed features from six sensors: Calls, Heart Rate (HR), Location, Screen, Sleep, and Steps, given their potential to inform depressive symptoms [216, 266, 32, 41, 273, 274], fatigue [255], MS symptom burden such as decreased mobility [231], and sleep quality [160, 222].

Location features captured mobility patterns. Steps and Heart Rate captured the extent of physical activities. Calls features captured communication patterns. Screen features might inform the ability for concentration [58, 133] and the extent of sedentary behavior [52], despite of potential caveats for pwMS and other chronic neurological disorders. Sleep features captured sleeping duration and patterns, which could indicate sleep disturbance (e.g., insomnia or hypersomnia) associated with depression [176].

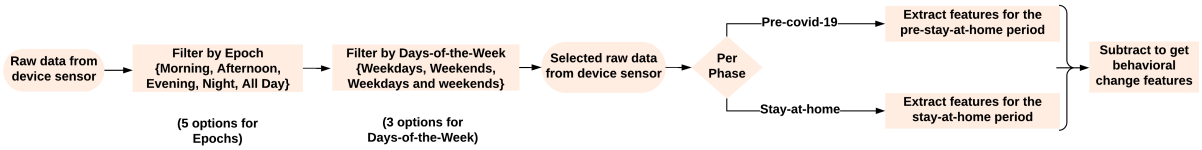
Calls Features: Calls features were calculated using the call logs from the smartphone. We extracted the following features: *Number and duration of all incoming, outgoing, and missed calls*, *number of correspondents overall*.

Heart Rate Features: Heart rate features were calculated from the heart rate over time using the Fitbit API. The following features were calculated: *Mean heart rate*, *total time spent in the "fat burn" heart rate zone*, *total time spent in the "cardio" heart rate zone*, *total time spent in the "peak" heart rate zone*, and *total time spent below any heart rate zones indicating exercise (i.e. the out of the range zone).*. The fat burn, cardio, and peak heart rate zones were calculated by Fitbit for each person separately.

Location Features: Location features were derived from the Location 'virtual' sensor of the smartphone which provided the best estimate of location based on available GPS, WiFi and cellular tower signals. We extracted the following Location features: *Location variance* (sum of the variance in latitude and longitude coordinates), *log of location variance*, and *total distance traveled*. *Circadian movement* [216] was calculated using the Lomb-Scargle method [195]. It encoded the extent to which a person's location patterns followed a 24-hour circadian cycle.



(a) Machine learning pipeline for predicting depression (PHQ-9), global MS symptom burden (MSRS-R), fatigue (MFIS-5), and sleep quality (PSQI) using passive sensors from smartphones and fitness trackers.



(b) For each sensor during the "pre-stay-at-home" period and the "stay-at-home" period, each feature was extracted from 15 time slices. The "pre-stay-at-home" features were subtracted from the "stay-at-home" features to obtain the behavioral change features. First, raw data from the device sensor were preprocessed and then filtered by a time-of-the-day epoch and a days-of-the-week option. Features were then extracted from the selected raw data.

Figure 5.1: Data Processing and Analysis Pipeline

Next, we labeled location samples as "static" or "moving" and clustered the "static" samples as described by Chikersal et. al. [41]. We then extracted: *number of significant places*, *radius of gyration* [32], *percentage of time spent at top-3 (most frequented) places*,

percentage of time spent moving, and *percentage of time spent in insignificant or rarely visited locations*. We further calculated the *average and standard deviation of length of stay at significant places* as well as *Location entropy* and *normalized location entropy* across significant places (using previously described method [216]). Higher location entropy occurred when time is spent evenly across significant places.

We assumed the place most visited by the participant at late night (between 00:00 hrs to 06:00 hrs) to be their home location and the place most visited by the participant during the afternoon (between 12:00 hrs to 18:00 hrs) to be their work location. To compute home location, we clustered the location coordinates from all nights and assumed the center of the most frequented cluster to be the participant's home location center. To compute work location, we clustered the location coordinates from all afternoons and assumed the center of the most frequented cluster to be the participant's work location center. We used the self-reported home addresses of the participant to verify their home location, and switch home and work locations if the computed home location was found to not match with their self-reported home addresses. We calculated the *time spent at home, assuming home to be within 100 meters of the home location center* based on the default geofencing radius used by automation systems like HomeKit and https://www.home-assistant.io/.

Screen Features: Screen features were calculated using the screen status sensor in the smartphone, which recorded screen status (on, off, lock, unlock) over time. We extracted the following phone usage features: *total number of unlocks*, *mean number of unlocks per minute*, *total time spent interacting with the phone*, and the *median length of bouts (or continuous periods of time) during which the participant was interacting with the phone and when the screen was unlocked*. A participant was noted to be "interacting" with the phone during the interval between "unlocked" (on) and "locked" (off) screen status.

Sleep Features: Sleep features were calculated from the daily sleep summaries and minute-to-minute sleep inferences (asleep, restless, awake, unknown) using the Fitbit API. Sleep captured by Fitbit is accurate +/- 45 minutes [147, 50, 279]. The following features were calculated from all sleep data: *total minutes asleep*, *total time in bed*, and *total sleep records*. The following features were calculated from the sleep data for the longest sleep record or "main" sleep period *i.e.*, excluding any short naps: *total time spent sleeping during main sleep*, *total time spent in bed during main sleep*, *sleep efficiency during main sleep*, which was calculated as (time asleep / (total time in bed - time to fall asleep)), *total time spent restless during main sleep*, *total number of times restless during main sleep*, *start time of main sleep in hours from midnight*, and *end time of main sleep in hours from midnight*.

Steps Features: Steps features were calculated from the step counts over time using the Fitbit API. The following features were calculated: *total number of steps*, the *number of minutes labeled as "sedentary" by Fitbit*, the *number of minutes labeled as "lightly active" by*

Fitbit, the *number of minutes labeled as "fairly active" by Fitbit*, and the *number of minutes labeled as "very active" by Fitbit*.

Features from the six sensors were extracted over a range of temporal slices (Figure 5.1b) preceding and during the stay-at-home period. For each period, we obtained the daily averages of these features by computing the average of the daily feature values. We computed features of behavioral changes by subtracting the daily averages of features during the pre-stay-at-home period from the stay-at-home period for the machine learning models.

Temporal Slicing: The temporal slicing approach extracted sensor features from different time segments (Figure 5.1b). Past work showed that this approach can better define the relationship between a feature and depression. For example, Chow *et al.* [46] found no relationship between depression and time spent at home during 4-hour time windows, but they found that people with more severe depression tended to spend more time at home between 10:00 AM and 6:00 PM. Similarly, Saeb *et al.* [217] found that the same behavioral feature calculated over weekdays and weekends could have a very different association with depression. Here, we obtained all available data (spanning multiple days of the study) from a specific epoch or time segment of the day (all day, night *i.e.*, 00:00-06:00 hrs, morning *i.e.*, 06:00-12:00 hrs, afternoon *i.e.*, 12:00-18:00 hrs, evening *i.e.*, 18:00-00:00 hrs) and for specific days-of-the-week (all days of the week, weekdays only *i.e.*, Monday-Friday, weekends only *i.e.*, Saturday-Sunday) to achieve 15 data streams or temporal slices. To extract features from each of the 15 temporal slices, we first computed daily features, and averaged daily features from the pre-stay-at-home period, and averaged daily features from the stay-at-home period. We then subtracted the pre-stay-at-home feature matrix from the stay-at-home feature matrix to obtain the behavioral change features. We concatenated the resulting 15 temporal slices of behavioral change features to derive the final feature matrix.

Feature Matrix: After feature extraction, each of the six sensors had a feature matrix, with each sample containing a participant's feature vector comprising behavioral change features from 15 different temporal slices.

Handling Missing Data

Missing sensor data can occasionally occur due to technical issues (*e.g.*, non-functioning phone/app/server, faulty or delayed data transfer) or compliance issues (*e.g.*, participant not carrying the smartphone or wearing the FitBit) but more often due to semantic reasons. For example, if a participant made and received 0 calls during a period, there would be no calls data. Thus, we encoded missing data into features since we could not differentiate whether such data were not collected or did not exist to be collected due to semantic reasons.

A missing feature during a time slice for many participants could indicate non-semantic issues such as non-functioning server. Further, a participant with many missing features could indicate non-semantic issues such as the non-functioning phone/app. To empirically determine the thresholds, we plotted the number of participants and features remaining for various thresholds and noted the largest differential in curves. Hence, we excluded all features (in a time slice) with missing values in more than 14 participants and likewise excluded participants missing more than 20% of all features. For each feature we calculated the minimum feature value, and imputed missing features as that value minus 1. As we handled missing data independently across feature time slices, the number of participants and features were different across sensors as missing features in each feature set.

Machine Learning using Nested Feature Selection

We built machine learning models to predict dichotomized outcomes using the dataset, building on a published approach [41], and validated our models using *leave-5-participants-out cross-validation* to minimize over-fitting. The model generation process followed these steps:

1. *Stable Feature Selection* using Randomized Logistic Regression, leveraging temporal slices.
2. *Training and Validating 1-Sensor Models* for each of the six feature sets: Calls, Heart Rate, Location, Screen, Sleep, and Steps.
3. *Obtaining Predictions from Combinations of Sensors* by combining detection probabilities from 1-sensor models to identify the best performing model.
4. *Classifying Different Outcomes* by running the pipeline for each outcome.

Stable Feature Selection:

To enable stable feature selection from a vast number of behavioral features, Chikersal et al. [41] proposed an approach called Nested Randomized Logistic Regression, which we deployed in this study. This method decomposed the feature space for each sensor by grouping features from the same time slices, and performed randomized logistic regression on each of these groups. The selected features from all groups (*i.e.*, all time slices) are then concatenated to give a *new and much smaller* set of features. Next, we performed randomized logistic regression again but on this *new* set of features to extract the final selected features for the sensor. We performed the nested feature selection for each of the six 1-sensor models, thereby *nesting* the process. This method was performed in a *leave-5-participants-out*

manner such that the model used to detect an outcome for a participant did not include that person during the feature selection process.

Training and Validating 1-Sensor Models: For each sensor, we built a model of the selected features from that sensor to detect an outcome. We used leave-5-participants-out cross-validation to choose the parameters for that model. We trained models using two machine learning algorithms: Logistic Regression and Gradient Boosting Classifier [41, 264]. We chose the model with the best f1-score for a given outcome, which provides the detection probabilities for the outcome. The process is independent of other outcomes.

Obtaining Predictions from Combinations of Sensors: The detection probabilities from all six 1-sensor models were concatenated into a single feature vector and given as input to an ensemble classifier, *i.e.*, AdaBoost with Gradient Boosting Classifier as a base estimator, which then outputted the final label for the outcome. For all outcomes, only the detection probabilities of the positive label "1" were concatenated. The positive label was the "presence of depression" for "depression", "high burden" for "global MS symptom burden", "severe fatigue" for "fatigue", and "poor sleep quality" for "sleep quality". The "n_estimators (The maximum number of estimators at which boosting is terminated.)" parameter was tuned during leave-5-participants-out cross-validation to achieve the best-performing combined model.

To analyze the usefulness of each sensor, we implemented a *feature ablation analysis* by generating detection results for *all possible combinations of 1-sensor models*. For six 1-sensor models, there were 57 combinations of feature sets, as the total combinations = combinations with 2 sensors + ... + combinations with 6 sensors = $\sum_{r=1}^{6} \binom{6}{r} = 57$.

Classifying Different Outcomes: This pipeline of training and validating six 1-sensor models and 57 combined models was run independently for each of the 4 outcomes. For each outcome, we reported the performance based on the best combination of sensors. We also reported the performance of baseline models (i.e., a simple majority classifier whereby every point is assigned to whichever is in the majority in the training set) as well as models containing all six sensors.

5.3 Results

5.3.1 Participant Characteristics

The characteristics of the 56 study participants were representative of the typical MS study (Median age = 43.5, 86% women). Table 5.1 shows the detailed participant characteristics.

Variable	Statistics	
Sex	**N**	**%**
Female	48	86
Male	8	14
Race	**N**	**%**
White	51	91
African or African American	5	9
Ethnicity	**N**	**%**
Non-Hispanic or Latino	55	98
Hispanic or Latino	1	2
Age	**Median**	**Interquartile Range**
In years	43.5	37 to 52
Disease Duration	**Median**	**Interquartile Range**
Years elapsed from age of first neurological symptom onset to study participation	13.0	6.7 to 17.4
Patient Determined Disease Steps (PDDS)	**Median**	**Interquartile Range**
PDDS score at the start of the study	1	0 to 3
Disease-modifying Treatment	**N**	**%**
Higher efficacy	38	68
Standard efficacy	12	21
Depression Diagnosis	**N**	**%**
Not diagnosed with clinical depression before study enrollment	39	70
Diagnosed with clinical depression before study enrollment	17	30
Pharmacotherapy for Depression	**N**	**%**
Not taking medication for depression before study enrollment	39	70
Taking medication for depression before study enrollment	17	30
Non-pharmmacotherapy for Depression	**N**	**%**
Not receiving non-medication therapy for depression before study enrollment	52	93
Receiving non-medication therapy for depression before study enrollment	4	7
Study Outcomes: Average Measures during the Stay-at-Home Period	**Median**	**Interquartile Range**
PHQ-9 (Depression)	3.7	0.0 to 7.4
MSRS-R (Global MS Symptom Burden)	7.5	3.4 to 10.3
MFIS-5 (Fatigue)	8.0	4.6 to 11.0
PSQI (Sleep Quality)	11.0	7.8 to 14.3

Table 5.1: Study participant characteristics.

5.3.2 Interrelated Outcomes

The main study outcome is patient-reported depression as well as associated neurological symptom burden, fatigue and sleep quality. We measured the Pearson correlations among the average values of the four outcomes during the stay-at-home period for the participants. Depression severity (PHQ-9) correlated with the global MS symptom burden (MSRS-R), fatigue severity (MFIS-5) and sleep quality (PSQI) (Figure 5.2)

To dissect the complex relationship among these outcomes to inform better patient mon-

	1	2	3	4
1. PHQ-9 Score (Depression)	1			
2. MSRS-R (Functional Disability)	0.602	1		
3. MFIS-5 (Fatigue)	0.713	0.73	1	
4. PSQI (Sleep Quality)	0.597	0.486	0.562	1

Figure 5.2: Correlations among the four clinically relevant patient-reported outcomes in this study. All correlations are p<0.01. N=56.

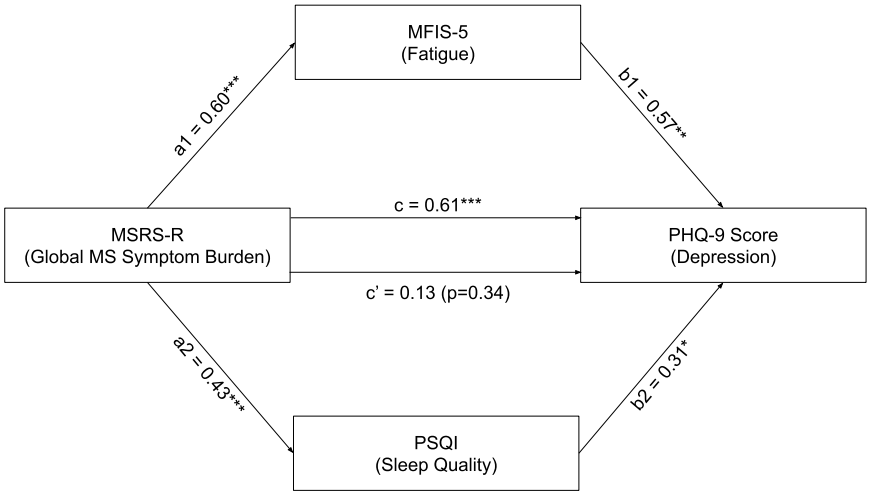


Figure 5.3: Parallel mediation analysis. Path model showing the effect of MSRS-R (measuring global MS symptom burden) on the PHQ-9 score (measuring depression) as mediated simultaneously by MFIS-5 (measuring fatigue) and PSQI (measuring sleep quality). Path c represents the effect of MSRS-R on PHQ-9 without mediators in the model. Path c' represents the effect of MSRS-R on PHQ-9 when MFIS-5 and PSQI mediators are included in the model. Paths a1b1 and a2b2 represent the effect of MSRS-R on PHQ-9 through MFIS-5 or PSQI respectively. The figure shows non-standardized β regression coefficients (*p<0.05, **p<0.001, ***p¡0.0001) as reported by PROCESS Macro in SPSS [102].

itoring and guide potentially more precise interventions, we performed mediation analysis (Figure 5.3). When MFIS-5 and PSQI were both included as mediators in the model (path c'), the association between MSRS-R and PHQ-9 was no longer significant (effect size = 0.13, and the bias-corrected bootstrap confidence intervals: -0.14 and 0.40). However, the

association between MSRS-R and PHQ-9 through MFIS-5 (path a1b1) remained significant (effect size = 0.34, and the bias-corrected bootstrap confidence intervals: 0.13-0.52). The association between MSRS-R and PHQ-9 through PSQI (path a2b2) also remained significant (effect size = 0.13, and the bias-corrected bootstrap confidence intervals are between 0.02 and 0.27). Hence, the relationship between the global MS symptom burden and depression might be mediated by both fatigue and sleep quality.

5.3.3 Predicting Outcomes during the Stay-at-Home Period

Figure 5.4 shows the performance of the machine learning pipeline for predicting each of the four outcomes using the best sensor combinations (*i.e.,* the set of sensors that had the best performance for each outcome).

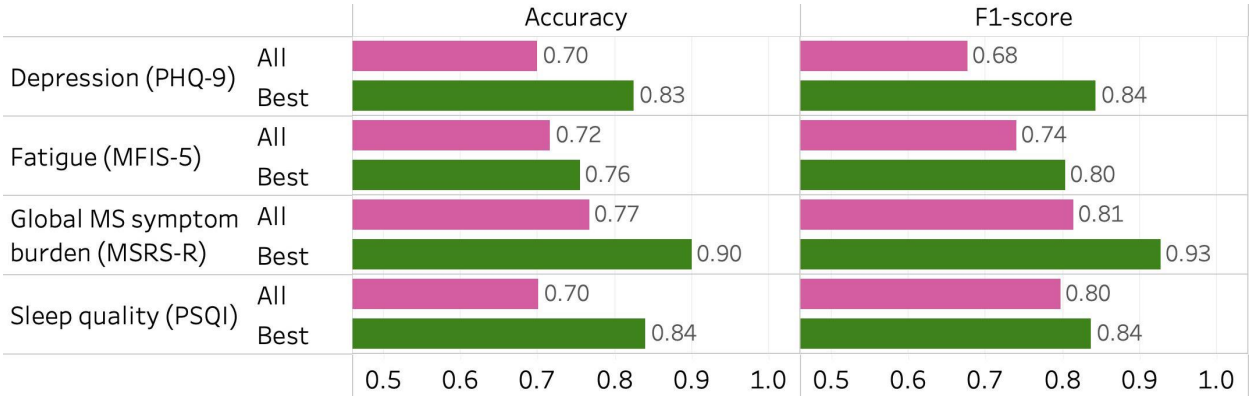


Figure 5.4: Performance of the machine learning pipeline using all sensors and the best sensor combination for predicting each of the four clinically relevant outcomes in pwMS during a state-mandated stay-at-home period. "Accuracy (All Sensors)" and "F1 Score (All Sensors)" are the accuracy ($\times$ 0.01) and F1-score obtained by combining all 6 sensors. "Accuracy (Best Sensors)" and "F1 Score (Best Sensors)" are the accuracy ($\times$ 0.01) and F1-score obtained by the best combination of sensors.

Depression

The baseline model (simple majority classifier) had an accuracy of 50.0% in predicting the presence of depression during the stay-at-home period. The model containing all sensors had an accuracy of 70% (40% improvement over the baseline). The model with the best

combination of sensors (calls, heart rate, and location) had an accuracy of 82.5% (65% improvement over the baseline).

Global MS Symptom Burden

The baseline model had an accuracy of 64.7% in predicting high global MS symptom burden (vs. "low burden") during the stay-at-home period. The model containing all sensors had an accuracy of 76.7% (18.5% improvement over the baseline). The model with the best combination of sensors (calls, heart rate, location, and screen) had an accuracy of 90% (39% improvement over the baseline).

Fatigue

The baseline model had an accuracy of 61.8% in predicting severe fatigue (vs. "mild fatigue") during the stay-at-home period. The model containing all sensors had an accuracy of 71.7% (16% improvement over the baseline). The model with the best combination of sensors (calls, heart rate, and location) had an accuracy of 75.5% (22% improvement over the baseline).

Sleep quality

The baseline model had an accuracy of 65.7% in predicting poor sleep quality (*i.e.*, "poor sleep quality" vs. "better sleep quality") during the stay-at-home period. The model containing all sensors had an accuracy of 70.2% (7% improvement over the baseline). The model with the best combination of sensors (location and screen) had an accuracy of 84% (28% improvement over the baseline).

5.4 Discussion

In this unique natural experiment conducted during the early wave of the COVID-19 pandemic, we reported the clinical utility of digital phenotyping for predicting clinically relevant outcomes for pwMS. Using only passively sensed data, our machine-learning models predicted the presence of depression, high global MS symptom burden, severe fatigue and poor sleep quality during the stay-at-home period with potentially clinically actionable performance.

The best models outperformed not only baseline models (simple majority classifier) but also models containing all sensors. The best sensor combinations for predicting depression and fatigue were the same (*i.e.* calls, heart rate, location), while these sensors were also included in the best sensor combination for predicting global MS symptom burden (*i.e.*

calls, heart rate, location, screen). Comparably, the best sensor combination for sleep quality (*i.e.* location, screen) had the smallest overlap with the sensor combinations for the other three outcomes. This observation was consistent with the finding that depression, fatigue, and global MS symptom burden were better correlated among themselves than with sleep quality (Figure 5.2). We also looked at the feature coefficients of the features selected by the best models. Examples of the best features of changed behavior selected by the best model for predicting depression (*i.e.* features with the highest absolute coefficients) include increase in number of incoming calls during evenings on weekdays, decrease in average heart rate when person is at rest or has low activity (outside exercise heart rate zones) during evenings on weekends, and increase in regularity in movement patterns in 24 hour periods with respect to nights on weekends.

Our findings built on a small body of prior work that explored the feasibility of passive sensing in pwMS and preliminary correlations between passively sensed behaviors and MS outcomes. For example, Newland *et al.* explored real-time depth sensors at home to identify gait disturbance and falls in 21 MS patients [169]. Other studies reported correlations between passively sensed physical activity and disability worsening in pwMS [231, 23, 245]. Chitnis *et al.* examined the gait, mobility, and sleep of 25 pwMS over 8 weeks using sensors mounted on their wrist, ankle and sternum, and reported correlations among gait-related features (*e.g.* turn angle, maximum angular velocity), sleep and activity, and disability outcomes [45].

Previous work on *predicting health outcomes for pwMS using passively sensed behaviors* is scarce. Tong *et al.* used passively sensed sleep and activity data collected from 198 pwMS over 6 months to predict their fatigue severity and overall health scores, and achieved good performance in line with acceptable instrument errors [255]. To our knowledge, our study is the first to use passively sensed behavior changes to predict multiple inter-related clinically relevant health outcomes in MS, including depression, disability, fatigue, and sleep quality. While several studies used passively sensed data from the general population to report behavioral changes during the COVID-19 pandemic [246, 182, 190, 113], our study provides the first real-world evidence of potential clinical utility of passively sensed behavioral changes to predict health outcomes during the unique stay-at-home period in a population with a chronic neurological disorder and complex health needs. From a methodological standpoint, the application of behavioral features computed over temporal slices to predict depression and other health outcomes in pwMS is novel. Our approach of using change in features between the period preceding the stay-at-home and stay-at-home periods to predict outcomes during the stay-at-home period is also novel. Finally, we included new heart rate features that can be computed using data from the Fitbit API.

Our approach has potential clinical utility, particularly during major stressful events

(beyond COVID-19) that worsen health outcomes and limit healthcare access. For instance, predictive models built using our approach could help patients self-monitor their health when access to in-person clinical care becomes suddenly limited and could encourage patients (or their caregivers) to actively seek medical attention sooner when the models predict adverse outcomes. Further, our models could help clinicians better monitor at-risk patients and make triage decisions for patients who require prioritization for interventions (*e.g.* medication, counseling), particularly in the setting of suddenly limited healthcare access and scarce resources.

Our study has two limitations. First, the COVID-19 pandemic started in the midst of our data collection for an ongoing larger study of pwMS. While it provided a unique opportunity to conduct a natural experiment to assess the utility of digital phenotyping to predict health outcomes in pwMS during the highly unusual stay-at-home period, we had a modest sample size of participants who happened to have sufficient sensor data collected both just before the sudden issue of the stay-at-home order and during the stay-at-home period and limited ability to seek external replication of the drastic behavior changes during the early stage of the pandemic since the stay-at-home order was lifted and has not been re-instated. To reduce the chance of over-fitting and improve validity of the findings, we used leave-5-participants-out-cross-validation, such that in each fold, the participants used for training and testing were different. Our approach performed well for not only one outcome, but all four clinically relevant outcomes pertaining to mental health and neurological disability in pwMS. We have reasonable confidence because of the consistently good model performance across all five folds and the consistently robust predictions for all four outcomes. We are not aware of other published studies with data from before and during the stay-at-home orders, particularly involving patient population with chronic neurological disorders such as MS who are at heightened risk for adverse health outcomes resulting from social isolation, reduced support, and limited healthcare access. Given the uniqueness of the data set, we believe the findings are clinically relevant despite the relatively modest sample size. Second, the study used patient-reported health outcomes. Given the restriction of in-person clinical visits during the stay-at-home period, rater-performed examination was not feasible. Importantly, these patient-reported outcomes are all validated for pwMS, highly correlated with rater-determined measures, interrelated among themselves, and clinically relevant.

5.5 Conclusion

In summary, we reported the potential clinical utility of digital phenotyping in predicting subsequent health outcomes in pwMS during a COVID-19 stay-at-home period. Specifically,

we predicted the presence of depression, high global MS symptom burden, severe fatigue, and poor sleep quality in pwMS during the stay-at-home period using passively sensed behavioral changes measured by smartphone and wearable fitness tracker. The predictive models achieved potentially clinically actionable performance for all four outcomes. This study paved the way for future replication studies during major stressful events, and has implications for future patient self-monitoring and clinician screening for urgent interventions in MS and other complex chronic diseases.

5.6 Addressing the Curse of Dimensionality

This section explains how this study addressed the curse of dimensionality challenge first introduced in chapter 1.

5.6.1 W.r.t. the feature space (C1)

In this study, we applied a novel feature selection technique (first developed in study 1) that allowed us to select meaningful features indicative of 4 health outcomes co-morbid in patients with multiple sclerosis, from longitudinal data by integrating data from multiple time slices and sensors while decomposing and reducing the dimensionality of the feature space. As explained in studies 1 and 2, this approach works by effectively reducing the size of the high dimensional feature space, thereby addresses the curse of dimensionality in the feature space. The results of this study show that our approach can be generalized to predict mental health outcomes in samples of participants from all walks of life (as opposed to the homogeneous sample of first year college students in this study).

5.6.2 W.r.t. multiple co-morbidities (C2)

Our approach is able to detect with high accuracy 4 health outcomes that are frequently co-morbid in patients with multiple sclerosis - depression, global MS symptom burden, fatigue, and poor sleep quality. By predicting each of these outcomes with high accuracy, our approach enables estimating a more complete picture of the patient's health, thereby addressing the curse of dimensionality w.r.t. multiple co-morbid outcomes.

Chapter 6

S4: Understanding Client Support Strategies to Improve Clinical Outcomes in an Online Mental Health Intervention

6.1 Introduction

Mental illness is increasing in occurrence [116]. It presents the largest cause of disability worldwide and is the strongest predictor of suicide [175, 259]. This makes the prevention and treatment of mental health disorders a public health priority [269] and has led to explorations of how the field of HCI, and the development of technology more broadly, can support access to, and increase the effectiveness of, mental health treatment [15, 200, 250, 219]. Over the last decade, this has brought forward developments of mobile apps [215, 90, 149], and computerized psycho-educational and psycho-therapeutic interventions [10, 53, 206, 271], or chat-based [84, 135, 227] programs to complement, and expand access to, psychotherapy.

Most existing digital mental health services are based on Cognitive Behavioral Therapy (CBT); the most widely applied and most extensively empirically tested psychotherapy in Western Healthcare [34]. CBT is solution-focused, teaches the person to attend to the relationships between their thoughts, feelings and behaviors, and is frequently used in treating depression, anxiety or post-traumatic stress. Its highly structured format makes it well suited for support by digital technology [**coyle2007**]. Further, extensive research has evidenced the clinical effectiveness of internet-delivered CBT (iCBT) with sustainable results comparable to face-to-face therapy [8, 10, 261, 271].

Despite these benefits, a key challenge for digital behavioral interventions like iCBT is sustaining the users' engagement with treatment [79, 123], where early disengagement and drop-out from the therapy program can mean users may not get the desired benefits. Thus, approaches to design an engaging online experience to sustain use and ensure beneficial health outcomes have become a deliberate focus for (HCI) research and development. This often means increasing opportunities for: (i) *interactivity*; (ii) *personalized experiences*; and (iii) *social support* through a community of (peer) moderators, trained supporters, or remote therapists [62, 138, 191, 204].

To aid engagement with online therapy, the involvement of a human supporter (*e.g.*, via text messages) has especially been shown to lead to more effective outcomes than unsupported interventions [118, 228, 271]. However, existing research on the effectiveness of supported-interventions has primarily assessed the impact of support duration and frequency [98, 128, 170, 252]; and to a lesser extent, different types of supporter behaviors [108, 188, 221]. Thus, it is less well understood how supporter behaviors impact program use, and therapeutic outcomes; and how this may differ between clients. Having a more nuanced

understanding of the impact of supporter behaviors on clients however could help to: better maximize the effect and outcomes of supporter involvement, assist in supporter training, and thus, increase quality of care for clients.

Simultaneously, the rise in the uptake of internet-delivered therapies and increase in the scale of automatically collected usage data from these treatment programs enables new methodological possibilities for improving our understanding of human behaviors and optimizing health outcomes (*e.g.*, [13, 120, 158]). Specifically, the fields of data mining and machine learning (ML) provide advanced computational methods to construct robust systems that can automatically learn from behavioral data [233]. These techniques have been successfully used in gaming and for recommender systems; and show great potential for advancing our understanding of health data [89] and to assist in the context of mental health (*e.g.* [76, 106, 197]).

Our work presents the first application of unsupervised machine learning, and statistical and data mining methods for analyzing complex, large-scale supporter-client interaction data to identify supporter behaviors that correlate with better clinical outcomes. Our analysis is based on a fully anonymized dataset of 234,735 supporter messages to clients (sent by 3,481 supporters to 54,104 clients) from an established iCBT program for depression and anxiety, on the SilverCloud platform (www.silvercloudhealth.com), that delivers treatment with regular feedback messages sent by a human supporter. More specifically, our work makes the following contributions:

1. We describe our approach and the challenges involved in developing computational methods for this analysis. This includes: (i) *clustering supporters* based on how the support messages they sent to clients correlate with client outcomes; (ii) *extracting linguistic features* in support messages indicative of supporter behaviors that correlate with "high" outcomes across clients in different contexts or situations (*e.g.* different levels of usage); and (iii) taking into account co-occurrent patterns of different context variables and individual support strategies, we leverage data mining to identify salient *context-specific patterns* of support.

2. Our work indicates that *concrete, positive*, and *supportive* messages from supporters that reference *social behaviors* are strongly associated with better outcomes; and that the importance of support strategies can vary dependent on a clients' specific context (*e.g.* their mental health, platform use). Based on these findings, we discuss: (i) design implications for personalized support in iCBT interventions; (ii) the need for human-centeredness in health data science; and (iii) ethical considerations for secondary data analysis.

6.2 Background for Human Support in Online Mental Health Therapy

In online mental health interventions, the role of supporters, who can be trainees or therapists, often differs from the responsibilities of a therapist in more traditional face-to-face therapy. While supporters encourage and facilitate client use of an iCBT program, the clients themselves have to learn the necessary self-management knowledge and skills that are collectively the active ingredients of the intervention [206, 228].

6.2.1 Modalities & Benefit of Human Support in iCBT

Human support in digital mental health interventions can take various forms, ranging from different communication modes (*e.g.* email, text, phone, or video chat [51, 142]), to variations in support frequency and duration [98, 128, 170, 252], and support behaviors [108, 188, 221]. Most studies on the effects of human supported iCBT apps, programs or platforms, assess the therapeutic or working *alliance* —a bond that facilitates collaborative engagement in therapy [24] —between supporters and clients. The research suggests that such an alliance can be formed within iCBT environments [178] with consistent evidence of the benefits of support in those interventions [118, 207, 228, 271]. For example, Richards & Timulak [207] studied what clients identified as helpful or hindering in their treatment, and found that clients rated the helpfulness of supporter behaviors equal to the core content of the intervention. The literature however is less conclusive on how differences in communication medium [142], frequency [51] and duration [252] of support impact outcomes. For example, Titov [252] found no difference between interventions with low-intensity support (¡3 hours) and high-intensity support (¿3 hours).

6.2.2 Human Support Behaviors & their Effectiveness in iCBT

To date, only a small number of works have explicitly studied iCBT support behaviors and their impact on client outcomes. This includes quantitative and qualitative content analysis of therapist emails to clients receiving iCBT for depression [108], anxiety [188] or bulimia [221]. Here, Sanchez-Ortiz et al. [221] analysed 712 emails and found that 95.4% of all support emails included *supportive comments*, but little *cognitive or behavioral guidance* (<15%). Paxling et al. [188] studied 490 emails and found four support behaviors to positively correlate with module completion: *task reinforcement* —making positive references to what a client has already done or achieved in the program; *self-efficacy shaping*

—prompting clients to engage in learned health promoting behaviors; *task prompting* —encouraging clients to complete the activities of the CBT program; and *empathetic utterances* —conveying an understanding of the person's suffering or life situation. Task reinforcement was further correlated with better client outcomes; whilst *deadline flexibility* (*e.g.* therapists postponing tasks) correlated negatively. Similar to task reinforcement, Holländare et al. [108] found (analysing 664 emails) *affirming* and *encouraging* behaviors (*e.g.*, validating and praising what the client did) most associated with immediate or longer-term improvements in outcomes, alongside therapist *self-disclosure*.

While previous research showed that the presence of a supporter correlates with better therapy outcomes, studies on the effectiveness of supporter behaviors remain sparse (*cf.* [108, 228]). Thus, there is a need for a deeper understanding of how supporter behaviors —as manifest in their online communications with clients —contribute to beneficial clinical outcomes.

6.3 The iCBT Intervention

SilverCloud is an established iCBT platform for the treatment of depression, anxiety, and functional impairments. Its development builds on both HCI [62] and clinical research, including randomized controlled trials that evidence the clinical effectiveness of offered treatment programs [208]. In this study, we focus on one of its most frequently used programs: treatment for depression and anxiety. Accessed online or via mobile, the program presents a self-guided intervention of seven core psycho-educational and psycho-therapeutic modules that are delivered using textual, video and audio contents as well as interactive activities, tools, quizzes and personal stories. Clients work through the program content at their own pace and time, with the recommendation to complete one module each week.

To encourage engagement and continued use, clients receive support from a trained supporter in the form of weekly reviews throughout their treatment journey. The supporters are graduate psychologists with further training in low-intensity interventions that are CBT based, including iCBT. Their support involves writing feedback messages to the client on their work, which usually takes 10-15 minutes to complete. Finally, to assess and monitor clients mental health throughout treatment, clients also complete clinical questionnaires each week, including the PHQ-9 for depression [131] and GAD-7 for anxiety [144]. Overall, the service aims to increase reach and provide effective care for those living with mental illness.

6.3.1 Frequency & Format of Supporter Interactions

Supporters typically review clients' work on a weekly basis over the 6-8 week treatment period. This serves to manage client expectations of support being restricted to certain times as opposed to immediate 24/7 availability. To this end, supporters can only see clients' activities and messages on, or after, the agreed review day. To review clients' progress, supporters have access to information about clients via usage metrics. These show: completed clinical scores of PHQ-9 and GAD-7; any client messages to the supporter; and how many content pages the client viewed, tools they used, and times they logged into the system. For each of these usage metric items, supporters can expand the view to retrieve more details about the content the client reviewed, and their responses in completing interactive tools. Clients have full transparency on what information is shared with their supporter [62].

In response to this client information, supporters compose a personalized feedback message. To facilitate this, they can select and adapt a messaging template from a drop-down menu within a specific supporter interface. These templates tend to be written by the supporters in their own words, and are then specifically tailored to each clients' situation. During training, supporters learn to personalize messages. Following prior research and guidelines this involves: referencing the clients name and things they did or said with a specific focus on activities of *task reinforcement* and *task prompting* [188]; *encouragement* [108, 207]; *guidance and advice* [207]; and effective communication using *simple language and explanations* [199]. As a final step, supporters can bookmark existing, and unlock additional therapy contents on the platform; and they select a subsequent review date. Once their message is submitted via the system, clients receive a notification email. They can view their supporter message when next accessing the program, at which time they will be automatically prompted to complete a new set of clinical questionnaires.

6.3.2 Dataset Description

Our dataset consists of information about: the supporter feedback messages; and the number and types of interactions that a client had with the platform (*e.g.*, how many CBT content pages they viewed); as well as the number and types of information that clients shared with their supporters (*e.g.*, number of journal entries, tool use data) in the time before and after each feedback message. Across the review period, we also have a subset of clinical scores indicative of the symptoms of the clients' depression (PHQ-9) and anxiety (GAD-7) before and after a supporter message. For our sample, initial PHQ-9 scores indicated that 32% of clients had minimal-to-mild symptoms of depression, 30% were moderate, 23% moderately-severe, and 15% severe. For GAD-7, initial scores showed that 36% of clients

had minimal-to-mild symptoms of anxiety, 31% were moderate, and 33% severe. Typically, each client is assigned only 1 supporter, but if that supporter becomes unavailable, they may be assigned a new one. Table 6.1 contains basic dataset statistics.

Dataset statistics	
Supporters	3,481
Clients	54,104
Clients with >1 supporters	5,967
Messages	234,735
Messages with Pre & Post PHQ-9 & GAD-7	77,006
Average message length (in words)	191
Average message length (in sentences)	9.5

Table 6.1: Overview of basic dataset statistics.

To protect full anonymity of both clients and supporters, only non-person identifiable, high-level interaction data was used for the analysis. For clients this included numbers, types and frequencies of interaction events, and aggregates of clinical scores. For supporters this meant the number, frequency and length of each feedback message. Features extracted from message texts were restricted to coarse-grained linguistics to preserve anonymity. This matched the terms and conditions of the service, and user consent, which permits the analysis of anonymous data for research purposes, and to improve the effectiveness and service tools of the treatment platform.

Our research employs ML and data mining methods to better understand what support strategies (*e.g.* use of encouraging words) characterize supporter messages that are correlated with better clinical outcomes for clients. As a first step, this requires us to identify what constitutes 'successful support messages' based on clinical outcomes. To this end, we next describe: (i) how we defined clinical outcomes as change and improvement rates in clients over time; and (ii) then used these measures in clustering to achieve three clusters of supporters whose messages correlate with either 'high', 'medium', or 'low' success rates in improving clients' clinical scores; we use the 'high' and 'low' clusters in further analysis.

6.4 Identifying Clusters of Successful Support

To better understand how different support strategies in support messages correlate with better clinical outcomes for clients, we need to identify what constitutes an appropriate 'outcome measure' in relation to supporter messages. While clients are typically asked to complete PHQ-9 and GAD-7 questionnaires immediately following the receipt of a new supporter message, clients can complete the questionnaires at any time before their next scheduled supporter review. Further, for some supporter messages, an impact on client outcomes may not be apparent immediately (week-on-week); and instead requires a focus on how outcomes develop over time.

To measure clinical outcomes over a period of time, Randomized Controlled Trials (RCTs) tend to assess the difference between the clients' initial and final clinical scores [253, 119]. While feasible in study set-ups where clients complete an 'exit survey' after a fixed period of time, this is more complicated for real-world observational data that can include client drop-out. Further, client clinical outcomes can be highly dependent on their situation (*e.g.*, loss of a loved one), symptomatology (*e.g.* seasonal or cyclic depression), personality or learning style, meaning that some clients will not improve despite the use of good support strategies in supporter messages. However, good strategies used consistently with a set of clients, should result in improvement for the majority of clients in that set. Hence, in our analysis, we focus on the 'actors' (supporters) who employ the support strategies in their messages, instead of the 'receivers' (clients). To this end, we propose computing message-level and client-level clinical outcomes for each supporter in our dataset; we describe this next.

6.4.1 Change & Improvement Rates as Clinical Outcomes

We compute the following clinical outcomes by averaging post-message change in PHQ-9 and GAD-7 clinical scores across the messages sent by each supporter. Clients should complete these questionnaires in between messages. Messages with incomplete before or after questionnaires are excluded.

1. *Message-level Change (MC):* The clinical score after a message is highly dependent on the clinical score before the message, as clients that have more severe symptoms before the message also tend to improve more on an average after the message. Hence, we measure Message-level Change as the difference between actual change and the expected change given the client score before the message. That is, for each supporter

S with NM messages, compute:

$$\frac{1}{NM}\sum_{r=1}^{NM}(actual\_change_m - expected\_change_m)$$
$$actual\_change_m = score\_before(m) - score\_after(m)$$
$$expected\_change_m = score\_before(m) - E(score\_after(m)|score\_before(m))$$

2. *Message-level Improvement Rate (MR):* If Message-level Change > 0, then the client improved more than expected post-message, and we label the message as "improved (1)". Otherwise, we label the message as "not improved (0)". For each supporter S with NM messages, we average these labels across all messages to compute this outcome.
3. *Client-level Change (CC):* While MC captures changes in clinical scores across all messages by a supporter, CC normalizes these changes across all clients of the supporter. For each supporter S, we first compute the MC for each client of S separately using the messages that S sent to them. Then, we average the MCs per client across all clients of S to get CC. *E.g.*, if a supporter sends 6 messages to client A whose change is +1 after each messages and 4 messages to client B whose change is always 0, the MC will be $\frac{6}{6+4} = 0.6$ while the CC will be $\frac{(6/6)+(0/4)}{2} = 0.5$. Thus, MC can be high even when only a few clients improve, whereas CC will only be high when these improvements are consistent across all/ many clients. This makes CC more robust to a single client's changing situations or symptoms.
4. *Client-level Improvement Rate (CR):* For each supporter S with clients NC, we first compute the average Message-level Improvement Rate using messages S sent to each client separately, and then sum these rates across all clients and divide by the total number of clients.

6.4.2 Clustering Supporters Based on Support Outcomes

When computing the above 4 outcome measures for both PHQ-9 and GAD-7, we achieve 8 outcome measures per supporter. As a next step, we use these 8 measures as 'features' for clustering. We apply K-means clustering to obtain K=3 clusters of supporters whose messages are generally linked with either 'high', 'medium' or 'low' improvements in client outcomes. The number of k=3 clusters was determined by plotting the sum of squared distances of samples to their closest cluster center against the number of clusters, and visually

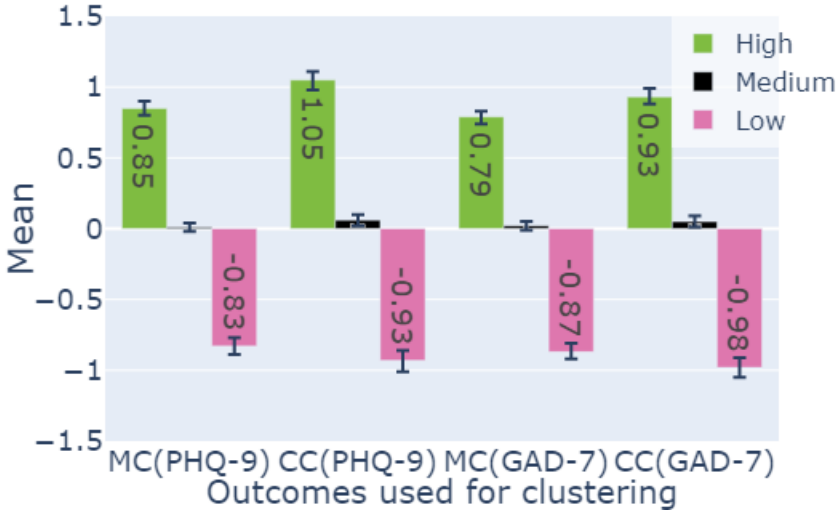
inspecting the elbow obtained. *We hypothesize that there are differences in the support messages sent by supporters in the 'high' versus 'low' outcome clusters; and that these differences will help us identify what may constitute more effective support strategies.* Table 6.2 reports the summary statistics of the three obtained clusters of supporters. Figure 6.1 shows the mean values of all 8 outcomes measures in these clusters, along with the 95% bootstrapped confidence intervals as error bars. Given the mean values of the outcomes and the narrow 95% confidence intervals, we can see that our clustering has reliably divided the supporters into 3 robust groups where these outcomes are *typically*: high, medium, and low. We did additional statistical analysis that confirmed the results in Figure 1 and showed how, independent of clients initial clinical scores, the differences in mean PhQ-9 and GAD-7 scores between the high and low clusters was significant ($p<0.05$).

Cluster	**#Supporters**	**#Clients**	**#Messages**	**#Messages Labeled**
High	438	11068	42734	14519
Medium	767	31789	123303	42740
Low	393	10828	47023	14266

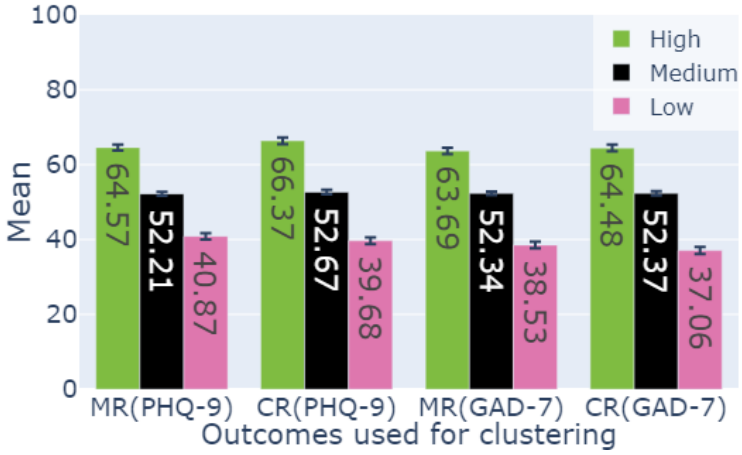
Table 6.2: Statistics for the 'high', 'medium' and 'low' outcome clusters. Only supporters with >9 (Median value) labeled messages for PHQ-9 and GAD-7 were included in the clustering.

6.5 Identifying Successful Support Strategies

As a next step, we want to identify what semantic or linguistic strategies are frequently associated with 'high' client outcomes. For this purpose, we analyze the differences between the messages sent by supporters in the 'high' outcomes cluster and those in the 'low' outcomes cluster. We are interested in identifying support strategies that occur significantly more often in the messages of supporters in the 'high' outcome cluster. Further, this difference needs to be consistent across different client contexts; meaning that the result has to be independent from variations in individual client context variables such as: the extent to which a client engages with the iCBT program ($ContentViews$), shares content with their supporter ($Shared$), the sequence number of the message *i.e.* the number of messages received so far



(a) Message-level (MC) and Client-level (CC) Changes



(b) Message-level (MR) and Client-level (CR) Improvement Rates

Figure 6.1: Means of all 8 outcomes in each supporter cluster with the 95% bootstrapped confidence intervals as error bars.

plus one: $MessageNumber$), or clients' current mental health state ($CurrentPHQ-9$ and $GAD-7$).

6.5.1 Methodology: <Strategy> Across ¡Context¿ Bins

We are mindful that the actions of supporters, as manifested in their feedback messages to clients, present a direct response to what they know about their clients' situation (*e.g.* symptom severity, level of platform use). To disentangle the clients' context from, whilst understanding the role that context can play in the use of, specific support strategies, we decided to first divide the messages of supporters in the 'high' and 'low' outcome clusters into different 'data bins' for each of 5 client context variables. This allows us to compare the differences in strategies found in support messages of the 'high' and 'low' clusters separately for different client contexts; and thereby to assess, if identified significant differences between the two groups are consistent across, and independent of, variations in client contexts. We found this approach to be more feasible than the use of context as a control variable in linear regression, which due to the large size of the data sample frequently resulted in statistically significant results, but whose effect sizes were difficult to interpret.

Next, we detail on: (i) the client *context* variables that we defined; (ii) semantic or linguistic features we extracted from supporter messages as *strategy* variables; and (iii) how we analyzed what semantic or linguistic support strategies were salient in the 'high' versus 'low' support clusters across various bins of each client *context* variable.

Client <Context> Variables

For each message i of every client, we extracted 5 *context* variables that describe the client's situation *just before the message i.e. between the* $(i-1)^{th}$ *and* i^{th} *messages*:

(1) $ContentViews$: Number of content views *i.e.* number of times the client clicked on a module or topic in the iCBT program (10 Bins: 0 views, 8 bins of 1-80 views in increments of 10 views, and >=81 views).

(2) *Shared*: Number of data instances shared with the supporter *i.e.* the total number of tools and journal entries shared, and messages sent to the supporter (4 Bins: 0 shared instances, 3 bins of 1-15 in increments of 5. We never had ¿=16.).

(3) $MessageNumber$: clients are expected to received up to 8 reviews. For the message number i we have 9 bins (8 bins for $i = 1$ to 8, and 1 bin for $i >= 9$).

(4-5) $CurrentPHQ-9$ and $CurrentGAD-7$: Clinical scores measuring the client's depressive and anxious symptoms, as seen by the supporter before the i^{th} message. Higher scores indicate more severe symptoms (7 Bins: score <= 9, 5 bins for scores 10-19 in increments of 2, and score >= 20).

Increments of bins were heuristically chosen by rounding up half the standard deviation, after excluding the first and last bins.

Semantic or Linguistic <Strategy> Variables in Messages

Similar to text-mining approaches for online mental health content analysis [36, 125, 171, 186, 218], a lexicon-based approach was used to extract comprehensive, high-level language characteristics from the supporter messages without risking to identify any text content. For each supporter message i, NLP techniques were used to extract 23 features indicative of a support strategy (*e.g.*, use of encouraging phrases). The features were achieved by mapping each word in the message i to a word category of a lexicon. We defined 23 *strategy* variables based on the literature and available lexicons:

(1-3) Sentiment: To capture the overall *sentiment* of a message, we used the NRC Emotion Lexicon [161] to extract the percentages of positive (Pos) and negative words (Neg), and the difference between them ($Pos - Neg$). Word percentage were used instead of absolute word counts, to better compare messages of different lengths [236].

(4-11) Emotions: As an indicator of the *emotional tone* of a message, we extracted the percentages of words related to the 8 emotion categories of the same lexicon [161]: $Anger$, $Anticipation$, $Disgust$, $Fear$, Joy, $Sadness$, $Surprise$, $Trust$.

(12-13) Pronouns: We extracted the percentage of first person plural pronouns (*e.g.* We, us, our), assuming that supporters will use these to convey a supportive alliance [24]. Assuming second person pronouns (*e.g.* you) are used more often to direct clients to engage in specific program activities (e.g., task prompting behaviors [188]), we calculated the difference between the percentages of first person plural and second person pronouns ($We - You$) as indicator of *supportive alliance*.

(14) Encouraging Phrases: Based on a series of support message examples used in supporter training, we derived a list of 15 commonly used encouraging phrases (*e.g.* 'good job', 'well done'). As an indicator of the *motivational tone* of a message, we calculated the ratio of the number of encouraging phrases to the number of sentences overall ($Encouragement : Sentences$).

(15-22) Mental Processes and Behaviors: We used the Regressive Imagery Dictionary [148], an extensively used lexicon in mental health research (*e.g.* to analyze psychotherapy dialogue [205, 110, 241]), for analyzing different categories of conceptual thought in text. Specifically, the analysis includes the seven categories of secondary cognition that are *logical, reality-based, and focused on problem solving.* This includes the percentages of words related to mental processes of: *Abstraction* (*e.g.*, know, may, thought), $InstrumentalBehavior$ (*e.g.*, make, find, work), $MoralImperative$ (*e.g.*, should, right, virtue), $Order$ (*e.g.*, simple, measure, array), *Restraint* (*e.g.* must, stop, bind), $SocialBehavior$ (*e.g.*, say, tell, call), and $TemporalReferences$ (*e.g.*, when, now, then). As additional behavioral cues, the percentage of $ActionVerbs$ in the text were also extracted, as specified by the Moby Project [196].

(23) Quantity: As a measure of *quantity of support*, we assess the length of the text messages via number of words ($WordCount$) and report number of sentences where relevant.

Analysis of Strategies Across Contexts

We identify a semantic or linguistic support strategy as successful, if: (i) it occurs significantly more often in messages by supporters in the 'high' outcome cluster (compared to the 'low' outcome cluster); and (ii) this difference is consistent across different client contexts. We operationalize this definition by considering each $< context, strategy >$ pair separately in our analysis. Since we have extracted 5 *context* variables and 23 *strategy* variables, we analyze 115 $< context, strategy >$ pairs. For each pair, we do the following:

1. We divide the messages from the 'high' and 'low' supporter clusters into multiple bins according to the client's *context* before the supporter composed a message.
2. For each bin, we then compute the mean of the *strategy* variable for composed messages in the two clusters *separately* along with their 95% bootstrapped confidence intervals.
3. For each bin, we also compare the means of the *strategy* variable across the two clusters using a bootstrapped resampling test.
4. Relevant support strategies should have statistically significant differences in means between the two clusters ($\alpha = 0.05$) and 95% confidence intervals that rarely overlap, across most bins.

Since the messages in each bin can belong to the same client or supporter, they are not independent. To address this, we use hierarchical bootstrapping for clustered data by randomly sampling supporters with replacement, and their clients and reviews without replacement 10000 times [203]. For comparing the means of the messages in each bin across the two clusters, we reject H_0: Difference in means = 0 (two-tailed) at level α if and only if a valid confidence interval with coverage probability $1 - \alpha$ does not cover 0 (α=0.05 in our analysis).

While only messages with pre and post PHQ-9 and GAD-7 scores were used in the initial clustering, we used all messages by supporters in the 'high' and 'low' clusters for this analysis, so as to not miss out on clients with low levels of usage.

6.5.2 Results: Strategies Used in Successful Messages

What follows are the main findings of significant differences in the linguistic strategies that were used in supporter messages associated with 'high' versus 'low' outcomes (improvements

in client clinical scores) across different client contexts.

Sentiment and Emotion in Support Messages

Across *all* supporter messages in our dataset, the sentiment analysis showed that positive words were generally used more frequently than negative words ($Mean_{pos} = 6.2\%$, $SD_{pos} = 2.8$; $Mean_{neg} = 1.7\%$, $SD_{neg} = 3.4$; $Mean_{pos\text{-}neg} = 4.5\%$, $SD_{pos\text{-}neg} = 1.3$). Further, **more successful supporter messages consistently used more positive and less negative words**. This effect remained consistent (*i.e.* $p < 0.05$ for all 3 sentiment *strategy* variables across all bins of the 5 *context* variables).

We also found that **more successful messages had less occurrences of negative emotions conveying sadness and fear than less successful messages** (*i.e.* $p < 0.05$ for these two emotion-related *strategy* variables across all bins of the 5 *context* variables). In addition, more successful messages used more frequently words that expressed joy, yet the difference in means was non-significant for several bins. Since anger and disgust rarely featured in the messages, we did not include them in further analysis, and there were no statistically significant results for any of the other emotions.

Pronouns: Sense of Supportive Alliance

Our results show that, across all messages, second person pronouns (*e.g.* you) were used more frequently than first person plural pronouns (*e.g.* we, us, our). However, we found that **more successful messages consistently employed first person plural pronouns more frequently than less successful messages, and had greater differences between uses of first person plural and second person pronouns** (*i.e.*, $p < 0.05$ for both pronoun *strategy* variables across all bins of the 5 *context* variables). Consistent with previous work, first person plural pronoun use may reflect a sense of supportive affiliation [91] or increased presence in computer-mediated communication [129].

Encouraging Phrases: Motivational Tone

To assess the motivational tone of a message, we calculated the ratio of number of common encouraging phrases to sentences ($Encouragement : Sentences$). Across all dataset messages, the mean ratio is 0.04 ($SD = 0.09$). Given the average length of a support message of 9.5 sentences, we can estimate that an encouraging phrase is used, on average, once in every 2.6 messages. We found that **more successful messages consistently contained significantly more encouraging phrases compared to less successful messages** ($p < 0.05$

for this *strategy* variable across all bins of the 5 *context* variables except the 61-70 bin for $ContentViews$).

Mental Processes & Behaviors: Action Orientation

For our our mental process variables, we found that **more successful messages consistently employed more words associated with social behavior and less words associated with abstraction, when compared to less successful messages** ($p < 0.05$ for these *strategy* variables across all bins of the 5 *context* variables). In order to better interpret the results of our analysis, we visualize each of these $< context, strategy >$ pairs. Figure 2 (left) shows the percentage of words associated with $SocialBehavior$ that are used in 'more' and 'less' successful support messages. More specifically it plots the mean of this *strategy* variable for all messages in each bin for the 'high' and 'low' outcome clusters on the Y-axis, and for each bins of the $ContentViews$ context variable on the X-axis. The error bars are 95% bootstrapped confidence intervals of the means. Some of the most frequently used social behavior words were: *help, call, discuss*, and *share*.

Figure 2 (right) further shows how 'more' successful messages contained on average less words associated with the strategy variable *Abstraction* than 'less' successful messages; and this finding was consistent independent, *e.g.*, of the clients' depressive symptoms prior to the supporter review (*cf.* the bins of the $CurrentPHQ-9$ variable). Frequently used abstraction words were: *think/thought, know, understand*, and *learn*.

Restraint, $MoralImperative$, and $Order$ occurred rarely in our dataset (as indicated by their mean occurrence percentages), and thus were not included for further analysis. For words associated with $InstrumentalBehavior$, the results were not statistically significant. While the results for words associated with $TemporalReferences$ were statistically significant across most bins of the *context* variables, the 95% confidence intervals overlapped and did not show any consistent trend.

Lastly, we found that, on average, about a quarter of every message comprised of action verbs ($Mean = 24.4\%$, $SD = 4.5$), and that **more successful messages consistently employed more action verbs compared to less successful ones** (*i.e.* $p < 0.05$ across all bins of the 5 *context* variables).

Quantity: Length of Supporter Messages

Finally, we found that messages were on average 9.5 sentences long ($SD = 6.7$) and included 191 words ($SD = 147$). We found that **more successful messages were consistently shorter** ($Mean_{words} = 177$, $Mean_{Sentences} = 9$) **than less successful ones** ($Mean_{Words} =$

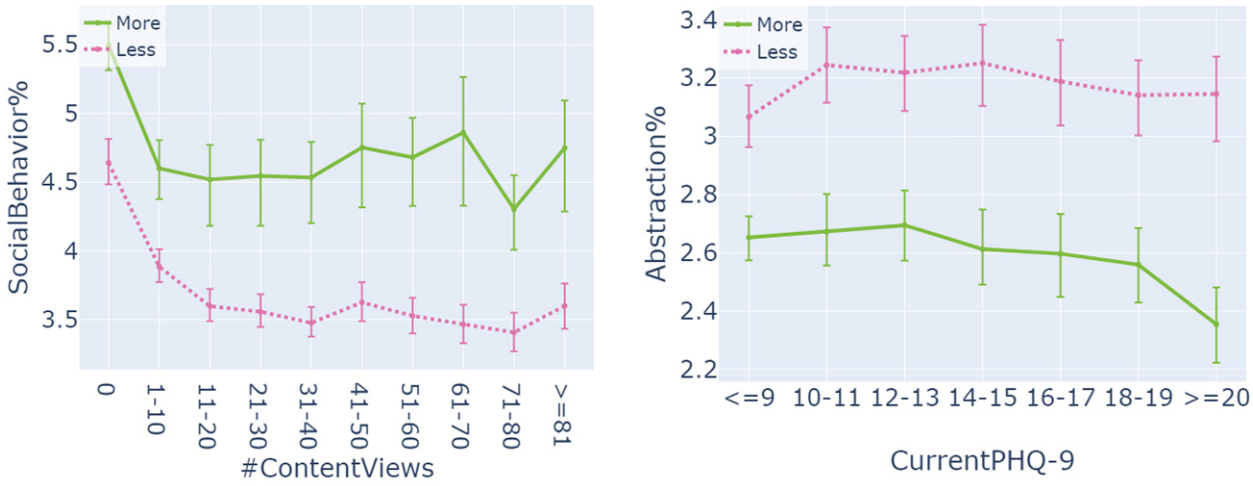


Figure 6.2: Mean percentage of words in 'more' and 'less' successful support messages that are associated with: *SocialBehavior* across each bin of the $ContentViews$ context variable (left); and *Abstraction* across each bin of the $CurrentPHQ-9$ context variable (right).

214, $Mean_{Sentences} = 10.4$); with $p < 0.05$ for $Word\ Count$ across all bins of the 5 *context* variables with very few exceptions). This surprisingly contradicts previous findings by [6], and thus, requires further research.

6.6 Context-specific Patterns of Support Success

In the previous section, we identified support strategies that were consistently associated with better clinical outcomes across 5 types of contexts, which we each had treated in isolation. In this final analysis, we want to better understand the more complex relationship that likely exists between the use of support strategies and different client context variables. In other words, how may considering the combination of multiple context variables (rather than each client *context* variable by itself) shift how salient a specific support strategy is in messages associated with either 'high' or 'low' client outcomes. We believe that identifying such relationship patterns could enable a more effective tailoring of support strategies to specific client contexts. Next, we therefore describe our approach to identifying <multidimensional context $\rightarrow$ strategy> patterns.

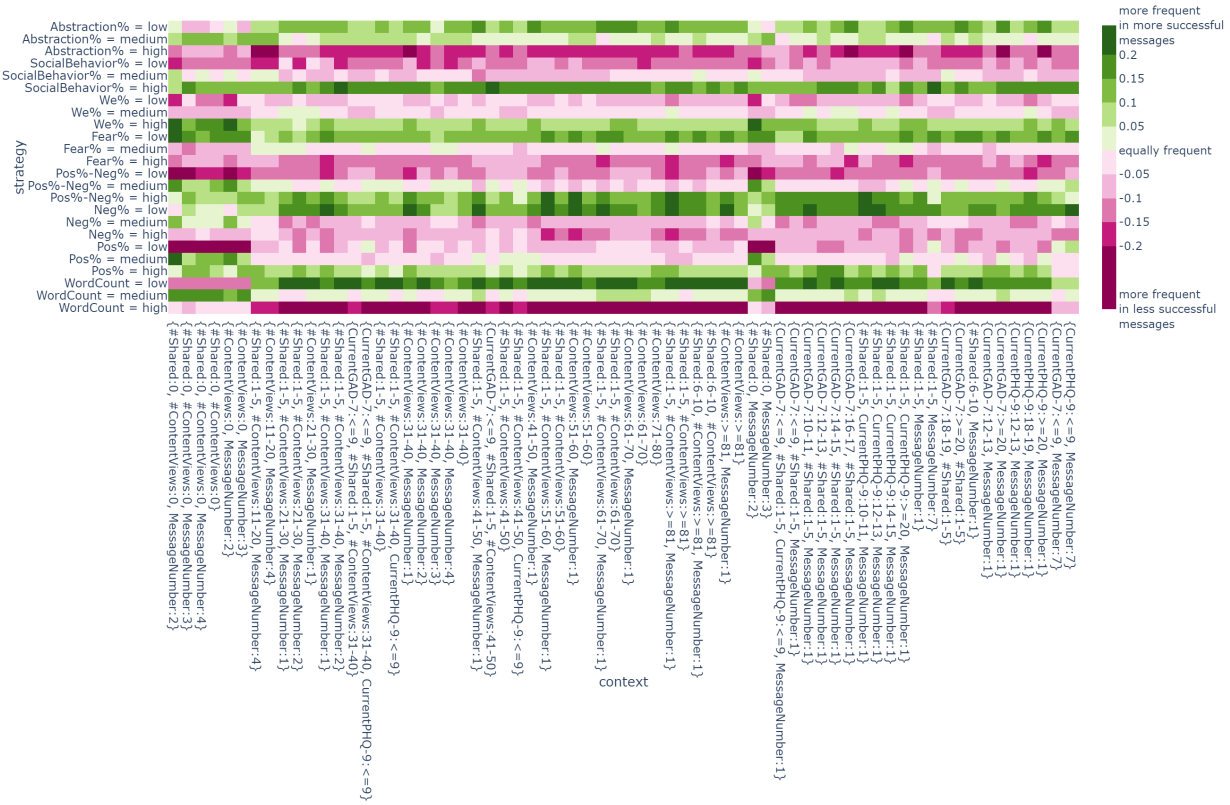


Figure 6.3: The salience of the most interesting <multidimensional context (A) → strategy (B)> patterns are visualized. Each cell represents the salience of a pattern $A \rightarrow B$. Context A is on the X-axis while strategy B is on the Y-axis. Salience is the difference between percentages of rules containing B given A across the two clusters, and is multiplied by "-1" when the rule occurs more frequently in less successful messages. Thus, *rules in green occur more frequently in more successful messages while rules in pink occur more frequently in less successful messages. Darker colors imply greater salience.* For readability, the contexts are sorted on $ContentViews$ followed by $Shared$ and then $MessageNumber$. The figure is best read *strategy*-wise from left to right.

6.6.1 Extracting Multidimensional Context → Strategy Patterns

To identify ¡multidimensional context → strategy¿ patterns, we needed to discover associations between client contexts containing two or more contextual variables, and each support

strategy. To find associations among multiple items in datasets, we used the well-known frequent item set mining *Apriori* algorithm. It first generates a set of frequent items that occur together, and then extracts association rules that explain the relationship between those items. Given two sets of items A and B, an association rule $A \rightarrow B$ exists if A and B frequently appear together in transactions. In our case, "transactions" are supporter messages that occur in client context A and employ strategy B. We consider each strategy separately in various client contexts, such that context A is multidimensional (it contains two or more of the 5 *context* variables) while strategy B will always be one singular *strategy* variable.

We extracted association rules separately for both our 'high' and 'low' outcome clusters; and then compared the rules that occurred in both groups using two "interestingness" measures: Confidence and Coverage [99, 136]. Confidence is the percentage of transactions containing A that also contain B ($P(B|A)$), whereas coverage is the percentage of transactions containing A ($P(A)$), chosen as a measure of the applicability of a rule. For example, if a context A occurs very frequently (*e.g.* coverage is high), rules associated with it are more interesting as they can be applied more often when recommending support strategies personalized to the client's context. We set minimum coverage to 0.01 and minimum confidence to 0 to extract a large number of rules, that we can then filter using a different metric that compares the salience of the rules.

For this analysis, we excluded the strategy variables: *Anger*, *Disgust*, *Restraint*, *Order*, and $MoralImperative$, as they rarely occurred in our data. The remaining 18 *strategy* variables were discretized to 'high', 'medium', and 'low', using equal frequency binning. Then, Apriori was applied separately on messages from our 'high' and 'low' outcome clusters. We found 22599 rules in both clusters. The salience for each of these rules was calculated as the absolute confidence difference between the two clusters. That is, **the salience of a rule $A \rightarrow B$ is measured as the difference between percentage of strategy B occurring across the two clusters, when the client is in context** A, such that more salient rules are used more frequently by supporters in either the more or less successful cluster. We select a subset of rules to interpret by first choosing interesting contexts and strategies that are used in at least one rule with salience >= 0.20, and then selecting all rules associated with them (giving us 1584 rules associated with 66 contexts and 8 *strategy* variables with 3 levels each), allowing us us to interpret the most salient rules in context.

The arules R package was used to set up Apriori in this way [112].

Messages from the 'high', 'low', and 'average' clusters were used for discretization.

Salience = abs($P(B_{high}|A_{high})$-$P(B_{low}|A_{low})$)

Rules containing $ActionVerb$% were excluded, because action verbs occur too frequently and are not very informative

6.6.2 Results: Salient Context-Specific Support Strategies

Figure 6.3 shows a heat map of the salience of these rules. Next, we present a few examples of interesting results. Taking the third row from the bottom, read horizontally, we see the rules associated with low $WordCount$. The row is mostly green, apart from the first 6 context rules. The green indicates that the support strategy "$WordCount$ = low" is more salient in more successful support messages, with a darker shade indicating higher salience. We see that for almost all contexts, a lower word count is more salient in more successful messages. However, this is flipped for the first 6 context rules that show no client engagement prior to a review (*e.g.* $SharedContent$ with supporter = 0, $ContentViews = 0$). Thus, for disengaged clients, writing shorter messages is more strongly associated with less successful outcomes (shown as color pink).

Outside of 'salience flips' between the two groups (color change); interesting patterns can also be identified through variations in the shade of the same color. For example, for strategies "$Fear$ = low" and "We = high", we see high salience (dark green) in more successful messages for the same first 6 client contexts; and reduced effects thereafter. This means that writing messages with less words related to fear, and more first person plural pronouns are stronger associated with more successful support messages, and this effect is particularly salient in situations where clients are disengaged.

Aligned with our previous results, we further find for highly engaged clients that successful messages reference more social behavior ($SocialBehavior$%=high) and less abstraction ($Abstraction$%=low), while less successful messages reference less social behavior ($SocialBehavior$%=low or medium) and more abstraction ($Abstraction$%=high) in the same contexts. There remain more rules and patterns to unpack. Here, we see our analysis as contributing initial insights on how identified support strategies can be more or less salient or successful depending on a specific client context. So far, our results imply that **for less engaged clients, writing longer, more positive and more supportive reviews is linked with greater outcomes**; whilst **more engaged clients appear to benefit more from messages with less negative words, less abstraction, and more references to social behaviors**.

6.7 Discussion

Our work presents the first application of unsupervised ML, and statistical and data mining methods to analyze complex, large-scale supporter-client interaction data of an established iCBT intervention for the treatment of depression and anxiety. We focused on developing a better understanding of how the behaviors of supporters, who assist clients' engagement with

this service, may correlate with better clinical outcomes for these clients; which presents a largely under-explored area. Below, we discuss the main implications of our work for future research that intersects the fields of HCI, ML and healthcare.

6.7.1 Identifying Effective Context-Specific Support Strategies

We described our approach to identifying 'more' and 'less' successful support behaviors that are manifest in communications to clients. Using semantic and linguistic feature extraction methods, our results indicate that supporter messages that typically achieve higher client outcomes contain more words that are positive, supportive, related to social behaviors, and less abstract; and those messages tend to be shorter than less successful messages. Largely, these findings align well with previous qualitative studies of iCBT support that emphasize the prevalence of supportive language [221], and importance of affirmations and encouragement [108, 188] for client outcomes. Extending this research in iCBT, our work presents novel findings of how the success of identified support strategies can vary dependent on a specific client context. Next, we discuss how having a better understanding of each persons' context for support enables new opportunities for personalized care; which, in turn, can improve client engagement with iCBT interventions and benefit their mental health outcomes.

6.7.2 Data-Enabled, Personalized Mental Health Support

So far, only few works have explored the design space and use of ML to personalize the treatment or delivery of (mental) health interventions (*e.g.* [73, 185, 167]). Most prominently, Paredes et al. [185] applied reinforcement learning for recommending stress coping interventions tailored to the person. Other recent trends include the development of just-in-time adaptive interventions (JITAIs) that utilize algorithms to provide the right type or amount of support, at the right time, in response to an individuals' internal and contextual state [167].

Design Implications for Personalized Human Support in iCBT

For guided iCBT interventions, there are multiple ways in which gathered data insights about context-specific support strategies can inform supporter interface design. For example, as supporters review a client, they may be presented in their feedback interface with recommendations of what strategies specific to this clients' situation may be strongly correlated with successful client outcomes, providing them additional input to their feedback practices.

To help translate linguistically-derived support strategies more directly into feedback messages, strategy-related words (*e.g.* positive words, certain pronouns) could be highlighted in real-time as part of the message editor that supporters use. Especially for training purposes and to support skills acquisition of novice supporters, it may also be helpful to integrate examples of 'support messages that were successful for similar clients' as guidance.

Human-Centered ML in (Mental) Healthcare

The above design examples further illustrate our orientation to human-centred ML and the integration of data insights into healthcare practice. Rather than promoting the use of templates or automatizing client feedback away from the human supporter, we suggest designing interventions that seek to *enhance supporter agency* by enabling them to personalize their feedback more effectively for each person (*cf.* [242]), and to better understand how their actions make (ideally) a positive difference to their clients. Thus, while advanced data tools that can identify complex patterns are often seen to generate more accurate, objective and less biased insights (*cf.* [106, 257, 276]), it is important that we (i) do not take away from, but help foster, the important relationship and genuine *human connection* that is formed between supporter and client and crucial to their alliance and positive outcomes [24, 162]; and (ii) ensure supporters feel that their input and expertise is valued rather than made redundant or replaced in favor of data science.

6.7.3 Understanding (Big) Data in Digital Health Interventions

Next, we discuss identified challenges and opportunities for working with complex, large-scale observational data.

Trade-Offs between Data Access and Use & Ethical Conduct

Although we had unprecedented, privileged access to large-scale supporter-interaction data, we made necessary trade-offs as to what kind of analysis we could conduct to protect the full anonymity of both client and supporter data. This meant much of our analysis was restricted to coarse-grained linguistic features and high-level usage data. While our research captured individual word associations, the use of other linguistic features such as n-grams (*e.g.* [2, 125]) could expand the analysis to word pairings, or even short word sequences that could enable a richer contextual understanding of identified support behaviors. At the same time, however, such explorations need to be done with care so as to not risk making too much content, or the people who produced it, identifiable.

Caution is also required for secondary data analysis for which additional user consent is likely unfeasible to collect (for every subsequent analysis). As is common in ML approaches for mental health, user privacy should be carefully addressed to preserve anonymity (cf. phone data [33, 266], sensors [202], social media [186, 218, 276]), and analysis should occur in a context where there is a clear case for the prospective benefits that could arise, *e.g.*, from improved healthcare provision (see public health ethics [44] and recent work on user acceptance of digital phenotyping [143, 211]). Going forward, we need to continue developing feasible, privacy-preserving data methods; and, as researchers, need to remain critical of, and sensitive to, the extent to which our data analysis is really 'justifiable' with regards to how it comes to benefit users and health services.

Deriving Insight from Data: Interpretation & Future Directions

Due to the necessary data use restrictions, we acknowledge that derived data insights —whilst novel —are limited to the definitions chosen, and require further research to validate. Future work may also explore multiple additional data mining avenues, including: (i) analysis of supporters' use and adaptation of messaging templates (*cf.* [6]); (ii) studies into sequential routines of strategies (*e.g.* using Inverse Reinforcement Learning [14]); (iii) supporter clustering using 'engagement' as an outcome metric alongside clinical improvement (*cf.* [163]); or (iv) the combination of support strategies, supporter-features, and client engagement features to predict clinical outcomes.

Despite manifold possibilities for data mining, challenges remain for ensuring that derived insights are both *human interpretable* (*e.g.* [1, 107, 194]) and *clinically useful*. While we were deliberate in our choice of data tools and visualizing of their results to create representations that are comprehensive to laypeople, other research methods (*e.g.* qualitative studies [108, 188, 221]) are needed to help contextualize, validate and advance our understanding of support, or other data-derived health behaviors. More research is also needed to develop our understanding of the potential value that these new types of data insights could bring to actual healthcare practices.

6.8 Conclusion

Aiming to understand how the behaviors of iCBT supporters impact client outcomes, we presented our ML approach and the analysis of 234,735 supporter messages sent to an unprecedentedly large clinical sample of 54,104 clients with a mental health diagnosis. Using various computational methods we identified support behaviors associated with 'more' or

'less' improvements in clinical scores, and showed how their salience varied dependent on different client contexts. Our work enables a better understanding of best practices, and opens-up new opportunities for personalizing support provision within computer-delivered mental health services. We discussed: the implications of our findings for the design of iCBT supporter interfaces; the need for a human-centered approach for sensibly integrating data insights into healthcare practices; and ethical considerations for secondary data use.

6.9 Addressing the Curse of Dimensionality

This section explains how this study addressed the curse of dimensionality challenge first introduced in chapter 1.

6.9.1 W.r.t. multiple co-morbidities (C2)

In this study, we identified the most effective support strategies for improving two co-morbid conditions by defining a "combined" target outcome based on both those measures. We clustered depression and anxiety scores over time to get 1 target outcome *i.e.* the success of the supporter, allowing us to identify effective strategies for improving both depression and anxiety while reducing the computational complexity of the problem and resulting in interpretable insights. By assessing the effectiveness of the intervention based on two co-morbid conditions combined together as one outcome, this study address the curse of dimensionality w.r.t. multiple co-mordid outcomes. This approach is different from studies 1 and 3, where I addressed the curse of dimensionality w.r.t. multiple co-mordid outcomes by showing that the approach generalizes to multiple co-morbid outcomes.

6.9.2 W.r.t. diversity in patient characteristics (C3)

In this study, we found the best overall support strategies by identifying support strategies that are consistently associated with better clinical outcomes across multiple client context variables (patient characteristics). We also presented a method for identifying context-specific patterns of support. That is, we explore how considering a combination of multiple client context variables might change the relationship between support strategies and client outcomes, thereby addressing the curse of dimensionality w.r.t. diversity in patient characteristics.

Chapter 7

S5: Predicting Periodically Assessed Multiple Sclerosis Outcomes Using Passively Sensed Behaviors and Ecological Momentary Assessments

7.1 Introduction

Multiple Sclerosis (MS) is a neurological disorder that affects around 2.8 million people worldwide [263]. Approximately 900,000 people in the United States live with MS [239], and the annual economic burden is estimated to be $85.4 billion [238]. MS is one the most common causes of disability in young adults and primarily affects women [103]. In addition to disability symptoms such as vision and mobility issues, pwMS have a significantly higher burden of comorbidities like depression, fatigue, and sleep quality issues, than the general population [117, 85]. There is currently no cure for MS. Treatment involves managing symptoms related to disability, depression, fatigue, and sleep, through a combination of pharmacological and behavioral interventions. The complexity of the disorder and the associated comorbidities necessitate that patients actively participate in the monitoring and management of their symptoms [93].

Longitudinal tracking of symptoms is consequential to clinical decision-making. However, frequent longitudinal tracking by clinicians is costly and frequent self-tracking by patients is hard to consistently achieve as it adds to their burden. Motivated by this, we propose a machine learning approach that leverages data from the smartphones and fitness trackers of pwMS to predict their health outcomes on a biweekly basis for depression, and on a 4-weekly basis for global MS symptom burden (disability), fatigue, and sleep quality, with the goal of enabling low-cost longitudinal symptom tracking with minimal active input from the patient.

Our work capitalizes on a large body of research that focuses on using passively generated data from personal digital devices (e.g., smartphones and fitness trackers) to capture human behavior and predict health outcomes. This moment-by-moment, in situ quantification of the individual-level human phenotype using data from personal digital devices is known as digital phenotyping [114]. Previous works using passively sensed smartphone and wearable data to predict physical disability and fatigue in pwMS have been exploratory in assessing the feasibility of data collection and the preliminary association between sensed behaviors and outcomes. The clinical applicability of digital phenotyping to predict health outcomes in pwMS at frequent intervals to enable longitudinal tracking, has not yet been established [169, 231, 45].

Here, we present a machine learning approach that enables longitudinal monitoring of clinically relevant health outcomes for pwMS by leveraging passively sensed data from sen-

sors in smartphones and fitness trackers. We also explore whether the performance of these models can be improved by incorporating a minimal amount of active input from the patient in the form of short surveys called Ecological Momentary Assessments (EMAs). Further, we explore if using behavioral features from the previous time period (context features) in addition to behavioral features from the current time period (action features) can improve the predict of these models, by helping the model contextualize the patient's current behaviors. This study differentiates from prior studies by examining the clinical utility of digital phenotyping with passive sensors for longitudinal tracking of health outcomes. Our approach can also inform predictive longitudinal tracking of symptoms in health conditions beyond MS.

7.2 Methods

The aim of the presented study was to examine the clinical utility of using sensors in smartphones and fitness trackers in predicting clinically relevant outcomes in pwMS with the goal of enabling frequent monitoring. Data collection from 104 participants in this study was done between November 2019 and January 2021. To briefly summarize our approach, we used data from 3 sensors in participants' smartphones (calls, location, screen activity) and 3 sensors in participants' fitness trackers (heart rate, sleep, steps) to predict patient-reported outcomes of depression, global MS symptom burden, fatigue, and sleep quality. Depression was predicted biweekly i.e., every 2 weeks, while the other outcomes were predicted once every 4 weeks. We also explored if adding Ecological Momentary Assessments (EMAs) would boost our model's predictive performance. EMAs are short surveys that enable "repeated sampling of subjects' current behaviors and experiences in real time, in subjects' natural environments" [234]. In this study, we administered EMAs through a mobile app thrice a day. Each EMA had 2 questions that took less than 15 seconds to complete. We computed 2 types of behavioral features from the sensor data and EMAs – action and context features. Action features are the features that capture the participant's actions and behaviors that are most closely related to the predicted outcome, which is the sensor and EMA features from the time period immediately preceding the current prediction time point. Context features are the features that capture the context of the participant's action features, that is, the sensor and EMA features from the time period immediately preceding the previous prediction time point. We then used action features, and action and context features to predict the outcomes. All methods were performed in accordance with the IRB guideline and institutional regulation.

7.2.1 Participants

The study included adults 18 years or older with a neurologist-confirmed MS diagnosis who owned a smartphone (Android or iOS) and enrolled in the Prospective Investigation of Multiple Sclerosis in the Three Rivers Region (PROMOTE) study, a clinic-based natural history study at the University of Pittsburgh Medical Center (UPMC) [139, 146]. The institutional review boards of University of Pittsburgh and Carnegie Mellon University approved the study. All participants provided written informed consent.

7.2.2 Study Design

Participants downloaded a mobile application to capture sensor data from their own smartphones and additionally received a Fitbit Inspire HR to track steps, heart rate, and sleep. Data were continuously collected from smartphone and Fitbit sensors of 104 participants during the study period (16 November 2019 to 24 Jan 2021). Additionally, the mobile app directed the participants to a short Ecological Momentary Assessment (EMA) via a notification three times a day throughout the duration of the study.

All 104 participants completed data collection for a pre-defined period of 12 weeks while 44 agreed to extend data collection for an additional 12 weeks (for a total of 24 weeks).

7.2.3 Survey Response and Patient-Reported Outcomes

All participants completed a baseline questionnaire, which queried their demographics and baseline health outcomes, on the Saturday following enrollment. During the study, participants completed additional questionnaires as described below at intervals according to each questionnaire. All questionnaires for the overall study were administered online using the secure, web-based Research Electronic Data Capture (REDCap) system [101, 100].

Depression

We used the Patient Health Questionnaire (PHQ-9) to measure the severity of depression symptoms once every two weeks [131]. PHQ-9 contained 9 questions, with each answer being scored on a scale of 0-3. Higher scores indicated more severe depressive symptoms.

Global MS Symptom Burden

We used the Multiple Sclerosis Rating Scale - Revised (MSRS-R) to measure global MS symptom burden and neurological disability once every four weeks [270]. MSRS-R assessed

eight neurological domains (walking, upper limb function, vision, speech, swallowing, cognition, sensory, bladder and bowel function; each domain scored as 0 to 4, with 0 indicating the absence of symptom and 4 indicating higher symptom burden and more severe disability).

Fatigue

We used the 5-item version of the Modified Fatigue Impact Scale (MFIS-5) to measure the impact of fatigue on cognitive, physical and psychosocial function once every four weeks [152]. Each item in MFIS-5 was scored on a five-point Likert scale from 0 (never) to four (almost always). Higher scores indicated more severe fatigue.

Sleep Quality

We used the Pittsburgh Sleep Quality Index (PSQI) to measure sleep disturbances once every four weeks [30]. PSQI comprised 19 individual items, with seven component scores (each on a 0-3 scale) and one composite score (0 to 21, where higher scores indicating a poorer sleep quality).

For each outcome, we dichotomized the scores from the surveys above using certain thresholds to get binary outcomes once every two weeks for depression and once every four weeks for the other outcomes. The binary outcomes would likely have better clinical utility as they are more easily understood by patients (for self-monitoring), volunteers with limited mental health training, or even clinicians. For "Depression", PHQ-9 scores were dichotomized as ">=5: presence of depression" and "<5: absence of depression". For "Global MS symptom burden", MSRS-R scores were dichotomized as ">=6.4: higher burden" and "<6.4: lower burden". For "Fatigue", MSIF-5 scores were dichotomized as ">=8: high fatigue" and "<8: low fatigue". For "Sleep quality", PSQI scores were dichotomized as ">=9: poorer sleep quality" and "<9: better sleep quality". The thresholds for depression and sleep quality were based on previous works [131, 83]. Given the lack of consensus from the literature, we calculated the median scores of the global MS symptom burden and fatigue in the whole dataset and used the median scores as the thresholds.

7.2.4 Sensor and EMA Data Collection

Each participant installed a mobile application based on the AWARE framework [82] which provided backend and network infrastructure that unobtrusively collected from smartphones the location, screen usage (i.e. when the screen status changed to on or off and locked or unlocked), and call logs (for incoming, outgoing and missed calls). This mobile app also

sent participants notifications every time they were scheduled to complete an Ecological Momentary Assessment (EMA), which was three times a day throughout the duration of the study for all participants. The EMA notification directed participants to a short EMA survey within the app's user interface. EMA surveys typically took the participants less than 15 seconds to complete, and asked participants to answer two questions: "How depressed do you feel?" and "How tired do you feel?". The participants responded to each EMA question using a Likert scale from 0 to 4, with 4 being the most depressed/ tired and 0 being the least depressed/ tired. The EMA responses were then synced to our servers. Further, participants wore a Fitbit Inspire HR that captured the number of steps, sleep status (asleep, awake, restless, or unknown), and heart rate. Calls and screen usage were event-based sensor streams, whereas location, steps, sleep, and heart rate were time series sensor streams. We sampled location coordinates at 1 sample per 10 minutes, and steps, sleep, and heart rate at 1 sample per minute.

Data from AWARE were deidentified and automatically transferred over WiFi to a study server at regular intervals. Data from the Fitbit were retrieved using the Fitbit API at the end of the data collection. Participants were asked to always keep their devices charged and carry their phone and wear Fitbit.

To protect confidentiality, we removed identifiable information (e.g., names, contact information) from survey and sensor data prior to analysis. We followed the standard practice for sensor data security.

7.2.5 Data Processing and Machine Learning

The data processing and analysis pipeline (Figure 7.1) involved several steps:

1. Feature extraction from sensors and Ecological Momentary Assessments (EMAs) over time slices to extract action and context related behavioral features.

2. Handling missing features.

3. Machine learning to predict patient-reported health outcomes on a biweekly basis for depression, and on a 4-weekly basis for global MS symptom burden, fatigue, and sleep quality. We use action features and action + context features for the following:

    a) Using 1-sensor models (i.e., models containing features from one sensor).

    b) Combining 1-sensor models to obtain the best combined sensor model for each outcome.

c) Combining EMA-only models with 1-sensor models from the 6 sensors to obtain best model for each outcome. We try 2 types of EMA-only models.

Feature Extraction

We computed features from six sensors: Calls, Heart Rate (HR), Location, Screen, Sleep, and Steps, given their potential to inform depressive symptoms [41, 216, 266, 32, 273, 274], fatigue [255], MS symptom burden such as decreased mobility [231], and sleep quality [160, 222].

Location features captured mobility patterns. Steps and Heart Rate captured the extent of physical activities. Calls features captured communication patterns. Screen features might inform the ability for concentration [58, 133] and the extent of sedentary behavior [52], despite of potential caveats for pwMS and other chronic neurological disorders. Sleep features captured sleeping duration and patterns, which could indicate sleep disturbance (e.g., insomnia or hypersomnia) associated with depression [177]. Please see Chikersal et al. [42] for details of features extracted from each sensor.

Features from the six sensors were extracted over a range of temporal slices (Figure 7.1b) over biweekly and 4-weekly periods. For each biweekly and 4-weekly period, we obtained the daily averages of these features by computing the average of the daily feature values.

As previously discussed, the participants were asked 2 Ecological Momentary Assessment (EMA) questions three times a day throughout the study: "How depressed do you feel?" and "How tired do you feel?". They responded to these questions using a Likert scale of 0 to 4. To compute EMA features, the EMA responses were extracted over a range of temporal slices (Figure 7.1c) over biweekly and 4-weekly periods. For each biweekly and 4-weekly period, we obtained the two types of EMA features: (1) Average EMA by computing daily averages of the response to each EMA question, and (2) Pre-survey EMA by taking the last response of each EMA question from each temporal slice. Hence, pre-survey EMA features are based on the EMA collected on the last weekend day and the last weekday before the survey to measure the outcome is administered.

Temporal Slicing: The temporal slicing approach extracted sensor features from different time segments (Figure 7.1b and 7.1c). Past work showed that this approach can better define the relationship between a feature and depression. For example, Chow et al. [46] found no relationship between depression and time spent at home during 4-hour time windows, but they found that people with more severe depression tended to spend more time at home between 10:00 AM and 6:00 PM. Similarly, Saeb et al. [217] found that the same behavioral feature calculated over weekdays and weekends could have a very different association with depression. Here, we obtained all available data (spanning multiple days of the study) from

a specific epoch or time segment of the day (all day, night i.e., 00:00-06:00 hrs, morning i.e., 06:00-12:00 hrs, afternoon i.e., 12:00-18:00 hrs, evening i.e., 18:00-00:00 hrs) and for specific days-of-the-week (all days of the week, weekdays only i.e., Monday-Friday, weekends only i.e., Saturday-Sunday) to achieve 15 data streams or temporal slices. To extract features from each of the 15 temporal slices, we first computed daily features, and averaged daily features over biweekly periods to predict depression, and over 4-weekly periods to predict global MS symptom burden, fatigue, and sleep quality. For EMA features, in addition to computing "average EMA" features by averaging daily features, we also computed "pre-survey EMA" features as the last daily feature from each temporal slice over biweekly periods to predict depression and 4-weekly periods to predict other outcomes. We concatenated the resulting features from 15 temporal slices to derive the final feature matrix.

Feature Matrix: After feature extraction, each of the six sensors and EMA types (average and pre-survey EMA) had a feature matrix, with each sample containing features from 15 different temporal slices, from a participant's biweekly or 4-weekly period. We call this feature matrix the "action" feature matrix, as the features in each sample capture the participant's actions from the current biweekly or 4-weekly period at the end of which we are trying to predict the outcome. For each participant, we concatenate features from the previous biweekly or 4-weekly period with the "action" feature matrix to obtain the "action + context" feature matrix, as features from the previous biweekly or 4-weekly period capture the context for the participant's current actions. That is, to predict the outcome at the end of the ith period i.e., at time T = iP where P = 2 weeks or 4 weeks, the action feature matrix is made up of features from time (i-1)P and time iP, whereas the action + context feature matrix is made up of features from time (i-2)P and time iP (see Figure 7.1d).

Handling Missing Data

Missing sensor data can occasionally occur due to several reasons.

Missing sensor data can occasionally occur due to technical issues (e.g., non-functioning phone/app/server, faulty or delayed data transfer) or compliance issues (e.g., participant not carrying the smartphone or wearing the FitBit) but more often due to semantic reasons. For example, if a participant made and received 0 calls during a period, there would be no calls data. Thus, we encoded missing data into features since we could not differentiate whether such data were not collected or did not exist to be collected due to semantic reasons.

A missing feature during a time slice for many participants could indicate non-semantic issues such as non-functioning server. Further, a participant with many missing features could indicate non-semantic issues such as the non-functioning phone/app. To empirically determine the thresholds, we plotted the number of participants and features remaining

for various thresholds and noted the largest differential in curves. Hence, we excluded all features (in a time slice) with missing values in more than 14 participants and likewise excluded participants missing more than 20% of all features. For each feature we calculated the minimum feature value, and imputed missing features as that value minus 1. As we handled missing data independently across feature time slices, the number of participants and features were different across sensors as missing features in each feature set.

Machine Learning Pipeline Using Action and Context Behavioral Features

We built machine learning models using Support Vector Machines with RBF Kernels and validated our models using leave-5-participants-out cross-validation to minimize over-fitting. The model generation process followed these steps:

1. *Selecting Features* by deciding if the model being trained should use average EMA features or pre-survey EMA features, and action-only or action + context features.

2. *Training and Validating 1-Sensor and EMA-only Models* for each of the six feature sets: Calls, Heart Rate, Location, Screen, Sleep, and Steps, and average or pre-survey EMA features.

3. *Obtaining Predictions from Combinations of Sensors* by combining detection probabilities from 1-sensor models to identify the best performing combined sensor model.

4. *Obtaining Predictions from Combinations of Sensors* and EMA by combining detection probabilities from 1-sensor models and an EMA-only model to identify the best performing final model.

5. *Classifying Different Outcomes* by running the pipeline for each outcome.

Selecting Features: We want to compare models trained that utilize different types of EMA features: average EMA or pre-survey EMA, and action-only features or action + context features. Hence, after feature extraction, before we train our models, we decide whether the current pipeline should train the models in question using average EMA features or pre-survey EMA features, and whether it should include action-only features or action + context features (see Figure 7.1a). We use ALL the resulting features during training. No other feature selection method is used.

Training and Validating 1-Sensor and EMA-only Models: For each sensor and EMA feature matrix, we built a model of the selected features from that sensor or EMA type to detect an outcome. We trained models using a Support Vector Machine classifier with

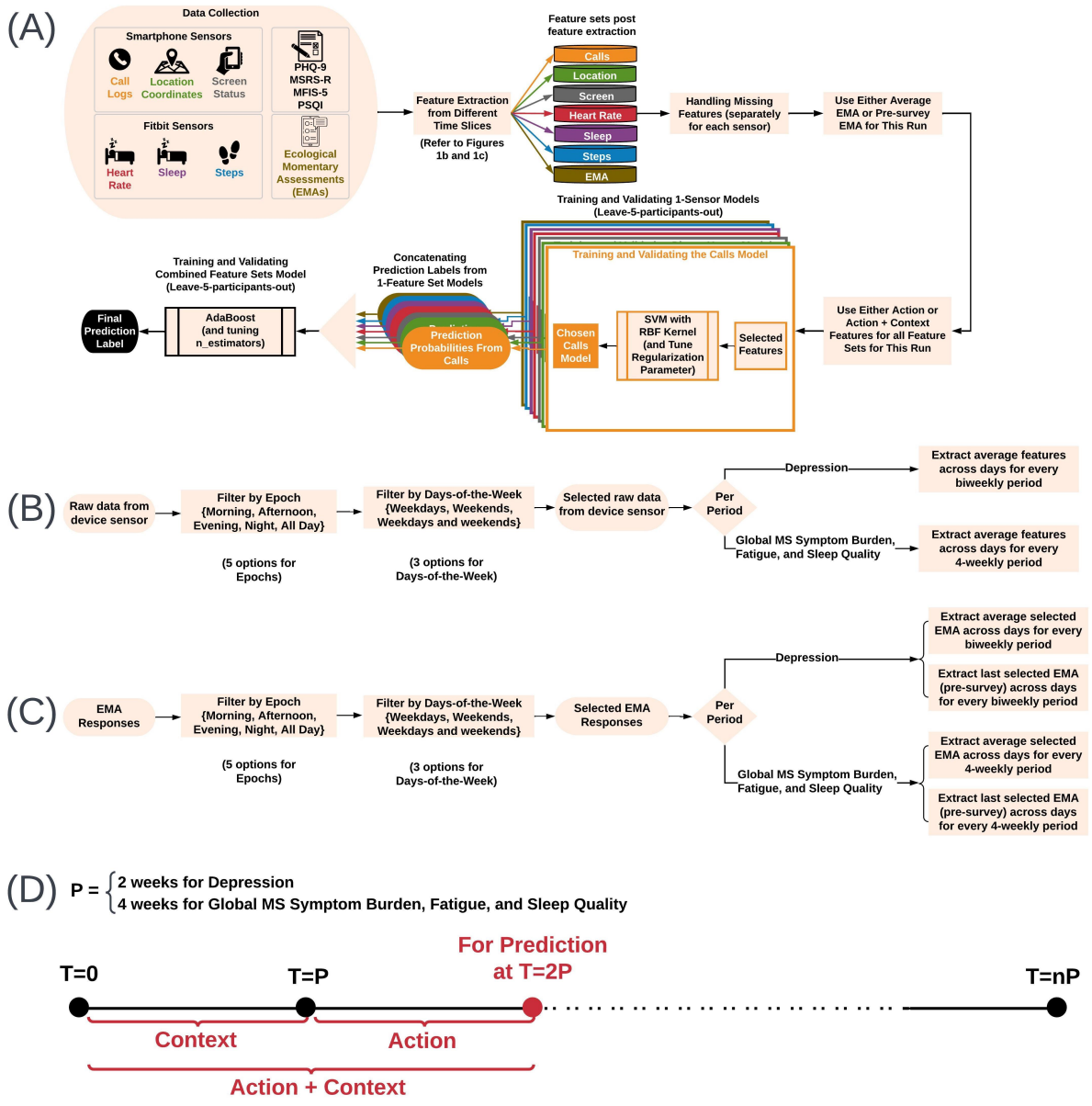


Figure 7.1: Data Processing and Analysis. (A) ML pipeline for predicting biweekly depression (PHQ-9), and global MS symptom burden (MSRS-R), fatigue (MFIS-5) and sleep quality (PSQI) every 4 weeks, using passive sensors from smartphones and fitness trackers, and EMAs. We run the pipeline for two types of EMA features (average and pre-survey EMA), and two types of feature matrices (action and action + context). (B) For each sensor, every feature was extracted from 15 time slices over biweekly or 4 weekly periods. First, raw data from the device sensor were preprocessed and filtered by a time-of-the-day epoch and a days-of-the-week option. Features were then extracted from the selected raw data. (C) For EMA, we used a similar approach to calculate average EMA and pre-survey EMA. (D) Action features are features from the period immediately preceding the prediction point, whereas context features are from the period preceding the "action period".

RBF Kernel (SVM-RBF). We used leave-5-participants-out cross-validation to choose the regularization parameter for SVM-RBF. The folds were split in a stratified manner and classes were balanced in the SVM-RBF, to ensure that positive and negative classes of the binary outcomes are adequately represented. We chose the model with the best f1-score for a given outcome, which provides the detection probabilities for the outcome. The process is independent of other outcomes.

We also tried other off-the-shelf ML algorithms such as Gradient Boosting Classifier, Random Forests, Logistic Regression, Lasso, and Linear SVM, but SVM-RBF outperformed them all. Further, SVM-RBF also outperformed the approach implemented by Chikersal et al. [41, 42] to predict depression every two weeks, and global MS symptom burden, fatigue, and sleep quality every four weeks.

Obtaining Predictions from Combinations of Sensors: The detection probabilities from all six 1-sensor models were concatenated into a single feature vector and given as input to an ensemble classifier, i.e., AdaBoost with Decision Tree Classifier as a base estimator, which then outputted the final label for the outcome. For all outcomes, only the detection probabilities of the positive label "1" were concatenated. The positive label was the "presence of depression" for "depression", "high burden" for "global MS symptom burden", "severe fatigue" for "fatigue", and "poor sleep quality" for "sleep quality". The "n_estimators (The maximum number of estimators at which boosting is terminated.)" parameter was tuned during leave-5-participants-out cross-validation to achieve the best-performing combined model.

To analyze the usefulness of each sensor, we implemented a feature ablation analysis by generating detection results for all possible combinations of 1-sensor models. For six 1-sensor models, there were 57 combinations of feature sets, as the total combinations = combinations with 2 sensors + ... + combinations with 6 sensors = $\sum_{r=2}^{6} \binom{6}{r} = 57$.

Obtaining Predictions from Combinations of Sensors and EMA-only Models: The detection probabilities from all six 1-sensor models and one EMA-only model were concatenated into a single feature vector and given as input to an ensemble classifier. We use the method described above to train this combined classifier.

To analyze the usefulness of each sensor, we implemented a feature ablation analysis by generating detection results for all possible combinations of 1-sensor models and EMA model. For six 1-sensor models and one EMA model, there were 57 combinations of feature sets, as the total combinations = combinations with 2 sensors + ... + combinations with 6 sensors = $\sum_{r=2}^{6} \binom{6}{r} = 57$.

Classifying Different Outcomes: We ran the following pipelined independently for each of the 4 outcomes, first using action-only features and then using action + context features:

1. Training and validating six 1-sensor models and 57 combined sensor models.

2. Training and validating six 1-sensor models and average EMA model and the resulting 120 combined models.

3. Training and validating six 1-sensor models and pre-survey EMA model and the resulting 120 combined models.

We obtain 6 final models for each outcome. For every final model of each outcome, we reported the performance based on the best combination of sensors or sensors and EMA. We also reported the performance of baseline models (i.e., a simple majority classifier whereby every point is assigned to whichever is in the majority in the training set) as well as models containing all six sensors or all six sensors and one EMA type.

Compare Machine Learning Models by Bootstrapping Predictions: We computed 95% confidence intervals for the F1-score and accuracy of each final model by bootstrapping the test predictions (iterations = 10000, alpha = 0.05). For each outcome, we also compared the resulting 6 final models in a pairwise manner (30 comparisons) using hierarchical bootstrapping by randomly sampling participant ID, prediction week with replacement over 10000 iterations. In each iteration, we took samples with the same participant ID, prediction week across the two models being compared and computed the difference in F1-score and difference in accuracy. After computing this for all iterations, we were able to the 95% confidence intervals of the difference in F1-score and difference in accuracy (alpha = 0.05). If one of the models in a pair is not statistically better than the other, we consider the model needing the least amount of data to be "best".

7.3 Results

Figure 7.2 shows the performance of the machine learning pipeline for predicting each of the four outcomes using the best sensor-EMA combinations (i.e., the set of sensors and/or average or pre-survey EMA, that had the best performance for each outcome). We repeat this for models trained on Action-only features and Action + Context features. Accuracy is the percentage of samples for which the outcome label was correctly predicted. We have multiple samples from each patient. F1-score is a metric of model performance that measures the harmonic mean of precision and recall. Precision is the positive predictive value, or the number of true positive labels divided by the number of all positive labels (true positive + false positive). Recall is sensitivity, or the number of true positive labels divided by the number of all samples which should have the positive labels (true positive + false negative). In this study, "positive" label refers to the outcome of interest (e.g., presence of depression is the positive label for depression).

Figure 7.2: Performance of the machine learning pipeline using the best sensor or sensor+EMA combinations for predicting each of the four clinically relevant outcomes in pwMS every 2 or 4 weeks. "Action-Only & NoEMA" is the best model that combines predictions of 1-sensor models trained on action-only features. "Action+Context & NoEMA" is the best model that combines predictions of 1-sensor models trained on action and context features. "Action-Only & AvgEMA" is the best model that combines predictions of 1-sensor models and the average EMA model trained on action-only features. "Action+Context & AvgEMA" is the best model that combines predictions of 1-sensor models and the average EMA model trained on action and context features. "Action-Only & PresurveyEMA" is the best model that combines predictions of 1-sensor models and the pre-survey EMA model trained on action-only features. "Action+Context & PresurveyEMA" is the best model that combines predictions of 1-sensor models and the pre-survey EMA model trained on action and context features. We use bootstrapping to report average Accuracy (X 0.01) and F1-score and the corresponding 95% confidence intervals (alpha=0.05) for each of these models.

Depression: The baseline model (simple majority classifier) had an accuracy of 59.5% in predicting the presence of depression every 2 weeks. The model containing all 6 sensors and no EMA had an accuracy of 74.7% with action-only features (25.5% improvement over the baseline), and 72.2% with action + context features (21.3% improvement over the baseline). The model with the best combination of sensors and no EMA had an accuracy of 74.7% with action-only features (Action-Only & NoEMA in Figure 7.2 gave a 25.5% improvement over the baseline. Best combination: calls, heart rate, location, screen, sleep, and steps), and 74.7% with action + context features (Action+Context & NoEMA in Figure 7.2 gave a 25.5% improvement over the baseline. Best combination: calls, heart rate, location, screen, and sleep).

Adding average EMA as a 7th sensor, led to the model with the best combination of sensors and average EMA as having an accuracy of 80.8% with action-only features (Action-Only & AvgEMA in Figure 7.2 gave a 35.8% improvement over the baseline. Best combination: heart rate, sleep, steps, and average EMA), and 81.3% with action + context features (Action+Context & AvgEMA in Figure 7.2 gave a 36.6% improvement over the baseline. Best combination: calls, heart rate, location, sleep, and average EMA).

Adding pre-survey EMA as a 7th sensor, led to the model with the best combination of sensors and pre-survey EMA as having an accuracy of 80.6% with action-only features (Action-Only & PresurveyEMA in Figure 7.2 gave a 35.5% improvement over the baseline. Best combination: heart rate, steps, and pre-survey EMA), and 81.4% with action + context features (Action+Context & PresurveyEMA in Figure 7.2 gave a 36.8% improvement over the baseline. Best combination: heart rate, location, screen, and pre-survey EMA).

We compared the six combination models whose performance is reported in Figure 7.2 in a pairwise manner by computing average accuracy and F1-score on bootstrapped test sample predictions and computed the corresponding 95% confidence intervals (alpha = 0.05). Action+Context & PresurveyEMA had the highest bootstrapped average Accuracy of 81.4% and the highest average F1-score is 0.77. This model was significantly better than both the NoEMA models (significantly better than Action-Only & NoEMA by 6.7% accuracy and 0.09 F1, and significantly better than Action+Context & NoEMA by 6.6% accuracy and 0.1 F1). Further, Action-Only & PresurveyEMA was significantly better than both the NoEMA models (significantly better than Action-Only & NoEMA by 6.0% accuracy and 0.09 F1, and significantly better than Action+Contect & NoEMA by 6.1% accuracy and 0.09 F1). However, there were no statistically significant differences between Action-Only & PresurveyEMA and Action+Context & PresurveyEMA. The models with average EMA (Action-Only & AvgEMA, and Action+Context & AvgEMA) were also significantly better than both the NoEMA models, but there were no statistically significant differences between any of the PresurveyEMA models and the AvgEMA models.

Hence, we can say that **for predicting the presence of depression on a biweekly basis, the Action-Only & PresurveyEMA model gives the best performance while requiring the least amount of EMA (accuracy: 80.6% - a 35.5% improvement over baseline; f1-score: 0.76)**. This model requires EMA from only two days before the biweekly survey for measuring depression is administered: the last weekday before survey, and the last weekend before survey. It contains heart rate, steps, and pre-survey EMA.

Global MS Symptom Burden: The baseline model had an accuracy of 51.1% in predicting high global MS symptom burden (vs. "low burden") every 4 weeks. The model containing all 6 sensors and no EMA had an accuracy of 70.7% with action-only features (38.4% improvement over the baseline), and 72.0% with action + context features (40.9% improvement over the baseline). The model with the best combination of sensors and no EMA had an accuracy of 77.3% with action-only features (Action-Only & NoEMA in Figure 7.2 gave a 51.3% improvement over the baseline. Best combination: heart rate, location, sleep, and steps), and 73.8% with action + context features (Action+Context & NoEMA in Figure 7.2 gave a 44.4% improvement over the baseline. Best combination: heart rate, location, and sleep).

Adding average EMA as a 7th sensor, led to the model with the best combination of sensors and average EMA as having an accuracy of 77.9% with action-only features (Action-Only & AvgEMA in Figure 7.2 gave a 52.4% improvement over the baseline. Best combination: heart rate, location, sleep, steps, and average EMA), and 79.7% with action + context features (Action+Context & AvgEMA in Figure 7.2 gave a 56.0% improvement over the baseline. Best combination: calls, heart rate, screen, sleep, and average EMA).

Adding pre-survey EMA as a 7th sensor, led to the model with the best combination of sensors and pre-survey EMA as having an accuracy of 78.0% with action-only features (Action-Only & PresurveyEMA in Figure 7.2 gave a 52.6% improvement over the baseline. Best combination: location, sleep, steps, and pre-survey EMA), and 75.1% with action + context features (Action+Context & PresurveyEMA in Figure 7.2 gave a 47.0% improvement over the baseline. Best combination: heart rate, location, screen, sleep, and pre-survey EMA).

On comparing the six combination models reported in figure 7.2, we found that no model was significantly better than Action-Only & NoEMA. Hence, we can say that **for predicting high global MS symptom burden every 4 weeks, the Action-Only & NoEMA model is the best model (accuracy: 77.3% - a 51.3% improvement over baseline; f1-score: 0.77)**. This model contains the best combination of sensors (i.e., heart rate, location, sleep, and steps) where each 1-sensor model is trained on action-only features. No EMA is needed for this model.

Fatigue: The baseline model had an accuracy of 50.9% in predicting severe fatigue (vs.

"mild fatigue") every 4 weeks. The model containing all 6 sensors and no EMA had an accuracy of 60.4% with action-only features (18.7% improvement over the baseline), and 69.7% with action + context features (36.9% improvement over the baseline). The model with the best combination of sensors and no EMA had an accuracy of 67.6% with action-only features (Action-Only & NoEMA in Figure 7.2 gave a 32.8% improvement over the baseline. Best combination: calls, heart rate, screen, and steps), and 73.8% with action + context features (Action+Context & NoEMA in Figure 7.2 gave a 45.0% improvement over the baseline. Best combination: heart rate, screen, and steps).

Adding average EMA as a 7th sensor, led to the model with the best combination of sensors and average EMA as having an accuracy of 72.2% with action-only features (Action-Only & AvgEMA in Figure 7.2 gave a 41.9% improvement over the baseline. Best combination: heart rate, screen, steps, and average EMA), and 76.1% with action + context features (Action+Context & AvgEMA in Figure 7.2 gave a 49.5% improvement over the baseline. Best combination: heart rate, screen, sleep, steps, and average EMA).

Adding pre-survey EMA as a 7th sensor, led to the model with the best combination of sensors and pre-survey EMA as having an accuracy of 68.3% with action-only features (Action-Only & PresurveyEMA in Figure 7.2 gave a 34.2% improvement over the baseline. Best combination: heart rate, screen, steps, and pre-survey EMA), and 77.1% with action + context features (Action+Context & PresurveyEMA in Figure 7.2 gave a 51.5% improvement over the baseline. Best combination: calls, heart rate, screen, steps, and pre-survey EMA).

On comparing the six combination models reported in figure 7.2 in a pairwise manner, we found that no model was significantly better than Action+Context & NoEMA. Hence, we can say that **for predicting severe fatigue every 4 weeks, the Action+Context & NoEMA model is the best model (accuracy: 73.8% - a 45% improvement over baseline; f1-score: 0.74)**. This model contains the best combination of sensors (i.e., heart rate, screen, and steps) where each 1-sensor model is trained on action and context features.

Sleep Quality: The baseline model had an accuracy of 56.2% in predicting poor sleep quality (vs. "better sleep quality") every 4 weeks. The model containing all 6 sensors and no EMA had an accuracy of 58.2% with action-only features (3.6% improvement over the baseline), and 68.7% with action + context features (22.2% improvement over the baseline). The model with the best combination of sensors and no EMA had an accuracy of 72.0% with action-only features (Action-Only & NoEMA in Figure 7.2 gave a 28.1% improvement over the baseline. Best combination: heart rate, location, sleep, and steps), and 69.5% with action + context features (Action+Context & NoEMA in Figure 7.2 gave a 23.7% improvement over the baseline. Best combination: calls, heart rate, sleep, and steps).

Adding average EMA as a 7th sensor, led to the model with the best combination of sensors and average EMA as having an accuracy of 74.4% with action-only features (Action-Only

& AvgEMA in Figure 7.2 gave a 32.4% improvement over the baseline. Best combination: heart rate, location, screen, sleep, and average EMA), and 72.7% with action + context features (Action+Context & AvgEMA in Figure 7.2 gave a 29.4% improvement over the baseline. Best combination: heart rate, location, sleep, steps, and average EMA).

Adding pre-survey EMA as a 7th sensor, led to the model with the best combination of sensors and pre-survey EMA as having an accuracy of 72.0% with action-only features (Action-Only & PresurveyEMA in Figure 7.2 gave a 28.1% improvement over the baseline. Best combination: heart rate, location, sleep, and steps. Pre-survey EMA was not selected), and 74.0% with action + context features (Action+Context & PresurveyEMA in Figure 7.2 gave a 31.7% improvement over the baseline. Best combination: calls, heart rate, sleep, and pre-survey EMA).

On comparing the six combination models reported in figure 7.2 in a pairwise manner, we found that no model was significantly better than any other model. Hence, we can say that **for predicting poor sleep quality every 4 weeks, the Action-Only & NoEMA model is the best model as it is as good as the other models while requiring the least amount of data (accuracy: 72.0% - a 28.1% improvement over baseline; f1-score: 0.70)**. The Action-Only & NoEMA contains heart rate, location, sleep, and steps, while the Action+Context & NoEMA model contains calls, heart rate, sleep, and steps.

7.4 Discussion

In this paper, we report the feasibility of monitoring clinically relevant health outcomes over time for pwMS by leveraging passively sensed data from sensors in smartphones and fitness trackers. For this purpose, we trained machine learning models of passive sensor data to predict the presence of depression, high global MS symptom burden, severe fatigue, and poor sleep quality. We also explored whether these models could be improved using Ecological Momentary Assessments (EMAs) that involve quick active input from the user. For this, we considered EMAs collected three times a day throughout the study (average EMA) or EMAs collected on the last weekday and the last weekend just before the survey to measure the outcomes was administered (pre-survey EMA). For each sensor and EMA model, we explored if using behavioral features from the previous time period (context features) in addition to behavioral features from the current time period (action features) can improve performance.

The best models containing a combination of sensors and no EMA outperformed the baseline models (simple majority classifier) for all 4 outcomes. For global MS symptom burden, fatigue, and sleep quality, best models containing a combination of sensors and

average or pre-survey EMA did not perform significantly better than the best models with no EMA. That is, passively collected sensor data was alone enough to predict these outcomes. This result is especially surprising for fatigue, as one of the two EMA questions asked the user to rate how tired they feel, and one could hypothesize that the users' responses to this question would significantly improve prediction of fatigue. However, it seems like the sensors are able predict fatigue with the highest accuracy and we do not explicitly need to ask the user to rate their level of tiredness by administering EMAs. For depression, the best models containing a combination of sensors and average or pre-survey EMA performed significantly better than the best models with no EMA, but there was no difference between the best models containing pre-survey EMA and those containing average EMA. Hence, to predict depression, passively collected sensor data and pre-survey EMA (i.e., EMA from the last weekday and last weekend before the survey that measure the outcomes), are needed to predict depression with the highest accuracy. This result can be explained by the fact that one of the two EMA questions asked the user to rate how depressed they feel. While sensor data predicted depression with good accuracy (74.7%) without EMA, adding pre-survey EMA resulted in a 8.8% increase in accuracy.

For each outcome, we trained 6 final models, based on whether they used Action, or Action and Context features, or contained no EMA, average EMA, or pre-survey EMA. We compare these models by computing the 95% confidence intervals of differences in their bootstrapped accuracy and F1-scores. If one model is not statistically better than the other, we consider the model needed the least amount of data to be best. The model containing action-only features and pre-survey EMA was the best model for depression, and the model containing action-only features and no EMA was the best model for global MS symptom burden. So, adding context features did not improve prediction of depression or global MS symptom burden. However, the model containing action and context features and no EMA was the best model for fatigue. Hence, including data from a longer period of time by adding context features improved the prediction of fatigue. For sleep quality, all 6 models had similar statistical performance but the model containing action-only features and no required the least amount of data. The sensors that are selected in the best models across all outcomes are heart rate and steps.

While there is a small body of prior work that explored the feasibility of passive sensing in pwMS and preliminary correlations between passively sensed behaviors and MS outcomes [169, 231, 45, 23, 245, 97], previous work on predicting health outcomes for pwMS using passively sensed behaviors is scarce. Tong et al. used passively sensed sleep and activity data collected from 198 pwMS over 6 months to predict fatigue severity and overall health scores, achieving good performance in line with acceptable instrument errors [255]. Then, Chikersal et al. used passively sensed behavior changes to predict depression, disabil-

ity, fatigue, and sleep quality in pwMS [42]. However, this work predicted average health outcomes for each participant during the Covid-19 stay-at-home period. That is, this work predicted outcomes once for each participant instead of predicting outcomes at specific intervals. To our knowledge, this study is the first to use passively sensed behavior features to predict multiple inter-related clinically relevant health outcomes in MS, including depression, disability, fatigue, and sleep quality, repeatedly at periodic intervals with the goal of enabling health monitoring over time. From a methodological standpoint, the application of behavioral features computed from the current (action-only), and current and previous (action and context) time periods, and the use of average and pre-survey Ecological Momentary Assessments (EMAs) to predict depression and other health outcomes in pwMS is novel.

Our approach explored the clinical utility of using digital phenotyping for predicting health outcomes in pwMS and enabling their frequent monitoring over time. For example, predictive models built using our approach could help patients self-monitor and track their health in between their doctor's appointments with minimal or no active input from them. Our models could also help clinicians better monitor at-risk patients and make triage decisions for patients who require prioritization for interventions (e.g., medication, counseling). This can be especially useful for healthcare services in places with limited healthcare access and scares resources.

Our study has two limitations. First, we had a modest sample size. While we make predictions for 700+ samples for depression and 300+ samples for the other outcomes, these samples come from 104 participants only. To reduce the chance of over-fitting and improve validity of the findings, we used leave-5-participants-out-cross-validation, such that in each fold, the participants used for training and testing were different. Our approach performed well for not only one outcome, but all four clinically relevant outcomes pertaining to mental health and neurological disability in pwMS. We have reasonable confidence because of the consistently good model performance across all five folds and the consistently robust predictions for all four outcomes. Second, the study used patient-reported health outcomes instead of clinician-reported outcomes. However, these patient-reported outcomes are all validated for pwMS, highly correlated with rater-determined measures, interrelated among themselves, and clinically relevant.

7.5 Conclusion

In summary, we explored the use of digital phenotyping in predicting health outcomes in pwMS, with the goal of enabling frequent monitoring. Specifically, we predicted the presence of depression every two weeks, and high global MS symptom burden, severe fatigue, and poor

sleep quality every four weeks in pwMS using passively sensed behavioral features measured by smartphone and wearable fitness tracker. The predictive models achieved potentially clinically actionable performance for all four outcomes. Our work can enable future patient self-monitoring and clinician screening for urgent interventions in MS and other complex chronic diseases.

7.6 Addressing the Curse of Dimensionality

7.6.1 W.r.t. multiple co-morbidities (C2)

Our approach is able to monitor over time with high accuracy 4 health outcomes that are frequently co-morbid in patients with multiple sclerosis - depression, global MS symptom burden, fatigue, and poor sleep quality. By predicting each of these outcomes with high accuracy, our approach enables estimating a more complete picture of the patient's health, thereby addressing the curse of dimensionality w.r.t. multiple co-morbid outcomes.

7.6.2 W.r.t. diversity in patient characteristics (C3)

Our approach accounted for diversity in dynamic patient characteristics, context, or histories by including behavioral features from the time period preceding the current time period (context features), and assessing its impact on model performance for the longitudinal monitoring of different health outcomes. We found that for certain health outcomes such as fatigue, contextual behavioral features significantly improve performance.

Chapter 8

Thesis Conclusion and Future Work

My PhD work focused on the development of computational methods and models that use user-generated data from multiple data sources including passively sensed smartphone and wearable sensor data, text messages exchanged between users, and the users' interaction logs with web or mobile apps, to analyze or predict mental health outcomes with the goal of making the diagnosis and treatment of mental health disorders more efficient and precise. More specifically, it addressed the the curse of dimensionality challenge in precision mental health care in terms of the feature space, outcomes, and patients.

Through 5 studies, my thesis made the following contributions: (1) Presented a machine learning based feature selection method that mitigates the curse of dimensionality in the feature space by decomposing and iterative reducing the feature space during feature selection. (2) Demonstrated the generalizability of the approach in detecting depression, change in depression, and loneliness, as well as forecasting these outcomes several weeks in advance. (3) Demonstrated that behavioral changes resulting from the stay-at-home mandates during the pandemic are predictive of health outcomes during the stay-at-home period for patients with multiple sclerosis. (4) Demonstrated how we can categorize supporters or identify patient phenotypes based on multiple co-morbid outcomes, thereby mitigating the curse of dimensionality with respect to multimorbidities. (5) Presented a method that visualizes and identifies support strategies that work best in an online mental health intervention for patients in a specific context or situation. (6) Demonstrated that accounting for the patient's history or behavioral context improves model performance for the longitudinal monitoring of some health outcomes such as fatigue.

8.1 Key Takeaways from the Thesis

Below are the key takeaways from my thesis:

1. Multimodal behavioral sensing for precision mental health care is a wicked problem due to the curse of dimensionality w.r.t. the feature space, the existence of co-morbidities, and the diversity in patient characteristics. It is important to consider this problem during the data collection (*e.g.* by collecting potentially confounding health measures and patient characteristics) and data analysis(*e.g.* by decomposing the feature space, computationally combining outcomes) phases of all projects.

2. Including features from multiple time slices is ideal as behaviors during different times of the day/ week can indicate different symptoms, but doing so greatly increases the

dimensionality of the feature space leading to poor results with off-the-shelf machine learning approaches. Hence, in the previous work, researchers have often used features from select time slices or averaged across multiple time slices, risking losing valuable information. In this thesis, I proposed an approach based on feature space decomposition and late fusion which was able to obtain good performance on a feature space containing data from many time slices. In the future, leveraging my method or other approaches based on feature space decomposition and/or late fusion can help researchers obtain good model performance while preserving valuable information.

3. Mental health conditions are frequently co-morbid with other mental or physical health conditions. Considering multimorbidities for modeling and analyzing can positively impact patient care by allowing for more precise interventions. For example, the symptoms of depression and anxiety frequently co-occur and effect each other. Clinicians may treat someone with high depression and high anxiety differently from someone with high depression and low anxiety. While both these groups will benefit from medication and/or therapeutic interventions for depression, the former group may not show significant improvement in depression until their anxiety is also treated. In this thesis, I demonstrate two ways in which researchers can account for multimorbidities – by modeling or analyzing each co-morbidity separately, or by combining multiple co-morbid outcomes into one target outcome using computational approaches such as clustering. In addition to providing a more holistic picture of the patient's health, modeling or analyzing each co-morbidity separately can help researchers evaluate the generalizability and validity of their approach. Whereas, combining multiple co-morbid outcomes into one target outcome can simplify analysis when we want to derive data-driven insights to assess the effectiveness of an intervention. Based on their research questions, researchers can choose between these methods and incorporate them into their own work.

4. My thesis shows that looking at the relationship between current behaviors and outcomes alone isn't sufficient, as patient contexts can often change the relationship between behaviors and outcomes. For example, in study 4, when analyzing the relationship between supporter behaviors and client outcomes, we found shorter messages to correlate with better outcomes. However, when we also factored in patient contexts into our analysis, we found that while shorter messages correlate with better outcomes for engaged clients, longer messages correlate with better outcomes when

Late fusion here refers to the technique in which we combined predictions from different sensor models 'late' in the pipeline instead of concatenating features from all sensors from the beginning.

the client has very low engagement. Hence, while analyzing the relationship between behaviors and outcomes can reveal interesting general insights, researchers should also factor in patient contexts to derive more precise or personalized insights. Study 5 also demonstrates that including contextual features improves model performances for the longitudinal monitoring of some health outcomes such as fatigue.

8.2 Future Work

Here are some ideas for future work.

8.2.1 Increasing sample sizes

Studies 1-3 and 5 had fairly modest sample sizes. In order to properly evaluate the robustness of our models, we need to carry out studies with larger sample sizes that attempt to replicate our work or deploy these models in real-time.

8.2.2 Deploying predictive models and studying their acceptability

Through preliminary interviews with clinical experts and users, I've realized that for these models to be widely accepted, explainability and active learning are the most important features to have beyond accuracy. That is, users should be able to understand what led to the prediction and the model should learn from the user's feedback over time. We could, for example, accomplish this by designing interactive visualizations of the feature importances learned by the model. Users could provide feedback by disagreeing with the model's prediction or by rejecting the model's inference about a certain feature's importance. The model can then be retrained using the user's feedback. While deploying the models, I would also like to conduct user studies and interviews to better understand user preferences for frequency of predictions, level of detail for model explanations, their feelings about false positives, etc.

8.2.3 Explore the feasibility and utility phenotyping patients based on multiple co-morbidities

In studies 1 and 3, we achieved high accuracy for detecting multiple co-morbid outcomes *separately*. In study 4, we identify the most effective support strategies for improving two co-

morbid conditions by defining a "combined" target outcome based on both those measures. Combining outcomes related to multiple co-morbidities to get one combined target outcome or phenotype can be useful for a number of reasons. For example, in study 4, clustering depression and anxiety scores over time to get 1 target outcome *i.e.*, the success of the supporter, allowed us to identify effective strategies for improving depression and anxiety while reducing the computational complexity of the problem and resulting in interpretable insights. For detection and monitoring, combining outcomes by, say clustering them to get phenotypes, can allow clinicians to develop treatments that cater to specific phenotypes. For example, say the goal is to treat depression, and we can compute the patient's health phenotype at a given time based on their depressive symptoms, neurological symptoms, sleep quality, and fatigue at that time. Through practice and data collection, clinicians can learn what treatments result in improvement in depressive symptoms for patients in different health phenotype states. We can then train a model that detects the patient's health phenotype state based on their passively sensed behaviors, and the clinician can use the predictions of the model along with their knowledge of the best treatments for specific health phenotypes, to offer the patient the best data-driven treatment.

8.2.4 ML-driven personalized interventions and user-driven self-experimentation

In the future, researchers could leverage our feature extraction, select and modeling approaches to build on the early work on personalizing interventions to improve user outcomes in high-impact application areas. However, beyond personalized interventions, I'm very excited about enabling self-experimentation. Self-experimentation is an emerging field that involves giving the users access to tools that allow them to run self-experiments for better decision-making. We could use data-driven insights based on the user's historical data, and incorporate them in a interactive interface to help them make decisions about what their next actions should be. For example, in the online mental health intervention I worked on study 4, supporters can be shown statistics about how their potential actions can impact the client given their situation – "This client is not engaged. After receiving support messages of over 100 words, 40% of similar clients were able to re-engage. Whereas, after receiving short support messages, only 10% of similar clients were able to re-engage." We can also use a similar approach to help users plan their day. For example, each day the user has a set of actions to choose from such as spend time with family, exercise, and practice mindfulness. If the user has limited time, the UI they're using to plan their day can help them choose between tasks based on the outcome they care about the most on that day – "So you want

to focus on mood today. Your data shows that the days you do mindfulness, your mood improves by 50%. Whereas the days you exercise, your mood improved by 20%. The days you don't spend time with your family, your mood declines by 50%. What would you like to do next?"

This work has the potential to minimize suffering, by enabling early diagnosis and frequent monitoring of health outcomes using passively sensed longitudinal behavioral data. My work also had implications for more effective treatments through personalization, and improving the patients' awareness about their own health and treatment. Future work can build on my methods and the key takeaways of my work to further advance the state of precision mental health care.

References

[1] Ashraf Abdul et al. "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda". In: *Proceedings of the 2018 CHI conference on human factors in computing systems*. ACM. 2018, p. 582.

[2] Marios Adamou et al. "Mining free-text medical notes for suicide risk assessment". In: *Proceedings of the 10th hellenic conference on artificial intelligence*. ACM. 2018, p. 47.

[3] Sharifa Alghowinem et al. "Cross-cultural detection of depression from nonverbal behaviour". In: *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. Vol. 1. IEEE. 2015, pp. 1–8.

[4] Sharifa Alghowinem et al. "Eye movement analysis for depression detection". In: *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE. 2013, pp. 4220–4224.

[5] Sharifa Alghowinem et al. "Head pose and movement analysis as an indicator of depression". In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE. 2013, pp. 283–288.

[6] Tim Althoff, Kevin Clark, and Jure Leskovec. "Large-scale analysis of counseling conversations: An application of natural language processing to mental health". In: *Transactions of the Association for Computational Linguistics* 4 (2016), pp. 463–476.

[7] American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub, 2013.

[8] Gerhard Andersson et al. "Internet-delivered psychological treatments: from innovation to implementation". In: *World Psychiatry* 18.1 (2019), pp. 20–28.

[9] Laura Helena Andrade et al. "Barriers to mental health treatment: results from the WHO World Mental Health surveys". In: *Psychological medicine* 44.6 (2014), pp. 1303–1317.

[10] G Andrews et al. "Computer therapy for the anxiety and depression disorders is effective, acceptable and practical health care: an updated meta-analysis". In: *Journal of anxiety disorders* 55 (2018), pp. 70–78.

[11] Sangwon Bae, Anind K Dey, and Carissa A Low. "Using passively collected sedentary behavior to predict hospital readmission". In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2016, pp. 616–621.

[12] R Bakshi et al. "Fatigue in multiple sclerosis and its relationship to depression and neurologic disability". In: *Multiple Sclerosis Journal* 6.3 (2000), pp. 181–185.

[13] Nina B Baltierra et al. "More than just tracking time: complex measures of user engagement with an internet-based health promotion intervention". In: *Journal of biomedical informatics* 59 (2016), pp. 299–307.

[14] Nikola Banovic et al. "Modeling and understanding human routine behavior". In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, pp. 248–260.

[15] Jakob E Bardram et al. "Designing mobile health technology for bipolar disorder: a field trial of the monarca system". In: *Proceedings of the SIGCHI conference on human factors in computing systems*. ACM. 2013, pp. 2627–2636.

[16] Aaron T Beck. *Cognitive therapy of depression*. Guilford press, 1979.

[17] Aaron T Beck and Keith Bredemeier. "A unified model of depression: Integrating clinical, cognitive, biological, and evolutionary perspectives". In: *Clinical Psychological Science* 4.4 (2016), pp. 596–619.

[18] Aaron T Beck, Robert A Steer, and Gregory K Brown. "Beck depression inventory-II". In: *San Antonio* 78.2 (1996), pp. 490–8.

[19] Richard E Bellman. *Adaptive control processes*. Princeton university press, 2015.

[20] Dror Ben-Zeev et al. "Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health." In: *Psychiatric rehabilitation journal* 38.3 (2015), p. 218.

[21] Visar Berisha et al. "Digital medicine and the curse of dimensionality". In: *NPJ digital medicine* 4.1 (2021), pp. 1–8.

[22] Leonard Bickman, Aaron R Lyon, and Miranda Wolpert. *Achieving precision mental health through effective assessment, monitoring, and feedback processes*. 2016.

[23] Valerie J Block et al. "Association of continuous assessment of step count by remote monitoring with disability progression among adults with multiple sclerosis". In: *JAMA network open* 2.3 (2019), e190570–e190570.

[24] Edward S Bordin. "The generalizability of the psychoanalytic concept of the working alliance." In: *Psychotherapy: Theory, research & practice* 16.3 (1979), p. 252.

[25] Mehdi Boukhechba, Anna N Baglione, and Laura E Barnes. "Leveraging Mobile Sensing and Machine Learning for Personalized Mental Health Care". In: *Ergonomics in Design* 28.4 (2020), pp. 18–23.

[26] Mehdi Boukhechba et al. "DemonicSalmon: Monitoring mental health and social interactions of college students using smartphones". In: *Smart Health* 9 (2018), pp. 192–203.

[27] Mehdi Boukhechba et al. "Predicting social anxiety from global positioning system traces of college students: feasibility study". In: *JMIR mental health* 5.3 (2018), e10101.

[28] Robert Bryll, Ricardo Gutierrez-Osuna, and Francis Quek. "Attribute bagging: improving accuracy of classifier ensembles by using random feature subsets". In: *Pattern recognition* 36.6 (2003), pp. 1291–1302.

[29] Jonathan Kenneth Burns. "Mental health and inequity: a human rights approach to inequality, discrimination, and mental disability". In: *Health & Hum. Rts.* 11 (2009), p. 19.

[30] Daniel J Buysse et al. "Quantification of subjective sleep quality in healthy elderly men and women using the Pittsburgh Sleep Quality Index (PSQI)". In: *Sleep* 14.4 (1991), pp. 331–338.

[31] Lihua Cai et al. "An integrated framework for using mobile sensing to understand response to mobile interventions among breast cancer patients". In: *Smart Health* 15 (2020), p. 100086.

[32] Luca Canzian and Mirco Musolesi. "Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis". In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM. 2015, pp. 1293–1304.

[33] Bokai Cao et al. "Deepmood: modeling mobile phone typing dynamics for mood detection". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 747–755.

[34] Alan Carr. *What works with children, adolescents, and adults?: a review of research on the effectiveness of psychotherapy*. Routledge, 2008.

[35] Carol K Chan et al. "Depression in multiple sclerosis across the adult lifespan". In: *Multiple Sclerosis Journal* 27.11 (2021), pp. 1771–1780.

[36] Stevie Chancellor. "Computational Methods to Understand Deviant Mental Wellness Communities". In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. DC05.

[37] Rosa Chaves et al. "Association rule-based feature selection method for Alzheimer's disease diagnosis". In: *Expert Systems with Applications* 39.14 (2012), pp. 11766–11774.

[38] Annie T Chen et al. "A multi-faceted approach to characterizing user behavior and experience in a digital mental health intervention". In: *Journal of biomedical informatics* 94 (2019), p. 103187.

[39] Zhenyu Chen et al. "ContextSense: unobtrusive discovery of incremental social context using dynamic bluetooth data". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM. 2014, pp. 23–26.

[40] Zhenyu Chen et al. "Inferring social contextual behavior from bluetooth traces". In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM. 2013, pp. 267–270.

[41] Prerna Chikersal et al. "Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing: A Machine Learning Approach With Robust Feature Selection". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 28.1 (2021), pp. 1–41.

[42] Prerna Chikersal et al. "Predicting multiple sclerosis outcomes during the COVID-19 stay-at-home period: Observational study using passively sensed behaviors and digital phenotyping". In: *JMIR Mental Health* 9.8 (2022), e38495.

[43] Prerna Chikersal et al. "Understanding client support strategies to improve clinical outcomes in an online mental health intervention". In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–16.

[44] James F Childress et al. "Public health ethics: mapping the terrain". In: *The Journal of Law, Medicine & Ethics* 30.2 (2002), pp. 170–178.

[45] Tanuja Chitnis et al. "Quantifying neurologic disease using biosensor measurements in-clinic and in free-living settings in multiple sclerosis". In: *NPJ digital medicine* 2.1 (2019), pp. 1–8.

[46] Philip I Chow et al. "Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students". In: *Journal of medical Internet research* 19.3 (2017), e62.

[47] Sarah Clement et al. "What is the impact of mental health-related stigma on help-seeking? A systematic review of quantitative and qualitative studies". In: *Psychological medicine* 45.1 (2015), pp. 11–27.

[48] Jeffrey F Cohn et al. “Detecting depression from facial actions and vocal prosody”. In: *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference on*. IEEE. 2009, pp. 1–7.

[49] Benjamin Lê Cook, Colleen L Barry, and Susan H Busch. “Racial/ethnic disparity trends in children’s mental health care access and expenditures from 2002 to 2007”. In: *Health services research* 48.1 (2013), pp. 129–149.

[50] Jesse D Cook, Michael L Prairie, and David T Plante. “Utility of the Fitbit Flex to evaluate sleep in major depressive disorder: a comparison against polysomnography and wrist-worn actigraphy”. In: *Journal of affective disorders* 217 (2017), pp. 299–305.

[51] Jonathan E Cook and Carol Doyle. “Working alliance in online therapy as compared to face-to-face therapy: Preliminary results”. In: *CyberPsychology & Behavior* 5.2 (2002), pp. 95–105.

[52] Sarah A Costigan et al. “The health indicators associated with screen-based sedentary behavior among adolescent girls: a systematic review”. In: *Journal of Adolescent Health* 52.4 (2013), pp. 382–392.

[53] David Coyle and Gavin Doherty. “Clinical evaluations and collaborative design: developing new technologies for mental healthcare interventions”. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2009, pp. 2051–2060.

[54] Mark É Czeisler et al. “Mental health, substance use, and suicidal ideation during the COVID-19 pandemic—United States, June 24–30, 2020”. In: *Morbidity and Mortality Weekly Report* 69.32 (2020), p. 1049.

[55] Ewa K Czyz et al. “Self-reported barriers to professional help seeking among college students at elevated risk for suicide”. In: *Journal of American College Health* 61.7 (2013), pp. 398–406.

[56] Munmun De Choudhury et al. “Predicting depression via social media.” In: *ICWSM* 13 (2013), pp. 1–10.

[57] Orianna DeMasi and Benjamin Recht. “A step towards quantifying when an algorithm can and cannot predict an individual’s wellbeing”. In: *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. ACM. 2017, pp. 763–771.

[58] Kadir Demirci, Mehmet Akgönül, and Abdullah Akpinar. "Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students". In: *Journal of behavioral addictions* 4.2 (2015), pp. 85–92.

[59] Bruce J Diamond et al. "Relationships between information processing, depression, fatigue and cognition in multiple sclerosis". In: *Archives of clinical neuropsychology* 23.2 (2008), pp. 189–199.

[60] Thomas G Dietterich. "Statistical tests for comparing supervised classification learning algorithms". In: *Oregon State University Technical Report* 1 (1996), pp. 1–24.

[61] Lisa B Dixon, Yael Holoshitz, and Ilana Nossel. "Treatment engagement of individuals experiencing mental illness: review and update". In: *World Psychiatry* 15.1 (2016), pp. 13–20.

[62] Gavin Doherty, David Coyle, and John Sharry. "Engagement with online mental health interventions: an exploratory clinical study of a treatment for depression". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2012, pp. 1421–1430.

[63] Christopher M Doran and Irina Kinchin. "A review of the economic impact of mental illness". In: *Australian Health Review* 43.1 (2017), pp. 43–48.

[64] Afsaneh Doryab. "Identifying Symptoms Using Technology". In: *Technology and Adolescent Mental Health*. Springer, 2018, pp. 135–153.

[65] Afsaneh Doryab et al. "Detection of Behavior Change in People with Depression." In: *AAAI Workshop: Modern Artificial Intelligence for Health Analytics*. 2014.

[66] Afsaneh Doryab et al. "Identifying Behavioral Phenotypes of Loneliness and Social Isolation with Passive Sensing: Statistical Analysis, Data Mining and Machine Learning of Smartphone and Fitbit Data". In: *JMIR mHealth and uHealth* 7.7 (2019), e13209.

[67] Afsaneh Doryab et al. "Impact factor analysis: combining prediction with parameter ranking to reveal the impact of behavior on health outcome". In: *Personal and Ubiquitous Computing* 19.2 (2015), pp. 355–365.

[68] David JA Dozois, Keith S Dobson, and Jamie L Ahnberg. "A psychometric evaluation of the Beck Depression Inventory–II." In: *Psychological assessment* 10.2 (1998), p. 83.

[69] Rakkrit Duangsoithong and Terry Windeatt. "Bootstrap feature selection for ensemble classifiers". In: *Industrial Conference on Data Mining*. Springer. 2010, pp. 28–41.

[70] David J Duffy. "Problems, challenges and promises: perspectives on precision medicine". In: *Briefings in bioinformatics* 17.3 (2016), pp. 494–504.

[71] Nathan Eagle, Alex Sandy Pentland, and David Lazer. "Inferring friendship network structure by using mobile phone data". In: *Proceedings of the national academy of sciences* 106.36 (2009), pp. 15274–15278.

[72] Kirstin Early, Stephen E Fienberg, and Jennifer Mankoff. "Test time feature ordering with FOCUS: interactive predictions with minimal user burden". In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2016, pp. 992–1003.

[73] Malin Eiband et al. "Bringing transparency design into practice". In: *23rd International Conference on Intelligent User Interfaces*. ACM. 2018, pp. 211–223.

[74] Daniel Eisenberg, Ezra Golberstein, and Sarah E Gollust. "Help-seeking and access to mental health care in a university student population". In: *Medical care* (2007), pp. 594–601.

[75] Daniel Eisenberg, Ezra Golberstein, and Justin B Hunt. "Mental health and academic success in college". In: *The BE Journal of Economic Analysis & Policy* 9.1 (2009).

[76] Sindhu Kiranmai Ernala et al. "Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 134.

[77] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[78] Catherine K Ettman et al. "Prevalence of depression symptoms in US adults before and during the COVID-19 pandemic". In: *JAMA network open* 3.9 (2020), e2019686–e2019686.

[79] Gunther Eysenbach. "The law of attrition". In: *Journal of medical Internet research* 7.1 (2005), e11.

[80] Asma Ahmad Farhan et al. "Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data." In: *Wireless Health*. 2016, pp. 30–37.

[81] Anthony Feinstein et al. "The link between multiple sclerosis and depression". In: *Nature Reviews Neurology* 10.9 (2014), pp. 507–517.

[82] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. "AWARE: mobile context instrumentation framework". In: *Frontiers in ICT* 2 (2015), p. 6.

[83] Norman L Fichtenberg et al. "Insomnia screening in postacute traumatic brain injury: utility and validity of the Pittsburgh Sleep Quality Index". In: *American journal of physical medicine & rehabilitation* 80.5 (2001), pp. 339–345.

[84] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. "Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial". In: *JMIR mental health* 4.2 (2017), e7785.

[85] W Elon Fleming and Charles P Pollak. "Sleep disorders in multiple sclerosis". In: *Seminars in neurology*. Vol. 25. 01. Copyright© 2005 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New ... 2005, pp. 64–68.

[86] Helen Ford, Peter Trigwell, and Michael Johnson. "The nature of fatigue in multiple sclerosis". In: *Journal of psychosomatic research* 45.1 (1998), pp. 33–38.

[87] Mary Jane Friedrich. "Depression is the leading cause of disability around the world". In: *Jama* 317.15 (2017), pp. 1517–1517.

[88] Yusong Gao et al. "How smartphone usage correlates with social anxiety and loneliness". In: *PeerJ* 4 (2016), e2197.

[89] Marzyeh Ghassemi et al. "Opportunities in machine learning for healthcare". In: *arXiv preprint arXiv:1806.00388* (2018).

[90] Martin Gjoreski et al. "Continuous stress detection using a wrist device: in laboratory and real life". In: *proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing: Adjunct*. 2016, pp. 1185–1193.

[91] Erving Goffman. *Forms of talk*. University of Pennsylvania Press, 1981.

[92] Dion H Goh and Rebecca P Ang. "An introduction to association rule mining: An application in counseling and help-seeking behavior of adolescents". In: *Behavior Research Methods* 39.2 (2007), pp. 259–266.

[93] Marvin M Goldenberg. "Multiple sclerosis review". In: *Pharmacy and therapeutics* 37.3 (2012), p. 175.

[94] Sarah Goodell et al. "Mental disorders and medical comorbidity". In: *Robert Wood Johnson Foundation* 2 (2011).

[95] Jack M Gorman. "Comorbid depression and anxiety spectrum disorders". In: *Depression and anxiety* 4.4 (1996), pp. 160–168.

[96] Darcy Gruttadaro and Dana Crudo. "College students speak: A survey report on mental health". In: *Retrieved from National Alliance on Mental Illness website: https://www.nami. org/About-NAMI/Publications-Reports/Survey-Reports/College-Students-Speak_A-Survey-Report-on-Mental-H. pdf* (2012).

[97] Gabriel Guo et al. "MSLife: Digital Behavioral Phenotyping of Multiple Sclerosis Symptoms in the Wild Using Wearables and Graph-Based Statistical Analysis". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.4 (2021), pp. 1–35.

[98] Heather D Hadjistavropoulos et al. "Therapeutic alliance in internet-delivered cognitive behaviour therapy for depression or generalized anxiety". In: *Clinical psychology & psychotherapy* 24.2 (2017), pp. 451–461.

[99] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[100] Paul A Harris et al. "Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support". In: *Journal of biomedical informatics* 42.2 (2009), pp. 377–381.

[101] Paul A Harris et al. "The REDCap consortium: Building an international community of software platform partners". In: *Journal of biomedical informatics* 95 (2019), p. 103208.

[102] Andrew F Hayes. *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. Guilford publications, 2017.

[103] National Institute of Health. *Multiple Sclerosis*. `https://www.nccih.nih.gov/health/multiple-sclerosis` [Accessed: 2023].

[104] Eric Heiligenstein et al. "Depression and academic impairment in college students". In: *Journal of American College Health* 45.2 (1996), pp. 59–64.

[105] Tim Bodyka Heng, Ankit Gupta, and Chris Shaw. "FitViz-Ad: A Non-Intrusive Reminder to Encourage Non-Sedentary Behaviour". In: *Electronic Imaging* 2018.1 (2018), pp. 332–1.

[106] Tad Hirsch et al. "It's hard to argue with a computer: Investigating Psychotherapists' Attitudes towards Automated Evaluation". In: *Proceedings of the 2018 Designing Interactive Systems Conference*. ACM. 2018, pp. 559–571.

[107] Fred Hohman et al. "Gamut: A design probe to understand how data scientists understand machine learning models". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 579.

[108] Fredrik Holländare et al. “Therapist behaviours in internet-based cognitive behaviour therapy (ICBT) for depressive symptoms”. In: *Internet Interventions* 3 (2016), pp. 1–7.

[109] Julianne Holt-Lunstad et al. “Loneliness and social isolation as risk factors for mortality: a meta-analytic review”. In: *Perspectives on psychological science* 10.2 (2015), pp. 227–237.

[110] Michael Hölzer et al. “Vocabulary measures for the evaluation of therapy outcome: Re-studying transcripts from the Penn Psychotherapy Project”. In: *Psychotherapy Research* 6.2 (1996), pp. 95–108.

[111] Wallace J Hopp, Jun Li, and Guihua Wang. “Big data and the precision medicine revolution”. In: *Production and Operations Management* 27.9 (2018), pp. 1647–1664.

[112] Kurt Hornik, Bettina Grün, and Michael Hahsler. “arules-A computational environment for mining association rules and frequent item sets”. In: *Journal of Statistical Software* 14.15 (2005), pp. 1–25.

[113] Jeremy F Huckins et al. “Mental health and behavior of college students during the early phases of the COVID-19 pandemic: Longitudinal smartphone and ecological momentary assessment study”. In: *Journal of medical Internet research* 22.6 (2020), e20185.

[114] Kit Huckvale, Svetha Venkatesh, and Helen Christensen. “Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety”. In: *NPJ digital medicine* 2.1 (2019), pp. 1–11.

[115] Alketa Hysenbegasi, Steven L Hass, and Clayton R Rowland. “The impact of depression on the academic productivity of university students”. In: *Journal of mental health policy and economics* 8.3 (2005), p. 145.

[116] Spencer L James et al. “Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017”. In: *The Lancet* 392.10159 (2018), pp. 1789–1858.

[117] Vallabh Janardhan and Rohit Bakshi. “Quality of life in patients with multiple sclerosis: the impact of fatigue and depression”. In: *Journal of the neurological sciences* 205.1 (2002), pp. 51–58.

[118] Robert Johansson et al. “Internet-based psychological treatments for depression”. In: *Expert review of neurotherapeutics* 12.7 (2012), pp. 861–870.

[119] Robert Johansson et al. "Tailored vs. standardized internet-based cognitive behavior therapy for depression and comorbid symptoms: a randomized controlled trial". In: *PloS one* 7.5 (2012), e36905.

[120] Michael I Jordan and Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects". In: *Science* 349.6245 (2015), pp. 255–260.

[121] Jyoti Joshi et al. "Can body expressions contribute to automatic depression analysis?" In: *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE. 2013, pp. 1–7.

[122] Harmeet Kaur Kang et al. "Prevalence of mental health disorders among undergraduate university students in the United States: A review". In: *Journal of psychosocial nursing and mental health services* 59.2 (2021), pp. 17–24.

[123] Evangelos Karapanos. "Sustaining user engagement with behavior-change tools". In: *Interactions* (2015).

[124] Raghavendra Katikalapudi et al. "Associating internet usage with depressive behavior among college students". In: *IEEE Technology and Society Magazine* 31.4 (2012), pp. 73–80.

[125] Ramakanth Kavuluru et al. "Classification of helpful comments on online suicide watch forums". In: *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM. 2016, pp. 32–40.

[126] Ronald C Kessler et al. "Trends in suicide ideation, plans, gestures, and attempts in the United States, 1990-1992 to 2001-2003". In: *Jama* 293.20 (2005), pp. 2487–2495.

[127] Jeremy Kisch, E Victor Leino, and Morton M Silverman. "Aspects of suicidal behavior, depression, and treatment in college students: Results from the Spring 2000 National College Health Assessment Survey". In: *Suicide and Life-Threatening Behavior* 35.1 (2005), pp. 3–13.

[128] Britt Klein et al. "A therapist-assisted cognitive behavior therapy internet intervention for posttraumatic stress disorder: pre-, post-and 3-month follow-up results from an open trial". In: *Journal of anxiety disorders* 24.6 (2010), pp. 635–644.

[129] Adam DI Kramer, Lui Min Oh, and Susan R Fussell. "Using linguistic features to measure presence in computer-mediated communication". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM. 2006, pp. 913–916.

[130] Kurt Kroenke and Robert L Spitzer. "The PHQ-9: a new depression diagnostic and severity measure". In: *Psychiatric annals* 32.9 (2002), pp. 509–515.

[131] Kurt Kroenke, Robert L Spitzer, and Janet BW Williams. "The PHQ-9: validity of a brief depression severity measure". In: *Journal of general internal medicine* 16.9 (2001), pp. 606–613.

[132] Autumn Kujawa et al. "Exposure to COVID-19 pandemic stress: Associations with depression and anxiety in emerging adults in the United States". In: *Depression and anxiety* 37.12 (2020), pp. 1280–1288.

[133] Min Kwon et al. "Development and validation of a smartphone addiction scale (SAS)". In: *PloS one* 8.2 (2013), e56936.

[134] Gionet Kylie. *Meet Tess: The Mental Health Chatbot that Thinks like a Therapists.* `https://www.theguardian.com/society/2018/apr/25/meet-tess-the-mental-health-chatbot`. 2018.

[135] Gionet Kylie. *Meet Tess: The Mental Health Chatbot that Thinks like a Therapists.* 2018. URL: `https://www.theguardian.com/society/2018/apr/25/meet-tess-the-mental-health-chatbot` (visited on 07/03/2018).

[136] Tien-Duy B Le and David Lo. "Beyond support and confidence: Exploring interestingness measures for rule-based specification mining". In: *2015 IEEE 22nd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*. IEEE. 2015, pp. 331–340.

[137] Catherine Lebel et al. "Elevated depression and anxiety symptoms among pregnant individuals during the COVID-19 pandemic". In: *Journal of affective disorders* 277 (2020), pp. 5–13.

[138] Reeva Lederman et al. "Moderated online social therapy: Designing and evaluating technology for mental health". In: *ACM Transactions on Computer-Human Interaction (TOCHI)* 21.1 (2014), p. 5.

[139] Seth N Levin et al. "Association of social network structure and physical function in patients with multiple sclerosis". In: *Neurology* 95.11 (2020), e1565–e1574.

[140] Seth N Levin et al. "Manifestations and impact of the COVID-19 pandemic in neuroinflammatory diseases". In: *Annals of clinical and translational neurology* 8.4 (2021), pp. 918–928.

[141] Elle Levit et al. "Worsening physical functioning in patients with neuroinflammatory disease during the COVID-19 pandemic". In: *Multiple Sclerosis and Related Disorders* (2022), p. 103482.

[142] Philip Lindner et al. "The impact of telephone versus e-mail therapist guidance on treatment outcomes, therapeutic alliance and treatment engagement in Internet-delivered CBT for depression: A randomised pilot trial". In: *Internet Interventions* 1.4 (2014), pp. 182–187.

[143] Melanie Lovatt and John Holmes. "Digital phenotyping and sociological perspectives in a Brave New World". In: *Addiction (Abingdon, England)* 112.7 (2017), p. 1286.

[144] Bernd Löwe et al. "Validation and standardization of the Generalized Anxiety Disorder Screener (GAD-7) in the general population". In: *Medical care* 46.3 (2008), pp. 266–274.

[145] Tommaso Manacorda et al. "Impact of the COVID-19 pandemic on persons with multiple sclerosis: Early findings from a survey on disruptions in care and self-reported outcomes". In: *Journal of Health Services Research & Policy* (2020), p. 1355819620975069.

[146] Ashika Mani et al. "Applying Deep Learning to Accelerated Clinical Brain Magnetic Resonance Imaging for Multiple Sclerosis". In: *Frontiers in neurology* 12 (2021).

[147] Janna Mantua, Nickolas Gravel, and Rebecca Spencer. "Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography". In: *Sensors* 16.5 (2016), p. 646.

[148] Colin Martindale. *English Regressive Imagery Dictionary (RID)*. 1975. URL: `https://rdrr.io/cran/lexicon/man/key_regressive_imagery.html` (visited on 09/18/2019).

[149] Mark Matthews et al. "In Situ Design for Mental Illness: Considering the Pathology of Bipolar Disorder in mHealth Design". In: *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. MobileHCI '15. Copenhagen, Denmark: ACM, 2015, pp. 86–97. ISBN: 978-1-4503-3652-9. DOI: 10.1145/2785830.2785866. URL: http://doi.acm.org/10.1145/2785830.2785866.

[150] Mark Matthews et al. "In situ design for mental illness: Considering the pathology of bipolar disorder in mhealth design". In: *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*. 2015, pp. 86–97.

[151] J Michael McGinnis, Pamela Williams-Russo, and James R Knickman. "The case for more active policy attention to health promotion". In: *Health affairs* 21.2 (2002), pp. 78–93.

[152] Virginia Meca-Lallana et al. "Assessing fatigue in multiple sclerosis: Psychometric properties of the five-item Modified Fatigue Impact Scale (MFIS-5)". In: *Multiple Sclerosis Journal–Experimental, Translational and Clinical* 5.4 (2019), p. 2055217319887987.

[153] Abhinav Mehrotra and Mirco Musolesi. "Using Autoencoders to Automatically Extract Mobility Features for Predicting Depressive States". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.3 (2018), p. 127.

[154] Nicolai Meinshausen and Peter Bühlmann. "Stability selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4 (2010), pp. 417–473.

[155] National Alliance on Mental Illness. *College students speak: A survey report on mental health*. 2012.

[156] *Mental Illness – NIMH*. https://www.nimh.nih.gov/health/statistics/mental-illness. Accessed: 2022-01-31.

[157] Krista Merry and Pete Bettinger. "Smartphone GPS accuracy study in an urban environment". In: *PloS one* 14.7 (2019), e0219890.

[158] Susan Michie et al. "Developing and evaluating digital interventions to promote behavior change in health and health care: recommendations resulting from an international workshop". In: *Journal of medical Internet research* 19.6 (2017), e232.

[159] Jan Mielniczuk and Paweł Teisseyre. "Using random subspace method for prediction and variable importance assessment in linear regression". In: *Computational Statistics & Data Analysis* 71 (2014), pp. 725–742.

[160] Jun-Ki Min et al. "Toss'n'turn: smartphone as sleep and sleep quality detector". In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM. 2014, pp. 477–486.

[161] Saif M Mohammad and Peter D Turney. "Nrc emotion lexicon". In: *National Research Council, Canada* (2013).

[162] David Mohr, Pim Cuijpers, and Kenneth Lehman. "Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions". In: *Journal of medical Internet research* 13.1 (2011), e30.

[163] David C Mohr et al. "Comparison of the Effects of Coaching and Receipt of App Recommendations on Depression, Anxiety, and Engagement in the IntelliCare Platform: Factorial Randomized Controlled Trial". In: *Journal of medical Internet research* 21.8 (2019), e13609.

[164] Scott M Monroe et al. "Major life events and major chronic difficulties are differentially associated with history of major depressive episodes." In: *Journal of abnormal psychology* 116.1 (2007), p. 116.

[165] Stuart A Montgomery and MARIE Åsberg. "A new depression scale designed to be sensitive to change". In: *The British journal of psychiatry* 134.4 (1979), pp. 382–389.

[166] Francesco Motolese et al. "The psychological impact of COVID-19 pandemic on people with multiple sclerosis". In: *Frontiers in neurology* 11 (2020), p. 1255.

[167] Inbal Nahum-Shani et al. "Just-in-time adaptive interventions (JITAIs) in mobile health: key components and design principles for ongoing health behavior support". In: *Annals of Behavioral Medicine* 52.6 (2017), pp. 446–462.

[168] Rokas Navickas et al. "Multimorbidity: what do we know? What should we do?" In: *Journal of comorbidity* 6.1 (2016), pp. 4–11.

[169] Pamela Newland et al. "Exploring the feasibility and acceptability of sensor monitoring of gait and falls in the homes of persons with multiple sclerosis". In: *Gait & posture* 49 (2016), pp. 277–282.

[170] Michelle G Newman et al. "A review of technology-assisted self-help and minimal contact therapies for anxiety and depression: is human contact necessary for therapeutic efficacy?" In: *Clinical psychology review* 31.1 (2011), pp. 89–103.

[171] Thin Nguyen et al. "Using linguistic and topic analysis to classify sub-groups of online depression communities". In: *Multimedia tools and applications* 76.8 (2017), pp. 10653–10676.

[172] Tom Nicolai and Holger Kenn. "Towards detecting social situations with Bluetooth". In: *Adjunct Proceedings Ubicomp*. 2006.

[173] Alicia L Nobles et al. "Identification of imminent suicide risk among young adults using text messages". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 413.

[174] Matthew K Nock and Ronald C Kessler. "Prevalence of and risk factors for suicide attempts versus suicide gestures: analysis of the National Comorbidity Survey." In: *Journal of abnormal psychology* 115.3 (2006), p. 616.

[175] Matthew K Nock et al. "Mental disorders, comorbidity and suicidal behavior: results from the National Comorbidity Survey Replication". In: *Molecular psychiatry* 15.8 (2010), p. 868.

[176] David Nutt, Sue Wilson, and Louise Paterson. "Sleep disorders as core symptoms of depression". In: *Dialogues in clinical neuroscience* 10.3 (2008), p. 329.

[177] David Nutt, Sue Wilson, and Louise Paterson. "Sleep disorders as core symptoms of depression". In: *Dialogues in clinical neuroscience* (2022).

[178] Emma O'Brien. "Therapist behaviours, the working alliance and clinician experience in iCBT for depression and anxiety". PhD thesis. Trinity College Dublin, 2018.

[179] Rory C O'Connor et al. "Mental health and well-being during the COVID-19 pandemic: longitudinal analyses of adults in the UK COVID-19 Mental Health & Wellbeing study". In: *The British Journal of Psychiatry* 218.6 (2021), pp. 326–333.

[180] Mark Olfson et al. "Dropout from outpatient mental health care in the United States". In: *Psychiatric Services* 60.7 (2009), pp. 898–907.

[181] Graziano Onder et al. "Facing multimorbidity in the precision medicine era". In: *Mechanisms of ageing and development* 190 (2020), p. 111287.

[182] Ju Lynn Ong et al. "COVID-19-related mobility reduction: heterogenous effects on sleep and physical activity rhythms". In: *Sleep* 44.2 (2021), zsaa179.

[183] Theodor Chris Panagiotakopoulos et al. "A contextual data mining approach toward assisting the treatment of anxiety disorders". In: *IEEE transactions on information technology in biomedicine* 14.3 (2010), pp. 567–581.

[184] Nirmita Panchal et al. "The implications of COVID-19 for mental health and substance use". In: *Kaiser family foundation* 21 (2020).

[185] Pablo Paredes et al. "PopTherapy: Coping with stress through pop-culture". In: *Proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and Technology. 2014, pp. 109–117.

[186] Albert Park, Mike Conway, and Annie T Chen. "Examining thematic similarity, difference, and membership in three online mental health communities from Reddit: a text mining and visualization approach". In: *Computers in human behavior* 78 (2018), pp. 98–112.

[187] Scott B Patten, Ruth Ann Marrie, and Mauro G Carta. "Depression in multiple sclerosis". In: *International Review of Psychiatry* 29.5 (2017), pp. 463–472.

[188] Björn Paxling et al. "Therapist behaviours in internet-delivered cognitive behaviour therapy: analyses of e-mail correspondence in the treatment of generalized anxiety disorder". In: *Behavioural and cognitive psychotherapy* 41.3 (2013), pp. 280–289.

[189] Jonathan Pearson-Stuttard, Majid Ezzati, and Edward W Gregg. "Multimorbidity—a defining challenge for health systems". In: *The Lancet Public Health* 4.12 (2019), e599–e600.

[190] Jean Louis Pépin et al. "Wearable activity trackers for monitoring adherence to home confinement during the COVID-19 pandemic worldwide: data aggregation and analysis". In: *Journal of Medical Internet Research* 22.6 (2020), e19787.

[191] Ramesh P Perera-Delcourt and Gemma Sharkey. "Patient experience of supported computerized CBT in an inner-city IAPT service: a qualitative study". In: *the Cognitive Behaviour Therapist* 12 (2019).

[192] John P Pestian, Pawel Matykiewicz, and Jacqueline Grupp-Phelan. "Using natural language processing to classify suicide notes". In: *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing*. Association for Computational Linguistics. 2008, pp. 96–97.

[193] Yongjun Piao et al. "A new ensemble method with feature space partitioning for high-dimensional data classification". In: *Mathematical Problems in Engineering* 2015 (2015).

[194] Forough Poursabzi-Sangdeh et al. "Manipulating and measuring model interpretability". In: *arXiv preprint arXiv:1802.07810* (2018).

[195] William H Press and George B Rybicki. "Fast algorithm for spectral analysis of unevenly sampled data". In: *The Astrophysical Journal* 338 (1989), pp. 277–280.

[196] Moby Project. *Public-domain lexical resources; word lists, thesaurus, hyphenation, pronunciation*. 2014. URL: `https://github.com/Hyneman/moby-project` (visited on 09/18/2019).

[197] Yada Pruksachatkun, Sachin R Pendse, and Amit Sharma. "Moments of Change: Analyzing Peer-Based Cognitive Support in Online Mental Health Forums". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 64.

[198] Gauri Pulekar and Emmanuel Agu. "Autonomously sensing loneliness and its interactions with personality traits using smartphones". In: *2016 IEEE Healthcare Innovation Point-Of-Care Technologies Conference (HI-POCT)*. IEEE. 2016, pp. 134–137.

[199] PWP. *Psychological Wellbeing Practitioner. Overview of Role*. 2019. URL: `https://www.instituteforapprenticeships.org/apprenticeship-standards/psychological-wellbeing-practitioner/` (visited on 09/17/2019).

[200] Olga Pykhtina et al. “Magic land: the design and evaluation of an interactive tabletop supporting therapeutic play with children”. In: *Proceedings of the Designing Interactive Systems Conference*. ACM. 2012, pp. 136–145.

[201] I Qualtrics. “Qualtrics”. In: *Provo, UT, USA* (2013).

[202] Mashfiqui Rabbi et al. “Passive and in-situ assessment of mental and physical well-being using mobile sensors”. In: *Proceedings of the 13th international conference on Ubiquitous computing*. ACM. 2011, pp. 385–394.

[203] Shiquan Ren et al. “Nonparametric bootstrapping for hierarchical data”. In: *Journal of Applied Statistics* 37.9 (2010), pp. 1487–1498.

[204] Stefan Rennick-Egglestone et al. “Health Technologies’ In the Wild’: Experiences of Engagement with Computerised CBT”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, pp. 2124–2135.

[205] Robert Reynes, Colin Martindale, and Hartvig Dahl. “Lexical differences between working and resistance sessions in psychoanalysis”. In: *Journal of Clinical Psychology* 40.3 (1984), pp. 733–737.

[206] Derek Richards, Angel Enrique, and Jorge E. Palacios. *Internet-delivered Cognitive Behaviour Therapy*. Ed. by Sarah Parry. London: Sage, 2019.

[207] Derek Richards and Ladislav Timulak. “Client-identified helpful and hindering events in therapist-delivered vs. self-administered online cognitive-behavioural treatments for depression in college students”. In: *Counselling Psychology Quarterly* 25.3 (2012), pp. 251–262.

[208] Derek Richards et al. “A randomized controlled trial of an internet-delivered treatment: its potential as a low-intensity community intervention for adults with symptoms of depression”. In: *Behaviour research and therapy* 75 (2015), pp. 20–31.

[209] Derek Richards et al. “Internet-delivered cognitive behaviour therapy”. In: *Cognitive behavioral therapy and clinical applications* (2018), pp. 223–238.

[210] Jane M Rondina et al. “SCoRS—A method based on stability for feature selection and mapping in neuroimaging”. In: *IEEE transactions on medical imaging* 33.1 (2013), pp. 85–98.

[211] John Rooksby, Alistair Morrison, and Dave Murray-Rust. “Student Perspectives on Digital Phenotyping: The Acceptability of Using Smartphone Data to Assess Mental Health”. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM. 2019, p. 425.

[212] Stephanie Rude, Eva-Maria Gortner, and James Pennebaker. "Language use of depressed and depression-vulnerable college students". In: *Cognition & Emotion* 18.8 (2004), pp. 1121–1133.

[213] Daniel W Russell. "UCLA Loneliness Scale (Version 3): Reliability, validity, and factor structure". In: *Journal of personality assessment* 66.1 (1996), pp. 20–40.

[214] Michael Rutter. "The interplay of nature, nurture, and developmental influences: the challenge ahead for mental health". In: *Archives of General Psychiatry* 59.11 (2002), pp. 996–1000.

[215] Marco de Sá and Luis Carriço. "Fear therapy for children: a mobile approach". In: *Proceedings of the 4th ACM SIGCHI symposium on Engineering interactive computing systems*. 2012, pp. 237–246.

[216] Sohrab Saeb et al. "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study". In: *Journal of medical Internet research* 17.7 (2015).

[217] Sohrab Saeb et al. "The relationship between mobile phone location sensor data and depressive symptom severity". In: *PeerJ* 4 (2016), e2537.

[218] Koustuv Saha and Munmun De Choudhury. "Modeling stress with social media around incidents of gun violence on college campuses". In: *Proceedings of the ACM on Human-Computer Interaction* 1.CSCW (2017), p. 92.

[219] Pedro Sanches et al. "HCI and Affective Health: Taking Stock of a Decade of Studies and Charting Future Research Directions". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: ACM, 2019, 245:1–245:17. ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300475. URL: http://doi.acm.org/10.1145/3290605.3300475.

[220] Wendy Sanchez et al. "Inferring loneliness levels in older adults from smartphones". In: *Journal of Ambient Intelligence and Smart Environments* 7.1 (2015), pp. 85–98.

[221] VC Sánchez-Ortiz et al. "A randomized controlled trial of internet-based cognitive-behavioural therapy for bulimia nervosa or related disorders in a student population". In: *Psychological Medicine* 41.2 (2011), pp. 407–417.

[222] Akane Sano et al. "Recognizing academic performance, sleep quality, stress level, and mental health using personality traits, wearable sensors and mobile phones". In: *2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*. IEEE. 2015, pp. 1–6.

[223] Shekhar Saxena et al. "Resources for mental health: scarcity, inequity, and inefficiency". In: *The lancet* 370.9590 (2007), pp. 878–889.

[224] Sara E Schaefer et al. "Wearing, thinking, and moving: testing the feasibility of fitness tracking with urban youth". In: *American Journal of Health Education* 47.1 (2016), pp. 8–16.

[225] Stefan Scherer, Giota Stratou, and Louis-Philippe Morency. "Audiovisual behavior descriptors for depression assessment". In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM. 2013, pp. 135–140.

[226] Stefan Scherer et al. "Investigating voice quality as a speaker-independent indicator of depression and PTSD." In: *Interspeech*. 2013, pp. 847–851.

[227] Jessica Schroeder et al. "Pocket skills: A conversational mobile web app to support dialectical behavioral therapy". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 2018, pp. 1–15.

[228] Stephen M Schueller, Kathryn Noth Tomasino, and David C Mohr. "Integrating human support into behavioral intervention technologies: the efficiency model of support". In: *Clinical Psychology: Science and Practice* 24.1 (2017), pp. 27–45.

[229] Mohammed Senoussaoui et al. "Model fusion for multimodal depression classification and level detection". In: *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM. 2014, pp. 57–63.

[230] Christopher W Seymour et al. "Precision medicine for all? Challenges and opportunities for a precision medicine approach to critical illness". In: *Critical Care* 21.1 (2017), pp. 1–11.

[231] Layal Shammas et al. "Home-based system for physical activity monitoring in patients with multiple sclerosis (Pilot study)". In: *Biomedical engineering online* 13.1 (2014), pp. 1–15.

[232] Eva Sharma and Munmun De Choudhury. "Mental Health Support and its Relationship to Linguistic Accommodation in Online Communities". In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM. 2018, p. 641.

[233] Adrian BR Shatte, Delyse M Hutchinson, and Samantha J Teague. "Machine learning in mental health: a scoping review of methods and applications". In: *Psychological medicine* 49.9 (2019), pp. 1426–1448.

[234] Saul Shiffman, Arthur A Stone, and Michael R Hufford. "Ecological momentary assessment". In: *Annu. Rev. Clin. Psychol.* 4 (2008), pp. 1–32.

[235] Richard J Siegert and Darrell A Abernethy. "Depression in multiple sclerosis: a review". In: *Journal of Neurology, Neurosurgery & Psychiatry* 76.4 (2005), pp. 469–475.

[236] Amit Singhal, Chris Buckley, and Manclar Mitra. "Pivoted document length normalization". In: *ACM SIGIR Forum*. Vol. 51. 2. ACM. 2017, pp. 176–184.

[237] Karen L Smarr and Autumn L Keefer. "Measures of depression and depressive symptoms: Beck Depression Inventory-II (BDI-II), Center for Epidemiologic Studies Depression Scale (CES-D), Geriatric Depression Scale (GDS), Hospital Anxiety and Depression Scale (HADS), and Patient Health Questionnaire-9 (PHQ-9)". In: *Arthritis care & research* 63.S11 (2011).

[238] National MS Society. *Cost of Multiple Sclerosis*. `https://www.nationalmssociety.org/Living-Well-With-MS/Work-and-Home/Cost-of-Multiple-Sclerosis` [Accessed: 2023].

[239] National MS Society. *How Many People Live With MS?* `https://www.nationalmssociety.org/What-is-MS/Who-Gets-MS/How-Many-People` [Accessed: 2023].

[240] Claudio Solaro, Giulia Gamberini, and Fabio Giuseppe Masuccio. "Depression in multiple sclerosis: epidemiology, aetiology, diagnosis and treatment". In: *CNS drugs* 32.2 (2018), pp. 117–133.

[241] Michael Stigler and Dan Pokorny. "Emotions and primary process in guided imagery psychotherapy: Computerized text-analytic measures". In: *Psychotherapy Research* 11.4 (2001), pp. 415–431.

[242] William B Stiles, Lara Honos-Webb, and Michael Surko. "Responsiveness in psychotherapy". In: *Clinical psychology: Science and practice* 5.4 (1998), pp. 439–458.

[243] Giota Stratou et al. "Automatic nonverbal behavior indicators of depression and PTSD: Exploring gender differences". In: *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE. 2013, pp. 147–152.

[244] Lauren B Strober and Peter A Arnett. "An examination of four models predicting fatigue in multiple sclerosis". In: *Archives of Clinical Neuropsychology* 20.5 (2005), pp. 631–646.

[245] Charlotte M Stuart et al. "Physical activity monitoring to assess disability progression in multiple sclerosis". In: *Multiple Sclerosis Journal–Experimental, Translational and Clinical* 6.4 (2020), p. 2055217320975185.

[246] Shaoxiong Sun et al. “Using smartphones and wearable devices to monitor behavioral changes during COVID-19”. In: *Journal of Medical Internet Research* 22.9 (2020), e19992.

[247] Ashleigh Sushames et al. “Validity and reliability of Fitbit Flex for step count, moderate to vigorous physical activity and activity energy expenditure”. In: *PloS one* 11.9 (2016), e0161224.

[248] Tan-Hsu Tan et al. “Indoor activity monitoring system for elderly using RFID and Fitbit Flex wristband”. In: *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. IEEE. 2014, pp. 41–44.

[249] John D Teasdale. “The relationship between cognition and emotion: The mind-in-place in mood disorders.” In: (1997).

[250] Anja Thieme et al. “Challenges for designing new technology for health and wellbeing in a complex mental healthcare context”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, pp. 2136–2149.

[251] Murali Thyloth, Hemendra Singh, Vyjayanthi Subramanian, et al. “Increasing burden of mental illnesses across the globe: current status”. In: *Indian Journal of Social Psychiatry* 32.3 (2016), p. 254.

[252] Nickolai Titov. “Internet-delivered psychotherapy for depression in adults”. In: *Current opinion in psychiatry* 24.1 (2011), pp. 18–23.

[253] Nickolai Titov et al. “Transdiagnostic internet treatment for anxiety and depression: a randomised controlled trial”. In: *Behaviour research and therapy* 49.8 (2011), pp. 441–452.

[254] Laura Toloşi and Thomas Lengauer. “Classification with correlated features: unreliability of feature ranking and solutions”. In: *Bioinformatics* 27.14 (2011), pp. 1986–1994.

[255] Catherine Tong et al. “Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Connnected Wellness Devices”. In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.3 (2019), pp. 1–19.

[256] A Trajman and RR Luiz. “McNemar $\chi 2$ test revisited: comparing sensitivity and specificity of diagnostic examinations”. In: *Scandinavian journal of clinical and laboratory investigation* 68.1 (2008), pp. 77–80.

[257] Truyen Tran et al. “An integrated framework for suicide risk prediction”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 1410–1418.

[258] Jean M Twenge and Thomas E Joiner. "US Census Bureau-assessed prevalence of anxiety and depressive symptoms in 2019 and during the 2020 COVID-19 pandemic". In: *Depression and anxiety* 37.10 (2020), pp. 954–956.

[259] Daniel Vigo, Graham Thornicroft, and Rifat Atun. "Estimating the true global burden of mental illness". In: *The Lancet Psychiatry* 3.2 (2016), pp. 171–178.

[260] Andre C Vogel et al. "Impact of the COVID-19 pandemic on the health care of¿ 1,000 people living with multiple sclerosis: a cross-sectional study". In: *Multiple sclerosis and related disorders* 46 (2020), p. 102512.

[261] Birgit Wagner, Andrea B Horn, and Andreas Maercker. "Internet-based versus face-to-face cognitive-behavioral intervention for depression: a randomized controlled non-inferiority trial". In: *Journal of affective disorders* 152 (2014), pp. 113–121.

[262] Fabian Wahle et al. "Mobile sensing and support for people with depression: a pilot trial in the wild". In: *JMIR mHealth and uHealth* 4.3 (2016).

[263] Clare Walton et al. "Rising prevalence of multiple sclerosis worldwide: Insights from the Atlas of MS". In: *Multiple Sclerosis Journal* 26.14 (2020), pp. 1816–1821.

[264] Rui Wang et al. "Predicting Symptom Trajectories of Schizophrenia using Mobile Sensing". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1.3 (2017), p. 110.

[265] Rui Wang et al. "SmartGPA: how smartphones can assess and predict academic performance of college students". In: *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*. ACM. 2015, pp. 295–306.

[266] Rui Wang et al. "StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM. 2014, pp. 3–14.

[267] Rui Wang et al. "Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2.1 (2018), p. 43.

[268] Yilun Wang et al. "A novel approach for stable selection of informative redundant features from high dimensional fMRI data". In: *arXiv preprint arXiv:1506.08301* (2015).

[269] Harvey A Whiteford et al. "Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010". In: *The lancet* 382.9904 (2013), pp. 1575–1586.

[270] Paul Wicks, Timothy E Vaughan, and Michael P Massagli. "The multiple sclerosis rating scale, revised (MSRS-R): Development, refinement, and psychometric validation using an online community". In: *Health and quality of life outcomes* 10.1 (2012), pp. 1–12.

[271] Jesse H Wright et al. "Computer-assisted cognitive-behavior therapy for depression: a systematic review and meta-analysis". In: *The Journal of clinical psychiatry* 80.2 (2019), p. 3573.

[272] Po-Yen Wu et al. "–Omic and electronic health record big data analytics for precision medicine". In: *IEEE Transactions on Biomedical Engineering* 64.2 (2016), pp. 263–273.

[273] Xuhai Xu et al. "Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5.1 (2021), pp. 1–27.

[274] Xuhai Xu et al. "Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students". In: *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3.3 (2019), p. 116.

[275] Zhixian Yan, Jun Yang, and Emmanuel Munguia Tapia. "Smartphone bluetooth based social sensing". In: *Proceedings of the 2013 ACM conference on Pervasive and ubiquitous computing adjunct publication*. ACM. 2013, pp. 95–98.

[276] Amir Hossein Yazdavar et al. "Semi-supervised approach to monitoring clinical depressive symptoms in social media". In: *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*. ACM. 2017, pp. 1191–1198.

[277] Broche-Pérez Yunier et al. "Fear of COVID-19, problems accessing medical appointments, and subjective experience of disease progression, predict anxiety and depression reactions in patients with Multiple Sclerosis". In: *Multiple Sclerosis and Related Disorders* (2021), p. 103070.

[278] Farzaneh Zahedi and Mohammad-Reza Zare-Mirakabad. "Employing data mining to explore association rules in drug addicts". In: *Journal of AI and Data Mining* 2.2 (2014), pp. 135–139.

[279] Massimiliano de Zambotti et al. "A validation study of Fitbit Charge 2™ compared with polysomnography in adults". In: *Chronobiology international* 35.4 (2018), pp. 465–476.

[280] Aurora Zanghı et al. “Mental health status of relapsing-remitting multiple sclerosis Italian patients returning to work soon after the easing of lockdown during COVID-19 pandemic: A monocentric experience”. In: *Multiple Sclerosis and Related Disorders* 46 (2020), p. 102561.

[281] Jie Zhang et al. “A feature sampling strategy for analysis of high dimensional genomic data”. In: *IEEE/ACM transactions on computational biology and bioinformatics* 16.2 (2017), pp. 434–441.

[282] Yan Zhang et al. “Feelings of depression, pain and walking difficulties have the largest impact on the quality of life of people with multiple sclerosis, irrespective of clinical phenotype”. In: *Multiple Sclerosis Journal* 27.8 (2021), pp. 1262–1275.