Spring 3-29-2023

TOWARDS SCALABLE MENTAL HEALTH: LEVERAGING DIGITAL TOOLS IN COMBINATION WITH COMPUTATIONAL MODELING TO AID IN TREATMENT AND ASSESSMENT OF MAJOR DEPRESSIVE DISORDER

Matthew D. Nemesure
Dartmouth College, matthew.d.nemesure.gr@dartmouth.edu

TOWARDS SCALABLE MENTAL HEALTH: LEVERAGING DIGITAL TOOLS IN COMBINATION WITH COMPUTATIONAL MODELING TO AID IN TREATMENT AND ASSESSMENT OF MAJOR DEPRESSIVE DISORDER

A Thesis
Submitted to the Faculty
In partial fulfillment of the requirements for the
Degree of

Doctor of Philosophy
in
Quantitative Biomedical Sciences

by Matthew D. Nemesure

Guarini School of Graduate and Advanced Studies
Dartmouth College
Hanover, New Hampshire

March 2023

Examining Committee:

(chair) ***Nicholas Jacobson***

James O'Malley

Paul Barr

Varun Mishra

F. Jon Kull, Ph.D.
Dean of the Guarini School of Graduate and Advanced Studies

ABSTRACT

Major depressive disorder (MDD) is a debilitating disorder that impacts the lives of nearly 280 million individuals worldwide, representing 5% of the overall adult population. Unfortunately, these statistics have been both trending upward and are also likely an underestimate. This can be primarily attributed to lack of screening paired with a lack of providers. Worldwide, there are roughly 450 individuals living with MDD per mental health care provider. Adding to this burden, approximately half of affected individuals that do receive care of any kind will fail to remain in remission. The goal of this thesis work is to leverage statistical and machine learning models to help close these gaps in both MDD assessment and treatment. The data used in this thesis comes from a variety of sources including cross-sectional data from a physician wellness visit, randomized controlled trial (RCT) data from various digital interventions for MDD, and longitudinal data assessing individual's depressive symptoms over time from the Tracking Depression Study. Supervised machine learning methods were applied to the wellness visit data to predict MDD presence and the RCT data to predict treatment response. The implication of these approaches is that in practice, they could enable passive assessments of MDD followed by personalized treatment planning using scalable interventions. As an addition to these machine learning approaches, statistical models were used to analyze longitudinal MDD symptom data to further understand individual changes in symptom dynamics. This work lays the foundation for dynamic treatment allocation that adapts as an individual's experience changes.

PREFACE

I have been involved in research across a variety of domains since the very start of my undergraduate degree. Regardless of the field and scope of my work, however, I was always most interested in the unique analytical approach for each problem. At the beginning of my PhD, my goal was first and foremost to join a team that was at the cutting edge of data science application. I knew there was a strong need for this work in the mental health space and combining the two was exactly where I wanted to position myself for my dissertation work. Lucky for me, my advisor Dr. Nicholas Jacobson began building his team at Dartmouth at the same time that I was beginning my dissertation work. There are no words that truly capture how fortunate I feel to have worked with Dr. Jacobson over these past few years. He has afforded me the opportunity to work on a wide variety of projects spanning the applied data science in mental health space. Beyond that, and even more important, Dr. Jacobson has been nothing but supportive in my research, career, and personal development. No matter how busy he got, he always made time to talk about whatever topic I was interested in discussing.

Prior to joining QBS and coming to Dartmouth, I had the opportunity to work with a number of amazing people, each of whom helped shape my approach to research in one way or another. I would love to thank Dr. Jennie Williams at Stony Brook, Dr. Gerardo Mackenzie at UC Davis and Dr. Corinne Kiessling at King's College who each played a role in introducing me to the research process. Additionally, I want to thank Dr. Christine DeLorenzo at Stony Brook for introducing me to machine learning in the mental health space and Dr. Jeffrey Pu at Upstate Medical University who supported my interest in the analytical part of research despite my primary role being in the wet lab.

After joining QBS, I was introduced to so many people who would end up playing a significant role in positively impacting my time as a Dartmouth student. First and foremost, I would like to give special thanks to the QBS administration, Dr. Kristine Giffin, Dr. Susan Diesel, Rosemary White, Shaniqua Jones, Dr. Robert Frost, and Dr. Scott Gerber. Every member of the QBS administration has gone above and beyond (more than once) to make my experience truly special. I would also like to give special shout

outs to all of the faculty that I have personally worked with either through TAing, rotations, or research collaborations. It was a pleasure to work with each of you and even if not directly, every single individual has had an influence on this work in one way or another. Thank you to my rotation advisors Dr. Erika Moen and Dr. Aaron Mckenna, the professors who I have worked with as a TA: Dr. James O'Malley, Dr. Ramesh Yapalparvi, Dr. Jennifer Emond, and Dr. Diane Gilbert-Diamond, and all of my collaborators of which there are far too many to name.

This entire work would not be possible without my qualifying committee as well as my dissertation committee. Special thanks to Dr. Nicholas Jacobson, Dr. Paul Barr, Dr. James O'Malley, Dr. Varun Mishra and Dr. Soroush Vosoughi. Scheduling this many people all together can always be quite challenging and I appreciate that each time we met, every member of my committee made the most of that time by being prepared to give me well thought out feedback that made my work the best it could possibly be.

Finally, I would like to give a huge shout out to my support system of friends and family. Without them, this work would probably never have been started, let alone finished. First, I want to thank all of the friends I have made through being a part of the QBS program. Whether we were going out to eat, running a 5k or just getting together to chat, their constant support made my experience truly enjoyable. To close this preface, I would love to thank my family and all the support they've given me throughout my PhD. To my mom, dad, brother and sister, I would not have been able to do this without you. To my fiancé's family, I appreciate you so very much. To my fiancé, Julia, thank you for everything and I love you! And finally to my dog, Ollie, thanks for being the best coworker out there!

Table of Contents

CHAPTER 1: INTRODUCTION

1.1 Background on major depressive disorder

1.1.1 Description of major depressive disorder

Major depressive disorder (MDD) is characterized by a collection of psychologic and somatic symptoms that meet a specific threshold of severity.[1] These symptoms include: depressed mood, anhedonia, weight loss/gain, sleep abnormality, motor issues, fatigue, worthlessness, lack of concentration and suicidal ideation.[2] To be diagnosed with MDD, a person would need to endorse either depressed mood and/or anhedonia. Any combination of at least four of the remaining MDD symptom components that occur most of the day, nearly every day for two weeks, could potentially lead to an MDD diagnosis.[3] Given these broad diagnostic guidelines, it is clear that there is substantial room for heterogeneity under the umbrella diagnosis of MDD. In fact, within the aforementioned constraints, there are 227 possible permutations of the MDD symptoms that meet the diagnosis criteria.[4] This leads to substantial between-person heterogeneity in the experience of MDD.

1.1.2 Prevalence of major depressive disorder

MDD currently impacts nearly 300 million individuals worldwide which accounts for five percent of the global population.[5] Demographically, females, young adults and white or Native American adults experienced MDD at higher rates than their counterparts.[6] This differential prevalence is hypothesized to be due to genetic and hormonal differences, though this topic is complex and still requires further investigation.[7]

Longitudinally, the mean lifetime prevalence of MDD is around twelve percent.[2] For a given person, the average episode of MDD lasted for greater than six months with the majority of episodes falling in the moderate or severe category.[6] Adding to this already heavy burden, psychological comorbidities tend to be the rule as opposed to the exception. Among individuals with lifetime MDD, over 70% met criteria for at least one other psychiatric disorder.[8] These included many forms of anxiety disorders such as

generalized anxiety disorder (GAD), panic disorder, and post-traumatic stress disorder (PTSD) in addition to substance use disorders, psychotic disorders and suicidality.[9]

1.1.3 Impact of major depressive disorder in daily life

Given the complex, debilitating symptomatology of MDD combined with the high prevalence, it is not surprising that it is the leading cause of global disease burden.[10] From an economic perspective, just in the United States, direct and indirect costs related to MDD accounted for over 200 billion dollars.[11] For an individual suffering from MDD, it can impact nearly every aspect of daily life. It can have adverse social, professional and financial effects that not only support a negative feedback loop for MDD but also lead to the increased likelihood of comorbidities.[12] In turn, these direct and indirect effects have long-term impacts related to interpersonal functioning, emotional regulation, and even all-cause mortality.[13,14]

1.2 Current state of diagnosis in major depressive disorder

1.2.1 Tools for diagnosis in major depressive disorder

Considering the impact of MDD, both at the individual and societal levels, it is imperative that there are effective tools to diagnose individuals as the first step on the path to treatment. Fortunately, there are a number of self-report questionnaires that are validated to assess MDD presence and or severity. These include the Patient Health Questionnaire-9 (PHQ-9), the Center for Epidemiologic Studies Depression Scale (CES-D), the Beck Depression Inventory (BDI-II) and others.[2,15] These surveys are further discussed in Chapter 2.3 and 3.1.3. Despite being self-report surveys, these are often the first step towards receiving an MDD or other mental health diagnosis (based on DSM-5 criteria).[15,16] Typically, this gold-standard comes from a psychiatric interview assessing not only the MDD symptoms but other potential factors that may rule out an MDD diagnosis.[17]

1.2.2 Current practices in major depressive disorder diagnosis

Unfortunately, despite the abundance of self-report tools and their capability to assess MDD, there is still a large proportion of individuals suffering from this disorder without a diagnosis. A study in 2022 recruited a cross-sectional convenience sample of US adults via an anonymous online survey. This investigation found that over 30% of individuals without a clinical MDD diagnosis had indicated moderate to severe levels of depression via a PHQ-9.[18] Although the limited number of mental healthcare providers is a contributing factor to this underdiagnosis, it is only part of the issue. Primary care physicians are also essential in first recognizing MDD and either providing some level of treatment or a referral.[19] A large study in 2020 recognized this fact and set out to test the impact of systematic screening for MDD in primary care.[20] The results of this work were clear, additional screening resulted in increased rates of both diagnosis and subsequent care. A secondary outcome, however, was that even after physicians were *specifically* instructed to deliver a PHQ-9 to all patients, only 59% of individuals were actually screened.[20] The investigators indicate that this lack of screening could be due to a number of issues from administration to distribution. This percentage is likely a lot smaller outside of this network and represents a large gap in current methods of assessment.

1.2.3 Opportunity for digital methods to improve diagnosis

Supervised machine learning tools provide a unique and non-burdensome method for closing the assessment gap in primary care. While there is some likelihood that a primary care clinician may not have the opportunity to screen for MDD during a wellness visit, there is other information collected at each encounter that may be used to detect MDD. This includes demographic information, biometric information (heart rate, blood pressure, etc.), and some lifestyle information. Paired with a machine learning approach, there is the opportunity to leverage this data alongside a predictive model to passively assess MDD. This approach has been shown to be successful in other domains as well as in the mental health space.[21] A lot of this work, however, is limited by the ground truth being self-report.[21] Additionally, the timing of a predictive model in this space is essential. It needs to be implemented in real-time at the time of

information collection when a self-report survey wouldn't otherwise be given. The model's prediction can then be used to provide a special alert to the physician in cases where screening or further testing may be necessary. This type of work could vastly improve the rate of diagnosis in primary care, thus increasing rates of treatment and reducing overall burden for both the patient and the provider. For the patient, this could substantially reduce the amount of time lived with the disorder. Related to the physician, this model could act as a both a passive indicator as well as a confirmation tool reducing the overall visit burden.

1.3 Current state of treatment in major depressive disorder

1.3.1 Current best methods for treatment

A recent guide for clinicians treating MDD cited pharmacotherapeutics (selective serotonin reuptake inhibitors, tricyclic tertiary amines, atypical antidepressants etc.) and psychotherapy (cognitive behavioral therapy, interpersonal therapy, behavioral therapy, etc.) as the two main treatment options.[22] While these treatments show consistent efficacy at the group level (i.e. treatment vs. control), they don't work for all individuals. The reason for this lack of efficacy can vary by individual and is not yet fully understood.[23] Among all adults across 165 placebo controlled trials for various pharmacotherapeutics, only 54% experienced improvement in depressive symptoms.[24,25] In the psychotherapy setting, across 35 randomized controlled trials, only 62% of adults no longer met MDD criteria via diagnostic interview following treatment.[25] Given the substantial impact MDD can have on an individual's day-to-day life, these rates of efficacy for gold standard treatment leave room for improvement.

1.3.2 Problems with current treatment scalability.

Lacking efficacy, however, is not the only burdensome issue in the MDD treatment space. Both pharmacotherapeutics and traditional psychotherapy require the input of a clinician. This is problematic given that globally, there is only one mental health care provider of any kind (including psychologists, psychiatrists, social workers etc.) for every 450 individuals suffering from MDD.[26] Adding to this already

insurmountable burden, this ratio does not even include the individuals, representing nearly a quarter of the global population, who are seeking mental health care for something other than MDD.[27] This lack of scalability is yet another barrier to treatment for those who have received an MDD diagnosis and are seeking help.

1.3.3 Digital Interventions as a potential solution.

To address the scalability issue, there have been major strides taken in the psychological intervention space by leveraging technology to create treatments that can be delivered via a digital medium.[28] Importantly, depending on the intervention type, these treatments can either reduce or completely remove the provider burden. An individual could access intervention materials on the web via their smartphone, tablet or computer.[29–32] Not only is this model of care substantially more accessible than in person treatment, it is also comparable in terms of efficacy.[33,34] Across multiple review articles as well as meta-analyses, digital interventions have been shown to be a substantial improvement over waitlist controls and, in many cases, can be show similar group level efficacy to treatment as usual.[35–37] Compared to no-intervention controls, the pooled effect size for digital interventions was small (Cohen's d = 0.33) but improved with supervision to a medium effect size (Cohen's d = 0.52).[35]

1.3.4 Paring digital interventions and machine learning

While digital interventions provide a solution to the scalability problem in treating MDD, they still fail to address the gap in treatment efficacy. The majority of effort in dealing with this problem, until now, has been put towards developing new digital interventions that are tested via randomized controlled trials. Not a single one of these treatments, however, has worked for every person in the treatment group.[35–37] This can be largely attributed to the heterogeneity of MDD and is the reason that the National Institute for Mental Health (NIMH) has specifically called for the development of strategies to personalize treatment in this domain (NIMH strategy 3.2.A and 3.2.B).

Taking into account the diversity among individuals with MDD paired with the varied treatment response, this problem is uniquely poised for a supervised machine learning solution. In this context, machine learning models could learn the characteristics of individuals who respond well to different types of digital interventions. This type of approach has been shown to be effective in allocating traditional in-person treatment (pharmacotherapy vs. psychotherapy) and thus can likely extend to digital interventions.[38,39] In this way, individuals seeking treatment for MDD could simply answer questions about themselves and the model would be able to predict whether or not a given treatment would work for them. Using this approach specifically in digital interventions would solve both the scalability problem and the treatment efficacy problem by helping individuals narrow down an accessible intervention that would actually reduce their MDD symptoms and severity.

1.4 Shifting paradigms on major depressive disorder characterization

1.4.1 Within-person heterogeneity

As previously noted, MDD is typically defined and evaluated as experiencing a combination of symptoms most of the day for at least two weeks.[3] The development of this framing, however, occurred back in the 1980s in the DSM-III based on the assumptions [rather than longitudinal data] of psychiatrists and psychologists about symptom course and duration.[40,41] This conceptualization has persisted over the years and while across persons, symptom construction can be highly heterogenous, there is no opportunity for within-person variation over time. This leaves a substantial gap in the understanding and conceptualization of MDD, especially given that studies have shown that symptoms are often far from stable.[42,43] In fact, ecological momentary assessment (EMA) studies that collect data during daily life through repeated surveys have shown that symptoms can vary within weeks and even across hours within a day.[44–46]

1.4.2 Dynamic symptom tracking to aid in treatment

Considering the limitations of the current characterization of MDD along with evidence of dynamic symptom changes from EMA studies, the need for a novel MDD

conceptualization is apparent. Rather than viewing the experience of MDD as a static collection of unchanging symptoms over a two-week period, these components could instead be allowed to vary within this timeframe. Additionally, these symptoms would have the ability to impact each other thus creating a dynamic system that can evolve over time. Conceptualizing MDD as a system of symptoms that have the capability to rapidly change across and within days has strong implications on the interaction between intervention performance and symptomatology.[47] The broad lack of treatment efficacy may not just be attributed to interindividual differences but rather could also be affected by the timing of treatment within an individual. To begin to understand this likely complex and dynamic relationship, however, the first step is to actually model intraindividual MDD symptom dynamics.

This work in understanding the trajectory of MDD as a dynamic system of symptoms, in combination with the personalized treatment allocation models, could drastically change the intervention experience. Taken together, these two items could be used for just-in-time interventions that are able to deliver the right intervention at the right time based on an individual's experience in that moment.

1.5 Current Work

1.5.1 Conceptual Model

The conceptual model for this work follows the assessment to treatment trajectory of an individual with MDD. Before attempting to receive any treatment, an individual must first be assessed for MDD. This can happen via a mental healthcare provider but wait-times can often be on the scale of months.[48] Given this, these assessments often happen in the primary care space. Unfortunately, even with broad guidelines from the US Preventative Task Force, implementing broad screening proved to be a challenge with studies showing as low as 59% implementation after specific screening instruction.[20] There are instances with better screening rates like 88.8% in the California health system, but this improvement took multiple years to implement and required medical assistants and extensive workflow integration, something many areas

would be unable to accomplish based on limited time and resources.[49] Following assessment, the second part of the conceptual model related to treatment and recovery. The idea was to leverage heterogeneity between individuals, as it relates to treatment outcomes, to pair individuals with a [digital] treatment that works for them. In this way, not only would we address the ***access to care*** issue via digital interventions, but also the ***treatment efficacy*** issue by pairing individuals with something that is likely to work.

1.5.2 Background Summary

Taken together, this dissertation work explores the impacts of passive computational tools to aid in both the assessment and treatment of MDD. As stated above, this work essentially follows the patient trajectory from diagnosis to recovery. Chapter 2 discusses a passive machine learning tool capable of aiding in the assessment of MDD, a necessary first step towards receiving treatment. Chapter 3 discusses a number of projects related to predicting treatment outcomes in advance with the ultimate goal of helping an individual choose an accessible treatment that works. Finally, chapter 4 is a direct follow-up on chapter 3 that challenges the conceptualization of MDD from the lens of trying to improve treatment. If we can understand how the dynamics of the MDD experience change over time, treatments can be modified and changed alongside MDD symptoms to improve overall outcomes.

Chapter 2: Leveraging machine learning to aid in the diagnosis of major depressive disorder

This below work was published in a peer reviewed journal, *Scientific Reports* on January 21, 2021. Minor changes have been made such that this work fits the format of the dissertation. The citation for this article is as seen below.

Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. Sci Rep. 2021;11(1):1980. doi:10.1038/s41598-021-81368-4

2.1 Abstract

Generalized anxiety disorder (GAD) and major depressive disorder (MDD) are highly prevalent and impairing problems, but frequently go undetected, leading to substantial treatment delays. Electronic health records (EHRs) collect a great deal of biometric markers and patient characteristics that could foster the detection of GAD and MDD in primary care settings. We approach the problem of predicting MDD and GAD using a novel machine learning pipeline to re-analyze data from an observational study. The pipeline constitutes an ensemble of algorithmically distinct machine learning methods, including deep learning. A sample of 4,184 undergraduate students completed the study, undergoing a general health screening and completing a psychiatric assessment for MDD and GAD. After explicitly excluding all psychiatric information, 59 biomedical and demographic features from the general health survey in addition to a set of engineered features were used for model training. We assessed the model's performance on a held-out test set and found an AUC of 0.73 (sensitivity: 0.66, specificity: 0.7) and 0.67 (sensitivity: 0.55, specificity: 0.7) for GAD, and MDD, respectively. Additionally, we used advanced techniques (SHAP values) to illuminate which features had the greatest impact on prediction for each disease. The top

predictive features for MDD were “Satisfied with living conditions” and “public health insurance.” The top predictive features for GAD were “vaccinations being up to date” and “marijuana use”. Our results indicate moderate predictive performance for the application of machine learning methods in detection of GAD and MDD based on EHR data. By identifying biomarkers of GAD and MDD, these results may be used in future research to aid in the early detection of MDD and GAD.

2.2 Background

Major depressive disorder (MDD) and generalized anxiety disorder (GAD) are prevalent psychiatric disorders that affect 16.2% and 13.3% of U.S. individuals, respectively, over their lifetimes.[8,50] MDD is the leading cause of disability worldwide,[10,51] and anxiety disorders are the sixth leading cause of disability.[5] MDD is characterized by persistent low mood, associated with disturbances with sleep, motivation, energy, appetite, and suicidal thoughts.[52] GAD represents a persistent, uncontrollable pattern of worry occurring in multiple domains of an individual’s life.[53] Left untreated, these syndromes often have devastating consequences for affected individuals, their families, and communities.[54,55]

Both MDD and GAD are prevalent in the college population. In a 2015 study, 23% of surveyed college students reported moderate to severe depressive symptoms.[56] Similarly, a 2019 study showed a 20% prevalence of GAD among college students in 2016, representing a 100% increase since 2008.[57] These syndromes negatively impact multiple domains of an individual’s functioning, and for college students, this may include interference with class attendance and learning retention.[58] Research among college students found that students with depression are more likely to report drinking-related harms and alcohol abuse.[59]

Two major challenges in adequately addressing MDD and GAD are identifying affected individuals and ensuring appropriate and timely treatment. Because MDD and GAD symptoms are internally experienced, MDD and GAD often go undetected.[60–62] There is an estimated 6 year and 14 year delay between disease onset and intervention

for MDD and GAD, respectively, during which time the disease may increase in severity, lowering student quality of life.[63,64]

Early detection and diagnosis is paramount to understanding and addressing mental illness on a populational level. With the rise in electronic health records (EHRs), spurred by initiatives like the Health Information Technology Act (Rights (OCR), 2009), there is increasing potential for addressing previously intractable clinical questions using computational analysis of large data sets. Multiple studies show promise in this area.[65–69]

A 2011 study by Trinh et al.[65] found that an EHR billing diagnosis of "depression" can serve as an effective proxy for identifying clinical depression. Although this study did not exploit advanced statistical models, it demonstrated prediction of psychiatric pathology using structured EHR data, albeit the clinical utility of these predictive models is questionable given that the predictors used were closely related to outcome. Perlis et al.[66] found improvements in prediction of MDD using unstructured clinical narrative features (extracted with NLP) and billing code data, compared with using billing code data alone. A more recent 2019 study by Wang et al.[67] utilized machine learning techniques for prediction of postpartum depression (PPD). The predictors were extracted from the EHR and the model ended up with a good predictive accuracy. Features found to be significant included *depression*, *anxiety*, *use of antidepressant drugs*, and *pain diagnoses*. Geraci et al.[68] used data extracted from psychiatric clinical texts to predict a diagnosis of depression, including both structured or unstructured psychiatric diagnoses. Huang et al.[69] exploit multiple structured features to predict depression, including diagnostic codes and patient prescriptions, which could include psychiatric medications.

Although promising early directions, a common limitation in these studies[65–69] is the use of features highly interdependent with MDD, including psychiatric billing codes or unstructured notes, likely containing explicit diagnostic information. This presents as a major limitation to the potential utility of using these prior studies to close the onset

to treatment gap among those with MDD and GAD. In particular, diagnostic codes could only be obtained from those whose MDD and GAD would have already been detected. Based on the limitations of prior studies that utilized psychiatric features to predict GAD and MDD, our study utilized an EHR dataset containing biometric and demographic data from 4,184 undergraduate students. Excluding all psychiatric features, we approach the problem of identification and diagnosis using a novel machine learning pipeline developed for the purpose of this study. The pipeline constitutes an ensemble of multiple algorithmically distinct machine learning methods, including deep learning methods. We trained the model to predict psychiatric illness using varied non-psychiatric input features such as blood pressure, heart rate, housing status, and public insurance. This is to say, unlike all prior studies, we did not use any psychiatric information in predicting diagnosis of GAD or MDD. We hypothesized that using such biomedical data, we could predict MDD and GAD with a level of certainty above chance. Our primary aim was to identify biomarkers for GAD and MDD risk.

2.3 Methods

Participants

Four thousand one hundred and eighty four undergraduate students from the University of Nice Sophia-Antipolis underwent a basic medical examination and participated in the current study. All data was publicly available on Dryad and completely de-identified and therefore this research does not meet the federal definition for human subjects research. Additionally, according to the original study, the National Data Protection Authority (NCIL) approved the study.[70] The methods of the study carried out in France were in accordance with the laws of non-interventional clinical research.[70] Due to this being an observational study in compliance with laws that regulate non-interventional clinical research in France (articles L.1121-1 and R.1121-2 of the Public Health Code), informed consent was not required.[70] Additionally, this study received institutional exemption from the Committee for the Protection of Human Subjects at Dartmouth College. These students were 57.4% female and 42.6% male and their ages were split into four categories: less than 18, 18, 19 and 20 or older. The

distribution among these categories was as follows: 5%, 36%, 28% and 31%. The outcomes of interest, MDD and GAD, had base rates of 12% and 8% respectively. [70]

Features

A total of 59 features were used including binary, ordinal and continuous variables. Specifically, features included age (4 levels: under 18, 18, 19, over 20), gender, French nationality, field of study, year of university, learning disabilities, difficulty memorizing lessons, professional objective (whether the student indicated an objective), informed about opportunities (whether the student indicated that they felt informed about opportunities at the university), satisfied with living conditions, living with a partner/child, parental home, having only one parent, at least one parent unemployed, siblings (yes/no), long commute, mode of transportation, financial difficulties, grant (yes/no), additional income (yes/no), public health insurance, private health insurance, universal health coverage, irregular rhythm of meals, unbalanced meals, eating junk food, on a diet, irregular rhythm or unbalanced meals, physical activity (3 levels: none, occasional, regular) , physical activity (2 levels: none or occasional, regular), weight (kg), height (cm), overweight and obesity, systolic blood pressure (mmHg), diastolic blood pressure (mmHg), prehypertension or hypertension, heart rate (bpm), abnormal heart rate, distant visual acuity of right eye (score/10), distant visual acuity of left eye (score/10), close visual acuity of right eye (score/10), close visual acuity of left eye (score/10), decreased in distant visual acuity, decreased in close visual acuity, urinalysis (glycosuria), urinalysis (proteinuria), urinalysis (hematuria), urinalysis (leukocyturia), urinalysis (positive nitrite test), abnormal urinalysis, vaccination up to date, control examination needed (whether the student needed a follow-up for any reason), cigarette smoker (5 levels: none, occasional, regular, frequent, heavy), cigarette smoker (3 levels: no, frequent, occasional), drinker (3 levels: no, occasional, regular), drinker (2 levels: no or occasional, regular or heavy), binge drinking, marijuana use, other recreational drugs.

Psychiatric Diagnoses

The outcomes of interest were MDD and GAD. MDD and GAD were each assessed in a multi-stage process. The first stage included a screening questionnaire that assessed four depressive items (anhedonia, loss of energy/fatigue, changes in activity and depressed mood) and four anxiety items (excessive worry, restlessness, fatigue, and irritability). If the assessment indicated possible presence of either disorder (positive answer to two of the four categories), the participants were assessed for full Diagnostic and Statistical Manual of Mental Disorders Fourth Edition (DSM IV) criteria by a medical provider[70].

Data Preprocessing

The preprocessing pipeline included creating dummy variables for ordinal outcomes, normalizing continuous variables, and single imputation for missing values using a Bayesian Ridge approach across features. A total of 20 of the 59 variables included NA values and the percentage missing ranged from <1% to 36%. Total missingness was 5% and median missingness across all variables was 0%

To enhance our model, we used feature engineering, informed by domain specific biomedical knowledge. Feature engineering as used in our study refers to the combination of distinct features into new "engineered" features, which have domain specific meaning and utility. Previous research has shown feature engineering to improve machine learning model performance.[71,72] By combining existing features, we created and used (1) Body Mass Index (BMI),[73] (2) Mean Arterial Pressure (MAP),[74] and (3) Pulse Pressure.[75] BMI is a function of an individual's height and weight. MAP and pulse pressure are clinically meaningful combinations of diastolic and systolic blood pressure.

Data analysis

The first step of analysis was dividing the data into 70% training (*N*=2929) and 30% (*N*=1255) held out testing (see *Figure 1*). The held out test set remained unseen throughout model training and was never used for hyperparameter tuning. The machine learning pipeline included six algorithmically unique machine learning classifiers to inform final predictions. These classifiers were XGBoost, Random Forest, Support Vector

Machine, K-nearest-neighbors and a neural network tuned using Bayesian hyperparameter optimization. A 5-fold validation technique was used to train each model. This allowed for each model type (e.g., logistic regression) to make one prediction for each subject in the training set. These predictions were saved to be used as inputs to a "higher level" model that would eventually make final predictions.

The aforementioned "higher level" model was an XGBoost classifier which was trained, again, using 5-fold validation, on the predictions of the original 6 models. Essentially, each "lower level" model made a prediction (i.e. probability of MDD) for each subject and the higher level model decided which model's predictions were most informative based on the true outcome. Using this information, the higher-level model made a final estimation for the probability of the outcome of interest.

These models were then used to make predictions on the held out test set to ensure there was no overfitting and that the results were meaningful and generalizable. To create the prediction matrix on the held out test set, all 5 saved models for each machine learning method made predictions on each subject. The predictions for each model type were then averaged and filled into the prediction matrix. The high level XGboost model then made final predictions. The area under the receiver operating characteristic curve (AUC) is a measure of how well the model can effectively distinguish between psychiatric diagnosis, reflecting the model performance in optimizing across both sensitivity and specificity. To guide interpretation of the results, please note that an AUC = 0.58 represents a small effect size, AUC = 0.69 represents a medium effect size, and AUC = 0.79 represents a large effect size, based on conversions to Cohen's d values of 0.2, 0.5, and 0.8 respectively.[76] This pipeline was used twice, once with the outcome being GAD and once with the outcome being MDD.

Model Explainability

SHAP (Shapley Additive Explanations) scores were utilized calculate and visualize feature importance this complex model.[77] The SHAP kernel explainer allows for a user to input data and a prediction function and it will return the relative importance for each feature for each subject. The prediction function, in this case, simply took the input data

and utilized the trained models from the pipeline to make predictions. These predictions were then averaged across the lower level models and fed into the upper level model. The upper level model returned the final prediction for each subject. With this setup, the kernel explainer would return the SHAP values for each of the features from the original input data based on how it informed the entire pipeline's prediction.

2.4 Results

Predictive Performance

The main results of this study are two-fold, the first is the prediction accuracy of the stacked machine learning models and the second is the important features driving those predictions. The validation and test-set AUC for MDD (see Figure 2) and GAD (see *Figure 3*) were (0.70, 0.67) and (0.70, 0.73) respectively. Thus, the ensemble model could predict diagnosis of MDD and GAD well above chance and with a medium effect size. Additionally, when compared to a simple standard logistic regression as run in the original study, the AUCs of the complex machine learning models were increased, on average, by 0.08 (*figure 2B and 3B)*. Given the AUC curve of the model, we can choose thresholds with higher sensitivity at the detriment of specificity. Given the non-invasive nature of secondary screening for each of these illnesses, it seems reasonable to allow a soft threshold for further diagnosis. Specifically, for MDD, the sensitivity and specificity were 55% and 70% respectively. Additionally, the positive predictive value was 20% and the negative predictive value was 92%. For GAD, the sensitivity and specificity were 70% and 66% respectively. The positive predictive value was 16% and the negative predictive value was 96%.

Model Explainability

The second and arguably more important set of results are the important features and how they inform predictions (Figures 4 and 5). The top features (figure 4a and 5a) are the most informative to the model but it is important to note that the impact of features on the outcome was distributed across a large number of features (i.e. the SHAP values for top features were small). This is likely indicative of the complex and heterogenous nature of the disease. To ascertain either MDD or GAD status, it

requires a not just a singular biomarker but rather a combination of features and feature interactions to accurately assess the disease state. This exemplifies the necessity for complex models to disentangle the relationships between variables and characterize and assess the disease in any given person.

MDD (See Figure 4)*:* The most important feature driving the prediction of MDD was whether the student was satisfied with their living conditions (4b). High diastolic blood pressure was also indicative of MDD and having public health insurance indicated, for the most part, non-MDD status (4c). In order, living in a parental home, mean arterial pressure and difficulty memorizing lessons made up the remaining important predictors from the top six. Additionally, after further assessing these top features, it was noted that many of them were predictive as part of two-way interactions, such that the relationship between a predictor and an outcome is conditional on another predictor. As seen in *Figure 4d*, typically individuals without public health insurance had lower predictions of MDD, but the extent was conditional on whether they were satisfied with their living conditions. Those who were satisfied with their living conditions seemed to be slightly more informative in telling the model that MDD was not apparent.

GAD (See Figure 5)*:* The most important predictor for GAD was having up to date vaccinations (4b). Another similar and important variable for prediction was the necessity for a control examination. This was essentially a binary indicator for whether or not the student needed to return to the doctor for something unrelated to the psychiatric outcome. The second most important predictor was marijuana use although the effect of this variable on model prediction was clearly impacted by interactions with other subject characteristics (4c). The remaining top six most important predictors were, in order, hypertension or prehypertension, systolic blood pressure and the use of other recreational drugs. These features, overall, were all much closer in importance than in MDD. This further indicated the model's reliance on all features, not just one true biomarker. Again, there were very clear two-way interactions between variables when the model was making predictions. Smoking marijuana was clearly more indicative of

predicted GAD if the individual was overweight or obese (4d). Other interactions included systolic blood pressure with prehypertension and hypertension and the necessity of a control examination with gender.

2.5 Discussion

Our objective was to evaluate the importance and effectiveness of standard clinical data on the prediction of MDD or GAD. We used state-of-the-art novel machine learning methodologies to make predictions. Additionally, SHAP values were generated to explain and clinically validate our findings. We trained our model with >2500 participants and assessed the model's performance on a held-out test set. Although our accuracy metrics are comparable to previous studies predicting psychiatric outcomes, ours is unique in its primary reliance on routine biomedical and demographic features, rather than features with a known correlation to psychiatric outcomes. Previous studies that have looked at EHR to detect MDD have had the significant limitation of including predictive variables that would nullify the clinical utility of the model by relying on features that are directly indicative of known psychiatric illness (e.g. including psychiatric billing codes, which are based upon clinician diagnosis). Thus, this study is the first known study to predict MDD and GAD using EHR data with potential for predictive validity in detecting unknown psychiatric diagnoses.

Studies using magnetic resonance imaging (MRI) have been able to achieve slightly higher predictive performances ranging from 67% to 94%.[78] Nevertheless, perhaps due to the considerable expense of collecting MRI data, a common limitation of these was their small sample sizes. These studies also had considerable range in performance, and the due to their small sample sizes the results are highly inconsistent.[79] Moreover, using MRI to predict MDD is unrealistic when there is no other reason to justify an MRI, especially in an otherwise physically healthy college-age patient.

In addition to the complex machine learning approach and our carefully curated feature set, we are providing insights to the complex clinical appearance of MDD. Our

pipeline, using SHAP values to visualize feature importance, provides not only the outcome prediction but the possible characteristics that a physician can identify when making a decision. These characteristics including mean arterial pressure, blood pressure, markers for low SES and general health markers have been shown to be previously associated with depression and anxiety.[80,81]

In further investigation of the predictors for generalized anxiety disorder, vaccination status may be reflective of overall poorer health outcomes in individuals with GAD [82]. Regarding the "marijuana use", prior research demonstrates high comorbidity between anxiety disorders and substance use disorders. [83] With regard to the most important features driving major depressive disorder, there is research supporting overall poorer life satisfaction in individuals with MDD, [84]which may certainly include dissatisfaction with living conditions. Low interest and energy, DSM criteria for MDD, may contribute to difficulties maintaining satisfactory living conditions. Though health insurance is mandatory in France, use of public health insurance may be an indicator of lower socioeconomic status. Robust research to date indicates that individuals of lower socioeconomic status are more likely to have MDD. [85] "Difficulty memorizing lessons" may be related to concentration difficulties, also identified by the DSM as a clinical feature of MDD. An additional top predictive feature for both MDD and GAD is hypertension. Research to date corroborates this finding by demonstrating that individuals with either MDD or GAD are more likely to have hypertension. [86,87]

This information has the potential to allow health care providers to make informed recommendations for further screening regardless of whether the patient discusses or even recognizes his or her symptoms. This is important because as previously mentioned, it can take on average 6 or 14 years from onset of illness until diagnosis for MDD and GAD respectively.[63] Our study is one of the first of its kind to tackle this issue by not relying on previous psychiatric diagnoses or expensive imagine techniques to capture the disease in an early stage.

This study has several important limitations which deserve mention. One is that the original screening for the outcomes of MDD and GAD may not have captured all

cases within the population. This, in addition to the study population, limits the generalizability of the results. Our dataset comes from French college aged students, who likely have baseline differences from other populations with psychiatric illness. Despite this limitation, our study still serves to show the predictive ability of mainly non-psychiatric variables for psychiatric illness. Such variables, further analyzed individually for their connection to psychiatric pathology, may prove the basis of further research. Another limitation of our study, which is fairly ubiquitous in mental health research is the low prevalence of anxiety and depression in our study population, as well as our sample size. Although this is a limitation in many studies of psychiatric nature, we were able to enhance our predictive power using a stacked ensemble model pipeline. Additionally, the lack of qualitative information (i.e., severity, subtype, etc.) regarding mental health diagnoses was not available to allow for a severity prediction analysis. Thus, future research should examine the potential for these biomarkers to predict severity and subtype of MDD and GAD.

This research is an important step in the direction towards identifying potentially difficult to diagnose illnesses with readily available and easy to obtain information. Our tool, using an optimal sensitivity/specificity split would be able to capture two out of every three subjects with GAD and one out of two MDD cases while only incurring a 30% false positive rate. Because there are detrimental outcomes to both the patient and provider in a false positive, looking at the efficacy of case identification while requiring 70% specificity gives a reasonable idea of how many cases would be captured if this model were to be deployed in a clinical setting. These findings have shown promise on multiple fronts: Ability to use easy to obtain information to inform possible detection of MDD and GAD, further understanding of the demographic and biological characteristics associated with illness, and both the success and necessity for computational tools to inform psychological medicine. We believe, given a larger and more heterogeneous sample, this modeling technique could be used to elucidate the drivers of psychological illness and provide a tool that indicates the necessity of treatment with high precision and accuracy.

2.6 Figures

Figure 1

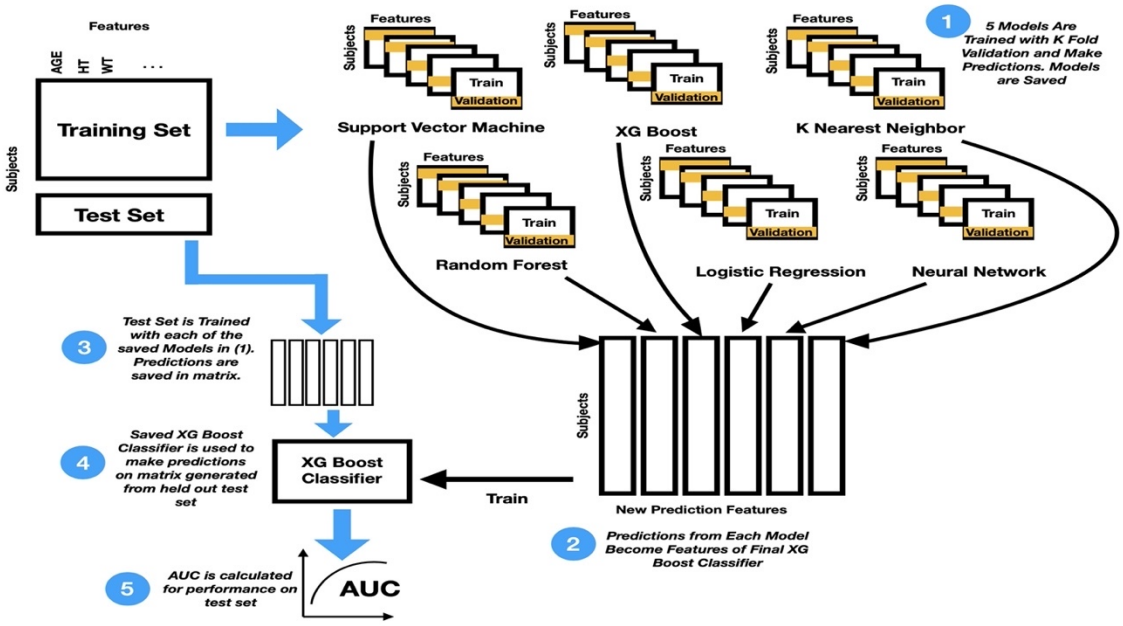


Figure 1. This is the pipeline used to train the machine learning models and generate predictions. The training set is sent through 5-fold training for each model type to generate a prediction for each training sample. These predictions are then used to train a higher level model to predict a final outcome given the predictions from the 5-fold training. Each of the 6 models from each fold then predicts on the held out test set and the average prediction for the probability of depression is stored. The higher level model then makes final predictions on the held out test set.

Figure 2

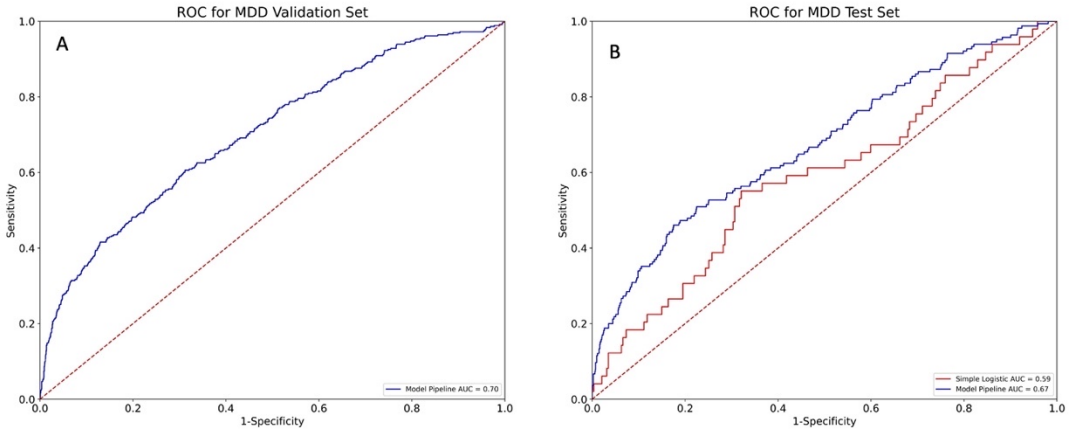


Figure 2. A: AUC for prediction of MDD in the training set. B: AUC for the prediction of depression in the held-out test set using both a simple logistic regression and our novel pipeline. These curves show the sensitivity and specificity at different thresholds for prediction.

Figure 3

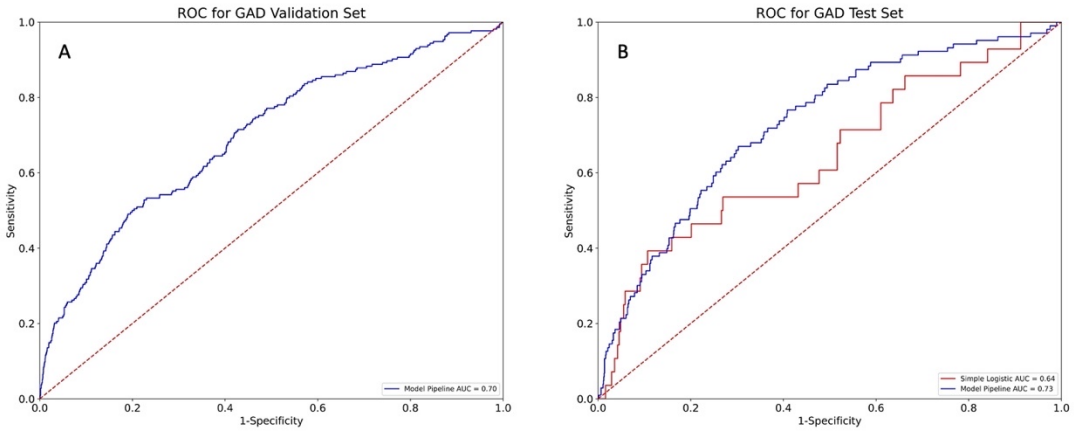


Figure 3. A: AUC for prediction of GAD in the training set. B: AUC for the prediction of anxiety in the held-out test set using both a simple logistic regression and our novel pipeline for prediction. These curves show the sensitivity and specificity at different thresholds for prediction.

Figure 4

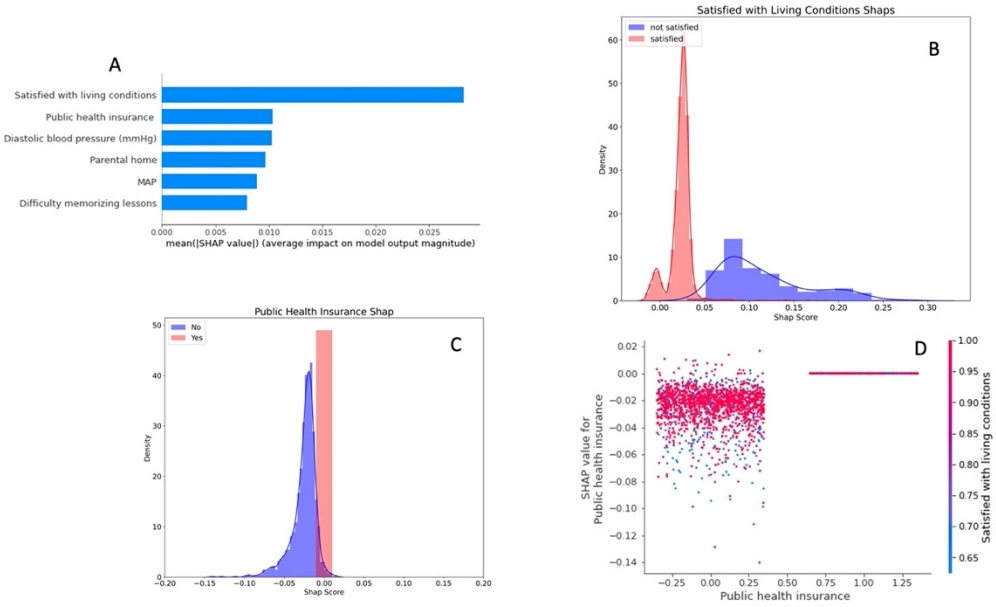


Figure 4. A: This plot shows the top six most important features for predicting MDD. This is displayed as the mean of the absolute value of SHAP scores across all subjects for that given feature. A higher SHAP value indicates that the feature was important in informing the models prediction. B: This plot displays the density distribution of SHAP values for the top performing feature in predicting depression. C: This plot also displays the density distribution of SHAP values for the second most important feature in predicting MDD. Positive SHAP score indicates that the feature was indicative of the subject having MDD. D: This is an interaction plot showing the effect of two features working together to inform the model. Here it is apparent that when a student does not have public health insurance, living conditions can partially inform prediction.

Figure 5

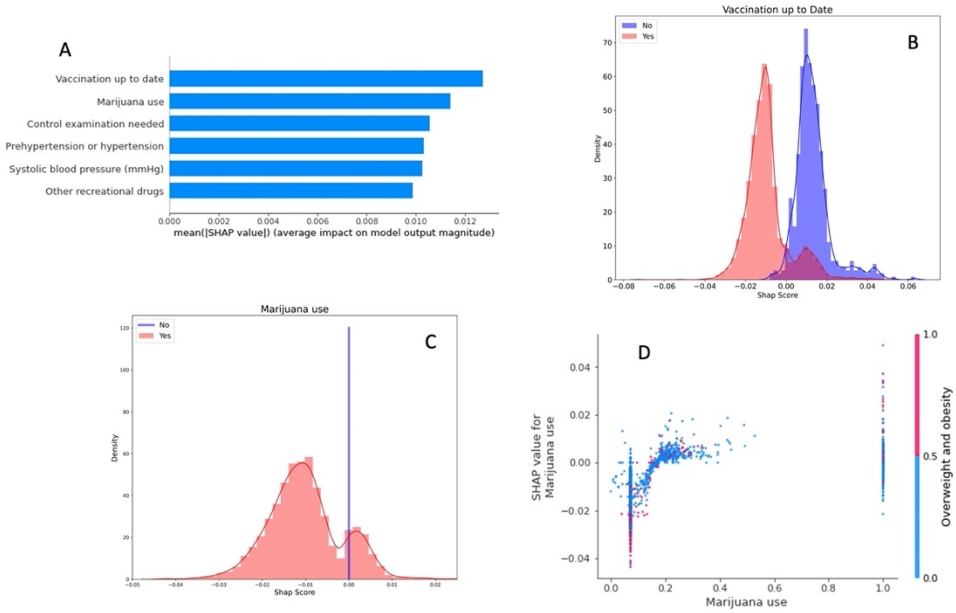


Figure 5. A: This plot shows the top six most important features for predicting GAD. This is displayed as the mean of the absolute value of SHAP scores across all subjects for that given feature. A higher SHAP value indicates that the feature was important in informing the models prediction. B: This plot displays the density distribution of SHAP values for the top performing feature in predicting GAD. C: This plot also displays the density distribution of SHAP values for the second most important feature in predicting GAD. Positive SHAP score indicates that the feature was indicative of the subject having GAD. D: This is an interaction plot showing the effect of two features working together to inform the model. Here it is apparent that marijuana use is more predictive of GAD in overweight students.

2.7 Acknowledgements

This work was funded by National Institute of Drug Abuse (NIDA P30DA02992610).

2.8 Author Contributions

MN – Had the original idea for the project, designed the machine learning pipeline, had a role in writing all sections of the manuscript.

MH – Helped with model hyperparameter tuning, engineering features and assisting in drafting multiple sections of the manuscript.

RH – Helped with the literature searching and did the majority of the writing for the introduction and some of the writing for the discussion.

NJ – Oversaw the entirety of the project, helped with design and modeling decisions and helped to write and edit all sections of the manuscript.

2.9 Financial Disclosures

None of the authors have any financial disclosures or conflicts of interest.

Chapter 3: Predicting treatment outcomes for digital intervention RCTs in major depressive disorder

3.1 Digital lifestyle intervention RCT focusing on physical activity

This work is currently under review at a peer-reviewed journal. The author list is shown below.

Nemesure, M.D., Heinz, M, McFadden, J., Huang, R., Klein, R., Jacobson, N.C. (2021) Predictive Modeling Approach to Evaluate Individual Response to a Physical Activity Digital Intervention for Subjects with Major Depressive Disorder.

3.1.1 Abstract

Worldwide, there is roughly one mental health care provider for every 400 people with major depressive disorder (MDD). Without including other disorders, it would be impossible for everyone suffering from MDD to get clinical assistance. One step towards closing this gap may be the development of digital interventions. These can be delivered via smartphone, personal computer or tablet and require a significantly decreased time commitment from a provider. Given these benefits, there has been an increasing number of new digital interventions being studied with varying results. This presents a need for evidence-based processes that select the right treatment for a given person. One digital intervention that has been widely studied is a physical activity intervention where subjects are encouraged, via the internet, to become more active as a method of reducing depressive symptoms. The goal of the present study was to evaluate whether baseline characteristics could be leveraged to determine whether individuals would be likely to respond to this form of digital intervention. Machine learning models were trained to predict all individuals' changes in Beck Depression Inventory-II (BDI-II) score and whether or not an individual had clinically significant change in depression. The correlation between predicted values and true values for

change in BDI-II was r = 0.399 and the AUC for predicting clinically significant change was 0.75. Important predictors included marital status, gender, and pre-intervention anxiety and depression severity. These models may facilitate precision medicine in the digital era by enabling personalized treatment planning of digital interventions.

3.1.2 Background

Major depressive disorder (MDD) is a serious psychiatric illness that affects an estimated 20.6% of US individuals over their lifetimes.[55] It is the leading cause of disability worldwide, with significant costs for both individuals and society.[52] MDD is characterized by persistent low mood, associated with disturbances in sleep, motivation, energy, appetite, and suicidal thoughts.[54] Left untreated, these symptoms can often have severe consequences for affected individuals, their families and communities.[55,88]

In a representative sample from 21 countries via the World Health Organization World Mental Health Surveys, of all individuals with MDD, only 16.5% received appropriate treatment.[89] There are numerous barriers that may limit access to treatment, these include cost to the patient in time and money as well as the fear of social stigma.[88,90] In an effort to address these barriers, researchers have increasingly turned to low-cost, scalable, digital interventions.[91] Digital interventions typically involve delivering either a targeted treatment such as cognitive behavioral therapy or a lifestyle intervention such as diet changes via a technological medium (i.e. smartphone, laptop, tablet etc.). Research to date has shown promising results using these types of interventions.[92,93] One specific form of a digitally-based mental health intervention has been physical activity interventions.[94]

These exercise interventions have continuously been found to decrease MDD symptoms. [94] For instance, a meta-analysis of 14 studies found that exercise intervention had a pooled effect size d = -1.1 (a significant negative effect size) and contributed to a mean difference in Beck Depression Inventory score of -7.3 between the treatment and control group.[95] Importantly, this literature indicates that exercise interventions work at the *sample* level. Unfortunately, like all existing interventions for

MDD, exercise interventions are likely to have varying degrees of efficacy across persons.[96]

One way to address this variation in individual responses to treatment is to leverage the pool of available interventions and match the individual patients to a specific intervention that has the highest likelihood of success for that person. To accomplish this matching, individual differences such as depressive symptom severity and demographic factors (e.g., age, gender) could potentially be used to train machine learning models that can predict which treatments will be the most effective for an individual. In theory, researchers could begin this process by building a digital profile for participants in need of treatment that could subsequently be used to predict which intervention might work best for a given individual.[97,98] Such an approach to personalized medicine has been shown to be clinically useful, but thus far has been limited to standard in-person treatments.[38]

The Present Investigation

The present research applies machine learning models to leverage data from a prior randomized controlled trial (RCT) targeting MDD.[94] The goal of the study was to re-analyze the previously collected data and evaluate patient-level differences in baseline characteristics to identify target subjects who would be best served by the specific digital intervention. The idea is that this work can act as both a foundation as well as a proof of concept for machine learning treatment recommendations in the digital psychological intervention space. The machine learning approach is well suited for this task as it is able to capture non-linear relationships and high order interactions between the participant's characteristics and the outcome of interest.[99–101] This approach to personalized medicine can be expanded to include an endless amount of other treatments. For our purposes, we look to predict the efficacy of a digital exercise intervention for patients with MDD; however, further development and expansions of this pipeline will choose from a variety of treatments, tailored to the specific needs of each individual, delivering superior patient-centered care.

3.1.3 Methods

Participants

Forty-eight participants living in Sweden were recruited by Ström and colleagues[10] via online and newspaper advertisements. All subjects were mildly to moderately depressed, lived a sedentary lifestyle and were physically capable of becoming more active. People with comorbid mental illness as well as somatic issues preventing activity were excluded from the study. During the recruitment process, MDD was assessed via screening questionnaires as well as a full diagnostic interview based on DSM-IV. Demographics collected included gender (40 female, 8 male), age (24-67 years old), marital status (22 married, 26 unmarried), education level, medication status (7 medication, 41 no medication) and psychotherapy status (only prior use, no current use).[94]

Three outcome measurements were recorded for each subject both before and after intervention: depression and anxiety status, physical activity, and quality of life. For the purposes of this study, the main outcome of interest was the Beck Depression Inventory (BDI-II), however, all scales measured prior to randomization were utilized for prediction.

Depression and anxiety status were evaluated using (i) the Montgomery-Asberg Depression Rating Scale (MADRS-S), (ii) the Beck Anxiety Inventory (BAI), and (iii) the Beck Depression Inventory (BDI-II). The MADRS-S is a robust diagnostic questionnaire used to gauge the severity of depressive episodes in patients with major depressive disorder; the scale consists of ten items (e.g., mood, appetite, ability to concentrate), and higher scores on each item indicate greater levels of impairment.[102] Previous studies have confirmed the clinical validity of the MADRS-S and examined its psychometric properties.[103] Specifically, this research found that the scale showed "good to excellent" internal consistency (Cronbach's alpha = 0.85; 0.94), strong concurrent validity (r = 0.81; 0.91), high sensitivity to change, and reasonable generalizability among patients with moderate to severe MDD.[103]

The BAI consists of twenty-one items used to analyze anxiety levels in adults and adolescents; the items describe common symptoms of anxiety (e.g., numbness or

tingling, feeling hot, wobbliness in legs), and the patient reports how often he/she felt bothered by that symptom during the past month.[104,105] Researchers attribute the high clinical efficacy of the BAI to its strong discriminative power and convergent validity when compared to similar anxiety questionnaires.[106]

The BDI is a twenty-one-item questionnaire used to measure the severity of depressive episodes in adolescents and adults.[107] Patients self-report the severity of their depressive symptoms by choosing a statement that most closely aligns with their true symptomatology (e.g., 0 for "I do not feel sad" versus 3 for "I am so sad an unhappy that I can't stand it"). The BDI has been modified numerous times, but the most updated version shows high construct validity and high clinical efficacy within depressive populations.[108]

The second outcome measure recorded for patients was physical activity via the International Physical Activity Questionnaire (IPAQ). This is a twenty-seven-item self-reported measure of physical activity. Questions in the IPAQ assess how often patients engage in physical activity (e.g., "How much time do you spend walking in your leisure time?"). Overall, the IPAQ has acceptable reliability and validity.[109]

The final outcome measure recorded for patients was quality of life. For this study the researchers used the Quality of Life Inventory (QOLI), which assesses positive mental health and happiness among sixteen distinct categories (e.g., love, work, play) that comprise one's overall quality of life. Previous studies have found significant positive correlations between QOLI scores and related measures of subjective well-being and significant negative correlations between QOLI scores and general psychopathology [110], suggesting a promising future for the QOLI as a robust measure of quality of life.

Intervention

Ström et al. (2013) conducted a randomized controlled trial that took place over a nine-week period. Throughout the study, subjects in the treatment group received the intervention at weekly intervals. This internet-delivered treatment consisted of nine sections from a 72-page self-help manual that highlighted the importance of physical

activity for both mental and physical health. At the beginning of each week, subjects digitally received one of the sections and were required to respond to computer administered essay questions designed to solidify their understanding of the material. The original hypothesis was that participation in both reading this manual and responding to essay questions would motivate subjects to exercise by promoting its importance.[94]

Data Preprocessing

The present analysis began by pre-processing the original intervention data for secondary analysis. First, we removed all individuals in the control group, given that the primary aim of the current re-analysis was to predict individual differences in how subjects responded to treatment. Second, treatment outcomes were defined using both continuous and binary outcome variables. The continuous outcome was a pre-post difference score calculated using the BDI-II self-report scale. By calculating the difference as post subtracted from pre, a positive value indicated an improvement in depressive symptoms.

A binary outcome was also calculated to differentiate participants who reported clinically significant BDI-II change following treatment from those who did not. Clinically significant change was defined by the original study[94] as having met three criteria: a) a decrease of at least five points, representing one standard error of the mean (SEM), on the BDI-II scale following treatment, b) having a BDI-II score prior to treatment of greater than 14, and c) reporting a BDI-II score (post treatment) of less than fourteen. In this way, all subjects classified as reporting clinically significant change necessarily started at relatively high levels of depression and ended at substantially lower levels with at least moderate improvement (1 SEM). All variables that were measured following treatment allocation were removed for predictive analysis.[94]

Data Analysis

Our goal was to use machine learning to evaluate whether key individual differences measured prior to treatment would predict treatment efficacy. Thus, we first trained several machine-learning models to evaluate the extent to which gender,

age, education, marital status, depression severity, anxiety severity, activity level, quality of life index, prior psychotherapy, and prior medication could be used to predict treatment efficacy for each participant in the experimental group. All analyses, including both the binary and continuous outcome prediction model, were conducted using a nested leave-one-subject-out cross-validation framework which has been shown to give unbiased performance estimates.[111] In this way, all final treatment outcome predictions were generated for subjects that were not involved in training or tuning the models. This is a major advantage over more traditional statistical procedures, which estimate key coefficients (e.g., regression slopes) using all available data and then evaluate the fit of these coefficients using the same data that were used to "train" them.[112] In predicting the continuous BDI-II change scores, the final model employed a combination of a regularized linear model (Lasso), a random forest model, and a boosted trees model (XGBoost).[99,100] For each fold, the model predicted change in depression following treatment on the validation subject, as well as on the fully held out subject. The collection of validation predictions and their associated accuracy was used to tune the model hyperparameters. The test set predictions were not evaluated until the tuning of the hyperparameters was complete.

In addition to the three base predictive models, a second model was trained exclusively on the predictions of the base models to determine a consensus prediction. The resulting output of this model was the final prediction. Once all hyperparameters were tuned, a correlation coefficient was calculated to quantify the relationship between the true change in depression scores of the completely held-out subjects and the predicted change in depression scores for these same held-out subjects. This correlation summarized the accuracy of the model in predicting actual change in BDI-II.

The model predicting binary clinically significant change scores was trained similarly to the aforementioned continuous model. For this model, the validation framework indicated that the best combination of base models was as follows: gaussian naïve bayes, a boosted decision tree classifier and logistic regression. The process for training these classifiers followed the same routine as continuous framework. They

were trained in a nested cross validation framework and validation predictions from the consensus model were used to tune hyperparameters. When that process was completed the held-out predictions were used to calculate and plot area under the receiver operating curve (AUROC) to measure model accuracy.

Model Explainability

The prediction pipelines used in both the continuous and binary predictions were complex combinations of multiple models, which can make it difficult to determine the most important features informing model prediction. To solve this problem, we rely on Shapley (Shap) scores to indicate which features (i.e., predictors) have the most significant impact on the entire pipeline's prediction of a given outcome, as well as the direction of effect.[113] Shap scores rely on a game theory approach where model training inputs are varied slightly, and feature importance is calculated based on how those fluctuations effect the model's predictions. By using this tool, we can determine feature importance for the pipeline as a whole by wrapping Shap around the entire prediction framework. To implement Shap, the hyperparameters were first optimized and the test set predictions were made. Next, the model was retrained on all of the data with the same tuning parameters. These parameters were then passed through into Shap's kernel explainer to produce feature importance plots in both the binary and continuous prediction framework. Using these plots, one can visualize the magnitude and directionality of each subject's impact on each feature's prediction of the model outcome.[113]

3.1.4 Results

Predicting Change in BDI-II Score

In the held-out test-set, the correlation between the predicted and the true change in BDI-II scores was r = 0.398 (see figure 1 and 2). The correlation within standard cross validation was r = 0.451. The relatively small difference between validation and test set correlation indicates that it is unlikely the model is overfit. More importantly, a correlation of nearly r = 0.4 demonstrates moderate predictive accuracy

and suggests that substantial variance in treatment efficacy was explained using our predictors.

The Shap scores indicate that marital status was the most important feature (i.e., variable) in predicting treatment efficacy, followed by pre-treatment BDI-II scores and pre-treatment BAI scores. More specifically, participants who were married reported more favorable treatment responses, as did those reporting lower baseline depression and higher baseline anxiety (see Figure 3). A positive Shap value indicates an individual whose data caused the model to predict a *higher* score on the outcome variable for a given feature value. Thus, a blue dot in the positive Shap bifurcation indicates that the lower feature value (i.e. lower baseline depressive symptoms) indicated to the model that the participant would have a more favorable outcome (i.e. had a decrease in depressive symptoms following treatment).

In addition to evaluating our model in the treatment group, we looked at the hypothetical case in which the control group subjects had received the treatment. This was achieved by allowing the model to make predictions on each control subject and displaying the projected change in BDI-II score if they would have received the intervention. Figure 3 shows that a substantial number of control subjects would have done better on the treatment but not all control subjects would have had symptom improvement and some would have actually done worse (Figure 4).

Predicting Clinically Significant Change

The AUC for the held-out test set in predicting clinically significant change in BDI-II depression scores was 0.75. The AUC within standard cross validation was 0.73. Similar to the continuous model, these parameters indicate moderate predictive accuracy of the present model. The individual models had an average AUC of 0.66 indicating that the consensus model is improving predictive accuracy.

In the binary prediction models, the most important feature was pre-intervention BDI score followed by pre-intervention MADRS score and gender. Typically, those who started less severe on the depression scales were more likely to have clinically significant change and females were also more likely to have better outcomes.

Overall, there were less variables that made substantial contributions to the binary prediction model than in the continuous prediction model.

3.1.5 Discussion

The symptoms exhibited by patients with MDD vary considerably, but the DSM-5's criteria for an MDD diagnosis include broad descriptions such as *depressed mood* and *fatigue or loss of energy* .[114] As a result of these broad categorizations all falling under one diagnosis, patients who display different combinations of symptoms with varying severity will ultimately be given the same diagnosis. This phenomenon of *diagnosis homogeneity* despite *symptom heterogeneity* underscores the need for a more robust modeling methodology when predicting which MDD treatments will be most effective for a particular patient.[115] This idea holds even in large meta-analyses of MDD treatment. Researchers acknowledge the heterogeneous nature of depression--and, more importantly, that response to a treatment varies according to individual factors.[116]

Unfortunately, despite these findings, the typical evaluation of treatment continues to happen at the *sample* level. This method of group analysis (i.e. treatment group versus control group) is effective for assessing the efficacy of novel treatments as well as for demonstrating whether these interventions will show similar levels of efficacy at the population level. But as stated earlier, MDD is a heterogeneous mental disorder that cannot be treated using a one-size-fits-all approach.

Our study presents a novel, machine learning-based approach to predicting how patients will respond to MDD treatment, with the eventual goal of matching specific patients to their best-suited treatments. Using an ensemble learner that can assess higher-order interactions, in this sample we can identify key differences between patients under the common diagnosis of MDD in order to (a) predict whether, and to what degree, they will respond to a particular form of MDD treatment and (b) determine the best treatment plan in the context of each patient's specific background.

The results of our study have shown that we can adequately predict which subjects, at the individual level, will respond best to this internet-based, therapist-led

exercise intervention for MDD. We defined treatment success using multiple metrics including *clinically significant* change as well as *overall* change, and these results suggest that an extension of the current machine learning models will allow us to match specific MDD patients to their best-suited treatments. The potential practical applications of this methodology include increased access to care, reduced cost of treatment, faster recovery times, and relief for patients for whom traditional MDD treatments have failed.

In addition to applications in treatment planning, the modeling approach with Shap allows for better understanding of the model's decision making. Specifically, for predicting *overall* change in depression level, the model identified marital status, prior depression level and prior anxiety level as top predictive features. Subjects that began the study married, with low depression and higher anxiety tended to have the best outcomes. This matches prior research showing that being married is related to higher activity in older adults and that exercise has a preventative effect for depression specifically in women (women made up the majority of the sample in this study).[117,118] Prior depression and anxiety levels being important in prediction also matches expectations given that this is a lifestyle intervention. The stepped-care model for treating individuals with MDD presents the idea of using increasingly more intensive treatments as depression severity worsens.[119] Given this framework, it follows that this less intensive treatment regimen would work best for those individuals starting with less severe depression.

In addition to assessing predictors for *overall* change in BDI-II, we evaluated the features of most importance for predicting *clinically significant* change. Here, it was interesting to note that there was some overlap between the most important independent predictors but not complete overlap. For *clinically significant* change, prior depression severity is the most important predictor via two different measurement scales (BDI and MADRS). This relationship between lessened severity prior to treatment and better outcomes matches the findings of the model predicting *overall* change. The relationship between marital status and *overall* outcome, however, becomes much less import for predicting *clinically significant* change vs. predicting *overall* change. One

potential reason that this may occur could be in the way that *clinically significant* change was defined. It was defined as a decrease in BDI-II scoring by at least one SEM and starting above a clinical threshold on the BDI-II. With this in mind, it is plausible that being married was strongly associated with decreases in depression severity that may not have been *clinically significant*, i.e. at least one SEM change. In this way, it was very important for predicting *overall* change but not *clinically significant* change.

Despite these strengths, however, this study has several limitations that must be considered when interpreting results. First, the small, 48-person sample size both limits the generalizability of these results and offers low statistical power. Within these 48 subjects, only 24 received the treatment and therefore could be used to assess the accuracy of the model based on ground truth data. Another limitation is that the population for this study came from adults in Sweden who were predominantly female. Although the prevalence of MDD is higher in women than men, this limits generalizability to other populations. In order to confirm the clinical utility of these results, this study should be repeated using a larger, more heterogeneous sample. Nonetheless, these findings indicate that utilizing machine learning methods for personalized medicine are promising and can be a powerful tool in removing treatment barriers by making it more accessible and affordable while maintaining the effectiveness of the treatment itself.

3.1.6 Figures and Tables

Figure 1

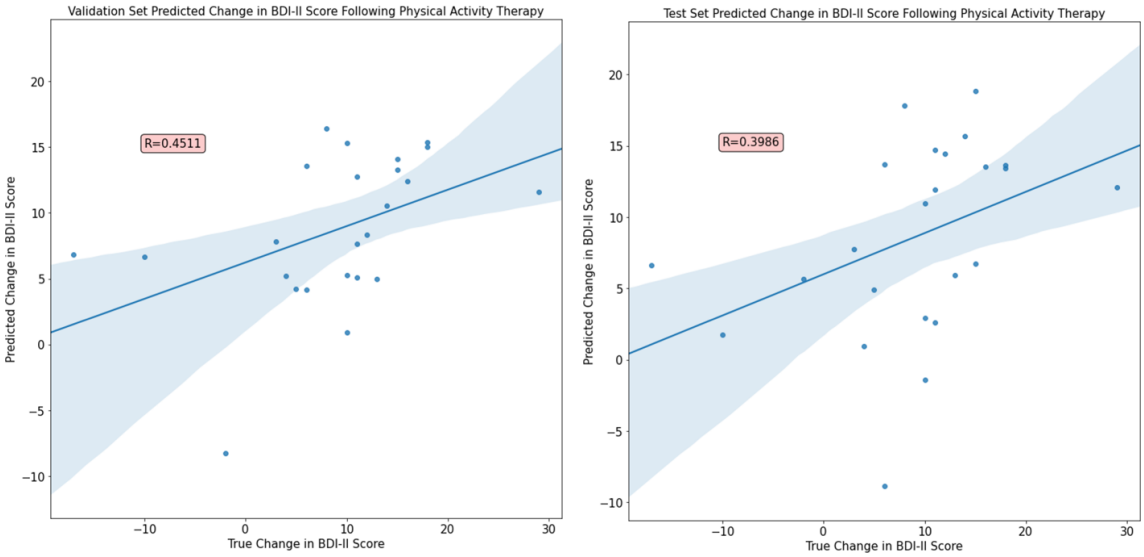


Figure 1: (Left) Within the validation set, each point represents an individual in the treatment group. The plot shows the correlation between predicted and true values for change in BDI-II score. (Right) This plot represents predicted vs. true change in BDI-II for subjects when they were part of the completely held out test set.

Figure 2

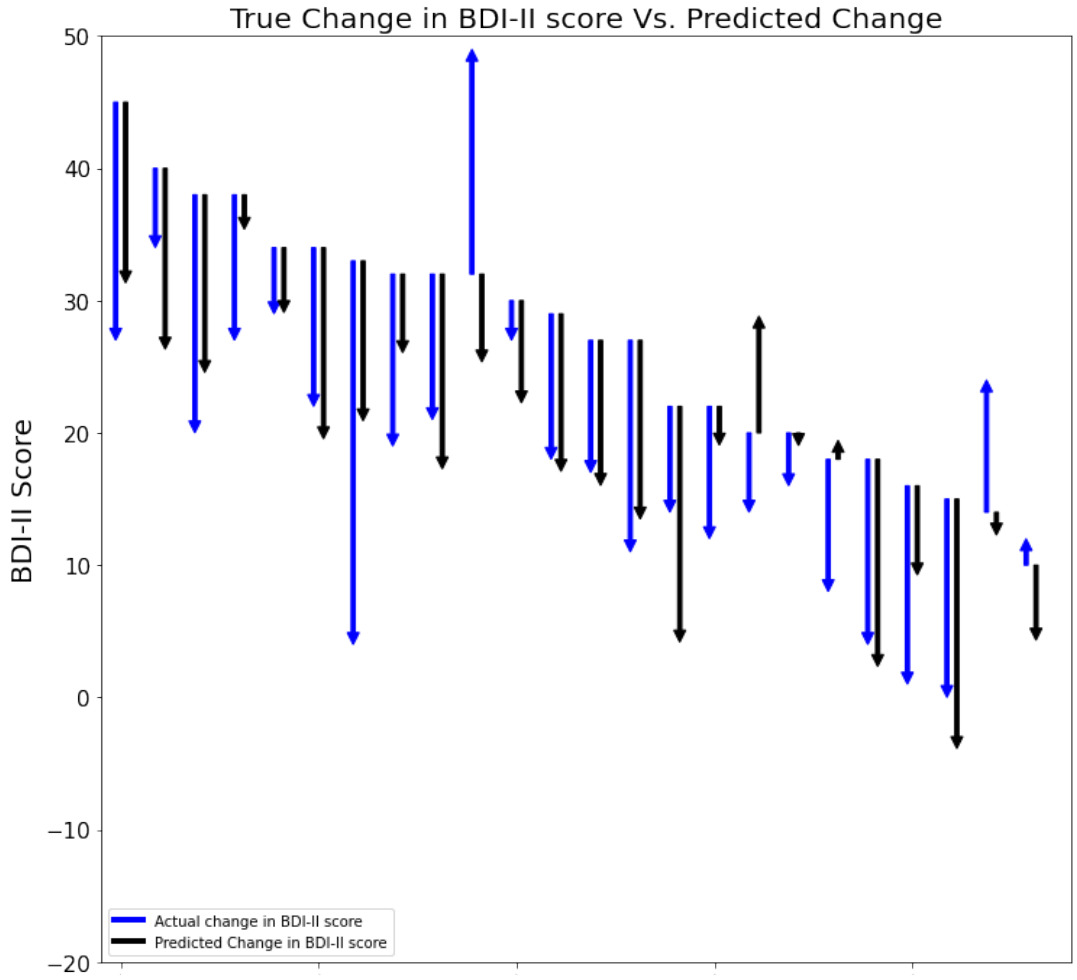


Figure 2: This plot shows the test set predictions, ordered by pre-intervention BDI-II. The beginning of the arrow is the starting point (prior to intervention) and the pointed end of the arrow represents their BDI-II at the end of the study. The blue arrows represent true change and the black arrows represent the model's prediction for that subject.

Figure 3

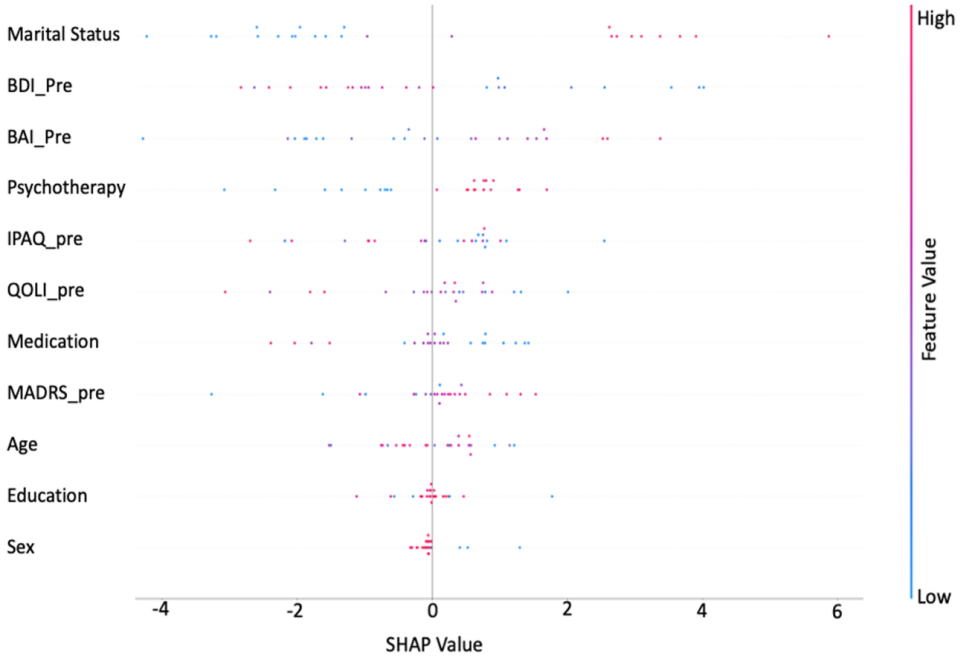


Figure 3: This Shap plot indicates the overall feature importance for the continuous model. The interpretation of a given point is as follows. In "BDI_Pre" a blue dot with a positive SHAP value indicates for that individual that a low depression severity prior to receiving the treatment indicated to the model that person would have a more favorable outcome.

Figure 4

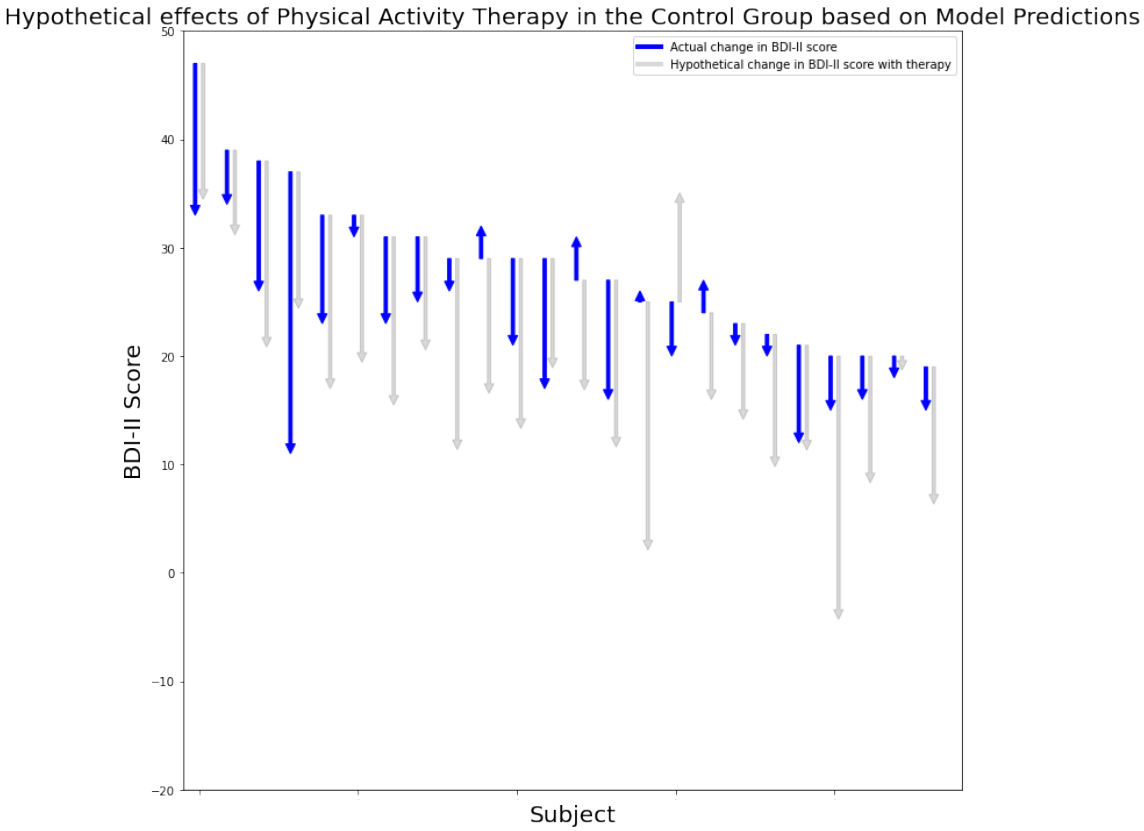


Figure 4: This plot shows the hypothetical predictions in the control group, ordered by pre-intervention BDI-II. The beginning of the arrow is the starting point (prior to intervention) and the pointed end of the arrow represents their BDI-II at the end of the study. The blue arrows represent true change after being part of the control group and the grey arrows represent the model's prediction for that subject if they were hypothetically to receive the treatment.

Figure 5

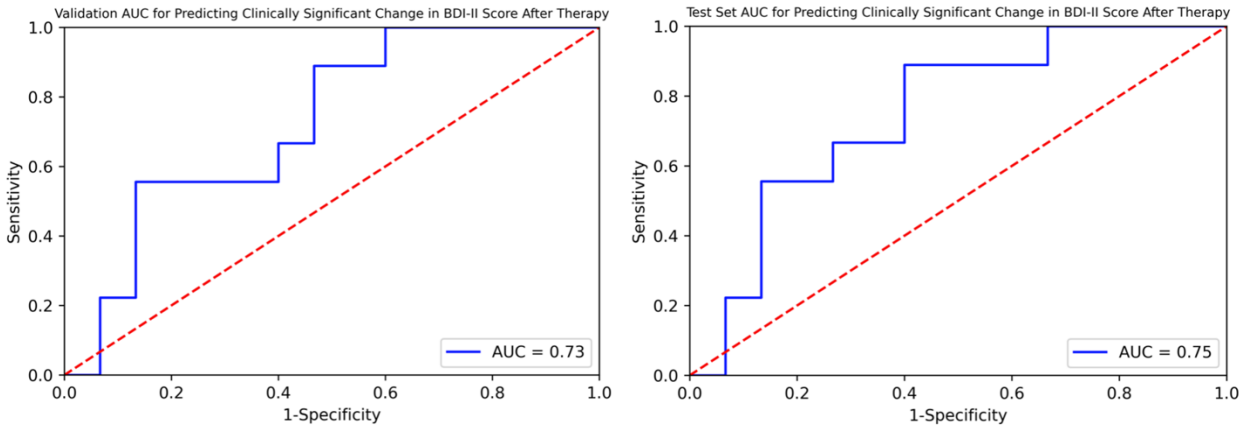


Figure 5: (Left) Within the validation set, this represents the tradeoff between sensitivity (proportion of those who had clinically significant change that were correctly identified and 1-specificity (the proportion of those without clinically significant change that were correctly identified, subtracted from one). (Right) This plot represents the same thing as the left plot but for subjects when they were part of the held out test set.

Figure 6

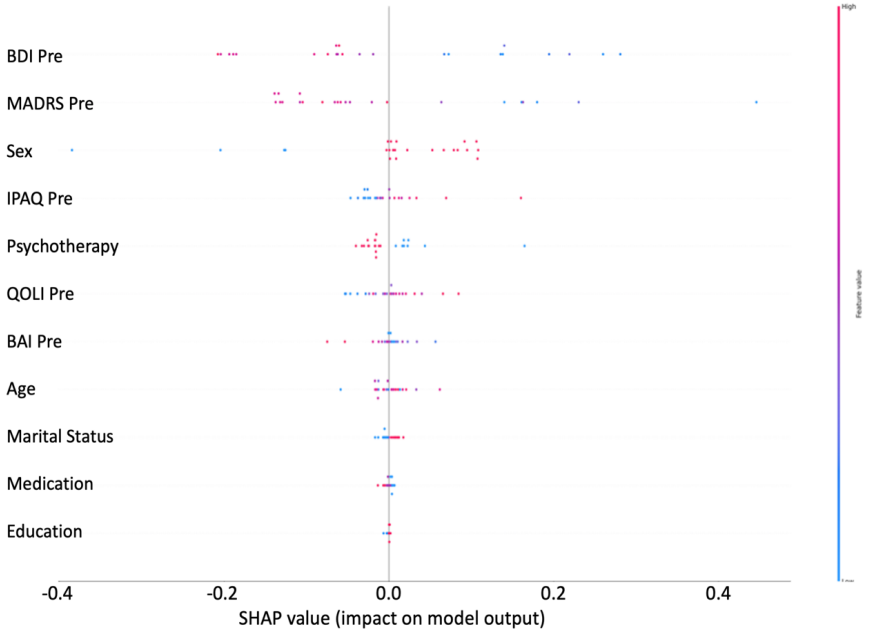


Figure 6: This Shap plot indicates the overall feature importance for the binary model predicting clinically significant change. The interpretation of a given point is as follows. In "BDI_Pre" a blue dot with a positive SHAP value indicates for that individual that a low depression severity prior to receiving the treatment indicated to the model that person would be more likely to have clinically significant change.

3.1.7 Acknowledgements

This research is supported in part by T32 (T32DA037202-07) and P30 (P30DA029926) grants provided by the National Institute of Drug Abuse.

3.1.8 Author Contributions

Each contributing author has read and approved the manuscript for submission. MN and NJ constructed the idea for the manuscript. MN and MH analyzed the data. All authors contributed substantially to writing and editing the manuscript.

3.1.9 Financial Disclosures

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

3.2 Digital lifestyle intervention RCT focusing on diet.

This work is currently in preparation for submission. The author list can be found below.

Nemesure MD, Park C, Jacobson NC. (2022) Utilizing an Ensemble Machine Learning Pipeline to Reliably Predict Change in Depression Following a Video Diet Intervention

3.2.1 Abstract

Major depressive disorder (MDD) is a leading cause of disease burden worldwide given its strain on both individuals and society. The components of this come not only from MDD severity but from the number of years lived with the disorder. In part, the longevity of major depressive episodes can be attributed to a mental health treatment system that cannot operate at the scale that is currently needed. Average wait times to see a provider can be upwards of six months and for many, this is further limited by finances and/or availability. To address these issues, many recent developments have occurred in the digital intervention space. Digital interventions can be delivered via a technological medium (i.e. computer or smartphone) and substantially reduce the burden on a provider. Additionally, these treatments are often less expensive and can be done at a time and place that is convenient for the user. One such digital lifestyle intervention is a video diet intervention designed to help individuals adjust their eating to combat depressive symptoms. Like many mental health treatments, this intervention works at the group level compared to waitlist controls but does not work for everyone. The goal of this work was to pair machine learning with this digital intervention to try and predict which individuals the treatment would work for, prior to administering the treatment. Predictive models were able to achieve this goal with an average AUC of 0.87 for predicting improvement and a correlation of 0.46 for predicting difference in pre-post severity. This type of approach lays the foundation for predictive models aiding individuals with selecting the digital intervention that works for them, thus effectively improving the efficacy of the intervention.

3.2.2 Background

Major depressive disorder (MDD) is a common and debilitating disorder worldwide, with approximately 300 million people diagnosed.[120] It is characterized first and foremost by two core symptoms: depressed mood and/or anhedonia.[3] Beyond this, there are a number of other psychologic and somatic components that can contribute to MDD including issues with sleep/fatigue, motor skills, concentration, eating/weight loss, and suicidal ideation.[3] Unfortunately, among those with MDD, severity in the moderate to severe range was far more common than mild.[6] Taken together, the prevalence and severity of MDD make it one of the leading global causes of disability, and the main cause in middle and high income countries.[120]

The global impact of MDD has led to an abundance of research in the treatment space. Currently, clinical guidelines suggest that the gold-standard treatment for MDD is either pharmacotherapeutics, psychotherapy or a combination of both.[22] At the group level, both of these interventions have been shown to be effective vs. waitlist controls across a number of trials and meta-analyses.[116,121] These treatments, however, require the input of a provider to either prescribe the anti-depressant or to deliver therapeutic content. Unfortunately this alone is inhibitive of individuals receiving treatment as the number of individuals seeking treatment for MDD outnumber providers 400 to one.[26] The obstructing factors, however, are not limited to a lack of trained health-care providers, there are also issues with high treatment costs and individuals not having the time to see a provider due to work or other external circumstances.[122]

To address the *access to care* issue, digital interventions for MDD are emerging as treatment options that have been shown to be nearly as efficacious as in-person treatment, and can, if possible, be combined with in-person treatment as well.[35,123] Digital interventions, at their core, are a form of treatment that can be delivered via a technological medium such as a smartphone, tablet or computer. Often, these interventions require little to no provider input and can be accessed at a time convenient for the user. The content of digital interventions can vary widely from digital

therapeutics (digitally delivered therapeutic content i.e. cognitive behavioral therapy) to interventions with social components (forums and support systems), to digital lifestyle interventions (changing diet, exercise etc.).[124] Similar to in-person treatments, these interventions tend to work at a group level compared to waitlist controls but do not work for everyone. The differential efficacy across persons is likely due to the heterogeneity of MDD and the individuals who are suffering from the disorder. Outside of efficacy, there is also a high dropout rate for many digital MDD interventions unless the intervention is personalized to the patient/user.[123] Given the differential efficacy and vast number of digital interventions, there is substantial space for the development of methods capable of pairing an individual with an intervention that is likely to work for them. This work focuses in on this area by leveraging machine learning in combination with randomized controlled trial (RCT) data from a digital diet intervention for MDD.[30]

Diet has been shown in previous studies as a lifestyle factor that plays into the risk of onset and maintenance of MDD, with specific foods and nutrients identified as better or worse for those at risk or suffering from MDD.[125,126] Digital diet-based interventions targeting MDD have resulted in success for both reducing symptoms and preventing exacerbation, along with being cost-effective.[30,127,128] The goal of the current work was to investigate the predictive ability of machine learning models in determining the change in MDD status and severity of individuals receiving a video-diet intervention as part of a randomized controlled trial. The main idea being that if the model is capable of predicting which individuals would respond well to the intervention, this could be a step towards personalization of digital intervention.

3.2.3 Methods

Participants

The data used in this work comes from a previous randomized controlled trial assessing the group-level effectiveness of a digital diet intervention.[30] The original trial was registered and the protocol was approved prior to beginning the work. For the purposes of this study, only de-identified data was used. The original investigators recruited 76 participants aged 17 to 35 from an undergraduate psychology class. Each

individual in the study met criteria for moderate depression and poor diet via the Depression, Anxiety and Stress-21 Scale and the Dietary Fat and Sugar Screener.[30] Each of these surveys demonstrated test-retest reliability as well as validity for the versions used in the current work.[129,130(p21)] Participants were excluded if they were pregnant, using illegal substances, or had other psychological comorbidities that might effect outcomes such as eating disorders or psychosis.[30]

Intervention

For this study, the intervention was a simple informational video delivered via an online medium that was created by a registered dietician. Participants had continuous access to the instructional content over the three-week intervention period. The video discussed the overall and mental health benefits of a proper diet. In particular, the focus of the content was on the Mediterranean diet and provided specific guidelines on both types of foods to eat as well as nutritional categories and their benefits/detriments. Additionally, participants received digital recipes as well as a small startup kit with basic ingredients.[30] Participants were briefly followed up with at the end of week one and two via phone call but were otherwise left with only the digital content. Follow-up occurred at the three-week mark and this demarcated the end of the study.[30]

Data Processing

As stated previously, only the de-identified trial data was used in the current investigation. The first step of processing was to split the data between individuals in the treatment group and the control group. To train the machine learning model, only individuals in the treatment group have true outcomes so these participants were used for the remainder of the training/testing pipeline. Within the treatment group, the next steps were to identify the predictors and outcomes in the model. Given that the ultimate goal is to predict treatment outcome prior to beginning treatment, the features used in training were limited to those collected prior to the beginning of the intervention. These included demographics and pre-intervention questionnaire data. A full list of features used for prediction can be found in table 1.

Two outcomes were predicted as part of the current work. The first outcome was change in severity score from pre to post assessment. This was operationalized as the difference of pre CES-D – post CES-D such that a positive score represented a favorable outcome. The second outcome was a binary outcome meant to be akin to clinical improvement. To be considered an improver an individual needed to have a pre-study CES-D score of greater than 16 (indicating clinical MDD) and then a post study CES-D score that had decreased from baseline.[131] This binarization yields a nearly even split of participants with 16 indicating improvement and 22 indicating no improvement based on this criteria.

Model Development

Due to the limited sample size, the validation framework for both the continuous and binary modeling approach was nested 10-fold cross validation. Nested cross validation allows each subject to be leveraged during training, validation and testing without any data leakage. To accomplish this, at first 10% of subjects are withheld entirely from the process acting as the test set. Then an additional 10% of remaining subjects are withheld as validation while the model trains on the remaining 80% of subjects. After training, the model makes predictions on the validation subjects' outcomes and then retrains on a different 80% of the training/validation pool withholding a new 10% as validation. After repeating this process 10 times, the model will have made a single validation prediction on each subject in the validation/training pool. This can be used to calculate validation accuracy and tune the model. This iterative process then repeats itself 10 times each time holding out a different test set. In this way, every subject is part of the test set at least once. As long as model tuning is only guided by the validation accuracy metrics, you can still get true test set accuracy metrics at the end by allowing the final models to make predictions on the test set at each iteration.

Given the tabular nature of the data, for both the continuous and binary prediction problems, a combination of tree-based and linear models were used.[132] Using multiple models that view the data in different ways, or ensemble learning, has

consistently shown improved performance in the supervised machine learning space.[133–135] In this case, lasso for the continuous modeling approach and regularized logistic regression are capable of uncovering linear relationships between predictors and outcomes as well as minimizing the coefficients of predictors with weaker relationships.[136] Tree-based models, in this case, random forests, are capable of uncovering non-linear relationships and interactions between predictors as they relate to the outcome.[99] To ensemble these two models together, the average prediction for each subject between the two models was taken as the final prediction.

Feature Importance

Another important component of this approach in addition to model performance is developing a knowledge base about how individual characteristics are predictive of both success and failure. Typically, this is evaluated by model coefficients that give group-level associations between predictors and outcomes. Assessing this at a group level, however, ignores that there may be individual differences in how a given feature impacts model performance. To address this issue, we leverage Shapley Additive Explainers (SHAP) to uncover how, for each individual, their characteristics are predictive of their outcomes.[137] Importantly, SHAP is model agnostic meaning it can generate these values for any complex or ensemble approach used to make predictions. The way it works is by generating "synthetic digital twins." Essentially, the SHAP algorithm will iteratively vary the input features for a given person and assess the models predicted output. For example, if a person is 34 years, old, the algorithm would hold all other variables constant and just vary age to determine how, for that person, age impacts the predicted output. It does this iteratively across one and many features to determine one SHAP value per feature per person. That value indicates how, for that individual, a specific feature is driving the model prediction.[137]

3.2.4 Results

Predicting Change in Severity

Across the ten validation folds, the average correlation between predicted and true outcomes was 0.503. In the held-out test set, the correlation between true and

predicted values was 0.465. The distribution of correlations across iterations of validation as well as the true vs. predicted values in the test set are reported in figure 1. These values are further reported in figure 2 as they relate to the baseline CES-D score for each individual. Taken together, these results are indicative of moderate predictive performance and little model overfitting.

In terms of feature importance, pre-study MDD severity, stress and diet were the strongest predictors across persons (figure 3). The most important feature was pre-study CES-D which indicated that greater pre-intervention severity informed the model of a higher likelihood of improvement. The second and third most important features were the stress and depression components of the Depression, Anxiety, Stress Scales (DASS). Higher stress also led to more favorable outcomes whereas lower DASS measured depression was also indicative of favorable outcomes. The opposite directions of CES-D depression and DASS depression likely means that the model was picking up on the unique components of MDD that each measure considers.

For the binary prediction model, the average validation AUC across folds was 0.88. The test set AUC was 0.87 indicating that there was little overfitting. The important predictors for binary prediction were nearly identical to those in the continuous prediction. This support across modeling types further indicates the validity and true signal found within this data. The main difference between binary and continuous feature importance was that the dietary questionnaire and stress component of the DASS switched places (figure 5).

3.2.5 Discussion

MDD is the global leading cause of disability and can severely impact the lives of those suffering from the disorder.[120] Despite this global burden, many individuals still cannot access gold-standard in-person treatments due to a lack of providers, financial concerns and time constraints.[122] Furthermore, even when individuals are able to receive in person care, many of them will still not have remission without relapse.[121,138] Digital treatments provide a unique solution to this *access to care* issue by making intervention content available without or with limited need for a provider. Additionally,

these materials are accessed via technological mediums so that they are accommodating to both finances and scheduling. An important similarity between in-person and digital interventions, however, is that both tend to work at the group level but not for every individual.[124]

The primary aim of this work was to leverage supervised machine learning to predict individual treatment outcomes for a digital intervention using only information collected at baseline. Applying this approach in the digital treatment space can help address issues related to both *access to care* and *individual treatment efficacy.* The idea is that this type of study lays the foundation for the potential of these methods to aid in the allocation of *accessible* digital interventions that are likely to work for the people who engage in them.

For this video diet intervention, supervised machine learning models were able to predict, with moderate performance, both the change in MDD severity as well as a binary marker for symptom improvement. Given that performance remained stable across validation folds and in the test set, for both modeling approaches, it is likely that the model was learning true signal relating baseline information to the outcome. Additionally, the most influential predictors remained stable across both models and, importantly, had logical relationships with the outcome of interest. Across both models the top predictors were baseline MDD severity and comorbid symptoms of anxiety such as stress. Prior research has shown that both of these items are strongly related to digital treatment outcomes.[139]

Another finding within the important predictive features that may be counterintuitive was that in both models (continuous and binary), increased baseline CES-D depression indicated favorable outcomes whereas increased baseline DASS-depression indicated unfavorable outcomes. One possible explanation for this could be that although both assessments are validated to assess MDD, they capture different components of MDD.[140] The DASS-depression evaluates more of the psychological symptoms of depression including feelings of anhedonia, depression, and worthlessness.[141] The CES-D evaluates these items but additionally assesses more of the

somatic components of MDD such as sleep, fatigue, concentration, eating etc.[142] Given the ability of the machine learning approach to model interactions and non-linearities in the data, it was likely able to parse the differences between the two surveys. With this in mind, the model potentially learned that having greater severity in the somatic symptoms of MDD and lower psychologic symptoms was related to favorable outcomes. This explanation makes sense given the focus of the digital intervention on diet.

This work, however, is not without limitations. The main limiting factor in this work is the sample size and generalizability. The study was originally conducted on undergraduates aged 17-35 and there was no assessment for how well the current model would be able to predict treatment success outside of this population. Additionally, the small sample size likely inhibited some of the model's ability to learn more complex relationships and potential outliers could have had a greater effect. Regardless, this work is not meant to exist in isolation but rather as part of a larger body of evidence that, when taken together, can show the effectiveness of machine learning in predicting treatment outcomes. As more trials for digital interventions take place with larger samples, there will be opportunity to leverage this work as the foundation for a model that can, in a more generalizable manner, predict digital treatment outcomes. This would help individuals suffering from MDD find an intervention that is both available and would likely work for them substantially quicker than current best practice.

3.2.6 Figures

Table 1

Features	Description
Age	Age
Gender	Gender
D1BMI	Pre-study BMI
D1Spectrophotometer	Pre-study Fruit intake estimation
D1_HVLTpercent_recall	Pre-study Verbal Learning Assessment
D1_HVLT_Linear_Curve	Pre-study Verbal Learning Assessment
D1_HVLT_quadratic_curve	Pre-study Verbal Learning Assessment
D1DSFS	Pre-study Dietary Questionnaire
D1MatrixReasoning	Pre-study cognitive score
D1GSESTotal	Pre-study Self-Efficacy Scale
D1CESDTotal	Pre-study MDD Scale
D1DASSdepression	Pre-study depression DASS
D1DASSanxiety	Pre-study Anxiety DASS
D1DASSStress	Pre-study Stress DASS
D1POMSAnger	Pre-study Anger
D1POMSDepression	Pre-study Depression
D1POMSConfusion	Pre-study Confusion
D1POMSTension	Pre-study Tension
D1POMSVigour	Pre-study Vigor
D1POMSFatigue	Pre-study Fatigue
D1DFSTotal	Pre-study Dietary Questionnaire
D1DCSTotal	Pre-study Dietary Questionnaire

Table 1 – Features used in predictive modeling

Figure 1

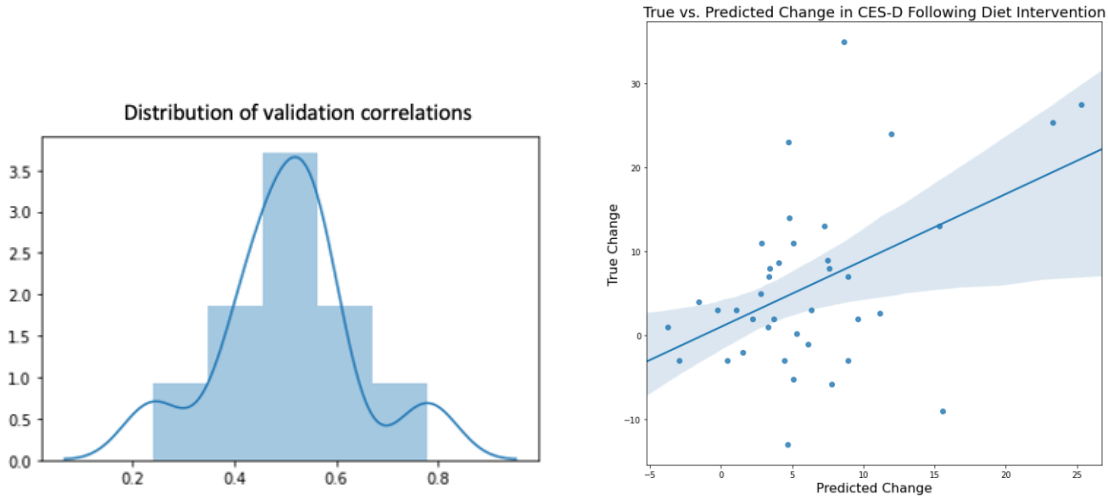


Figure 1 – *Left:* Distribution of correlations across iterations of model validation. *Right:* True vs. predicted change in CES-D from pre-post in the test set.

Figure 2

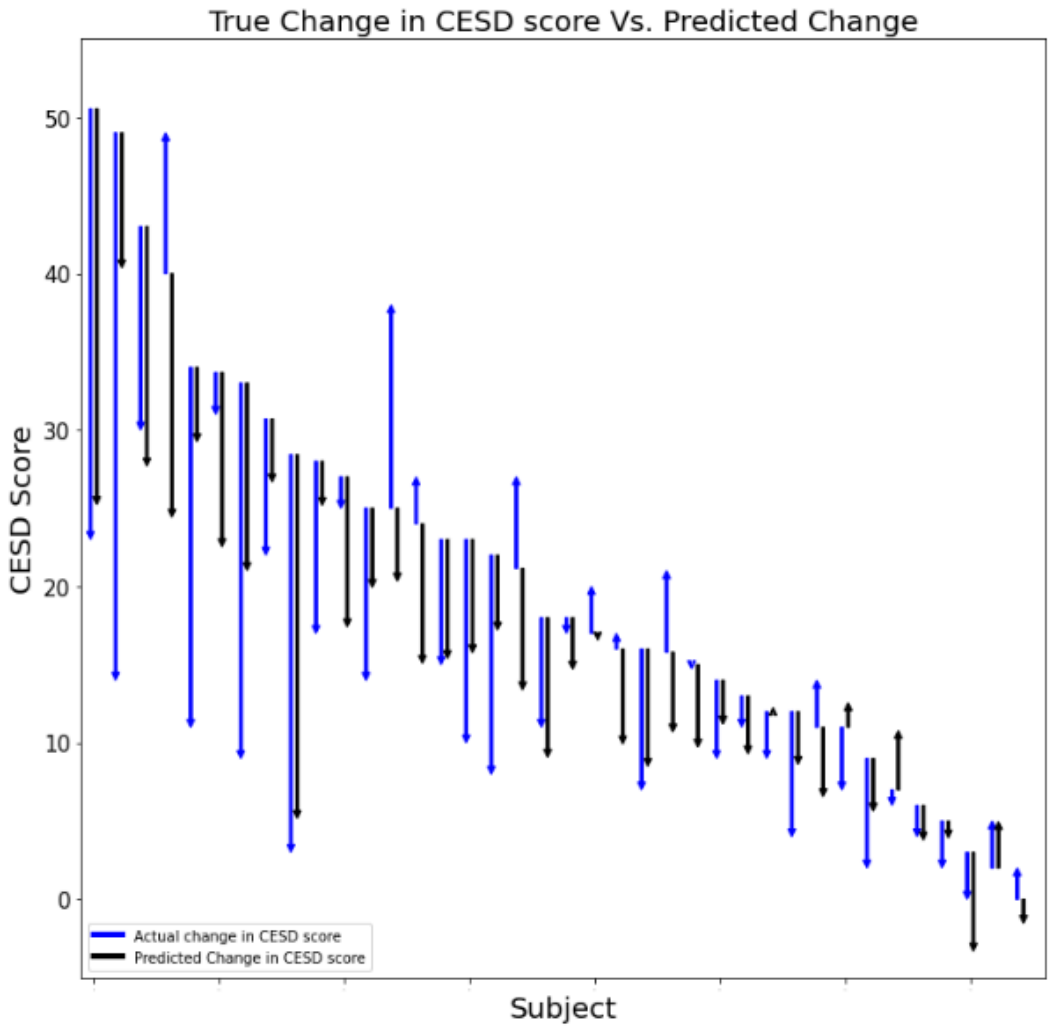


Figure 2 – True vs. predicted change in CES-D score from baseline value ordered by decreasing baseline severity.

Figure 3

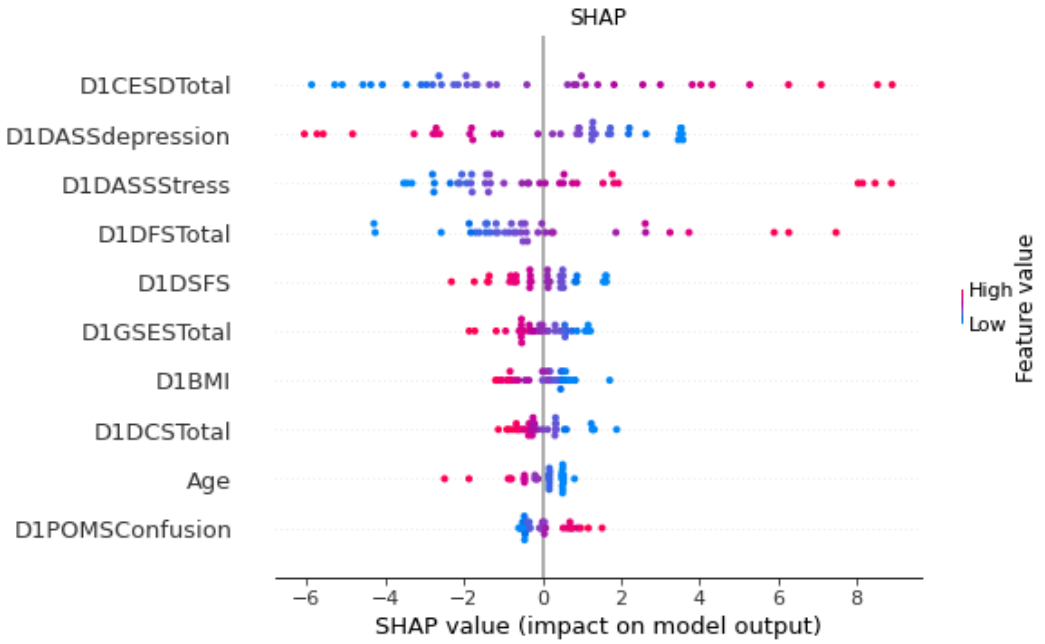


Figure 3 – This is a SHAP plot showing feature importance. Features are ordered in decreasing importance from top to bottom. Each point represents an individual and the color of the point is their value for the corresponding feature. The X-axis indicates how that feature for that point pushed the model's prediction. For example, a red point on the right side of the plot for CES-D indicates that for that individual, a high pre-intervention MDD severity led the model to predict a favorable outcome.

Figure 4

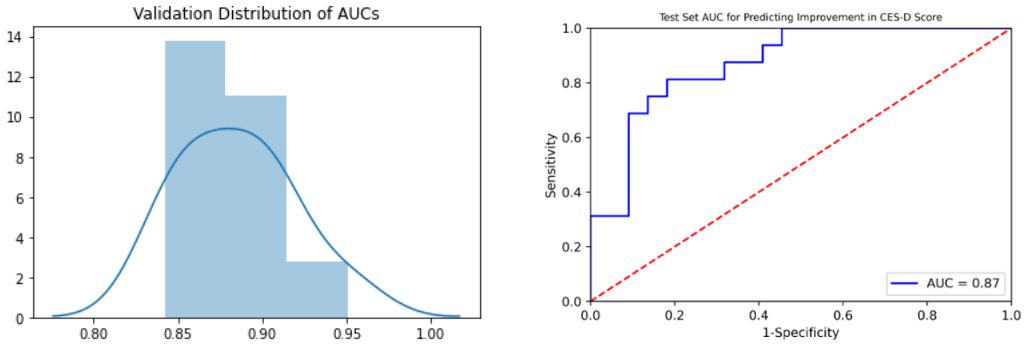


Figure 4 – *Left:* Distribution of AUCs across validation folds. *Right:* Test set AUC.

Figure 5

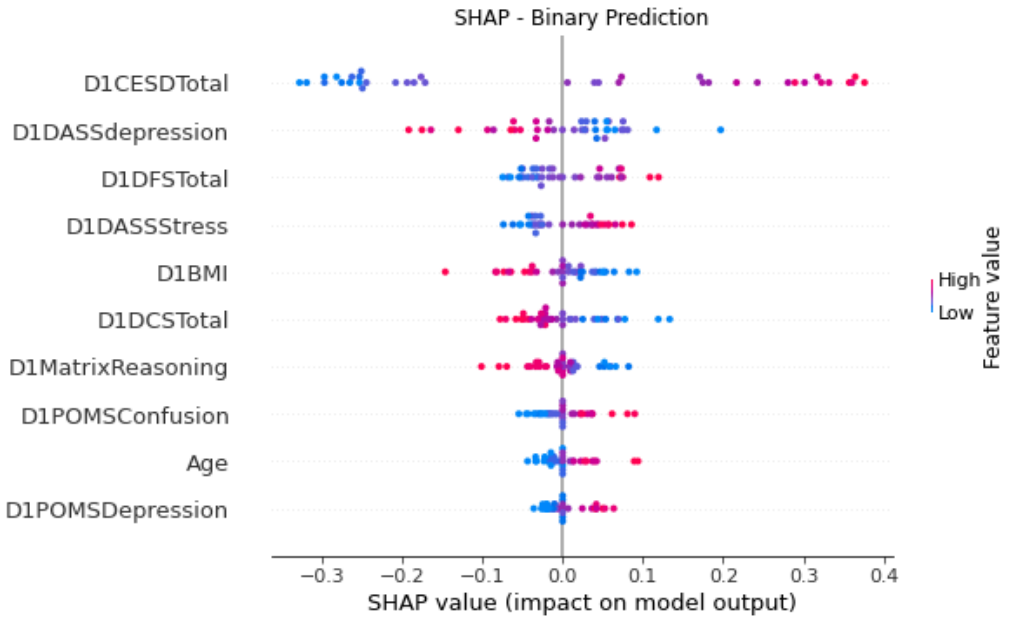


Figure 5 – This is a SHAP plot showing feature importance. Features are ordered in decreasing importance from top to bottom. Each point represents an individual and the color of the point is their value for the corresponding feature. The X-axis indicates how that feature for that point pushed the model's prediction. For example, a red point on the right side of the plot for CES-D indicates that for that individual, a high pre-intervention MDD severity led the model to predict a favorable outcome.

3.2.7 Acknowledgements

This research is supported in part by T32 (T32DA037202-07) and P30 (P30DA029926) grants provided by the National Institute of Drug Abuse.

3.2.8 Author Contributions

Each contributing author has read and approved the manuscript for submission. MN and NJ constructed the idea for the manuscript. MN analyzed the data. All authors contributed substantially to writing and editing the manuscript.

3.2.9 Financial Disclosures

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

3.3 Digital therapeutic intervention focusing on online supportive therapy and stress management and coping

This work was published as a peer reviewed short report in *Psychiatry Research.* The formatting has been minorly modified to fit this dissertation. The citation for this work can be found below.

Jacobson NC, **Nemesure MD**. Using Artificial Intelligence to Predict Change in Depression and Anxiety Symptoms in a Digital Intervention: Evidence from a Transdiagnostic Randomized Controlled Trial. Psychiatry Research. 2021;295:113618. doi:10.1016/j.psychres.2020.113618

3.3.1 Abstract

While digital psychiatric interventions reduce treatment barriers, not all persons benefit from this type of treatment. Research is needed to preemptively identify who is likely to benefit from these digital treatments in order to redirect those people to a higher level of care. The current manuscript used an ensemble of machine learning methods to predict changes in major depressive and generalized anxiety disorder symptoms from pre to 9-month follow-up in a randomized controlled trial of a transdiagnostic digital intervention based on participants' (*N*=632) pre-treatment data. The results suggested that baseline characteristics could accurately predict changes in depressive symptoms in both treatment groups (*r*=0.482, 95% CI[0.394, 0.561]; *r*=0.477, 95% CI[0.385, 0.560]) and anxiety symptoms in both treatment groups (*r*=0.569, 95% CI[0.491, 0.638]; *r*=0.548, 95% CI[0.464, 0.622]). These results suggest that machine learning models are capable of preemptively predicting a person's responsiveness to digital treatments, which would enable personalized decision-making about which persons should be directed towards standalone digital interventions or towards blended stepped-care.

3.3.2 Background

Anxiety and depressive disorders occur in one-fifth of the population.[143] Traditional care systems are unable to meet the sizeable patient population—most persons with psychiatric diagnoses receive no specialized treatment and wait years before initiating care and then months before care begins.[144] Digital interventions may help to close this substantial treatment gap. Nevertheless, many persons may respond differently to digital interventions. Some persons may respond well to digital interventions, whereas others may be unresponsive to these digital interventions and require further traditional care. Consequently, research is needed to determine whether responses to these digital interventions can be predicted before digital treatment initiation to facilitate timely and efficient treatments. Three prior studies have examined predicting symptom changes in digital treatments using machine learning with early promising results in obsessive-compulsive disorder (*N* = 61),[145] depression (*N* = 283),[146] and social anxiety (*N* = 26).[147] Although promising, prior approaches to predicting responsiveness have examined a limited range of potential machine learning models and utilized smaller samples. The current trial examined a comprehensive ensemble of machine learning methods to predict changes in anxiety and depressive symptoms in a transdiagnostic digital interventions with a sample of 632 persons.

3.3.3 Methods

Randomized Controlled Trial. This study is based on a sample of 632 persons who received a digital intervention following discharge from an inpatient treatment for cardiologic, psychosomatic, and/or orthopedic rehabilitation surrounding work-related stress.[148] Participants were randomized to active intervention groups with one group receiving online supportive expressive therapy with feedback from a therapist (*N* = 303) and the other group receiving information related to stress management and coping (*N* = 329).[148] The current study predicted changes in depressive and anxiety symptoms separately in each treatment group based on the Patient Health Questionnaire-9 and Generalized Anxiety Disorder-7 item questionnaire completed before the intervention and after a 9-month follow-up. Predictors in the machine learning models included responses to a variety of self-reported questionnaires occurring before the initiation of

treatment (e.g. work capacity, perfectionism, calmness, need for occupational treatment, problem-solving, ambition, resignation, anxiety and depressive symptoms, social support, pension, education, sex, age, willingness to change, among others).[148]

Data Analysis. Data preprocessing was done in R and machine learning modeling was performed in python with sklearn. The current application was based on an ensemble of many machine learning models, which routinely demonstrate superiority and robustness compared to single-model solutions. The ensemble model consisted of two layers: (1) base models including a large variety of neural network models, ridge regression, random forests, general linear models, gaussian process, extreme gradient boosting, k-nearest neighbors, and support vector machine; and (2) an averaging layer that took the mean prediction for a given subject across models and validation folds. Notably, all models were trained on entirely out-of-sample nested cross-validation folds, which minimized overfitting (no hyperparameter tuning was done with subjects in the test fold).

3.3.4 Results

The ensemble predicted change in depressive symptoms in both treatment groups (r=0.482, 95% CI[0.394, 0.561]; r=0.477, 95% CI[0.385, 0.560]) and anxiety symptoms in both treatment groups (r=0.569, 95% CI[0.491, 0.638]; r=0.548, 95% CI[0.464, 0.622])with high accuracy (Figure 1). This accuracy was increased an average of 11.9% over a single base learner. Additionally, the ensemble accuracy was increased on average 6.00 % over a theoretically guided set of predictors (age, gender, medication status, and baseline severity) utilized in a GLM.

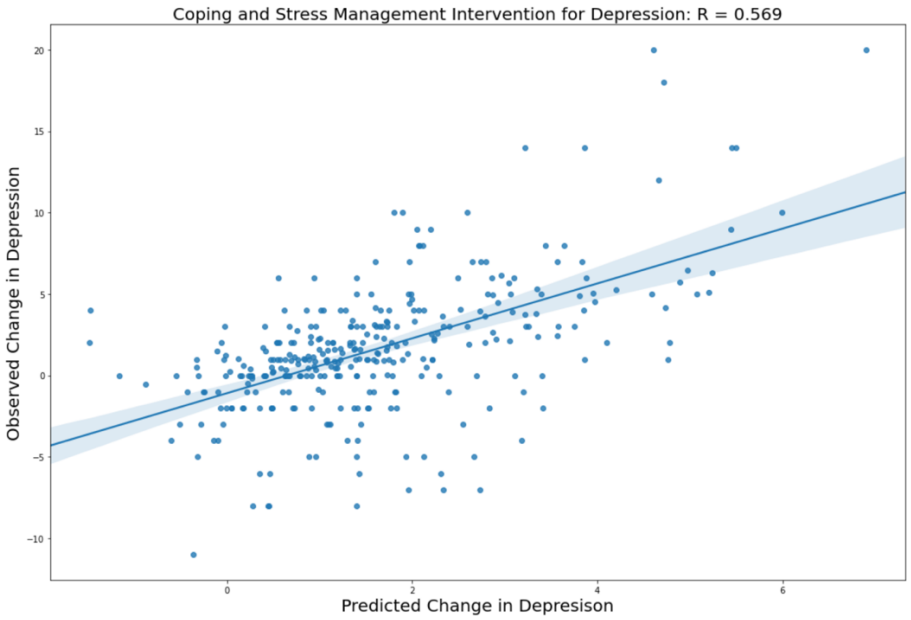
3.3.5 Discussion

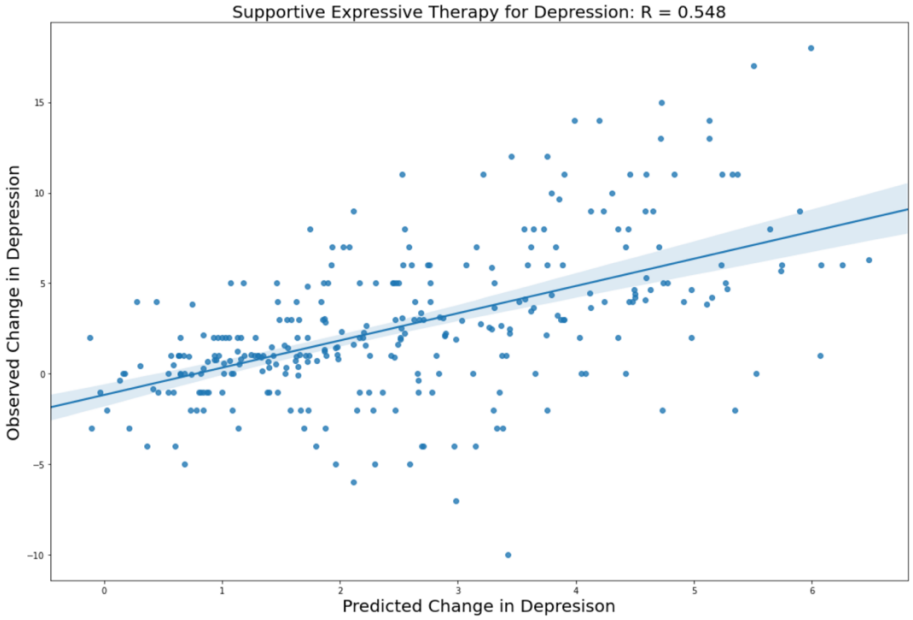
These results suggest that inferring treatment prognosis of changes in anxiety and depressive symptoms can be predicted using pre-treatment data alone. By delivering this information to clinicians, these machine learning models may be utilized to help guide future decision-making between low-resource digital interventions or higher levels of traditional care. The success of these predictions suggest that these models could guide recommendations about the appropriate level of digital or

traditional care before any care commences, saving patients 'limited time, energy, and resources while immediately triaging of patients to the appropriate levels of care. It may also save providers' time by only seeing those patients who are unlikely to benefit from digital interventions.

3.3.6 Figures

Figure 1





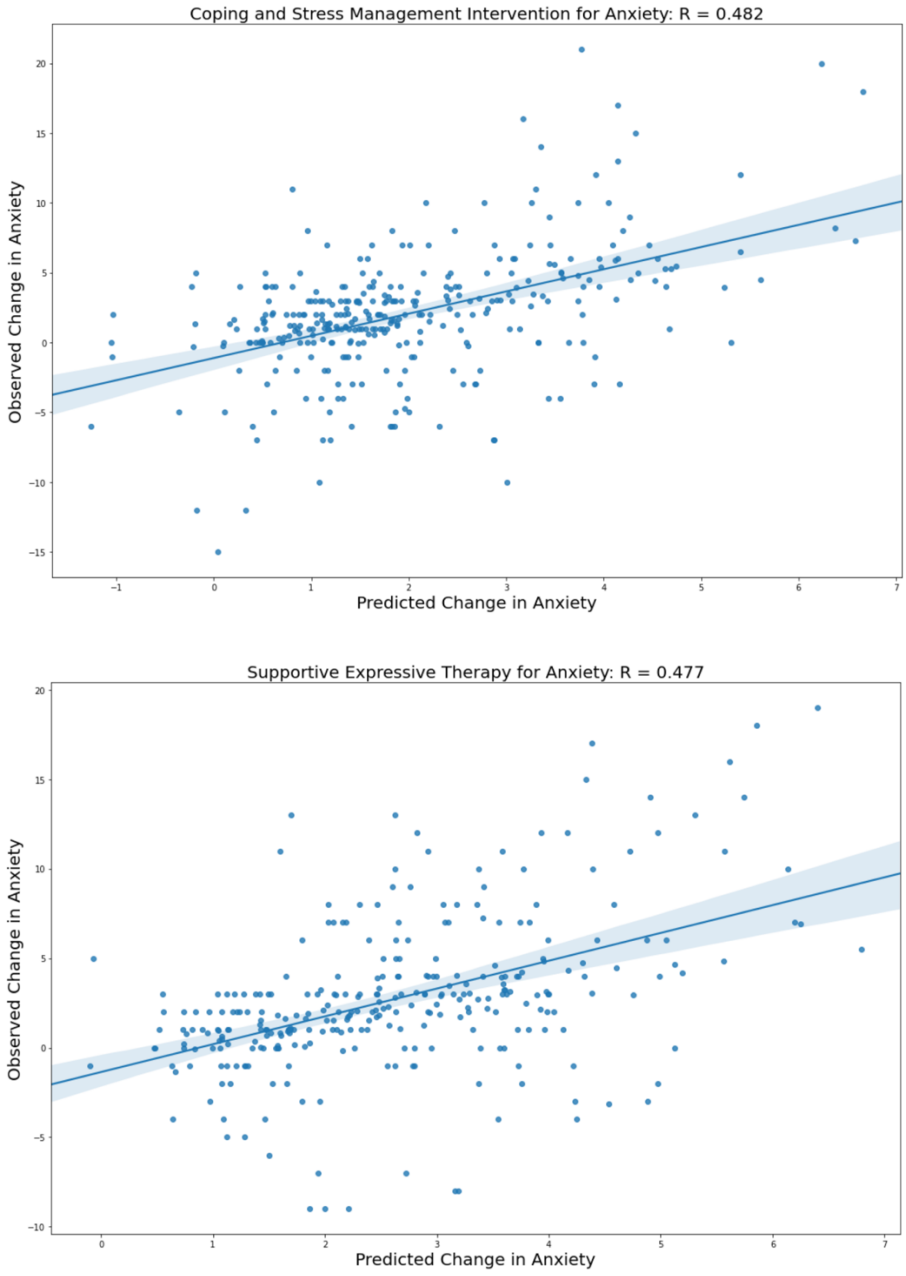


Figure 1. Observed and predicted change in the Patient Health Questionnaire-9 and Generalized Anxiety Disorder-7 symptoms in both treatments.

Chapter 4: Building a better understanding of the experience of major depressive disorder over time

This work is currently under review at a peer-reviewed journal. The citation for the *Psyarxiv* version is below.

Nemesure MD, Collins AC, Price G, et al. Depressive Symptoms as a Heterogeneous and Constantly Evolving Dynamical System: Idiographic Depressive Symptom Networks of Rapid Symptom Changes among Persons with Major Depressive Disorder. PsyArXiv; 2022. doi:10.31234/osf.io/pf4kc

4.1 Abstract

Major depressive disorder (MDD) is the leading cause of global disease burden. Diagnostically, major depressive episodes are conceptualized as a series of individual symptoms occurring most of the day for at least two weeks. Despite this operationalization, among those meeting criteria for MDD, symptoms are highly variable, showing greater variation within and across days than across weeks. Moreover, MDD is highly heterogeneous, varying considerably across people in both function and form. Recent efforts have examined this heterogeneity by studying MDD as a system where symptoms influence one another over time across individuals. In studying MDD symptom dynamics, however, most studies have made a strong assumption: that the symptom dynamics themselves are static and do not dynamically change over time. Nevertheless, there is a possibility that the individual MDD system dynamics change continuously across time. As a part of the Tracking Depression Study, funded by the National Institute of Mental Health and the National Institute of General Medical Sciences, participants (*N* = 105) completed ratings of MDD depressive symptoms three times a day for 90 days. In the current manuscript, we conducted time-varying vector autoregressive models to investigate the idiographic symptom networks of the entire sample of participants collected to date, and illustrate the finding with a case series of five persons with MDD. Aligned with prior research, the results indicate that there is high heterogeneity across

persons, such that the individual network composition is unique from person to person. Moreover, the results suggest that for most persons, individual MDD symptom networks change rather dramatically across the 90 days. This is supported by the fact that 84% of individuals experienced at least one change in their most influenced symptom with the median number of shifts being two over the 90 days. Additionally, the majority of individuals had a top predicted symptom fall into the bottom half of predicted symptoms over the 90-day study period. Our findings offer further insight into short-term symptom dynamics, suggesting that the composition and driving factors of MDD are not only heterogeneous across persons but also within-persons across time.

4.2 Background

Major depressive disorder (MDD) is one of the most common mental disorders and the leading cause of global disease burden [149]. Unfortunately, current diagnostic conceptualizations are faced with several difficulties. Firstly, MDD is a complex, heterogeneous system with over 1,000 unique symptom presentations [47,150]. However, current diagnostic conceptualizations of MDD treat different MDD symptoms as interchangeable (e.g., meeting 5 of 9 diagnostic criteria or using sum scores of individual symptoms of MDD occurring most of the day, nearly every day for at least two weeks to create a total score, reflecting overall depression severity). This conceptualization falls short of accurately capturing an individual's MDD given that presentations can vary substantially from person to person [47,151]. Indeed, conceptualizing MDD as a sum score leaves out crucial information as to which symptoms are the most important and influential in a person's overall diagnostic presentation [152]. A second problem with diagnostic conceptualizations is the assumption that within a major depressive episode, MDD is assumed to be chronic and unwavering (e.g., occurring "most of the day nearly every day for two weeks"). However, when persons with MDD are measured intensively within persons' daily lives, their symptoms are far from stable across days or weeks, but rather vary more substantially across hours within days as opposed to across weeks or months [43–46,153].

Rather than using traditional diagnostic conceptualizations, MDD may be better conceptualized as a constantly evolving, complex system for individual persons. As such, MDD symptoms dynamically interact with each other such that one symptom can influence the development and maintenance of other MDD symptoms, thus contributing to the overall, complex system of how MDD presents [150,152]. Moreover, symptom dynamics can vary from person to person, and, importantly, can highly vary within a single day for an individual person due to both internal (e.g., negative cognitions) and external factors (e.g., stressful life events; [154]. Thus, it is important to utilize methodological and statistical approaches to accurately capture the dynamic system of MDD.

Current Network Approaches to Symptom Variability

Ecological momentary assessments (EMAs) allow for researchers to collect several assessments throughout a single day for a longitudinal period and can be used to investigate short-term changes in symptom variability. Vector autoregressive (VAR) modeling is a statistical analysis that uses repeated measures regression to examine the temporal relationships between symptoms in psychopathology, including MDD [155], and often include data from EMAs. More recently, VAR modeling has been applied to network science to examine MDD as a symptom and investigate how symptoms change and influence each other over time [155,156]. When applied to network analysis, VAR models are able to detect whether certain symptoms at one time point (e.g., insomnia) directly lead to increases or decreases of other symptoms (e.g., anhedonia) or the same symptom (e.g., insomnia) at the next time point. Prior research that has investigated the dynamics of symptom networks within MDD with VAR models have commonly employed graphical ($N = 1$) or multilevel ($N > 1$) methods.

Utilization of these types of VAR models for network analysis, however, may not provide an accurate picture of how symptoms dynamically fluctuate over time. First, the multilevel VAR model takes a group-level approach to examine network structures over time, and cannot provide information as to how symptoms change at an individual level. Alternatively, graphical VAR takes an idiographic approach, although relatively few

studies have utilized this modeling for depression [45,46,157–159], so further research is needed to investigate how symptoms change over time within single individuals. Second, graphical and multilevel VAR models assume that the relationships between variables (e.g., insomnia and anhedonia) are static across time, and assume dynamic changes do not occur between variables across time [155,160–162], neglecting the possibility that MDD symptom networks themselves change over time. Thus, prior work has not yet been able to accurately capture the complex, dynamic system of MDD and has instead investigated MDD as a complex, static system.

Time-Varying Vector Autoregressive Modeling

To address this deficiency in capturing the dynamic nature of MDD symptoms over time, a time-varying auto and cross regressive modeling framework can be applied. The major benefit of this approach is it allows for both autoregressive and cross-regressive relationships to change with time (i.e., model nonstationary processes). This better reflects the reality of MDD in that it is still capable of capturing static symptom dynamics over time (which may be the case for some individuals), but can also model fluctuating dynamics as both internal and external factors arise.

Generalized additive models (GAMs) are uniquely poised to handle this modeling framework as they allow for non-linear smooths to estimate changing coefficients over time [163]. When applied in a single-lag auto and cross regressive approach, this methodology allows us to estimate the changing relationship between each symptom of MDD, as measured by the PHQ-9, and every other symptom (including the symptom of interest) at the next time point. For example, the relationship between anhedonia and itself (autoregressive) at the prior time point can be different on day two than it is on day nine, and there is a smoothed trajectory for this changing association over time. This same idea can also be applied to any cross-regressive association (e.g., the association between concentration difficulties and psychomotor difficulties at the time-point prior can be quite large at time-point 15 and quite small at time-point 30). This changing association may be due to changing external or internal factors, and importantly, using this approach we are able to capture these changing dynamics.

Rationale

As noted above, prior research has investigated MDD symptoms over time using a stationary approach (e.g., graphical and multilevel VAR), thus potentially leaving out important information about the heterogeneity of MDD symptoms and how they can dynamically change and influence each other over time. Moreover, relatively few studies have examined MDD symptoms with an idiographic network analysis approach. There is one existing study that has examined the dynamic changes of MDD with a time-varying VAR approach [164]. Their findings indicate that MDD symptoms can substantially vary both across and within persons over time, providing further evidence that it is important to investigate MDD as a nonstationary, dynamic system.

Although their findings provide important information about MDD symptom dynamics, Siepe et al. (2022) only assessed two depressive symptoms per day for 20 individuals. Thus, whereas they were able to examine variability across days, it is also important to examine variability within days given that symptoms can change over the course of hours [43–46,153]. In addition, using time-varying VAR modeling to examine the dynamics of all MDD symptoms may give further insight into how the symptoms influence each other over time. Their analyses also utilized block bootstrapping methods to assess the stability of their estimates, which suggested that the parameter estimates were unstable; however, this method may not be suitable to use with time-varying coefficient models. Thus, the purpose of the current case series is to examine the dynamics of MDD symptoms within individual persons using a time-varying VAR approach from ESM data and validate the findings through a qualitative analysis of individuals' written accounts.

4.3 Methods

Procedure

As a part of the Tracking Depression Study, an R01 study funded by the National Institute of Mental Health and the National Institute of General Medical Sciences, individuals over 18 years of age with current major depressive disorder (MDD) were

recruited remotely across the United States via Google Ads. Participants were required to use an Android-based phone as their primary mobile device. They were screened for current MDD and exclusion criteria through online surveys and virtual interviews, allowing for mental health assessments and for collection of demographic information. The mental health assessment included a clinician-administered Structured Clinical Interview for DSM-5 (SCID-5). Individuals were excluded from participation if, at any point during the screening process, they endorsed active suicidality, current or past psychotic symptoms, or bipolar disorder, or if they did not meet criteria for a major depressive episode within the past 30 days. This study was approved by the Dartmouth College Institutional Review Board (the Committee for the Protection of Human Subjects (STUDY00032081)) and participants were asked to provide both written and verbal consent prior to taking part in study.

Following screening, qualifying participants were asked to install the smartphone application, MLife, on their Android device. MLife is a mobile sensing application developed to collect passive sensing and ecological momentary assessment (EMA) data [165]. Participants were instructed to keep MLife running throughout the 90-day study period, and were prompted three times a day by the application to answer an EMA (i.e. a short survey), which included questions about depressive symptoms and a diary entry. EMA notifications were delivered starting four hours after participant self-reported wake time (morning EMA), and at four-hour intervals thereafter (afternoon and evening EMAs), with a total of 270 EMA prompts per participant over the 90 days. Upon study completion, participants were compensated $1 per EMA completed.

Patient Health Questionnaire-9 (PHQ-9)

At each EMA, participants completed a modified version of the Patient Health Questionnaire-9 (PHQ-9), a validated measure used to assess depression severity [166,167]. In the current study, we utilized a modified PHQ-9 to make the EMA questions more mobile-friendly (see Supplemental Materials). Participants were asked to use a sliding scale (rather than the original 4-item Likert-scale) to select a value ranging from 0 to 100 that best reflected how they felt, as done in prior work (Torous et al., 2015). Participants

were asked to think of the sliding scale as ranging from a day in their past when the relevant question was not an issue at all (i.e., “0” or “Not at All”) to a day in their past when the relevant question was the most applicable (i.e., “100” or “Constantly”), and to assess how they had felt over the past four hours within this range.

Diary Entry Question

The optional diary entry question prompted participants to describe how they had felt over the past four hours and why. Participants were provided the option of responding through either video, audio, or text, and were encouraged to use this space to provide any other relevant information including changes in medication or therapy.

Participants

At the time of this analysis, the Tracking Depression Study had a total of 105 participants that had completed the EMA portion of the larger trial and were included in the present analyses. The five participants included in the current analyses were selected as a representative sample based on overall symptom variability and number of diary entries. Participants with the greatest number of diary entries were selected in order to allow for qualitative validation of model findings. The five illustrative participants met criteria for MDD via the SCID-5 and were between the ages of 20-40 years old, with the majority of participants identifying as female (80% female, *n* = 4) and Non-Hispanic White (48%, *n* = 3) (see Table 1 for individual demographic data).

Data Preparation

The first step of data processing was to select those individuals with the most daily diary entries as described above. Using these subjects allowed for thorough qualitative validation of the modeling outputs. Data used for this analysis consisted of all PHQ-9 EMA data collected over the course of the 90-day study except for the question related to sleep difficulties, which we excluded as it was only presented to each participant once a day rather than three times per day. EMA entries were either fully completed or did not exist, and thus there were no EMAs that contained partial data. Across the five participants they each answered 277, 266, 273, 216, and 168 EMAs

respectively,[1] and no participant went more than three days without answering a survey across the 90 days of the study.

From this EMA data, eight data tables were generated per person, with each data table representing one MDD symptom as the outcome and all eight symptoms at a lag of one EMA as the predictors. These data tables had (t-1) rows with one row for each EMA excluding the first because there are no lag one predictors at t=1.

Modeling Approach

Following the data preparation, eight GAM models, representing each of the eight MDD symptoms as an outcome, were fit per person (*N* = 5) included in the case series, resulting in 40 total GAM models. An example GAM model formula looks as follows for the prediction of PHQ-9 Question 1 (Q1): Having little interest or pleasure. The rest of the PHQ-9 questions Q1-Q8 (excluding the sleep question) at the time prior to the outcome (t-1) are the predictors.

$$Q1_t = \beta_0 + f_1(t) * Q1_{t-1} + f_2(t) * Q2_{t-1} + \ldots + f_8(t) * Q8_{t-1} + \epsilon, \epsilon \sim N(0, \sigma^2)$$

In this formula, we evaluate the linear relationship between each MDD symptom as it predicts every other MDD symptom as a (non)linear smooth function (*f*) of time. Additionally, we add an L1 penalty term allowing predictors to be penalized to zero [168]. This penalization was put in place to prevent spurious results from estimates with high variability. From each of these models, we were able to obtain a coefficient for each lagged predictor at each time point (each EMA). These changing coefficients represented the dynamic, directional relationship between the lagged predictor and the outcome for a given symptom.

Model Outputs and Evaluation

Given the aforementioned modeling approach, the per-person outputs consisted of a coefficient for each symptom as it was predicted by every other symptom, including

1 Note that surveys beyond the 270 required were due to participants' entering surveys before the official start date or completing more than the required number of surveys in a given day.

itself, at the previous time point. These results were then constructed into an adjacency matrix of coefficients for each time-point (EMA). From these adjacency matrices, directed networks could be generated with nodes representing symptoms and edges representing the association for how well the starting node predicted the receiving node at the next time point. The primary outcome of interest was the changing indegree for each node in the network. In this case, indegree is essentially just the sum of the absolute value of the coefficient for every lag one predictor across each outcome. This value represents how impacted a given symptom is by all other symptoms, including itself at the previous time point. With this, we can evaluate the changing influence over time of any given symptom by all other PHQ-9 measured components of MDD.

This approach is similar to what would typically be seen in a Gaussian Graphical Model (GGM) where nodes are represented by symptoms and edges are represented as partial correlations between symptoms across persons [156,169]. Typically, these models are validated via bootstrapping to assess whether the partial correlations persist over bootstrap iterations. Unfortunately, with the idiographic approach, there is no appropriate method to bootstrap over time points and thus a quantitative validation becomes implausible. To address this issue, we chose to complete a qualitative evaluation using written diary entries from the participants. A trained predoctoral clinical psychology intern read through each participant's diary entries while qualitatively evaluating how well they corresponded to the dynamic symptom fluctuations. Entries that corresponded temporally to shifts in symptomatology were noted and mapped back to the model outputs to qualify the modeling results.

4.4 Results

The primary results of this analysis are the changing symptom dynamics across the five participants selected for this case series. These are represented in Figures 1-5, which display the changing indegree for each symptom as it is predicted by all other symptoms at the previous time point. Through this we can evaluate an individual's changing depression profile as represented by the variation in how impacted a given symptom is

by all prior symptoms at the previous time point. Note, an even more nuanced representation of symptom dynamics is reported in Supplemental Materials. These animated gifs are able to capture all symptom to symptom relationships over time instead of simply taking the sum of a symptom's influencers.

Participant 1

Participant 1 is a White, non-Hispanic transgender female in the 20-40 age range. Of the five participants, Participant 1 had the most variable symptom profile in this case series as evidenced by their top impacted symptom changing seven times (see Figure 1). Across the eight measured symptoms, this participant had four unique symptoms that were the most influenced, each for a given period of time. Of note, Participant 1 seemed to experience a periodic effect for "lack of concentration" where this symptom varied from high to low importance in an oscillatory manner. In addition, anhedonia and fatigue seemed to maintain relatively high importance, and were the most influenced symptom profile components when prior high impact symptoms had dampened effects.

Participant 2

Participant 2 is an Asian, non-Hispanic female in the 20-40 age range. The majority of this participant's symptom profile was characterized by anhedonia and fatigue, and they demonstrated only one change in their top impacted symptom (see Figure 2). Towards the end of the 90-day period, psychomotor difficulties quickly grew to be highly influenced by other symptoms despite starting as a symptom with the lowest amount of impact. Remaining PHQ-9 symptoms were relatively static with respect to their influence by this participant's MDD dynamics.

Participant 3

Participant 3 is a White, non-Hispanic female in the 20-40 age range. Their topmost impacted symptom changed three times (see Figure 3). For the first two months of the study, the symptom of psychomotor difficulties was the primarily influenced factor by their symptom dynamics, increasing in impact for the first month, and, while still dominant over other symptoms, slowly decreasing in importance for the

second month. Both anhedonia and feeling down and depressed maintained a constant impact to start with, and then both fell off in their importance in the second month alongside the psychomotor difficulties. Across the 90-days, the symptom of feeling bad about self remained a relatively constant impacted factor. In the final month, psychomotor difficulties returned as an being importantly driven by other symptoms but was not nearly as influenced as it had been previously. In addition, anhedonia also returned as the third most important symptom, while feeling down and depressed remained low.

Participant 4

Participant 4 is a White, Hispanic male in the 20-40 age range. This participant had the least variable depression profile with no changes in their top impacted symptom, concentration difficulties, which was maintained throughout the duration of the study (see Figure 4). Beyond this, however, all other symptoms also exhibited a relatively static level of predictability across the study with psychomotor difficulties and feeling down and depressed as the next two most important features. Participant 4's symptom variability profile exemplifies how MDD has been broadly defined in the past as a relatively static combination of symptoms. In context with the four other participants studied, this finding highlights the inherent flexibility of this methodology to capture not only dynamic MDD symptom fluctuations, but also more consistent MDD experiences.

Participant 5

Participant 5 is a White, non-Hispanic female in the 20-40 age range. Their top impacted symptom changed three times (see Figure 5). At the outset of the study, the symptom of concentration difficulties seemed to be impacted by their overall depression characterization. Over the first half of the study, however, this symptom importance decreased, followed by a sustained lack of impact starting midway into the study. Instead, fatigue and weight/appetite difficulties became the primarily influenced factors by this participant's symptom profile across the second part of the study.

Comparisons Between Time-Varying Networks and Qualitative Data

We also included selective diary entries to further validate the changes in symptom networks and investigate internal and external factors that may have influenced these changes. When examining this qualitative data (i.e., diary entries), a common theme emerged such that participants often wrote about symptoms, thoughts, or behaviors that were related to the most influenced depressive symptom at the time instead of directly writing about the actual depressive symptom. Thus, it is likely that some of these depressive symptoms are capturing more than the symptom itself, including anxiety and somatic symptoms. For example, Participant 1 provided more diary entries about having headaches when difficulties concentrating was the most impacted symptom in the network. Specifically, 20% of their diary entries included a headache between Oct-30 and Nov-11, compared to 4% of their diary entries from Oct-16 to Oct-29. Participant 2 endorsed having COVID-19 around Dec-20, which is when feelings of tired/no energy became highly impacted by other symptoms of the network. Participant 3 provided more diary entries regarding her physical activity (e.g., exercising more) and medical problems (e.g., blood sugar decreasing) and increased anxiety throughout her 90 days. Participant 4 did not have any changes in indegree in the network as difficulties concentrating remained the most impacted symptom in the network across the 90 days. However, they may have consistently endorsed concentration difficulties due to them ruminating daily on negative aspects of their life, including a recent breakup, hopelessness, and worthlessness.

All Participants

In addition to the primary case-series participants, we also used this modeling approach to analyze the 105 participants that had completed the larger study. This was done in order to evaluate the distribution of symptom dynamic variability as assessed by this method. In this broader analysis, and as a means of simplifying a more complex set of results, symptom dynamic variability was defined as “the number of times the most impacted feature changes” (see Figure 6 for symptom dynamic variability distribution). The average number of times the most predictive feature changed for an individual was 1.894 times with a median of two times. Furthermore, 88 out of the 105 individuals

included in the study had their most impacted symptom change at least once over the 90-day study period.

4.5 Discussion

In the current study, we conducted a novel investigation of the dynamics of depressive symptoms within 105 participants over the course of 90 days using EMA data and a time-varying VAR approach and used five individuals as exemplars to illustrate the approach. In line with prior research, our results indicate that there is high heterogeneity across persons, such that the individual network composition is unique from person to person [157,158,164]. Moreover, our results show that for most persons, individual depressive symptom networks can change dramatically in form across a three-month period, as evidenced by some participants exhibiting significant variability within their symptom networks. Further investigation of symptom changes in the larger sample (*N* = 105) also revealed heterogeneity across persons, as evidenced by variability across the sample in the number of times that the most influenced sample changed for a given individual (i.e., 0-8 times). Within the larger sample, 84% of individuals had their top symptom change at least once and 53% had this occur more than once. Furthermore, 30% of individuals had at least one symptom be both the most impacted and least impacted at some point over the course of the 90 days, and 70% of individuals had their most influenced symptom fall into the bottom half. Taken together, our findings suggest that the dynamics of depressive symptom networks vary from person to person and are highly variable across time.

Clinical Implications

Our findings hold important clinical implications for treatment as well. The field of network science has thus far provided important information about the development and maintenance of depressive symptoms. If reflecting causal relationships, centrality measures (e.g., indegree) can give us information about which depressive symptoms are the most influenced by others in a network and potentially suggest which symptoms can serve as important targets for clinical interventions. For example, for individuals where

anhedonia emerges as the most impacted symptom, interventions targeting this symptom (e.g., positive affect treatments) may be more beneficial than other treatments [159]. As currently explored with graphical and multilevel VAR models, the symptom that emerges as the most impacted may indicate that this symptom is a risk factor and an important intervention target overall, but these models do not assess time-sensitive changes in symptom dynamics and intervention needs. As evidenced by our findings, MDD is better represented as a heterogeneous, dynamic system, given that, for some individuals, symptoms and symptom dynamics change dramatically across time. Moreover, symptoms also dynamically change during treatment, often as a result of direct therapeutic change. Thus, investigating depressive symptoms with a dynamic, time-varying approach may provide better information as to how the symptoms dynamics change over time in response to psychological and pharmacological therapies [170,171]. This approach may help to bridge the gap between network science and clinical practice for providing personalized therapeutic care based on person-specific networks.

Based on the differences between the five case studies presented here and prior research [172,173], individuals likely benefit from different treatments depending on their initial presentation. For example, anhedonic depressed individuals may benefit more from positive affect treatments, and primarily depressed individuals may benefit more from cognitive-behavioral therapy. Thus, taking a “one size fits all” treatment approach across individuals can be potentially problematic and ineffective. Additionally, our findings indicate that the dynamic nature of depressive symptoms may be better suited for interventions that are more time-sensitive and fluid rather than traditional, weekly in-person interventions. Thus, a “one size fits all” treatment approach within an individual may also be potentially problematic as a patient’s therapeutic needs will most likely fluctuate over time in response to treatment or other internal (e.g., negative cognitions) or external factors (e.g., stressful life events).

Fortunately, digital interventions represent a growing field in the literature, with several interventions currently in use for MDD [34]. Digital interventions offer an

advantage over traditional in-person interventions as they are often cheaper, less time consuming, and available in the moment to individuals [174]. Given the range in variability of symptom changes from person to person, those who experience greater fluctuations in symptoms may benefit more from digital interventions that can be used in the moment than weekly in-person interventions. Just-in-time, adaptive interventions (JITAI) in particular can be utilized for those individuals whose symptoms tend to change dynamically over the course of hours or days [175,176]. Thus, being able to monitor individuals' symptom dynamics over time, and implement JITAIs in response to specific symptom changes, may help advance personalized treatment.

Limitations

Although our findings provide important, novel information as to how depressive symptom networks vary on an idiographic level, there are several limitations of the current study. First, due to space constraints, we were unable to include all of the participants in the current presentation and consequently only selected the five individuals with the most written diary entries to include in our qualitative analyses and illustration. While these individuals were more inclined to write diary entries and may not have been representative of the broader population, we picked them specifically so that we could validate whether the modeling approach was accurately detecting symptom changes. Moreover, given the time-series nature of the data, a quantitative validation would not have worked given that bootstrapping is not suitable with an idiographic approach. However, despite the selection process as a potential limitation, the symptom variability for these five individuals proved representative of the range of variability for all participants in the sample (i.e., Participant 1 had significant variability in symptoms over time and Participant 4 had no variability).

Second, given the nature of the time-varying vector autoregressive model, we were unable to include the symptom related to sleep difficulties as this symptom was only measured once per day (compared to three times per day for all other symptoms). Thus, it is possible that excluding this item impacted the variability of symptoms overall for some individuals. For example, sleep difficulties could indeed be the most influenced

for some individuals; however, we were unable to capture this phenomenon with the current sampling framework and inherent missingness [170].

Third, we recruited participants online via Google Ads, allowing us to sample participants more representative of the general population within the United States than had we used a community or clinical sample (e.g. from a local hospital). Given that we did not recruit from patients in a hospital or outpatient clinic, it is unclear whether our sample extends to a more specific clinical, treatment-seeking sample of depressed individuals. However, three participants endorsed receiving treatment for MDD (i.e., psychotherapy and/or psychotropic medication) at some point during their 90 days in the study, thus, it is possible that we would see similar results if investigated within a clinical setting.

Finally, the MLife app utilized for the current study was developed for use on Android devices. Thus, participants were required to own and use an Android phone as their primary device, resulting in exclusion of participants who used smartphones other than an Android (e.g. iPhones). Given that Android devices constitute 44% of smartphone usage in the United States [177], our sample does not accurately reflect the larger United States population with regards to smartphone usage.

Conclusions

In the current study, we conducted the first case series investigating the symptom dynamics of major depressive disorder using time-varying vector autoregressive models. Our findings support prior research that MDD is a dynamic, constantly-evolving system and suggest that the dynamics of depressive symptoms are person-specific and can dramatically change over time in response to both internal and external factors. Moreover, our findings suggest that digital interventions may be promising toward providing personalized, in-the-moment treatment for depressed individuals. Thus, monitoring depressive symptoms with intensive, longitudinal data may allow for better detection of symptom changes and for implementation of time-sensitive interventions.

4.6 Figures

Table 1

Table 1
Participant Demographics

	Overall Sample	Participant 1	Participant 2	Participant 3	Participant 4	Participant 5
Gender	86% Women	Transgender Woman	Woman	Woman	Man	Woman
Race	82% White	White	Asian	White	White	White
Ethnicity	92% Non-Hispanic	Non-Hispanic	Non-Hispanic	Non-Hispanic	Hispanic	Non-Hispanic
Treatment	–	None	None	None[a]	Psychotherapy/ Antidepressant[b]	Antidepressant

Note. Psychotherapy indicates that the participant was seeing a mental health clinician for therapy (i.e., not medication management). Antidepressant indicates that the participant was taking a psychotropic medication for MDD (i.e., SSRI or SNRI).
[a]Participant 3 started an SSRI in April
[b]Participant 4 started esketamine in March

Figure 1

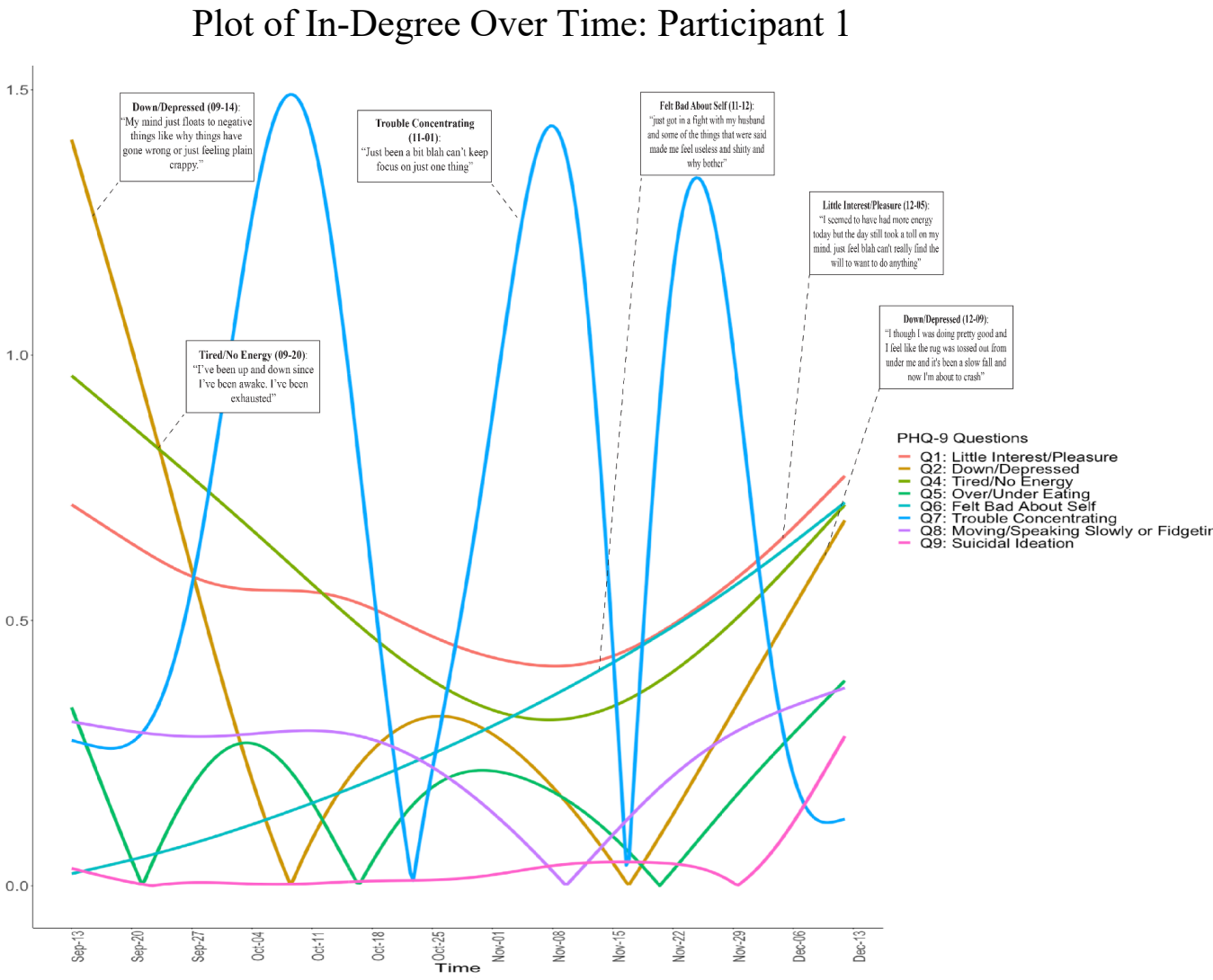


Figure 1. This figure shows the sum of the absolute value of in-degree for each symptom as it is predicted by every other symptom at the previous time point. As an example, the value for Q1 (Little Interest/Pleasure) for the last survey on September 14th is the sum of how predicted it is by all measured symptoms of the first survey on September 13th.

Figure 2

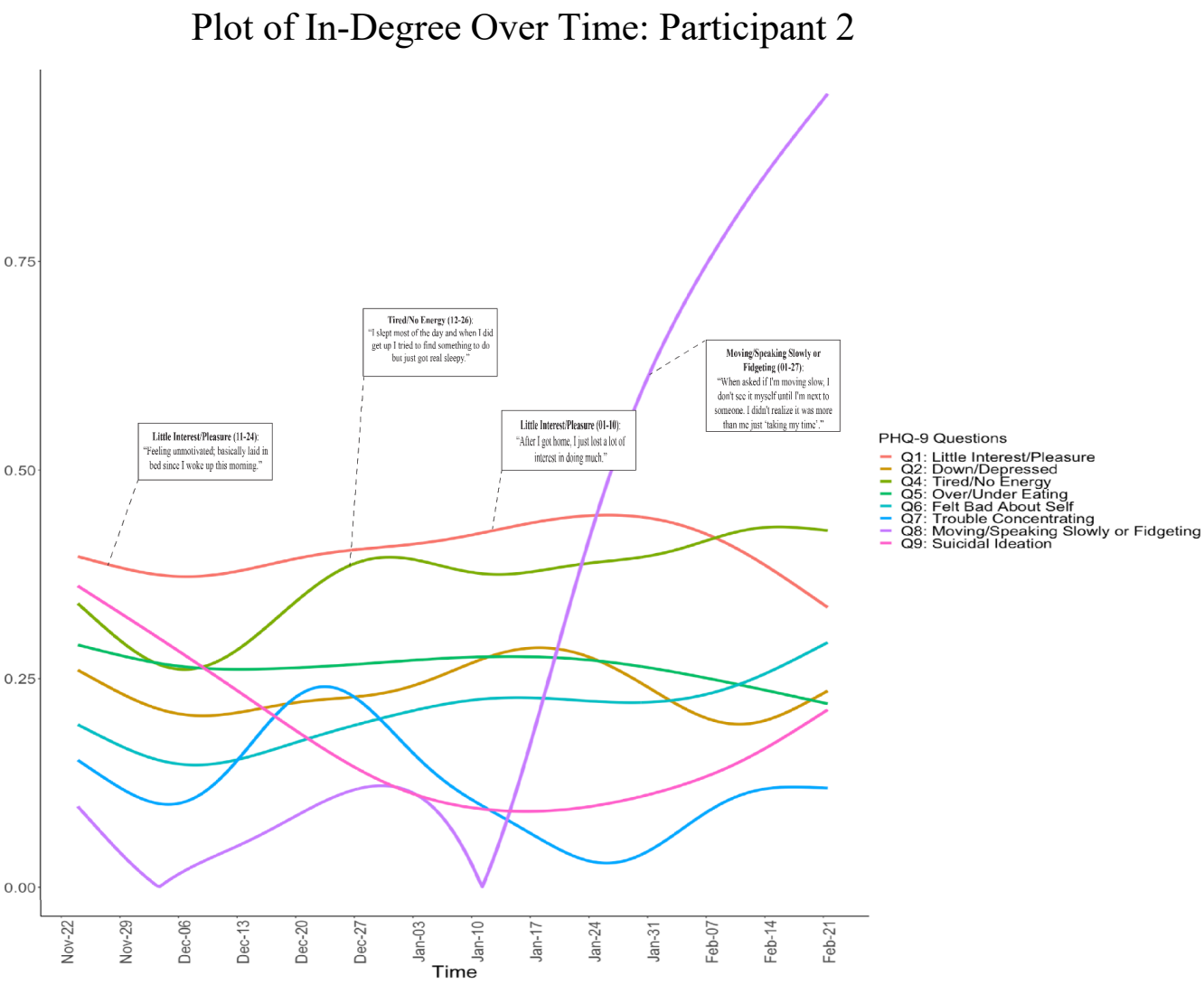


Figure 2. This figure shows the sum of the absolute value of in-degree for each symptom as it is predicted by every other symptom at the previous time point. As an example, the value for Q1 (Little Interest/Pleasure) for the last survey on November 30th is the sum of how predicted it is by all measured symptoms of the first survey on November 29th.

Figure 3

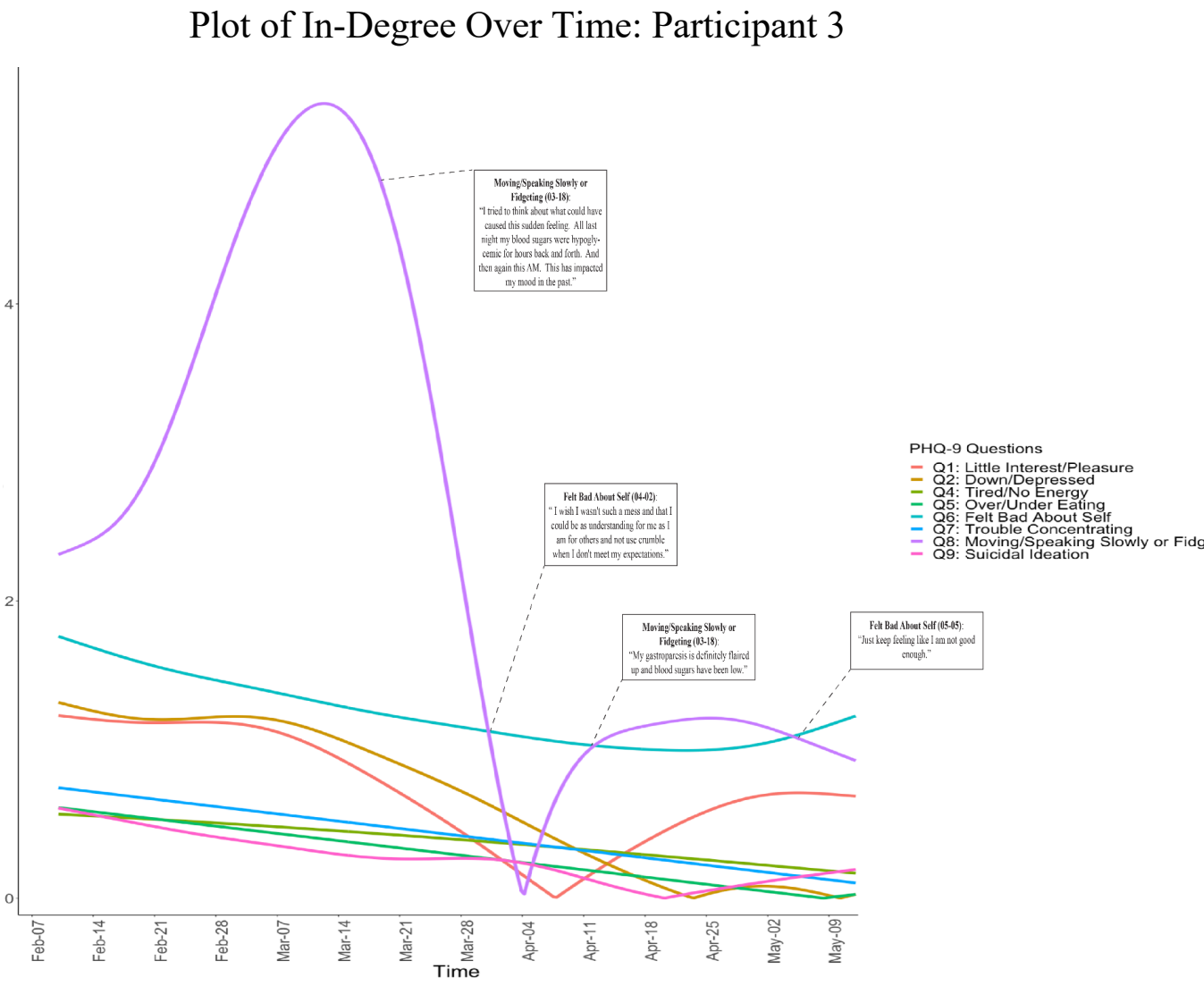


Figure 3. This figure shows the sum of the absolute value of in-degree for each symptom as it is predicted by every other symptom at the previous time point. As an example, the value for Q1 (Little Interest/Pleasure) for the last survey on February 22nd is the sum of how predicted it is by all measured symptoms of the first survey on February 21st.

Figure 4

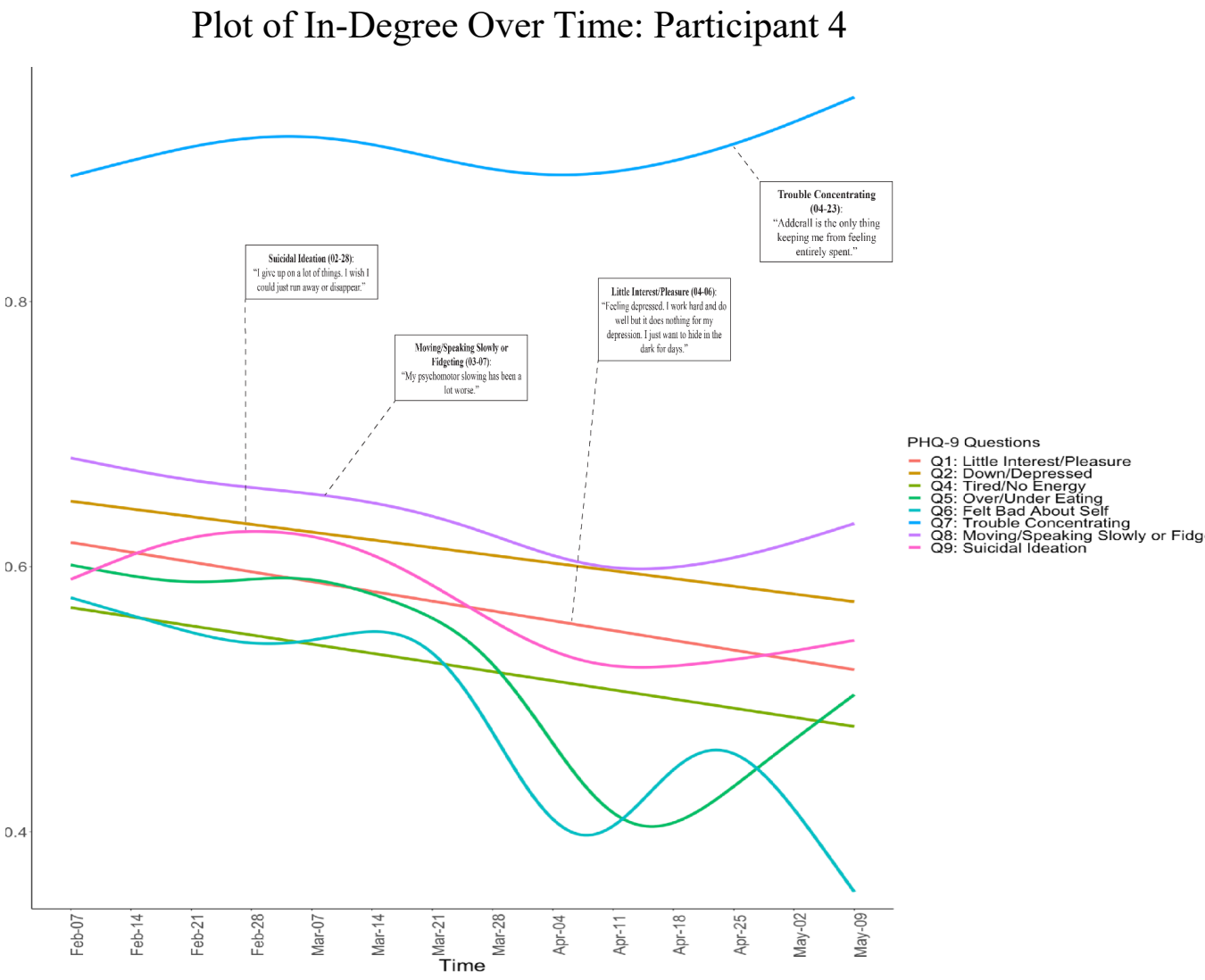


Figure 4. This figure shows the sum of the absolute value of in-degree for each symptom as it is predicted by every other symptom at the previous time point. As an example, the value for Q1 (Little Interest/Pleasure) for the last survey on February 22nd is the sum of how predicted it is by all measured symptoms of the first survey on February 21st.

Figure 5

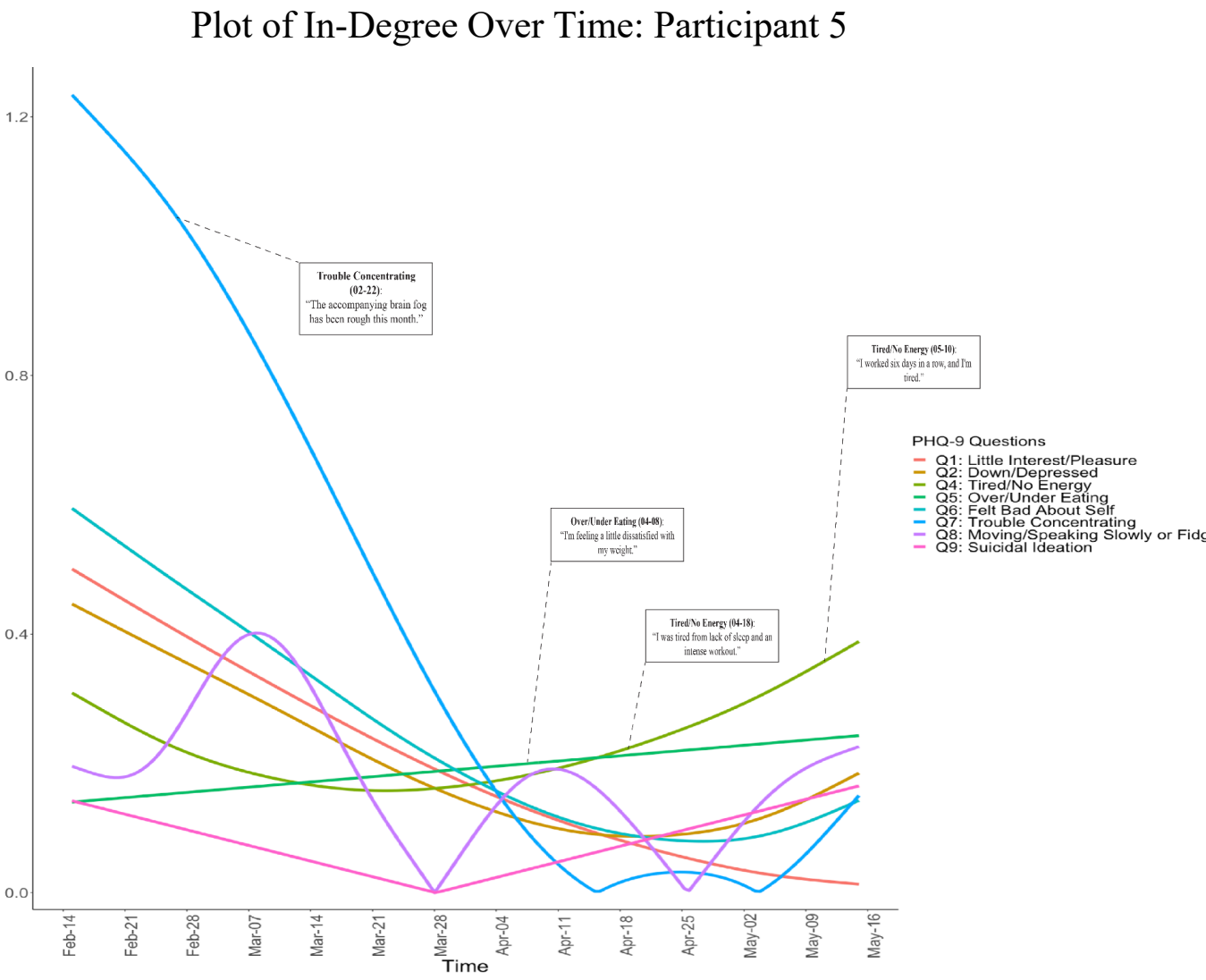


Figure 5. This figure shows the sum of the absolute value of in-degree for each symptom as it is predicted by every other symptom at the previous time point. As an example, the value for Q1 (Little Interest/Pleasure) for the last survey on February 22nd is the sum of how predicted it is for by all measured symptoms of the first survey on February 21st.

Figure 6

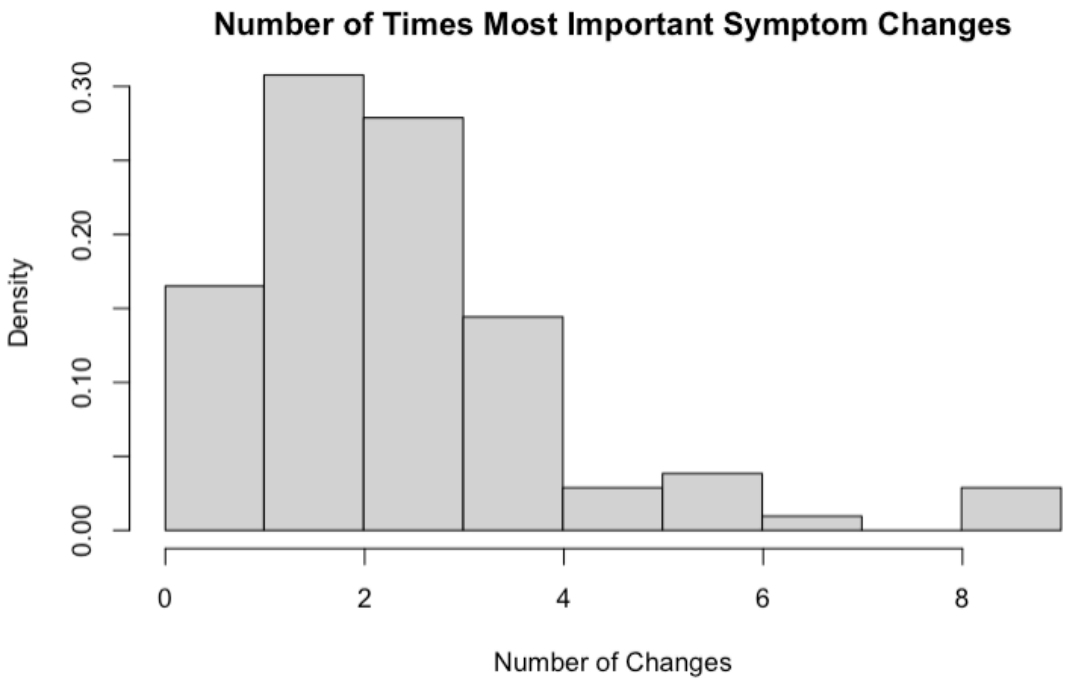


Figure 6. This histogram shows the distribution for the number of times the top most predicted symptom changed over the course of the study for the first 105 individuals to complete the study.

Supplemental Table 1

Daily EMA Questions

1. In the past 4 hours, I have had little interest or pleasure in doing things
2. In the past 4 hours, I have felt down, depressed, or hopeless
3. Last night I had trouble with sleep
4. In the past 4 hours, I have felt tired or have had little energy
5. In the past 4 hours, I have had a poor appetite or have been overeating
6. In the past 4 hours, I have felt bad about myself
7. In the past 4 hours, I have had trouble concentrating
8. In the past 4 hours, I have been moving or speaking slowly, or fidgeting more
9. In the past 4 hours, I have had thoughts of hurting myself or that I would be better off dead

Supplemental Table 1. This table shows the modified PHQ-9 Questionnaire given as part of the ecological momentary assessment.

4.7 Acknowledgements

This research is supported in part by R01 (MH12348201), T32 (T32DA037202-07) and P30 (P30DA029926) grants provided by the National Institute of Drug Abuse.

4.8 Author Contributions

MDN – Came up with the idea, ran all analyses, helped with manuscript writing.

ACC – Worked on qualitative analysis of diary entries and helped with manuscript writing. Also helped with participant screening.

GDP – Helped with custom figure generation and manuscript writing.

TZG – Oversaw participant recruitment and helped with manuscript writing.

AP – Helped design and maintain application for data collection.

SN - Helped design and maintain application for data collection.

MVH – Helped with participant screening and recruitment.

DL – Helped with data collection and participant recruitment.

ATC – Oversaw original study design data collection.

NCJ – Oversaw original study design and data collection, helped revise manuscript.

4.9 Financial Disclosures

None of the authors on this manuscript have any conflicts of interest to disclose.

Chapter 5: Discussion and Conclusions

5.1 Primary takeaways

The main focus of this thesis work was to build a foundation for leveraging machine learning and statistical models to aid in both assessment and treatment of Major Depressive Disorder (MDD). Computational methods are nicely positioned in this domain to support a mental healthcare system that is currently unable to scale and meet the needs of all individuals. The original conceptual model for this work involved thinking about the path to treatment/recovery for an individual suffering from MDD. The first step on this path is to identify individuals with the disorder. Already, there are challenges due to lack of providers in even getting this assessment. Looking at MDD alone, the relationship of individuals requiring mental health care far exceeds, by a factor in the hundreds, those that are able to provide help.[26] This results in an exceptionally limited amount of time that mental healthcare providers can devote to assessment. More often than not, as a result of the limited number of mental health care practitioners, this assessment is performed in a primary care setting. During this time, however, the provider is also trying to simultaneously assess most components of an individual's health. This leads to lower rates of mental health screening, even when it's specifically advised.

Chapter 2 focused on trying to combat this issue of low assessment prioritization in primary care. Given that active solutions, i.e. instructing physicians to screen at every visit, failed to completely address the problem, a passive supervised machine learning solution was uniquely poised to fill the gap. These models could leverage the data that was already being collected at a wellness visit to passively predict whether or not an individual was at risk for MDD. If the model determined that an individual might be at risk, it could alert the physician to perform an assessment. In this way, the model is not acting in isolation but rather as an additional tool for the provider to help prompt an assessment when otherwise it may not be given.

The benefit of this work, however, extends beyond the development of a population-specific model for MDD detection. Generating a body of evidence that this

type of passive approach to mental healthcare can be beneficial is crucial for starting to build a digital assessment system at scale. Additionally, beyond laying the groundwork for this type of assessment modeling, feature importance analysis can provide insight into how certain individual characteristics uniquely impact prediction outputs. Not only does this work inform the generation of better models in the future, it can also provide clinical insights into what aspects of an individual's life are likely associated with a potential MDD diagnosis.

Relating back to the conceptual model, a diagnosis is only the first step on the path to intervention and recovery. The next step involves actually finding and receiving treatment. At this stage, there were two main barriers preventing a successful trajectory from treatment to recovery. These inhibiting factors are ***access to care*** and ***care efficacy***. The first component, access to care, is primarily constrained by the lack of providers, but this is not the only limiting factor. In addition to wait lists being on the magnitude of months, even if an individual does have the opportunity to see a provider, they may be limited in their ability to follow through.[178] This could be due to financial concerns or time constraints, e.g. not being able to get out of work for an appointment. Unfortunately, the barriers to recovery continue even for those who are able to see a provider. The two best-practice interventions, pharmacotherapeutics and psychotherapy, whether used individually or in combination, are still only effective for just over half of people suffering from MDD as assessed via a meta-analysis of RCTs.[25] Taken together, the ***access to care*** and ***care efficacy*** barriers are cause for major concern when it comes to the mental health care system's ability to deal with the current global leading cause of disease burden.

Chapter 3 of this dissertation work is an effort to start building a solution to both of these limiting factors in MDD treatment. Each of the studies included in this section leverages machine learning paired with a digital intervention. The digital intervention component helps to address the issue of access to care given its ability to be delivered remotely via technological mediums. These treatments will typically require reduced provider input or, in some cases, remove the necessity for a provider altogether. In this

way, digital interventions have the opportunity to work at scale, reaching the people who need them at a time and place that is convenient and accessible.

Within this digital intervention domain, a machine learning approach could then be applied to predict differential treatment response across treatment types. The results of the work in this chapter are laying the early groundwork of feasibility in this space. If intervention outcomes can be determined prior to beginning the course of treatment, the number of failed attempts can be substantially reduced and the time to recovery greatly improved. This framework for treatment allocation would effectively increase the efficacy of all digital interventions given they would only be attempted when the chance for success is high.

To operationalize this approach, the models for determining treatment outcomes are trained on easy to collect information that could be obtained prior to beginning an intervention. These data typically include demographics, MDD severity, comorbidities and various lifestyle surveys. In the different sections of this chapter, three studies assessing digital treatment efficacy at the sample level were used as benchmarks to determine the success of this individualized treatment approach. These studies were split across two domains: one digital therapeutic (with two intervention styles) and two digital lifestyle interventions. In both cases, using only data collected at baseline, we were able to achieve moderate predictive performance. For both the digital therapeutics and the digital lifestyle interventions, correlations between predicted MDD severity post treatment and true severity post treatment ranged from 0.4-0.5. In all aspects, model performance was directly comparable across all intervention types.

Again, in a similar framing to the assessment modeling from chapter 2, these models were developed with a use-case in mind. The idea is that the models were trained on information that could easily be collected via an online survey. Importantly, this survey could be accessed independently without the necessity of having the time, money and patience to get off of a waitlist and see a provider. The individual would need to be seeking treatment but for those that were, if they chose to fill out the

survey, they could evaluate the model's predictions (and reasons for that prediction) on success likelihood across digital treatment domains. Instead of using these metrics in isolation, they could use this knowledge as a supplement to their own determinations about what may work for them. The final result of this ongoing line of research (with the need for more extensive testing) would be that an individual could select a digital intervention with the best possible chance of reducing MDD symptoms.

Additionally, outside of the prediction accuracy itself, these models can help guide our understanding of the individual characteristics related to treatment outcomes. Information like this can aid in the development of new digital interventions that are more tailored to the people they work best for. This idea lends itself well to a common theme across this dissertation work; in addition to building better approaches to handle MDD at scale, we are also trying to inform a better understanding of MDD and its differential manifestations in both assessment and treatment.

The final chapter gets more to this point about understanding the varied experience of MDD. This work came about as a direct result of the outcomes from the various studies in chapter 3. It was apparent that across digital intervention domains, there was a limit in how well a model could predict treatment response. While this moderate predictive performance is an improvement over an uninformed selection, the original hypothesis was that the approach would have performed more favorably. This overestimation of predictive accuracy can likely be attributed, in part, to the features used in training the models. Additional features such as individual item scores from self-report surveys as well as additional demographic information would have potentially improved outcome metrics, however, the model would have likely stayed in the range of moderate predictive ability. The question, then, was what was the limiting factor at play?

The models developed in this thesis primarily relied on the between person heterogeneity of MDD. They leveraged between-person differences to determine differential outcomes. None of these models, however, addressed within-person differences in the MDD experience. The common conceptualization of MDD is that they

experience a combination of symptoms, meeting a certain threshold, most of the day for at least two weeks. The goal of the work in this chapter was to investigate whether or not the MDD experience and dynamical system of MDD varied at a rate greater than that definition allows. If an individual's symptoms were to fluctuate within these time constraints, it could and likely would have a substantial impact on treatment response.

The outcome of this investigation indicated that for the majority of individuals with MDD, the most affected components of their overall symptoms shifted rapidly both across and within days. This work provides a new basis on which to approach MDD conceptualization and treatment. For many people, it's not as simple as enduring a static set of symptoms the majority of the time, but rather it's a dynamic and evolving experience. This understanding of MDD can be effectively paired together with the digital intervention and machine learning work done in chapter 3 to continue the path towards accessible, personalized treatment. Taken together, this work can be leveraged to provide individuals, in a scalable way, with an accessible, effective treatment at the right time.

5.2 Looking ahead – what's next?

The world is moving into a unique new digital age where most individuals own a smartphone and data collection is streamlined and ubiquitous. With the continued collection of more and more data, passive computational models, like the ones I have described in this work, will continue to become more personalized and accurate. While my work was specifically focused on the application of machine learning to MDD, the framework of personalized medicine can apply broadly to mental health and healthcare as a whole. Not only will this improve patient outcomes, but it will also help us better understand the complex dynamics of different healthcare conditions.

As this type of work gains more traction, the infrastructure for deployment and implementation will also continue to grow. That would mean models similar to the one generated in Chapter 2 for detecting MDD presence could be directly integrated into healthcare systems. The idea of these models, however, is that they are a tool that a

physician can actually use within their workflow. This means that future work is not only in the space of developing better, more generalizable models on the foundation that this thesis work provides; but also about implementing these models in a way that is actually clinically useful. Integration of something like this should be the result of a collaboration between physicians and researchers in both the computational and implementation science space. Ultimately, the goal would be that these models are embedded in a physician workflow and allows them to be just as efficient in normal practice while simultaneously improving screening and referral in the MDD space.

Looking further along the conceptual model of assessment to recovery in the treatment space, machine learning tools have a unique role to play here as well. There are many digital treatments available that have been shown to work at the group level, however, none of them work for all individuals. While my work specifically assessed how well a model could predict an individual's response to a given treatment, future work can leverage this information across treatments. I envision the deployment of this work as a simple, easy to use, online tool that collects a set of easy to answer questionnaires. This type of deployment would necessitate that an individual be seeking treatment, but it would remove the barriers to entry of in-person care. After answering the questions, the online tool would use the predictive model[s] to provide the individual with a rank-ordered list of digital interventions sorted by the probability of success. Not only that, but using tools for model introspection, it could also provide information as to why the individual would have a higher probability of success based on their questionnaire answers. This person could then synthesize the information in this list alongside their own feelings about the digital interventions and make an informed treatment decision that is accessible to them.

Looking even further into the future of personalized treatment in mental healthcare, I see the work in chapter 3 being paired with the work of chapter 4. There are already planned studies within my dissertation lab to assess longitudinal trajectories of MDD alongside different digital interventions. In this way, there is the opportunity to use the quantitative tools provided in chapter 4 to assess changing MDD symptom

dynamics and relate that to longitudinal changes in treatment outcomes. From this data, machine learning tools could be developed to predict how broad changes in symptom dynamics predict to different treatment outcomes. In turn this information could be used to inform real-time treatment changes, delivered directly to an individual's smartphone, to maximize patient outcomes.

Ultimately, the long-term goal of a lot of this work is passive, just-in-time treatments that are accessible to an individual who needs them when they need them. Through this work I have exemplified the potential for computational and machine learning models in the MDD assessment and treatment space. Eventually, the computational tools built in chapter 2 will help more individuals be screened and assessed for MDD. Once diagnosed, they will, without waiting, be able to download an app that can use passive sensing to monitor for changes in MDD symptom dynamics using the work from chapter 4. As the app detects these changes, leveraging the foundation of the work in chapter 3, it will be able to provide a real-time digital intervention that is likely to work for that person at the time it is provided. The hope is that this type of passive work will increase the scalability of mental healthcare such that in-person visits and the current gold-standard care will be accessible for those who really need it without having to get off a six-month waitlist.

Each of the individual studies that comprise this thesis have their own limitations related to generalizability, sample size, etc. That being said, the overall approach of applying computational tools to aid in the assessment and treatment of MDD (and other mental health disorders) is both effective and necessary in the effort towards scalable mental health.

Chapter 6: References

1. Belmaker RH, Agam G. Major Depressive Disorder. *N Engl J Med*. 2008;358(1):55-68. doi:10.1056/NEJMra073096

2. Bains N, Abdijadid S. Major Depressive Disorder. In: *StatPearls*. StatPearls Publishing; 2022. Accessed February 28, 2023. http://www.ncbi.nlm.nih.gov/books/NBK559078/

3. Kennedy SH. Core symptoms of major depressive disorder: relevance to diagnosis and treatment. *Dialogues in Clinical Neuroscience*. 2008;10(3):271-277. doi:10.31887/DCNS.2008.10.3/shkennedy

4. Zimmerman M, Ellison W, Young D, Chelminski I, Dalrymple K. How many different ways do patients meet the diagnostic criteria for major depressive disorder? *Comprehensive Psychiatry*. 2015;56:29-34. doi:10.1016/j.comppsych.2014.09.007

5. Friedrich MJ. Depression Is the Leading Cause of Disability Around the World. *JAMA*. 2017;317(15):1517-1517. doi:10.1001/jama.2017.3826

6. Hasin DS, Sarvet AL, Meyers JL, et al. Epidemiology of Adult *DSM-5* Major Depressive Disorder and Its Specifiers in the United States. *JAMA Psychiatry*. 2018;75(4):336. doi:10.1001/jamapsychiatry.2017.4602

7. Albert PR. Why is depression more prevalent in women? *jpn*. 2015;40(4):219-221. doi:10.1503/jpn.150205

8. Kessler RC, Berglund P, Demler O, et al. The Epidemiology of Major Depressive Disorder: Results From the National Comorbidity Survey Replication (NCS-R). *JAMA*. 2003;289(23):3095-3105. doi:10.1001/jama.289.23.3095

9. Ittasakul P, Thaipisuttikul P, Waleeprakhon P, Wisajun P, Jullagate S. Psychiatric comorbidities in patients with major depressive disorder. *NDT*. Published online November 2014:2097. doi:10.2147/NDT.S72026

10. Reddy MS. Depression: The Disorder and the Burden. *Indian J Psychol Med*. 2010;32(1):1-2. doi:10.4103/0253-7176.70510

11. Greenberg PE, Fournier AA, Sisitsky T, et al. The Economic Burden of Adults with Major Depressive Disorder in the United States (2010 and 2018). *PharmacoEconomics*. 2021;39(6):653-665. doi:10.1007/s40273-021-01019-4

12. Kamran M, Bibi F, ur. Rehman Asim, Morris DW. Major Depressive Disorder: Existing Hypotheses about Pathophysiological Mechanisms and New Genetic Findings. *Genes*. 2022;13(4):646. doi:10.3390/genes13040646

13. Park EH, Jung MH. The impact of major depressive disorder on adaptive function: A retrospective observational study. *Medicine*. 2019;98(52):e18515. doi:10.1097/MD.0000000000018515

14. Pratt LA, Druss BG, Manderscheid RW, Walker ER. Excess mortality due to depression and anxiety in the United States: results from a nationally representative survey. *General Hospital Psychiatry*. 2016;39:39-45. doi:10.1016/j.genhosppsych.2015.12.003

15. Ng C, How C, Ng Y. Major depression in primary care: making the diagnosis. *smedj*. 2016;57(11):591-597. doi:10.11622/smedj.2016174

16. Zimmerman M. A Self-Report Scale to Diagnose Major Depressive Disorder. *Arch Gen Psychiatry*. 1986;43(11):1076. doi:10.1001/archpsyc.1986.01800110062008

17. Dunstan DA, Scott N, Todd AK. Screening for anxiety and depression: reassessing the utility of the Zung scales. *BMC Psychiatry*. 2017;17(1):329. doi:10.1186/s12888-017-1489-6

18. Handy A, Mangal R, Stead TS, Coffee RL, Ganti L. Prevalence and Impact of Diagnosed and Undiagnosed Depression in the United States. *Cureus*. Published online August 14, 2022. doi:10.7759/cureus.28011

19. Park LT, Zarate CA. Depression in the Primary Care Setting. Solomon CG, ed. *N Engl J Med*. 2019;380(6):559-568. doi:10.1056/NEJMcp1712493

20. Pfoh ER, Janmey I, Anand A, Martinez KA, Katzan I, Rothberg MB. The Impact of Systematic Depression Screening in Primary Care on Depression Identification and Treatment in a Large Health Care System: A Cohort Study. *J GEN INTERN MED*. 2020;35(11):3141-3147. doi:10.1007/s11606-020-05856-5

21. Zulfiker MdS, Kabir N, Biswas AA, Nazneen T, Uddin MS. An in-depth analysis of machine learning approaches to predict depression. *Current Research in Behavioral Sciences*. 2021;2:100044. doi:10.1016/j.crbeha.2021.100044

22. Gautam S, Jain A, Gautam M, Vahia V, Grover S. Clinical Practice Guidelines for the management of Depression. *Indian J Psychiatry*. 2017;59(5):34. doi:10.4103/0019-5545.196973

23. Voineskos D, Daskalakis ZJ, Blumberger DM. Management of Treatment-Resistant Depression: Challenges and Strategies. *NDT*. 2020;Volume 16:221-234. doi:10.2147/NDT.S198774

24. Levkovitz Y, Tedeschini E, Papakostas GI. Efficacy of Antidepressants for Dysthymia: A Meta-Analysis of Placebo-Controlled Randomized Trials. *J Clin Psychiatry*. 2011;72(04):509-514. doi:10.4088/JCP.09m05949blu

25. Cuijpers P, Stringaris A, Wolpert M. Treatment outcomes for depression: challenges and opportunities. *The Lancet Psychiatry*. 2020;7(11):925-927. doi:10.1016/S2215-0366(20)30036-5

26. Saraceno B, Saxena S. Mental health resources in the world: results from Project Atlas of the WHO. *World Psychiatry*. 2002;1(1):40-44.

27. Nochaiwong S, Ruengorn C, Thavorn K, et al. Global prevalence of mental health issues among the general population during the coronavirus disease-2019 pandemic: a systematic review and meta-analysis. *Sci Rep*. 2021;11(1):10173. doi:10.1038/s41598-021-89700-8

28. Fairburn CG, Patel V. The impact of digital technology on psychological treatments and their dissemination. *Behaviour Research and Therapy*. 2017;88:19-25. doi:10.1016/j.brat.2016.08.012

29. StrÃ¶m M, Uckelstam CJ, Andersson G, HassmÃ©n P, Umefjord G, Carlbring P. Internet-delivered therapist-guided physical activity for mild to moderate depression: a randomized controlled trial. *PeerJ*. 2013;1:e178. doi:10.7717/peerj.178

30. Francis HM, Stevenson RJ, Chambers JR, Gupta D, Newey B, Lim CK. A brief diet intervention can reduce symptoms of depression in young adults – A randomised controlled trial. Matsuoka YJ, ed. *PLoS ONE*. 2019;14(10):e0222768. doi:10.1371/journal.pone.0222768

31. Schuster R, Fichtenbauer I, Sparr VM, Berger T, Laireiter AR. Data from: Feasibility of a blended group intervention (bGT) for major depression: uncontrolled interventional study in a university setting. Published online 2018:12872 bytes. doi:10.5061/DRYAD.3RP58

32. Ly KH, Topooco N, Cederlund H, et al. Smartphone-Supported versus Full Behavioural Activation for Depression: A Randomised Controlled Trial. Bockting CLH, ed. *PLoS ONE*. 2015;10(5):e0126559. doi:10.1371/journal.pone.0126559

33. Kambeitz-Ilankovic L, Rzayeva U, Völkel L, et al. A systematic review of digital and face-to-face cognitive behavioral therapy for depression. *npj Digit Med*. 2022;5(1):144. doi:10.1038/s41746-022-00677-8

34. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: A meta-analytic review. *Psychological Bulletin*. 2021;147(8):749-786. doi:10.1037/bul0000334

35. Garrido S, Millington C, Cheers D, et al. What Works and What Doesn't Work? A Systematic Review of Digital Mental Health Interventions for Depression and Anxiety in Young People. *Front Psychiatry*. 2019;10:759. doi:10.3389/fpsyt.2019.00759

36. Fu Z, Burger H, Arjadi R, Bockting CLH. Effectiveness of digital psychological interventions for mental health problems in low-income and middle-income countries: a systematic review and meta-analysis. *The Lancet Psychiatry*. 2020;7(10):851-864. doi:10.1016/S2215-0366(20)30256-X

37. Lattie EG, Adkins EC, Winquist N, Stiles-Shields C, Wafford QE, Graham AK. Digital Mental Health Interventions for Depression, Anxiety, and Enhancement of Psychological Well-Being Among College Students: Systematic Review. *J Med Internet Res*. 2019;21(7):e12869. doi:10.2196/12869

38. van Bronswijk SC, Bruijniks SJE, Lorenzo-Luaces L, et al. Cross-trial prediction in psychotherapy: External validation of the Personalized Advantage Index using machine learning in two Dutch randomized trials comparing CBT versus IPT for depression. *Psychotherapy Research*. 2021;31(1):78-91. doi:10.1080/10503307.2020.1823029

39. DeRubeis RJ, Cohen ZD, Forand NR, Fournier JC, Gelfand LA, Lorenzo-Luaces L. The Personalized Advantage Index: Translating Research on Prediction into Individualized Treatment Recommendations. A Demonstration. Cho WCS, ed. *PLoS ONE*. 2014;9(1):e83875. doi:10.1371/journal.pone.0083875

40. Roth SM. CLASSIFICATION OF AFFECTIVE AND RELATED PSYCHIATRIC DISORDERS. In: *Biological Aspects of Affective Disorders*. Elsevier; 1991:1-46. doi:10.1016/B978-0-12-356510-5.50007-8

41. Pichot P. [DSM-III: the 3d edition of the Diagnostic and Statistical Manual of Mental Disorders from the American Psychiatric Association]. *Rev Neurol (Paris)*. 1986;142(5):489-499.

42. Ebrahimi OV, Burger J, Hoffart A, Johnson SU. Within- and across-day patterns of interplay between depressive symptoms and related psychopathological processes: a dynamic network approach during the COVID-19 pandemic. *BMC Med*. 2021;19(1):317. doi:10.1186/s12916-021-02179-y

43. Fried EI, Flake JK, Robinaugh DJ. Revisiting the theoretical and methodological foundations of depression measurement. *Nat Rev Psychol*. 2022;1(6):358-368. doi:10.1038/s44159-022-00050-2

44. Lorenz N, Sander C, Ivanova G, Hegerl U. Temporal Associations of Daily Changes in Sleep and Depression Core Symptoms in Patients Suffering From Major Depressive Disorder: Idiographic Time-Series Analysis. *JMIR Ment Health*. 2020;7(4):e17071. doi:10.2196/17071

45. Wichers M, Smit AC, Snippe E. Early Warning Signals Based on Momentary Affect Dynamics can Expose Nearby Transitions in Depression: A Confirmatory Single-Subject Time-Series Study. *jpor*. 2020;6(1):1-15. doi:10.17505/jpor.2020.22042

46. Wichers M, Groot PC, Psychosystems, ESM Group, EWS Group. Critical Slowing Down as a Personalized Early Warning Signal for Depression. *Psychother Psychosom*. 2016;85(2):114-116. doi:10.1159/000441458

47. Fried EI, Nesse RM. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders*. 2015;172:96-102. doi:10.1016/j.jad.2014.10.010

48. Yang F, Wangen KR, Victor M, Solbakken OA, Holman PA. Referral assessment and patient waiting time decisions in specialized mental healthcare: an exploratory study of early routine collection of PROM (LOVePROM). *BMC Health Serv Res*. 2022;22(1):1553. doi:10.1186/s12913-022-08877-4

49. Garcia ME, Hinton L, Neuhaus J, Feldman M, Livaudais-Toman J, Karliner LS. Equitability of Depression Screening After Implementation of General Adult Screening in Primary Care. *JAMA Netw Open*. 2022;5(8):e2227658. doi:10.1001/jamanetworkopen.2022.27658

50. Bystritsky A, Khalsa SS, Cameron ME, Schiffman J. Current Diagnosis and Treatment of Anxiety Disorders. *P T*. 2013;38(1):30-57.

51. Mathers C, Fat DM, Boerma JT, World Health Organization, eds. *The Global Burden of Disease: 2004 Update*. World Health Organization; 2008.

52. Fava M, Kendler KS. Major Depressive Disorder. *Neuron*. 2000;28(2):335-341. doi:10.1016/S0896-6273(00)00112-4

53. Lader M. Generalized Anxiety Disorder. In: Stolerman IP, Price LH, eds. *Encyclopedia of Psychopharmacology*. Springer; 2015:699-702. doi:10.1007/978-3-642-36172-2_317

54. Bonari L, Pinto N, Ahn E, Einarson A, Steiner M, Koren G. Perinatal Risks of Untreated Depression during Pregnancy: *The Canadian Journal of Psychiatry*. Published online November 1, 2004. doi:10.1177/070674370404901103

55. Ghio L, Gotelli S, Cervetti A, et al. Duration of untreated depression influences clinical outcomes and disability. *Journal of Affective Disorders*. 2015;175:224-228. doi:10.1016/j.jad.2015.01.014

56. Beiter R, Nash R, McCrady M, et al. The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *Journal of Affective Disorders*. 2015;173:90-96. doi:10.1016/j.jad.2014.10.054

57. Scheffler R. Impact of Anxiety and Depression on Student Academic Progress. IBCCES. Published May 1, 2019. Accessed October 19, 2019. https://ibcces.org/blog/2019/05/01/impact-anxiety-depression-student-progress/

58. Alonso J, Liu Z, Evans-Lacko S, et al. Treatment gap for anxiety disorders is global: Results of the World Mental Health Surveys in 21 countries. *Depression and Anxiety*. 2018;35(3):195-208. doi:10.1002/da.22711

59. Weitzman ER. Poor Mental Health, Depression, and Associations With Alcohol Consumption, Harm, and Abuse in a National Sample of Young Adults in College. *The Journal of Nervous and Mental Disease*. 2004;192(4):269-277. doi:10.1097/01.nmd.0000120885.17362.94

60. Kessler D, Bennewith O, Lewis G, Sharp D. Detection of depression and anxiety in primary care: follow up study. *BMJ*. 2002;325(7371):1016-1017.

61. Kessler D, Lloyd K, Lewis G, Gray DP. Cross sectional study of symptom attribution and recognition of depression and anxiety in primary care. *BMJ*. 1999;318(7181):436-440.

62. Löwe B, Gräfe K, Zipfel S, et al. Detecting panic disorder in medical and psychosomatic outpatients: comparative validation of the Hospital Anxiety and Depression Scale, the Patient Health Questionnaire, a screening question, and physicians' diagnosis. *J Psychosom Res*. 2003;55(6):515-519. doi:10.1016/s0022-3999(03)00072-2

63. Kessler RC, Olfson M, Berglund PA. Patterns and Predictors of Treatment Contact After First Onset of Psychiatric Disorders. *AJP*. 1998;155(1):62-69. doi:10.1176/ajp.155.1.62

64. Thompson A, Issakidis C, Hunt C. Delay to Seek Treatment for Anxiety and Mood Disorders in an Australian Clinical Sample. *Behaviour Change*. 2008;25(2):71-84. doi:10.1375/bech.25.2.71

65. Trinh NHT, Youn SJ, Sousa J, et al. Using electronic medical records to determine the diagnosis of clinical depression. *International Journal of Medical Informatics*. 2011;80(7):533-540. doi:10.1016/j.ijmedinf.2011.03.014

66. Perlis RH, Iosifescu DV, Castro VM, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med*. 2012;42(1):41-50. doi:10.1017/S0033291711000997

67. Wang S, Pathak J, Zhang Y. Using Electronic Health Records and Machine Learning to Predict Postpartum Depression. Published 2019. Accessed December 4, 2019. http://ebooks.iospress.nl/publication/52116

68. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evid Based Mental Health*. 2017;20(3):83-87. doi:10.1136/eb-2017-102688

69. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. *J Am Med Inform Assoc*. 2014;21(6):1069-1075. doi:10.1136/amiajnl-2014-002733

70. Tran A, Tran L, Geghre N, et al. Health assessment of French university students and risk factors associated with mental health disorders. *PLOS ONE*. 2017;12(11):e0188187. doi:10.1371/journal.pone.0188187

71. Garla VN, Brandt C. Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics*. 2012;45(5):992-998. doi:10.1016/j.jbi.2012.04.010

72. Xu Y, Hong K, Tsujii J, Chang EIC. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *J Am Med Inform Assoc*. 2012;19(5):824-832. doi:10.1136/amiajnl-2011-000776

73. Stensland SH, Margolis S. Simplifying the calculation of body mass index for quick reference. *J Am Diet Assoc*. 1990;90(6):856.

74. MEANEY E, ALVA F, MOGUEL R, MEANEY A, ALVA J, WEBEL R. Formula and nomogram for the sphygmomanometric calculation of the mean arterial pressure. *Heart*. 2000;84(1):64. doi:10.1136/heart.84.1.64

75. Franklin Stanley S., Khan Shehzad A., Wong Nathan D., Larson Martin G., Levy Daniel. Is Pulse Pressure Useful in Predicting Risk for Coronary Heart Disease? *Circulation*. 1999;100(4):354-360. doi:10.1161/01.CIR.100.4.354

76. Salgado JF. Transforming the Area under the Normal Curve (AUC) into Cohen's d, Pearson's r pb , Odds-Ratio, and Natural Log Odds-Ratio: Two Conversion Tables. *The European Journal of Psychology Applied to Legal Context*. 2018;10(1):35-47. doi:10.5093/ejpalc2018a5

77. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9

78. Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. *NeuroImage: Clinical*. 2016;10:115-123. doi:10.1016/j.nicl.2015.11.003

79. Toenders YJ, van Velzen LS, Heideman IZ, Harrison BJ, Davey CG, Schmaal L. Neuroimaging predictors of onset and course of depression in childhood and adolescence: A systematic review of longitudinal studies. *Developmental Cognitive Neuroscience*. 2019;39:100700. doi:10.1016/j.dcn.2019.100700

80. Licht CMM, de Geus EJC, Seldenrijk A, et al. Depression Is Associated With Decreased Blood Pressure, but Antidepressant Use Increases the Risk for Hypertension. *Hypertension*. 2009;53(4):631-638. doi:10.1161/HYPERTENSIONAHA.108.126698

81. Defoe IN, Farrington DP, Loeber R. Disentangling the relationship between delinquency and hyperactivity, low achievement, depression, and low socioeconomic status: Analysis of repeated longitudinal data. *Journal of Criminal Justice*. 2013;41(2):100-107. doi:10.1016/j.jcrimjus.2012.12.002

82. Louise P, Siobhan O, Louise M, Jean G. The burden of generalized anxiety disorder in Canada. *Health Promot Chronic Dis Prev Can*. 2017;37(2):54-62.

83. Alegría AA, Hasin DS, Nunes EV, et al. Comorbidity of Generalized Anxiety Disorder and Substance Use Disorders: Results From the National Epidemiologic Survey on Alcohol and Related Conditions. *J Clin Psychiatry*. 2010;71(09):1187-1195. doi:10.4088/JCP.09m05328gry

84. Rissanen T, Viinamäki H, Lehto SM, et al. The role of mental health, personality disorders and childhood adversities in relation to life satisfaction in a sample of general population. *Nordic Journal of Psychiatry*. 2013;67(2):109-115. doi:10.3109/08039488.2012.687766

85. Everson SA, Maty SC, Lynch JW, Kaplan GA. Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. *Journal of Psychosomatic Research*. 2002;53(4):891-895. doi:10.1016/S0022-3999(02)00303-3

86. Härter MC, Conway KP, Merikangas KR. Associations between anxietydisorders and physical illness. *European Archives of Psychiatry and Clinical Neurosciences*. 2003;253(6):313-320. doi:10.1007/s00406-003-0449-y

87. Wu EL, Chien IC, Lin CH, Chou YJ, Chou P. Increased risk of hypertension in patients with major depressive disorder: A population-based study. *Journal of Psychosomatic Research*. 2012;73(3):169-174. doi:10.1016/j.jpsychores.2012.07.002

88. Cassano P, Fava M. Depression and public health: An overview. *Journal of Psychosomatic Research*. 2002;53(4):849-857. doi:10.1016/S0022-3999(02)00304-5

89. Thornicroft G, Chatterji S, Evans-Lacko S, et al. Undertreatment of people with major depressive disorder in 21 countries. *Br J Psychiatry*. 2017;210(2):119-124. doi:10.1192/bjp.bp.116.188078

90. Tylee A. Major depressive disorder (MDD) from the patient's perspective: overcoming barriers to appropriate care. *International Journal of Psychiatry in Clinical Practice*. 2001;5(1):37-42. doi:10.1080/13651500152048432

91. Garrido S, Millington C, Cheers D, et al. What Works and What Doesn't Work? A Systematic Review of Digital Mental Health Interventions for Depression and Anxiety in Young People. *Front Psychiatry*. 2019;10. doi:10.3389/fpsyt.2019.00759

92. Cheng P, Luik AI, Fellman-Couture C, et al. Efficacy of digital CBT for insomnia to reduce depression across demographic groups: a randomized trial. *Psychol Med*. 2019;49(3):491-500. doi:10.1017/S0033291718001113

93. Marcelle ET, Nolting L, Hinshaw SP, Aguilera A. Effectiveness of a Multimodal Digital Psychotherapy Platform for Adult Depression: A Naturalistic Feasibility Study. *JMIR mHealth and uHealth*. 2019;7(1):e10948. doi:10.2196/10948

94. Ström M, Uckelstam CJ, Andersson G, Hassmén P, Umefjord G, Carlbring P. Internet-delivered therapist-guided physical activity for mild to moderate depression: a randomized controlled trial. *PeerJ*. 2013;1:e178. doi:10.7717/peerj.178

95. Lawlor DA, Hopker SW. The effectiveness of exercise as an intervention in the management of depression: systematic review and meta-regression analysis of randomised controlled trials. *BMJ (Clinical research ed)*. 2001;322(7289):763-767. doi:10.1136/bmj.322.7289.763

96. Schuch FB, Deslandes AC, Stubbs B, Gosmann NP, Silva CTB da, Fleck MP de A. Neurobiological effects of exercise on major depressive disorder: A systematic review. *Neuroscience & Biobehavioral Reviews*. 2016;61:1-11. doi:10.1016/j.neubiorev.2015.11.012

97. Holzinger A. Trends in Interactive Knowledge Discovery for Personalized Medicine: Cognitive Science meets Machine Learning. Published online 2014:9.

98. Iniesta R, Stahl D, McGuffin P. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol Med*. 2016;46(12):2455-2465. doi:10.1017/S0033291716001367

99. Breiman L. Random Forest. *Machine Learning*. 2001;45(1):5-32. doi:10.1023/A:1010933404324

100. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM; 2016:785-794. doi:10.1145/2939672.2939785

101. Chekroud AM, Bondar J, Delgadillo J, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*. 2021;20(2):154-170. doi:10.1002/wps.20882

102. Fantino B, Moore N. The self-reported Montgomery-Åsberg depression rating scale is a useful evaluative tool in major depressive disorder. *BMC Psychiatry*. 2009;9(1):26. doi:10.1186/1471-244X-9-26

103. Bondolfi G, Jermann F, Rouget BW, et al. Self- and clinician-rated Montgomery–Åsberg Depression Rating Scale: Evaluation in clinical practice. *Journal of Affective Disorders*. 2010;121(3):268-272. doi:10.1016/j.jad.2009.06.037

104. Robert A. Steer, Aaron T. Beck. Beck Anxiety Inventory. In: *APA PsycInfo*. ; 1997:23-40.

105. Thomas Fydrich, Deborah Dowdall, Dianne L. Chambless. Reliability and validity of the beck anxiety inventory. *Journal of Anxiety Disorders*. 6(1):55-61. doi:https://doi.org/10.1016/0887-6185(92)90026-4

106. Mustafa Ulusoy, Nesrin H. Sahin, Husnu Erkmen. Turkish Version of the Beck Anxiety Inventory: Psychometric Properties. *Journal of Cognitive Psychotherapy*. 1998;12(2).

107. Gellman MD, Turner JR, eds. *Encyclopedia of Behavioral Medicine*. Springer; 2013.

108. CKW Schotte, M Maes, R Cluydts, D De Doncker, P Cosyns. Construct validity of the Beck Depression Inventory in a depressive population. *Journal of Affective Disorders*. 1997;46(2):115-125. doi:https://doi.org/10.1016/S0165-0327(97)00094-3

109. Kurtze N, Rangul V, Hustvedt BE. Reliability and validity of the international physical activity questionnaire in the Nord-Trøndelag health study (HUNT) population of men. *BMC Med Res Methodol*. 2008;8:63. doi:10.1186/1471-2288-8-63

110. Michael B Frisch, John Cornell, Michael Villanueva, Paul J Retzlaff. Clinical validation of the Quality of Life Inventory. A measure of life satisfaction for use in treatment planning and outcome assessment. *Psychological Assessment*. 1992;4(1):92-101. doi:https://doi.org/10.1037/1040-3590.4.1.92

111. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS One*. 2019;14(11):e0224365. doi:10.1371/journal.pone.0224365

112. Koul A, Becchio C, Cavallo A. Cross-Validation Approaches for Replicability in Psychology. *Front Psychol*. 2018;9:1117. doi:10.3389/fpsyg.2018.01117

113. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. ; 2017:4765-4774.

114. Uher R, Payne JL, Pavlova B, Perlis RH. MAJOR DEPRESSIVE DISORDER IN DSM-5: IMPLICATIONS FOR CLINICAL PRACTICE AND RESEARCH OF CHANGES FROM DSM-IV: Review: Major Depressive Disorder in DSM-5. *Depress Anxiety*. 2014;31(6):459-471. doi:10.1002/da.22217

115. Buchsbaum MS, Haier RJ. Biological Homogeneity, Symptom Heterogeneity, and the Diagnosis of Schizophrenia*. *Schizophrenia Bulletin*. 1978;4(4):473-475. doi:10.1093/schbul/4.4.473

116. Khan A, Faucett J, Lichtenberg P, Kirsch I, Brown WA. A Systematic Review of Comparative Efficacy of Treatments and Controls for Depression. Holscher C, ed. *PLoS ONE*. 2012;7(7):e41778. doi:10.1371/journal.pone.0041778

117. Wang F, DesMeules M, Luo W, Dai S, Lagace C, Morrison H. Leisure-time physical activity and marital status in relation to depression between men and women: A prospective study. *Health Psychology*. 2011;30(2):204-211. doi:10.1037/a0022434

118. Pettee KK, Brach JS, Kriska AM, et al. Influence of marital status on physical activity levels among older adults. *Med Sci Sports Exerc*. 2006;38(3):541-546. doi:10.1249/01.mss.0000191346.95244.f7

119. Broten LA, Naugle AE, Kalata AH, Gaynor ST. Depression and a Stepped Care Model. In: Draper C, O'Donohue WT, eds. *Stepped Care and E-Health*. Springer New York; 2011:17-43. doi:10.1007/978-1-4419-6510-3_2

120. Friedrich MJ. Depression Is the Leading Cause of Disability Around the World. *JAMA*. 2017;317(15):1517. doi:10.1001/jama.2017.3826

121. Cipriani A, Furukawa TA, Salanti G, et al. Comparative Efficacy and Acceptability of 21 Antidepressant Drugs for the Acute Treatment of Adults With Major Depressive Disorder: A Systematic Review and Network Meta-Analysis. *FOC*. 2018;16(4):420-429. doi:10.1176/appi.focus.16407

122. Saraceno B, van Ommeren M, Batniji R, et al. Barriers to improvement of mental health services in low-income and middle-income countries. *The Lancet*. 2007;370(9593):1164-1174. doi:10.1016/S0140-6736(07)61263-X

123. Moe-Byrne T, Shepherd J, Merecz-Kot D, et al. Effectiveness of tailored digital health interventions for mental health at the workplace: A systematic review of randomised controlled trials. König LM, ed. *PLOS Digit Health*. 2022;1(10):e0000123. doi:10.1371/journal.pdig.0000123

124. Moshe I, Terhorst Y, Philippi P, et al. Digital interventions for the treatment of depression: A meta-analytic review. *Psychological Bulletin*. 2021;147(8):749-786. doi:10.1037/bul0000334

125. Ljungberg T, Bondza E, Lethin C. Evidence of the Importance of Dietary Habits Regarding Depressive Symptoms and Depression. *Int J Environ Res Public Health*. 2020;17(5):1616. doi:10.3390/ijerph17051616

126. Grajek M, Krupa-Kotara K, Białek-Dratwa A, et al. Nutrition and mental health: A review of current knowledge about the impact of diet on mental health. *Front Nutr*. 2022;9:943998. doi:10.3389/fnut.2022.943998

127. Kalita N, Cooper K, Baird J, et al. Cost-effectiveness of a dietary and physical activity intervention in adolescents: a prototype modelling study based on the Engaging Adolescents in Changing Behaviour (EACH-B) programme. *BMJ Open*. 2022;12(8):e052611. doi:10.1136/bmjopen-2021-052611

128. Grosso G, Marventano S, Buscemi S, et al. Factors associated with adherence to the Mediterranean diet among adolescents living in Sicily, Southern Italy. *Nutrients*. 2013;5(12):4908-4923. doi:10.3390/nu5124908

129. Francis H, Stevenson R. Validity and test-retest reliability of a short dietary questionnaire to assess intake of saturated fat and free sugars: a preliminary study. *J Hum Nutr Diet*. 2013;26(3):234-242. doi:10.1111/jhn.12008

130. Ng F, Trauer T, Dodd S, Callaly T, Campbell S, Berk M. The validity of the 21-item version of the Depression Anxiety Stress Scales as a routine clinical outcome measure. *Acta Neuropsychiatr*. 2007;19(5):304-310. doi:10.1111/j.1601-5215.2007.00217.x

131. Henry SK, Grant MM, Cropsey KL. Determining the optimal clinical cutoff on the CES-D for depression in a community corrections sample. *Journal of Affective Disorders*. 2018;234:270-275. doi:10.1016/j.jad.2018.02.071

132. Grinsztajn L, Oyallon E, Varoquaux G. Why do tree-based models still outperform deep learning on tabular data? Published online 2022. doi:10.48550/ARXIV.2207.08815

133. Papini S, Pisner D, Shumake J, et al. Ensemble machine learning prediction of posttraumatic stress disorder screening status after emergency room hospitalization. *Journal of Anxiety Disorders*. 2018;60:35-42. doi:10.1016/j.janxdis.2018.10.004

134. Nemesure MD, Heinz MV, Huang R, Jacobson NC. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning

approach with artificial intelligence. *Sci Rep*. 2021;11(1):1980. doi:10.1038/s41598-021-81368-4

135. Gyorda JA, Nemesure MD, Price G, Jacobson NC. Applying ensemble machine learning models to predict individual response to a digitally delivered worry postponement intervention. *Journal of Affective Disorders*. 2023;320:201-210. doi:10.1016/j.jad.2022.09.112

136. Musoro JZ, Zwinderman AH, Puhan MA, ter Riet G, Geskus RB. Validation of prediction models based on lasso regression with multiply imputed data. *BMC Med Res Methodol*. 2014;14(1):116. doi:10.1186/1471-2288-14-116

137. Lundberg S, Lee SI. A Unified Approach to Interpreting Model Predictions. Published online 2017. doi:10.48550/ARXIV.1705.07874

138. Cuijpers P, Ebert DD, Acarturk C, Andersson G, Cristea IA. Personalized Psychotherapy for Adult Depression: A Meta-Analytic Review. *Behavior Therapy*. 2016;47(6):966-980. doi:10.1016/j.beth.2016.04.007

139. Venkatesan A, Forster B, Rao P, Miller M, Scahill M. Improvements in Depression Outcomes Following a Digital Cognitive Behavioral Therapy Intervention in a Polychronic Population: Retrospective Study. *JMIR Form Res*. 2022;6(7):e38005. doi:10.2196/38005

140. Weiss RB, Aderka IM, Lee J, Beard C, Björgvinsson T. A Comparison of Three Brief Depression Measures in an Acute Psychiatric Population: CES-D-10, QIDS-SR, and DASS-21-DEP. *J Psychopathol Behav Assess*. 2015;37(2):217-230. doi:10.1007/s10862-014-9461-y

141. Henry JD, Crawford JR. The short-form version of the Depression Anxiety Stress Scales (DASS-21): Construct validity and normative data in a large non-clinical sample. *British Journal of Clinical Psychology*. 2005;44(2):227-239. doi:10.1348/014466505X29657

142. Radloff LS. The CES-D Scale: A Self-Report Depression Scale for Research in the General Population. *Applied Psychological Measurement*. 1977;1(3):385-401. doi:10.1177/014662167700100306

143. Steel Z, Marnane C, Iranpour C, et al. The global prevalence of common mental disorders: A systematic review and meta-analysis 1980-2013. *International journal of epidemiology*. 2014;43(2):476-493. doi:10.1093/ije/dyu038

144. Wilhelm S, Weingarden H, Ladis I, Braddick V, Shin J, Jacobson NC. Cognitive-Behavioral Therapy in the Digital Age: Presidential Address. *Behavior Therapy*. Published online August 8, 2019. doi:10.1016/j.beth.2019.08.001

145. Lenhard F, Sauer S, Andersson E, et al. Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach. *Int J Methods Psychiatr Res*. 2018;27(1). doi:10.1002/mpr.1576

146. Pearson R, Pisner D, Meyer B, Shumake J, Beevers CG. A machine learning ensemble to predict treatment outcomes following an Internet intervention for depression. *Psychol Med*. 2019;49(14):2330-2341. doi:10.1017/S003329171800315X

147. Månsson KNT, Frick A, Boraxbekk CJ, et al. Predicting long-term outcome of Internet-delivered cognitive behavior therapy for social anxiety disorder using fMRI and support vector machine learning. *Transl Psychiatry*. 2015;5(3):e530-e530. doi:10.1038/tp.2015.22

148. Zwerenz R, Becker J, Gerzymisch K, et al. Evaluation of a transdiagnostic psychodynamic online intervention to support return to work: A randomized controlled trial. *PLoS ONE*. 2017;12(5):e0176513. doi:10.1371/journal.pone.0176513

149. Smith K. Mental health: A world of depression. *Nature*. 2014;515(7526):180-181. doi:10.1038/515180a

150. Cramer AOJ, van Borkulo CD, Giltay EJ, et al. Major Depression as a Complex Dynamic System. Branchi I, ed. *PLoS ONE*. 2016;11(12):e0167490. doi:10.1371/journal.pone.0167490

151. Fried EI, Nesse RM. Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. *BMC Med*. 2015;13(1):72. doi:10.1186/s12916-015-0325-4

152. Beard C, Millner AJ, Forgeard MJC, et al. Network analysis of depression and anxiety symptom relationships in a psychiatric sample. *Psychol Med*. 2016;46(16):3359-3369. doi:10.1017/S0033291716002300

153. Ebrahimi OV, Burger J, Hoffart A, Johnson SU. Within- and across-day patterns of interplay between depressive symptoms and related psychopathological processes: a dynamic network approach during the COVID-19 pandemic. *BMC Med*. 2021;19(1):317. doi:10.1186/s12916-021-02179-y

154. Bringmann LF, Ferrer E, Hamaker EL, Borsboom D, Tuerlinckx F. Modeling Nonstationary Emotion Dynamics in Dyads using a Time-Varying Vector-Autoregressive Model. *Multivariate Behavioral Research*. 2018;53(3):293-314. doi:10.1080/00273171.2018.1439722

155. Jordan DG, Winer ES, Salem T. The current status of temporal network analysis for clinical science: Considerations as the paradigm shifts? *J Clin Psychol*. 2020;76(9):1591-1612. doi:10.1002/jclp.22957

156. Epskamp S, Waldorp LJ, Mõttus R, Borsboom D. The Gaussian Graphical Model in Cross-Sectional and Time-Series Data. *Multivariate Behavioral Research*. 2018;53(4):453-480. doi:10.1080/00273171.2018.1454823

157. de Vos S, Wardenaar KJ, Bos EH, Wit EC, Bouwmans MEJ, de Jonge P. An investigation of emotion dynamics in major depressive disorder patients and healthy persons using sparse longitudinal networks. Waldorp LJ, ed. *PLoS ONE*. 2017;12(6):e0178586. doi:10.1371/journal.pone.0178586

158. Kaiser T, Laireiter AR. Process-symptom-bridges in psychotherapy: an idiographic network approach. *JPOR*. Published online December 27, 2018:49-62. doi:10.17505/jpor.2018.06

159. Wichers M, Riese H, Hodges TM, Snippe E, Bos FM. A Narrative Review of Network Studies in Depression: What Different Methodological Approaches Tell Us About Depression. *Front Psychiatry*. 2021;12:719490. doi:10.3389/fpsyt.2021.719490

160. Bringmann LF, Lemmens LHJM, Huibers MJH, Borsboom D, Tuerlinckx F. Revealing the dynamic network structure of the Beck Depression Inventory-II. *Psychol Med*. 2015;45(4):747-757. doi:10.1017/S0033291714001809

161. Haslbeck JMB, Bringmann LF, Waldorp LJ. A Tutorial on Estimating Time-Varying Vector Autoregressive Models. *Multivariate Behavioral Research*. 2021;56(1):120-149. doi:10.1080/00273171.2020.1743630

162. Lütkepohl H. *New Introduction to Multiple Time Series Analysis*. New York : Springer; 2005.

163. Hastie T, Tibshirani R. Generalized additive models for medical research. *Stat Methods Med Res*. 1995;4(3):187-196. doi:10.1177/096228029500400302

164. Siepe BS, Sander C, Schultze M, et al. Temporal dynamics of depressive symptomatology: An idiographic time series analysis applying network models to patients with depressive disorders. *PsyArXiv*. Published online 2022. doi:10.31234/osf.io/hnw69

165. Wang R, Chen F, Chen Z, 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; 2014:3-14. doi:10.1145/2632048.2632054

166. Kroenke K, Spitzer RL. The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*. 2002;32(9):509-515. doi:10.3928/0048-5713-20020901-06

167. Torous J, Staples P, Shanahan M, et al. Utilizing a Personal Smartphone Custom App to Assess the Patient Health Questionnaire-9 (PHQ-9) Depressive Symptoms in Patients With Major Depressive Disorder. *JMIR Ment Health*. 2015;2(1):e8. doi:10.2196/mental.3889

168. Wood SN. Mixed GAM computation vehicle with automatic smoothness estimation. Published online 2017.

169. Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*. 2007;94(1):19-35. doi:10.1093/biomet/asm018

170. Bringmann LF. Person-specific networks in psychopathology: Past, present, and future. *Current Opinion in Psychology*. 2021;41:59-64. doi:10.1016/j.copsyc.2021.03.004

171. Bringmann LF, Hamaker EL, Vigo DE, Aubert A, Borsboom D, Tuerlinckx F. Changing dynamics: Time-varying autoregressive models using generalized additive modeling. *Psychological Methods*. 2017;22(3):409-425. doi:10.1037/met0000085

172. Fisher AJ. Toward a dynamic model of psychological assessment: Implications for personalized care. *Journal of Consulting and Clinical Psychology*. 2015;83(4):825-836. doi:10.1037/ccp0000026

173. Jacobson NC, Nemesure MD. Using artificial intelligence to predict change in depression and anxiety symptoms in a digital intervention: Evidence from a transdiagnostic randomized controlled trial. *Psychiatry Research*. 2021;295:113618. doi:10.1016/j.psychres.2020.113618

174. Wilhelm S, Weingarden H, Ladis I, Braddick V, Shin J, Jacobson NC. Cognitive-behavioral therapy in the digital age: Presidential address. *Behavior Therapy*. 2020;51(1):1-14. doi:10.1016/j.beth.2019.08.001

175. Teepe GW, Da Fonseca A, Kleim B, et al. Just-in-Time Adaptive Mechanisms of Popular Mobile Apps for Individuals With Depression: Systematic App Search and Literature Review. *J Med Internet Res*. 2021;23(9):e29412. doi:10.2196/29412

176. Wang L, Miller LC. Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review. *Health Communication*. 2020;35(12):1531-1544. doi:10.1080/10410236.2019.1652388

177. statcounter. Mobile Operating System Market Share United States Of America. Published 2022. Accessed October 19, 2022. https://gs.statcounter.com/os-market-share/mobile/united-states-of-america

178. Reichert A, Jacobs R. The impact of waiting time on patient outcomes: Evidence from early intervention in psychosis services in England. *Health Economics*. 2018;27(11):1772-1787. doi:10.1002/hec.3800