MAKING AUGMENTED HUMAN INTELLIGENCE IN MEDICINE PRACTICAL: A CASE STUDY OF TREATING MAJOR DEPRESSIVE DISORDER

BY

ARJUN PRASANNA ATHREYA

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Professor Ravishankar Iyer, Chair
Professor Wen-Mei Hwu
Professor Gene Robinson
Professor William Sanders
Professor Dan Roth, University of Pennsylvania
Professor Richard Weinshilboum, MD, Mayo Clinic

ABSTRACT

Individualized medicine tailors diagnoses and treatment options on an individual patient basis. This is a paradigm shift from choosing a treatment based on highest reported efficacy in clinical trials, which is often not effective for all individuals. In this dissertation, we assert that treatment selection and management can be individualized when clinicians assessment of disease symptoms are augmented with a few analytically identified patient-specific measures (e.g., genomics, metabolomics) that are prognostic or predictive of treatment outcomes. Patient-derived biological, clinical and symptom measures are sufficiently complex, i.e., heterogeneous, noisy and high-dimensional. The question for research then becomes: "Which few among these large complex measures are sufficient to augment the clinician's disease assessment and treatment logic to individualize treatment decisions?"

This dissertation introduces, *ALMOND* — Analytics and Machine Learning Framework for Actionable Intelligence from Clinical and Omics Data. As a case study, this dissertation describes how ALMOND addresses the unmet need for individualized medicine in treating major depressive disorder — the leading cause of medical disabilities worldwide. The biggest challenge in individualizing treatment of depression is in the heterogeneity of how depressive symptoms manifest between individuals, and in their varied response to the same treatment.

ALMOND comprises a systematic analytical workflow to individualize antidepressant treatment by addressing the challenge of heterogeneity of major depressive disorder. First, "right patients" are identified by stratifying patients using unsupervised learning, that serves as a foundation to associate their disease states with multiple pharmacological (drug-associated) measures. Second, "right drug" selection is shown to be feasible by demonstrating that psychiatrists' depression severity assessments augmented with pharmacogenomic measures can accurately predict remission of depressive symptoms using supervised learning. Finally, probabilistic graphs provide early and easily interpretable prognoses at the "right time" to a psychiatrist by accounting for changes in routinely assessed depressive symptoms' severity. By choosing antidepressants that have the highest-likelihood of the patient achieving remission, the chances of persisting depressive symptoms are reduced, which is often the leading medical conditions in those who commit suicide or develop chronic illnesses.

To my parents, Vandana Prasanna and Prasanna S. Bidare.
For their limitless love, blessings and wonderful upbringing.

ACKNOWLEDGMENTS

I want to begin by thanking my mentor and adviser Professor Ravishankar K. Iyer (Ravi) for his complete support and confidence in my abilities and research interests. Ravi infused me with high levels of self-confidence that allowed me to think big, and to articulate and present research that looks at the bigger picture of our work, and extends beyond the scope of projects we undertaken for this dissertation. Additionally, Ravi always tells me that "we can agree to disagree in this office, but we both have very complementary thoughts that could make this work very strong," which allowed us to bring the best of both our thoughts in our research efforts. Besides his academic advising, Ravi and his wife Pamela Iyer, have offered wise counsel on issues I encountered in my personal life, and have been wonderful individuals I can count on in my life's journey beyond research.

I thank Professor R. K. Shyamasundar for introducing me to Ravi, and Heidi Leerkamp (our program manager) for scheduling a telephone meeting within a day of receiving the introduction. Without the timely connection with Ravi, I would not have come to Illinois.

My pursuit of multidisciplinary research to address complex and globally relevant problems like the one presented in this work reflects the vision and belief in "team science" encouraged by the University of Illinois at Urbana-Champaign (UIUC) and Mayo Clinic as institutions. My graduate school journey began with Professor Gene Robinson at the Institute for Genomic Biology (IGB), who hosted me in his lab meetings so that I could begin to articulate computational problems in life science research. I am very thankful for the support from UIUC Provost Professor Andreas Cangellaris (then Dean of the College of Engineering at UIUC) and the Advanced Digital Sciences Center in Singapore that allowed me to conduct some of this dissertations' work under Professor C. N. Lee's supervision at the National University Hospital, Singapore.

In multidisciplinary collaborative research, patience and camaraderie are needed among investigators to allow their trainees to thrive and deliver on groundbreaking research promises. In this context, I am all the more thankful for the friendship and camaraderie among Professors Ravi Iyer, Ditlev Monrad, Gene Robinson, Tamer Başar, Zbigniew Kalbarczyk, Janak Patel, Wen-mei Hwu, C. N. Lee, Rashid Bashir, William H. Sanders, Dan Roth, and Derek Wildman

and Drs. Richard Weinshilboum, Liewei Wang, William Bobo, and Konstantinos Lazaridis, each of whom taught me different facets of science and technology, that made me confident in addressing challenges in both fields. Regular interactions with all these faculty and their laboratory members allowed me to place our research in better context. I thank my mentors and friends in the technical industry, including Alexander Clemm, Rajeev Koodli, Jonathan DeMent, Pete Hofstee, Brad McCredie, Venky Ananth, Rajeshwari Ganesan, and Guru Deshpande, who offered excellent suggestions to make our work "industry-ready." I thank Drs. Elisabeth Binder, Mark Frye, and A. John Rush for helping me add more specificity to the clinical implications of our work.

To modify the proverb "it takes a village to raise a child," it took three villages to help complete my graduate school journey. First, my friends and staff in Urbana-Champaign, IL in the DEPEND and PERFORM groups, "swim for fitness," CSL (especially Mahanth, Nirupam, and Debjit), and the IGB have been wonderful for my overall development as an individual and researcher. Special thanks to Zachary Estrada and Cuong Pham, who were instrumental in improving my presentation skills, Don Armstrong for improving my figure-making skills, and Subho Banerjee for his critiques. Sincere thanks to Saurabh, Zach, Atul, Uttam, and Priya & Nikhil for hosting me several times for supper. Second, thanks to Mrs. Miriam Anderson, Chris Schad, and Drs. Wang, Weinshilboum, and Rani Kalari at Mayo Clinic in Rochester, MN for hosting me and welcoming me into the community. Third, I thank the Bobo family in Jacknsonville, FL for all their love and support during my visits to Mayo Clinic, Florida. I thank Carol Bosley, Donna Foley, Pauline Wee, Kathleen Atchley, and Carol Wisniewski for their help and administrative support, and Jenny Applequist in helping refine our manuscripts and proposals.

This dissertation would not have had aspects of clinical validity if Drs. Wang, Weinshilboum, and Bobo had not spent countless hours with me on weekends and holidays, and after work hours to review progress and provide domain knowledge to complement my engineering efforts. I am eternally thankful to Dr. Bobo for having spent that valuable time with me despite his hectic transition from Mayo Clinic, Minnesota to become Chair of the Department of Psychiatry and Psychology at Mayo Clinic, Florida. One special person who made it possible for Dr. Bobo to spend that much time with me was his wife, Tammy Bobo. She valued the commitment Dr. Bobo and our entire team were putting into an important cause, and never complained if I ever called Dr. Bobo at odd hours of the day to seek some clarifications. Tammy always brought us food when we worked long hours, and encouraged time-outs, during which we walked their lovely dogs on Jacksonville beaches while discussing a wide range of non-scientific topics. Similarly, in Champaign, Ravi's wife, Pamela Iyer encouraged some lovely discussions that allowed me to gain broad perspectives on a wide range of topics.

This friendship that evolved over time with the Bobo and Iyer families is special, and I am thankful for their encouragement and support.

My loving friends Saurabh Jha, Pavithra & Abhiram, Kaushik Nagraj, Anjali & Ajith, Subashini & Vinod, Dhanashree & Aditya, Shilpa Nandakumar, Nanda Kishore and Maruja Yoshimura ensured I was never alone in this graduate school journey. Sincere thanks to Saurabh, who has never said no to stepping out for coffee or lunch breaks with me to get a break from work. Subashini & Vinod's generous hospitality and affection made it possible for me to decompress on long weekends and holidays during my entire time at UIUC. I thank Uttam for all the wonderful discussions and laughter during our cooking ventures, and Keywhan Chung and Phuong Cao for their countless rides to various places in Champaign.

While I have been away in the United States, the families of Surya Prakash M. S., R.R. Purushotham, B. S. Shankarnarayan, K.S.G. Shankar, Dr. N. Suresh, Dr. Baliga, and Dr. H.V. Shankar, our neighbors in Bangalore, and all my mother's students have been extremely supportive and kind to me and my parents. I thank Arti & Vijay for their wonderful hospitality during my visits to Singapore.

I am grateful for the funding support as the material presented in this dissertation is based upon work partially supported by a Mayo Clinic and Illinois Alliance Fellowship for Technology-Based Healthcare Research; a CompGen Fellowship; an IBM Faculty Award; the National Science Foundation (NSF) under grants CNS 13-37732, CNS 16-24790, and CNS 16-24615; the National Institutes of Health (NIH) under grants U19 GM61388, R01 GM28157, R01 MH108348, RC2 GM092729, R24 GM078233, RC2 GM092729, and T32 GM072474; and the Mayo Clinic Center for Individualized Medicine. Any opinions, findings, and conclusions or recommendations expressed in this material are those of my own and those of my co-authors on manuscripts drawn on for this dissertation, and they do not necessarily reflect the views of the funding bodies or institutions. We thank all patients for consenting to the use of their data for research.

Finally, none of this work could have been possible without all the constant love, prayers, and wishes from my parents Vandana and Prasanna, based in Bangalore, India. They have loved and supported me during my highs and lows, and have always offered excellent suggestions and counsel whenever I needed them. I could not have asked for a better set of individuals as my parents. As a family, my parents and I are thankful for the blessings of our Gurus, especially His Holiness Shankara Bharathi, who has prayed for the well-being and prosperity of our family, our friends, and our colleagues so that we are a happy community.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Analytics for Augmenting Human Intelligence in Medicine

Individualized medicine tailors diagnoses and treatment options on an individual patient basis [1–4]. This is a paradigm shift from choosing a treatment based on aggregate therapeutic efficacy (i.e., percentage of patients achieving the desired therapeutic benefit in clinical trials). Individualized treatment is greatly needed because not all patients will respond to a particular drug, and drug response varies by patient [1].

Reduction of costs in assaying patients' genetic makeup (genome) among other biological measures (e.g., hormones) has accelerated advances in individualized medicine [5]. However, there are only a few well-characterized systemic diseases for which dramatically improved treatments have been achieved by using genomics data in routine clinical settings (e.g., targeted therapies in breast cancer) [1, 6]. To expand the benefit of individualized medicine to other widely prevalent complex diseases (e.g., mental health disorders and migraines), a key question to address is: "Can a clinician's assessments of disease severity, augmented with patient-specific biological measures, help identify a therapeutic agent (e.g., drug) with the highest likelihood of achieving the desired therapeutic benefit?"

The answer to that question needs high-quality data, rich clinical insights/annotations, and analytical approaches to combine heterogeneous patient-specific measures to generate "actionable intelligence." *Actionable intelligence* is a small set of patient-derived biological/clinical measures that provide clinicians novel insights on (1) treatment prognoses, (2) prediction of therapeutic efficacy, and (3) disease pathophysiology or drug mechanisms. Analytical approaches for generating actionable intelligence have to augment clinician's treatment selection criteria prior to the initiation of treatment, and during intermediate follow-ups that span the entire duration of the treatment. Therefore, methods have to account for static measures (e.g., genomics) and time-varying measures with conditional properties (e.g., symptom improvement in response to treatment intervention). Given the widely varying differences in disease manifestations and treatment practices, there are no universal theoretical foundations for generating relevant actionable intelligence [7].

1.2 ALMOND: Analytics and Machine Learning Framework for Actionable Intelligence from Clinical and Omics Data

This dissertation introduces *ALMOND*, the Analytics and Machine Learning Framework for Actionable Intelligence from Clinical and Omics Data. By combining patient-derived biological measures with rich insights provided by physicians (domain experts) as illustrated in Fig. 1.1, ALMOND addresses the broader clinically important question: "Is a given therapeutic agent (e.g., drug) effective for an individual?" The broader analytical challenge in addressing this question is in how ALMOND can augment a physician's symptom assessment (i.e., knowledge of disease physiology) with a few biological measures that are associated with treatment outcomes. Then the physician can match a patient with a treatment strategy that has the highest likelihood of enabling the patient to achieve the desired therapeutic benefit. Hence, predicting or forecasting treatment outcomes by combining a physician's assessments with patient-derived biological measures has to be conditioned upon (1) the state of the disease prior to treatment initiation, (2) patient history, and (3) how symptoms change between disease states in response to treatment.

ALMOND formalizes conditional dependencies between symptom variations during treatment by using probabilistic graphical models, as shown in Fig. 1.1. Using hidden Markov models, ALMOND models symptom variations (visible to the physician) in response to drug treatment in strata of patients identified at each time-point of the treatment. Then, a patient's eventual treatment response trajectories can be inferred using optimization techniques such as the Viterbi algorithm, for example. However, the challenge is that not all diseases have well-defined disease states from either a symptomatic or a biological perspective. In addressing this challenge, ALMOND uses unsupervised learning methods to learn disease states based on symptom assessments that serve as the foundation for developing a graph, and the probabilistic transitions between the states (i.e., nodes of the graph) are learned from input data. The inferred states can then be engineered to jointly consider clinical assessments and biological measures to derive accurate prognoses or predictions of treatment outcomes. Thus, in answering the question of whether a given therapeutic agent (e.g., drug) will be effective for an individual, this dissertation through the development of ALMOND shows that systematic data-driven innovations are needed that bring multiple analytical approaches together in a workflow. Finally, the success of machine learning or artificial intelligence approaches in healthcare science and delivery rests in the ability of physicians to relate their diagnoses to underpinnings of analytical methods to the degree that prediction results should be explainable with minimal complexity and maximum precision. For example, "in the patient's transition from disease state A to disease state B after treatment initiation, since X

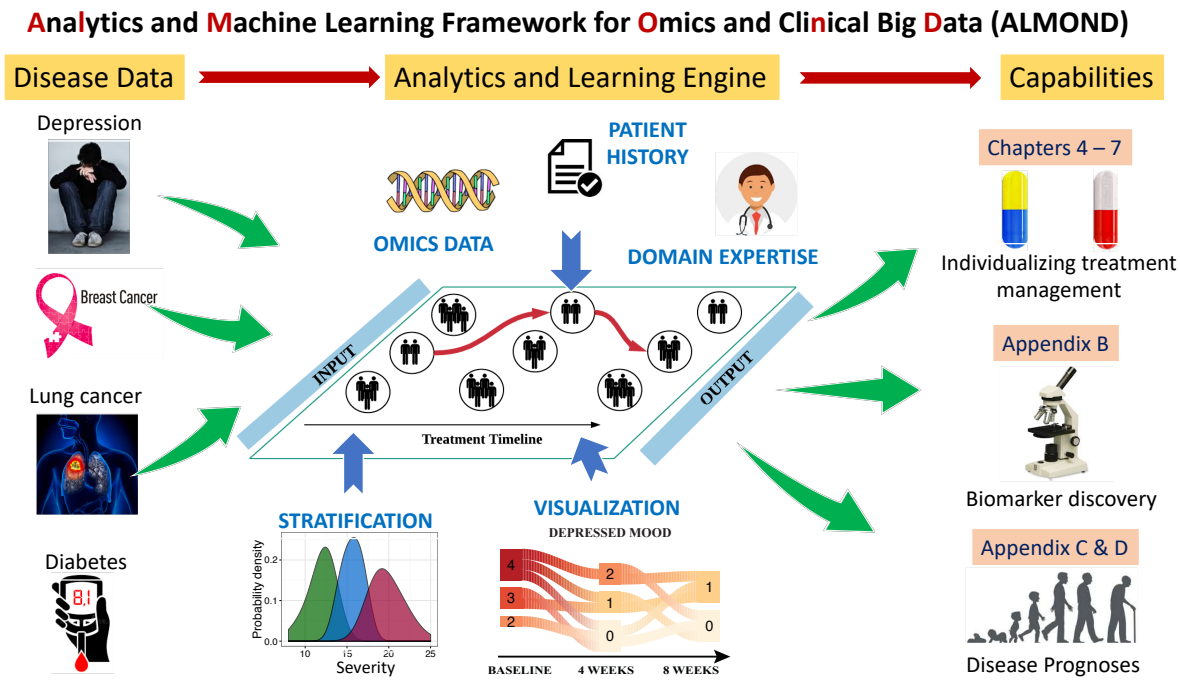


Figure 1.1: Analytics and Machine Learning Framework for Actionable Intelligence from Clinical and Omics Data (ALMOND).

symptoms have improved by $Y\%$, the likely treatment outcome is symptom remission with a $Z\%$ probability."

ALMOND is capable of analyzing data that range from the granularity of expressions of the human genome within each cell, to population-level diagnostic data, in order to generate actionable intelligence as illustrated in Fig. 1.1. The generated intelligence demonstrates the ability of analytical innovations to drive biomedical research for both discovery and translational science, which are important facets of individualizing medicine. The following are the capabilities of ALMOND that are embodied in its unique workflows, each of which brings to fore, a combination of probabilistic graphical models, and unsupervised and supervised learning methods.

1. **Prediction, and Prognoses of Treatment Outcomes:** By using a patient's genomic and longitudinal symptom assessment data, ALMOND can predict the outcome of antidepressant treatment prior to its initiation; it can also provide a psychiatrist with a prognosis of the long-term categorical outcome based on early changes in disease symptom severity after treatment initiation. The workflow is discussed in Chapters 4 – 7.

2. **Novel Drug Mechanisms:** A genome's biological functions in drug mechanisms can be studied in terms of variations in the gene expression levels (referred to collectively as

the *transcriptome*) [8]. Because the costs of sequencing the genome within each cell have been going down, it is now possible to identify groups of cells that react differently to drug treatment [9]. Such a grouping of cells is based on differential patterns of expressions of genes across cells. The mixture-model-based single-cell analysis (MiMoSA [10]) workflow (described in Appendix B) takes as input,the untreated and drug-treated single-cell transcriptome data and identifies genes whose expression is strikingly affected by drug treatment. As a case study, MiMoSA identified a gene *CDC42* as a candidate for laboratory experiments to study the anti-cancer properties of the commodity diabetic drug metformin. The laboratory experiments demonstrated a novel mechanism by which metformin inhibited cancer cell migration in triple-negative breast cancer patients [11].

3. **Disease Prognoses:** Transcriptome variations of genes with competing biological functions have the potential to explain the development of deadly diseases such as cancer. The game-theoretic transcriptome analysis (GiTA [12]) workflow (described in Appendix C) takes as input a patient's transcriptome data and describes the likelihood of disease development. For that, non-small cell lung adenocarcinoma was used as a case study.

4. **Forecasting Readmission for Surgery:** For aging patients with chronic diseases such as type II diabetes, frequent surgical interventions can lower quality of life and greatly increase care costs. The Singapore diabetes readmission graphical network (SINGA-DRAGN [13]) workflow (described in Appendix D) processes population electronic health records of diabetic patients in Singapore, first to identify longitudinal relationships between comorbidities that, when poorly controlled, warrant surgical interventions. Then, for a new patient's current diagnoses, SINGA-DRAGn can forecast the comorbid complexities that are likely to warrant surgical interventions in the future.

Validation Considerations: The following measures were taken to explain the biological significance of any biomarkers identified or predictive models developed in ALMOND. ALMOND-identified biomarker candidates were studied in the laboratory at Mayo Clinic in order to establish their biological relevance in drug mechanisms. Independent trial datasets (not used in training) were used to validate the performance of the predictive methods.

An important feature of ALMOND is its usability through a clinician-friendly interface. Data, results, and associated statistical significance information are presented, and are illustrated in a way that feeds naturally into a clinician's treatment guidelines. As a case study, this dissertation describes how ALMOND addresses the unmet need for individualized medicine in treating major depressive disorder.

1.3 Major Depressive Disorder and Treatment Challenges

Major Depressive Disorder (MDD) is the number-one psychiatric disease worldwide, and is expected to the be leading cause of medical disabilities by 2030 [14, 15]. Globally, 300 million people are affected by MDD, regardless of age, gender, ethnicity, race, or economic status, among other sociodemographic factors. MDD is a complex disease characterized by several symptoms. Common depressive symptoms are melancholy, loss of pleasure and energy, appetite variations, guilt feelings and delusions, inability to concentrate or sleep well, and the presence of suicidal thoughts. MDD severity is measured by summing the patient's categorical responses to individual symptom questionnaires assessed by a psychiatrist using a standardized depression rating scale. The higher the score, the higher the severity of the disease. MDD is treatable with antidepressant medication, psychotherapy or a combination of both [14].

MDD is a heterogeneous disease in the sense that depressive symptoms often manifest differently in different patients, even within the same family or community. While MDD is a heterogeneous disease, antidepressant treatment outcomes are also heterogeneous [16–20]. That is, patients with the same sociodemographic factors and pre-treatment MDD severity see different outcomes in response to the same dose and duration of antidepressant treatment.

Unlike biologically quantitative and distinct evidence of the presence/absence of cancer in a tissue biopsy, the marked heterogeneity in patient-reported MDD symptom severity has limited the ability to find genetic associations with disease or antidepressant response. Hence, to date, there is no mapping between a patient's genomic data and antidepressants that can help in choosing the medication with highest chance of achieving desired treatment benefit. In the absence of quantitative biological evidence to guide antidepressant treatment selection, current practice follows what Dr. Roy Perlis (a reknowned psychiatrist) describes as "artisanal medicine" [21]. By relying largely on a psychiatrist's experience, on a "try-and-try-again" basis, patients may go through several trials of antidepressants before they eventually achieve remission from MDD symptoms. Recent studies have shown that over 70% of MDD patients undergo 5 – 7 trials of antidepressants each lasting 8 – 12 weeks before they achieve remission from MDD symptoms [22]. With each failure of an antidepressant medication, persisting MDD symptoms hinder patients from functioning normally, and some patients commit suicide. Clearly, these patients need help and deserve the kind of precision currently achieved in individualized breast cancer therapeutics [23].

1.4 Toward Individualizing Antidepressant Treatment Management in Depressed Adults

1.4.1 ALMOND's Design Driver: Real-World Patient Data

With the intent of making ALMOND-generated "actionable intelligence" augment clinicians' treatment know-how in routine practice, ALMOND has been developed with data from over $3,000$ consenting patients with MDD. The data are richly characterized by diversity in race and geographical location (e.g., United States, Scandinavian Europe, East Asia), patient-derived biological measures (e.g., genomics, metabolomics, blood drug levels) and clinical assessments during the course of antidepressant treatments lasting at least 8 weeks. Furthermore, data were collected at three time-points, i.e., baseline (pre-treatment), at 4 and 8 weeks, and up to 30 weeks after treatment initiation. Data from trials conducted in both outpatient settings (i.e., no hospital stay) and inpatient settings (i.e., extended hospital stay for extremely sick patients with MDD and other severe chronic conditions) are used to develop and test the robustness of insights gained using ALMOND. Data used in this study are pooled from both single-site trials (i.e., all patients seen by the same set of physicians in the same hospital by following the same treatment protocol), and multi-site trials (i.e., data collected from patients treated in multiple hospitals following the same treatment protocol) settings. Reflecting on the marked heterogeneity in the manifestation of MDD symptoms and treatment outcomes, the translational success of ALMOND when used by clinicians rests in its capabilities to provide reliable insights. In this context insights from a trial's dataset are considered reliable if they replicate across independent trials wherein data are collected in from psychiatric clinics and hospitals, or that have evidence from laboratory experiments that relate to disease pathophysiology or treatment response. In this work, methods were developed using data from Mayo Clinic Pharmacogenomics Research Network Antidepressant Medical Pharmacogenomic Study (Mayo PGRN-AMPS, N = 900 subjects) [24]. Inferences and predictions gained using Mayo PGRN-AMPS were tested in two independent datasets, taken from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D [25], N = 1,800 subjects) and the International SSRI Pharmacogenomics Consortium (ISPC [26], N = 900 subjects).

1.4.2 Overview: ALMOND's Analytical Approach to Individualizing Antidepressant Treatment

Our goal is to move from the current *try-and-try-again* approach of *artisanal medicine* described by Dr. Perlis, to a *right patient, right drug, right time* approach of individualizing antidepressant treatment management in measurement-based psychiatry. Next, we elucidate a set of focused treatment-relevant questions that were the basis for our design of ALMOND's analytical workflow to achieve that goal. These questions were defined based on aspects of current clinical practices for treating MDD for which additional degrees of specificity were needed in order to individualize antidepressant treatment management. For each of these questions, we summarize the value of the *augmented actionable intelligence* either from the perspective of current treatment practices or for future clinical research.

1.4.3 Finding The "Right Patient" Through Patient Stratification

Stratification by Sex

Treatment-relevant Question: *Are there sex-differences in biological profiles of patients with MDD pre- and post-antidepressant treatment?* We began this stratification process by determining whether there were sex-differences in metabolomic profiles of patients with MDD at all three time-points of the trial. Metabolomics is a study of metabolites, which are products of metabolism (e.g., biochemical reactions in various parts of the body due to a drug/disease). Unlike genomics, where the DNA is stable during shorter durations of time (e.g., the duration of antidepressant treatment), metabolites offer an instantaneous snapshot of the biochemical state of the body/organism in which one is interested [27]. MDD patients treated with antidepressants who participated in a clinical trial at the Mayo Clinic, United States provided blood samples that were used to measure 35 metabolites of key neurological mechanisms associated with mental health disorders.

Approach and Significance: By using multivariate analyses of variance (MANOVA), we demonstrated significant differences in average concentration of metabolite concentrations between blood samples of men and women at all three time-points of the trial (discussed in Chapter 4). While it is well-known that the prevalence of MDD in women is twice that in men, clinicians often overlook sex as a key factor in antidepressant treatment selection. It is also important to note that prior machine learning work on predicting antidepressant treatment outcomes by using sociodemographic measures (including sex as a measure) as predictors have not shown sex to be a top predictor of treatment outcomes [28, 29]. Because

of the sex difference in quantitative biological measures of patients we stratified the rest of the analyses in ALMOND by sex, instead of testing whether sex is an important factor in predicting antidepressant treatment outcomes.

Augmented Actionable Intelligence: Sex-dependent biological factors could drive response to the same antidepressant treatment differently in men and women.

Stratification by Depression Severity

Treatment-relevant Question: *Can MDD patients be stratified by their depression severity?* Treatment decisions for many diseases, such as breast cancer (based on hormone characterization in a tumor) or hypertension (based on the severity of the blood pressure) are based on how the disease states are stratified biologically/symptomatically [23]. There are no strata of MDD severity based on ranges of total depression severity scores at any intermediate time-points of treatment. Definitions of antidepressant treatment outcomes such as *remission*, or *response* are empirically defined based on whether a patient's total depression severity had dropped below a threshold or quantum after 4 or 8 weeks of antidepressant treatment [30]. The lack of stratification has often limited our ability to understand pathological mechanisms of variation in MDD severity.

Approach: To establish such a stratification at baseline and at 4 and 8 weeks, we used an unsupervised machine learning approach to cluster patients based on depression severity at baseline, 4 weeks, and 8 weeks using patient data from Mayo Clinic (discussed in Chapter 4). We first observed that total depression severity distribution in both men and women, and at all treatment time-points, was a mixture of Gaussians. Hence, we used mixture-model based unsupervised learning with Gaussian mixture models (GMM) to algorithmically identify the minimum number of Gaussians that best approximated the actual distribution of depressive symptom severity at each time-point.

Significance: Our unsupervised learning approach algorithmically identified three distinct clusters of men and women based on their total depressive symptom severity at baseline and after 4 weeks and 8 weeks of antidepressant treatment, using data from Mayo PGRN-AMPS subjects. We replicated the clusters (and associated distributions in each cluster) at all time-points in multiple independent datasets (STAR*D and ISPC). Clustering methods always assign data-points into a specified number of clusters. However, the validity of the clusters in the context of the application is not guaranteed. In this instance, clusters inferred by ALMOND had two levels of significance. First, in a data-driven manner, we found that the three clusters after 8 weeks of antidepressant treatment in all trials comprised patients who were labeled as *remitters*, *responders without remission*, and *non-responders* based on clinical

definitions that use empirically defined thresholds of total depression severity. Second, given the acknowledged levels of heterogeneity in MDD manifestation, replication of patterns of clusters at baseline and 4 weeks in independent trials provides a foundation for investigating any associated differences in the disease biology and patterns of patient movement between clusters.

Augmented Actionable Intelligence: The replicating patterns of patient clusters increase the potential for learning the underlying pathological mechanisms of variation in MDD severity, and trajectories of patient movement between clusters.

1.4.4 Toward Identifying the "Right Drug" by Predicting Its Efficacy via Pharmacogenomic Measures

Treatment-relevant Question: *Can a psychiatrist's pre-treatment assessments of MDD severity, augmented with a patient's pharmacogenomic measures, predict eventual remission of MDD symptoms to antidepressant treatment?* We leveraged the inferred patient clusters to associate depression severity with pharmaco-omics measures. *Pharmaco-omics* is the study of how patients' -omics measures (in this work, genomics and metabolomics) affect in the patient's response to a drug [31]. In this work, we assessed the predictive capabilities of pharmacogenomic biological measures (biomarkers) associated with citalopram or escitalopram, commonly prescribed antidepressants that fall under the category of selective serotonin reuptake inhibitors (SSRIs).

Approach: In each of the inferred patient clusters, we achieved the multi-omic integration in the following two steps. In the first step, we identified metabolites whose concentrations were associated with depression severity scores in clusters at all time-points. In the second step, we conducted a genome-wide association study (GWAS) to identify a few genomic markers among 7 million single-nucleotide polymorphisms (SNPs, which are genetic variations), that are associated with concentrations of metabolites [27]. Of all the metabolites assayed in Mayo PGRN-AMPS samples (N = 290 of 900 subjects), we found that serotonin and kynurenine concentrations were the most highly associated with antidepressant treatment outcomes at 8 weeks or with baseline depressive symptom severity, respectively. The pharmacogenomic biomarkers (i.e., SNPs) associated with those metabolites are TSPAN5 (rs10516436), ERICH3 (rs696692), DEFB1 (rs5743467, rs2741130, rs2702877), and AHR (rs17137566) [32, 33]. A psychiatrist's depression severity assessments, augmented with the inferred biomarkers, were then used as predictors of antidepressant treatment outcomes as discussed in Chapters 5 and 6.

Significance: Supervised machine learning methods available in ALMOND were trained using pharmacogenomics biomarkers and total baseline depression scores of Mayo Clinic's

subjects. The trained models predicted sex-specific remission/response at 8 weeks with AUC ≥ 0.7 in Mayo Clinic's subjects, and with predictive accuracies $> 65\%$ ($p \leq 0.07$) in independent datasets. Furthermore, predictive accuracies obtained using pharcogenomic biomarkers were on average 12% better than those obtained when using only sociodemographic measures. The predictive performance with external replications demonstrated for the first time that pharmacogenomic biomarkers could reliably predict SSRI treatment outcomes prior to treatment initiation. If the multi-omic approach is extended to other antidepressants, drug-specific prediction models could potentially help individualize antidepressant treatment selection by identifying the treatment with the highest predicted likelihood of enabling the patient to achieve remission from MDD symptoms.

Augmented Actionable Intelligence: A few patient-derived pharmacogenomic biomarkers can augment a psychiatrist's MDD assessments to predict antidepressant treatment outcomes better than the use of sociodemographic factors alone.

1.4.5 Identifying Prognostic Changes in Depressive Symptoms at The "Right Time"

Treatment-relevant Question: *Is there a small set of individual depressive symptoms whose early change in assessment scores is prognostic of long-term treatment outcomes?* During the treatment management of MDD, clinicians often make decisions about whether to continue or alter antidepressant treatment plans at intermediate treatment time-points, before a full therapeutic trial is complete [34]. To operationalize the decision to continue or alter treatment plans at an intermediate treatment time-point, clinicians often focus on changes in individual depressive symptoms or total depression severity scores, measured using depression rating scales. Hence, there is still a need to obtain additional specificity by defining which individual depressive symptoms must change, and by how much, in order to accurately predict an eventual categorical treatment outcome, such as remission or non-response.

Approach: To address this question (discussed in Chapter 7), we first introduce the use of probabilistic graphs in ALMOND to algorithmically explore the most likely longitudinal variations (paths) of total depression severity as a patient progresses toward an eventual categorical treatment outcome. To define the graph, we used the inferred clusters of patients as nodes of the graph, and we defined the probabilistic edges based on the fractions of patients who moved between clusters of consecutive time-points of the trial (e.g., from the baseline cluster to a cluster at 4 weeks). Probabilistic graphs provide the mathematical foundation needed to model the conditional dependencies that follow a clinician's treatment logic, i.e., accounting for improvement in total depression severity, conditioned upon baseline

depression severity and changes in depressive symptoms at intermediate time-points, in a purely data-driven manner, without *a priori* specification of trajectories. Second, we used hierarchical clustering to identify a small set of individual depressive symptoms with homogeneous responses to a symptom assessment questionnaire, and for which, their early (e.g., at 4 weeks) changes were prognostic and predictive of categorical treatment outcomes at 8 weeks (remission, response, or non-response).

Significance: We identified a small set of individual depressive items comprising sad mood, anxiety, guilt feelings/delusions, and work/activities, whose changes at 4 weeks were prognostic of categorical outcomes at 8 weeks. Based on how many of these symptoms failed to improve or exceed specific thresholds at 4 weeks that were needed to achieve specific treatment outcomes, the probability of the most likely outcome for a specific patient was computed. In these analyses, there were also notable sex-differences in the longitudinal variation of depressive symptoms. The most compelling was that regardless of depression severity at baseline, women were more likely than men to achieve remission status in response to antidepressant treatment. This important sex-difference highlights a key aspect of antidepressant response that would have been overlooked had we not first separated the analyses by sex. A clinician can interact with ALMOND through a web interface, wherein patient's gender and symptom assessment data at baseline and at 4 weeks are collected. ALMOND then computes the probability and statistical significance of the most-likely treatment outcome at 8 weeks and informs the clinician through the interface.

Augmented Actionable Intelligence: ALMOND draws a clinician's attention to changes in a small subset of depressive symptoms after 4 weeks of treatment, and thus could accurately inform the prognoses of eventual treatment outcomes that is likely to be observed after 8 weeks of antidepressant treatment. If the prognoses forecasts a poor treatment outcome, the clinician could prompt a change in the treatment so as to avoid extended weeks of disease burden from the lack of response.

1.5 The Broader Impact of ALMOND

1.5.1 In Treating Depression: The Multi-Trillion Dollar Economics

The consequences of persisting MDD symptoms in the short term because of treatment failures are significant. MDD is a serious disability-causing health condition that leads to poor function at work [14, 35, 36]. In a study led by the World Health Organization (WHO) that was published in 2016, it was shown that depression and anxiety disorders cost the global

economy $ 2 – 8 trillion (U.S.) in lost productivity [37]. By 2030, the economic burden is expected to double if the current trend of increasing prevalence of MDD continues [37]. The same study also showed that in the United States of America (U.S.) alone, MDD translated into an average of 8 hours per week of lost productivity, and over 20 days of absence per year, costing $210 billion (U.S.) in lost productivity per year. Furthermore, early diagnoses and treatments of MDD are administered by general practitioners (GPs), who do not use standardized psychiatric metrics to assess the severity of MDD [38, 39]. The GPs' ability to assess MDD severity and adequately treat patients is called into question by the fact that a higher proportion of MDD patients achieve remission from depressive symptoms when treated by mental health practitioners (e.g., psychiatrists) [39]. ALMOND's web interface could be used by a GP to generate prognoses of eventual antidepressant treatment outcomes, using a patient's responses to a standardized symptom questionnaire which GPs to assess MDD symptoms. ALMOND's will be of significant value if it enables a GP to accelerate a patient's access to a psychiatrist on the basis of a poor prognosis. The timely recommendation to see a psychiatrist is crucial because even in developed countries, MDD is often not treated. For example, in the U.S., 37% of 16 million adults with MDD do not receive mental health treatment. [1]

In the long term, one's life is at risk if there is prolonged exposure to MDD (due to treatment failures) or if it is left untreated (if the presence of the disease is not acknowledged). On the extreme end is the risk of suicide, as more than 90% of individuals who commit suicide suffered from a history of mental health disorders. Another health risk is in the potential development of chronic diseases such as type II diabetes and coronary diseases [40, 41]. These chronic diseases develop because of poor management of stress, sleep, appetite, and reduced physical activities resulting from persisting MDD symptoms. Studies have also shown that persisting depression may alter brain function [42]. Brain inflammation is key to an individual's resilience and recovery from injury or illness. A study conducted with positron emission tomography scans (PET scans) revealed excessive inflammation in the brains of patients with persistent and/or untreated MDD symptoms. It is well known that excessive inflammation in the brain is associated neurodegenerative diseases like Alzheimer's and Parkinson's [43, 44].

Taken together, ALMOND's potential to accelerate therapeutic success in treating MDD can improve patients' quality of life and reduce their risk of developing other chronic illnesses.

[1] `https://www.nimh.nih.gov/health/statistics/major-depression.shtml`

1.5.2 In Medicine Beyond Psychiatry

From a biological mechanistic perspective, on one end of the clinical spectrum there are diseases such as breast cancer for which treatment is based on tumor subtypes. At the other end of the spectrum are neuropsychiatric diseases such as MDD for which subtyping of individual patients is possible based on their reported symptom severity, as demonstrated in this work. Between those two extremes are migraine headaches and inflammatory diseases, such as rheumatoid arthritis (RA), in which patients are subtyped by the degree of swelling of their joints (which does not directly reflect a specific mechanistic biomarker) and by pain ratings reported by the patients using validated scales that are similar to the QIDS-C scale used to rate depressive symptoms in this work [45–47]. Therefore, there is sufficient heterogeneity in RA symptoms to make treatment response phenotypes so complex that the methodological innovation presented in this work could be used to overlay biological measures that provide a significant mechanistic perspective. This approach could then be tested for the prediction of outcomes in response to the drug therapy for migraine headaches or RA, as additional examples of a possible broader application of the approach described here.

1.5.3 In Pharmacogenomics Research

Pharmacogenomics research focuses on understanding the interplay between drug effects and functions of the genome [1]. In that context, we reflect on the improvements in breast cancer therapeutics wherein treatment selection is based on molecular characteristics of the tumor. In diseases such as MDD, for which such biologically based subtyping is not yet possible, the approach described in this work for stratifying patients by using symptomatic characteristics will be of immense value for pharmacogenomics research. In particular, trials could be designed in which multi-omics (metabolomics, trancriptomics, genomics, etc.) and other biological measures (neuroimaging, electrophysiology, etc.) could be collected that help to establish biological associations with patients stratified using the proposed approach. Then, longitudinal effects of the drug on those biomarkers could be used to study why patients either respond well to the intervention or do not. Furthermore, as already demonstrated in this work, associations of biological markers with inferred patient stratification can provide improved predictability of treatment outcomes. Such patient clustering that can potentially identify underlying differences in pathophysiology or predisposition to treatment outcomes represents a significant advance in moving away from “artisanal medicine” practices.

1.5.4 In Advancing Analytics

Behavioral and environmental exposure are said to play an important role in brain activity, disease development, and therapeutic response. Thus, a framework like ALMOND that can identify strata of patients with homogeneous disease states and has the ability to characterize longitudinal symptom characteristics will be enhanced when one incorporates data with additional granularity such as functional neuroimaging and daily activity data from wearable technologies. With ALMOND, predictions or prognoses of treatment outcomes can incorporate both physician-assessed symptom data, and quantified data relating to a patient's lifestyle and daily activities. Such integration of image data along with physiological monitoring data will require analytical innovations that potentially extend beyond the traditional image analysis techniques embodied in the deep learning literature. Incorporation of improvements in daily activities recorded by wearable devices, and patient-reported side-effects and symptom improvements and lifestyle changes, could help clinicians individualize the dosing and frequency of medications via techniques such as reinforcement learning.

1.6 A Vision for Augmenting Human Intelligence in Medicine

During periods of residence at Mayo Clinic, National University Hospital, Singapore, and while presenting this research in biomedical/engineering conferences, whenever the author of this dissertation mentioned words such as "prediction," "machine learning" (ML), and "artificial intelligence" (AI), he was always met with the question, "Will psychiatrists become obsolete?" The answer to this question has always been "No, clinicians are essential to treating patients, and in working alongside engineers to develop technologies that addresses unmet needs in medicine." Thr role of computer engineering scientists will be to develop methods that provide a small set of analytically chosen, meaningful, relevant patient-derived information in a way that amplifies and augments a clinician's assessment of a patient's disease state. That is the thing that will transform what Dr. Perlis called "artisanal medicine" into "individualized medicine."

An analogy can be made based on the continued need for human pilots in commercial airplane cockpits that have complex automation systems. The current generation of airplanes (e.g., Boeing's Dreamliner or the Airbus A350) continues to offer historically high degrees of reliability in flight safety, and these airplanes collect tremendous volumes of high-throughput data during the course of each flight. Pilots were impressed by the highly simplified design of cockpit controls and data display in these airplanes, among other improvements. By analyzing data from millions of miles and hours of flying across the globe in many flying conditions,

airplane manufacturers have arrived at a reduced (if not simple) set of well-characterized measures that have been deemed sufficient to allow a trained pilot to fly planes safely, even in rapidly escalating adverse conditions. The continuing relevance of pilots in modern-day flying, however, has been emphasized by several recent aviation incidents such as the one in which Capt. Sullenberger safely landed a commercial passenger airplane in New York City's Hudson River after both engines of his plane were disabled by a bird strike. [2]

In fact, the National Transportation Safety Board of the United States found that over half of the flight simulations developed following the incident showed that, if Capt. Sullenberger had proceeded to any nearby airports,a safe landing would have been unrealistic. Furthermore, given rapidly escalating failures during that flight, the plane would have crashed into New York City had Capt. Sullenberger delayed his decision to land in the Hudson River by 35 seconds. The investigative findings of that incident highlight the fact that *human piloting skills augmented with simplified presentations of required real-time safety-critical data, have significantly reduced the chances of major catastrophes*. On can argue about whether an automated pilot trained with numerous emergency scenarios could have safely landed Capt. Sullenberger's plane in the Hudson River, but, for today and the foreseeable future, pilots' skills are still deemed essential to safe flying of airplanes.

In light of that analogy, we can return to the context of augmentation of human intelligence in medicine, broadly focusing on psychiatry and mental health. A century's worth of research has resulted in psychiatrists' ability to measure, quantify and diagnose the presence of depression using diagnostic criteria and more than a dozen symptom severity questions [48, 49]. Numerous statistical and machine learning approaches have identified a list of sociodemographic measures, with minimal concordance among them, as predictors of antidepressant treatment outcomes. None of the machine learning methods so far have seen adoption in clinics, in part because clinicians do not often use such a wide variety of sociodemographic measures to guide treatment selection. Furthermore, psychiatrists always knew and will continue to believe in the importance of patient-derived biological measures in furthering their understanding of mental health disorders and in predicting treatment outcomes. However, in the context of individualized antidepressant treatment management, the broader question to answer has been: "Which of the psychiatrists' assessments need to be augmented with which relevant biological measures to accurately predict antidepressant treatment outcomes?" Finally, in the event that the patient appears to a trained psychiatrist to be more suicidal than a questionnaire would have found, the patient must be immediately treated to prevent suicide, and not with an antidepressant selected by predictive analytics. The surrounding context of the event, be it treating a suicidal depressed patient for suicide first, or landing

[2] `https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR1003.pdf`

the airline in Hudson River as opposed to crashing into New York City needs a human in the loop. This is because, it is likely that the automated systems despite the sophisticated analytics would be unprepared to learn the surrounding context of impending catastrophes.

The future of individualized medicine, partly rests in the capability of engineers to further augment clinicians' symptom assessments with real-time data from patients' daily activities (e.g., hours of sleep, quality of sleep, number of steps, and resting heart rate) as measured by wearable technologies, and the patient's *exposome*. The exposome comprises environmental exposure external to the human body, such as air quality and types of surrounding vegetation. By combining exposome data with other biological measures, we could potentially observe whether diseases manifest in biologically different ways in different patients, while also considering where the patients live, how they live, and what they do during the course of a day. The optimal choice of medication may draw on all of the resulting insights. Toward that end, the vision of this dissertation in the context of making augmented actionable intelligence relevant to clinical practice, is that *analytical combination of vast, and rapidly growing highly complex, patient-derived data with the constantly evolving treatment strategies used by clinicians, means that both clinicians (because of their domain knowledge) and the augmented human intelligence enabled by AI/ML technologies will be essential to standard individualized medicine care giving in clinics/hospitals.*

1.7 Summary of Contributions

From a computer engineering perspective. The overarching goal is to analytically generate intelligence from real-world clinical data to advance biomedical research for both discovery and translational science. This dissertation presents ALMOND as an analytical framework that has the capability to analyze data that are high-dimensional and heterogeneous in both type and the time dimension (static vs. dynamic) to generate actionable intelligence. Incorporation of inputs and annotations from clinicians and other domain experts in analyzing data not only allows methods to be closer to clinical reality, but has allowed us to generate results with higher degrees of interpretability. Finally, all insights of ALMOND from the discovery perspective have either had their biological significance verified through laboratory experiments, or been replicated in independent trial datasets pooled from across the globe.

Through insights generated by ALMOND, this dissertation makes the following key contributions toward augmentation of clinicians' intelligence in various fields of medicine.

- **Psychiatry:** For the first time, we have demonstrated the ability of pharmacogenomic measures augmented with psychiatrist's depression severity assessments to robustly

predict antidepressant treatment outcomes. In addition to demonstrating significant sex-differences in antidepressant response profiles, we have established much-needed specificity in early symptom improvements for prognoses of treatment outcomes in multiple rating scales. In assessing ALMOND's ability to replicate disease states defined by patient strata at all time-points of treatment, we found that all aspects of the analyses, from longitudinal characteristics to predictive performance using pharmacogenomic markers, replicated across independent trial datasets and depression rating scales.

- **Oncology:** Triple-negative breast cancer is among the most aggressive breast cancers in women, and it does not have a targeted treatment yet. A novel mechanism of cancer-migration inhibition by the diabetic drug metformin identified by ALMOND has provided additional evidence for ongoing efforts to repurpose commodity drugs in treating cancer.

  Given that lung adenocarcinoma is often detected at later stages of tumor development and warrants invasive biopsies down to the lungs, trancriptomic variations in a few genomic biomarkers prognostic of disease development, obtained through saliva, could help inform physicians about the state of the lung's health.

- **Endocrinology:** Our ability to individualize the risk of deterioration in a diabetic patient's health in the long-term offers caregivers to optimize care (medication and lifestyle suggestions) to potentially minimize hospitalizations and surgical interventions.

CHAPTER 2

BACKGROUND, CHALLENGES, AND RELATED WORK

Individualization of breast cancer therapeutics has been successful due to genomic-guided treatment selection [23]. Major depressive disorder (MDD), on the other hand, despite its global footprint is yet to benefit from individualized medicine approaches. What characteristics of MDD limit the ability of genomics to successfully individualize antidepressant selection?

In measurement-based care in psychiatry, symptom questionnaires assess the severity of the spectrum of depression symptoms by using ordinal scores [49]. The total depression severity score is a sum of ordinal responses to individual depression symptom questions. It is well-known that higher depression severity at baseline correlates with poorer outcomes of antidepressant treatment. It is also known that early improvements in total depression severity scores correlate with eventual treatment outcomes. Why are the pre-treatment baseline total depression severity score and its improvement at 4 weeks insufficient to enable psychiatrists to accurately forecast a clinical outcome at 8 weeks? Where lies the challenge in transforming these widely used depression rating scales into prognostic scales?

The use of statistical and machine learning approaches in predicting antidepressant treatment outcomes is not new. An overwhelming majority of the existing methods have used routinely collected sociodemographic measures as predictors of categorical outcomes to antidepressant treatment. Why have these approaches yet to individualize antidepressant treatment selection in routine clinical settings?

Finally, access to high-quality data and clinical measurements of symptom severity is key to answering the broader question of "right patient, right drug, right time" to individualize choices of antidepressant medications. To that end, what characteristics of data from clinical trials of multiple routinely prescribed antidepressants can potentially drive analytical innovations towards individualization of antidepressant treatment management?

Contribution. In this chapter, we describe the opportunities for analytical innovation by describing the complexities faced by psychiatrists in the antidepressant treatment management of MDD patients. We reflect on the related work and describe the research gap addressed in this dissertation. We briefly introduce several multidisciplinary concepts with the intention of making this work accessible to readers from multiple disciplines.

2.1 Major Depressive Disorder

Major depressive disorder (MDD) is the number-one psychiatric disease worldwide [14, 15]. MDD is a complex disease characterized by several depressive symptoms.

Common depressive symptoms are melancholy (sad mood or depressed mood); loss of interest, pleasure, and energy; weight and appetite variations; guilt feelings and delusions; inability to concentrate or sleep well, and the presence of suicidal thoughts [14].

2.1.1 Diagnoses and Measurement-Based Care

Diagnoses: The diagnostic criteria for depression are specified in the *Diagnostic and Statistical Manual of Mental Disorders* (DSM) developed by the American Psychiatric Association [50, 51]. For an individual to be diagnosed with MDD, he or she must be experiencing five or more depressive symptoms during a two-week, period with at least one of the symptoms being depressed mood or loss of interest or pleasure. Furthermore, the individual must be experiencing sufficient impairment in daily activities, at work, or socially. Finally, these symptoms must not be a result of substance abuse or a side effect of another medication.

Measurement-Based Care: To evaluate the severity of MDD symptoms, several rating scales that measure the severity of individual depressive symptoms have been proposed [52–54]. Each rating scale comprise 10 – 17 questions , which measure the severity of individual depressive symptoms (also referred to as *individual scale items*) through ordinal responses. Depression severity could be "clinician-rated" wherein a clinician asks the patient scale's questions, and records the severity of the responses. Patients can also "self-report" the severity of depressive symptoms by rating the severity using their own judgment.

Popular rating scales that measure a full spectrum of depressive symptoms include the 17-item Hamilton Depression Rating Scale (HDRS [53]), the 16-item Quick Inventory of Depressive Symptomatology (QIDS [52]; there is also a *QIDS-C* for clinician severity rated, and a *QIDS-SR* for self-reported severity), and the 10-item Montgomery-Asberg Depression Rating Scale (MADRS [54]). The total depression severity score (TS) is a summation of individual depressive item rating scores (in HDRS and MADRS), and a summation of the highest rating scores derived from a group of depressive symptoms (in QIDS).

2.1.2 Treatment Options and Outcome Definitions

Treatment Options. Antidepressant medication and psychotherapy are two broad treatment options commonly available for treating MDD [34]. Psychotherapy, also commonly known as "talk therapy," is based on a constant dialogue between a psychologist and the patient. The patient can openly discuss his or her problems with a psychologist. In an objective, nonjudgemental, and neutral way, a psychologist will engage with patients to find ways to change the patient's thoughts and behavior in order to reduce the burden of MDD symptoms in his or her daily functioning. Antidepressant medication makes changes to brain signaling in order to reduce MDD symptom severity. Common classes of antidepressants include tricyclic antidepressants (TCAs), selective serotonin reuptake inhibitors (SSRIs), and selective serotonin noradrenaline reuptake inhibitors (SSNRIs).

There is no consensus on which treatment option is more reliably effective [21, 34]. Many psychiatrists may prefer that the MDD patient begin with psychotherapy if the depression severity is mild during diagnosis. Alternatively, a psychiatrist might start an MDD patient on an antidepressant medication from the start if the patient shows moderate or severe depression severity during diagnosis. In many instances, antidepressant medication and psychotherapy are combined in treating MDD. In this work, we will focus on predicting or deriving prognoses of response to SSRI treatment, which is the most commonly prescribed primary antidepressant.

Treatment Outcome Definitions. Antidepressant treatment management comprises two phases; the acute phase and the extended phase. The acute phase begins from the time of treatment initiation after patient has met the DSM-IV (or DSM-V) diagnosis criteria for MDD, and it lasts for about 8 weeks. The treatment comprises follow-ups at 2, 4, and 6 weeks after treatment initiation; during the followups, MDD symptoms are measured using the same rating scale chosen prior to treatment initiation. In the extended phase, if the acute phase treatment results in the patient's achievement of remission status (defined next), the same antidepressant is continued for upto 24 to 30 weeks, with followup visits at every 4 weeks. If the patient fails to achieve remission status by the end of the acute phase, another acute phase with a different antidepressant is initiated. Table 2.1 illustrates the treatment outcome definitions of *remission, response (without remission),* and *non-response* after 8 weeks of acute-phase treatment [30]. The outcome definitions are based on emperically derived thresholds based on either the total score after 8 weeks of treatment, or the improvement in the baseline (pre-treatment) total score (TS_b). If a patient continues to stay in remission or response status after an acute-phase treatment and during the entire duration of the extended phase, the extended-phase treatment's outcome is considered *extended remission* or *extended response*, respectively.

Table 2.1: Antidepressant treatment outcome definitions.

	Total score at 8 weeks		
Rating scale	QIDS	HDRS	MADRS
Remission at 8 weeks	≤ 5	≤ 7	≤ 10
Response (without remission) at 8 weeks	$\leq 0.5 * TS_b$	$\leq 0.5 * TS_b$	$\leq 0.5 * TS_b$
Non-response at 8 weeks	$> 0.5 * TS_b$	$> 0.5 * TS_b$	$> 0.5 * TS_b$

2.2 Pharmacology Primer

Pharmacology is the study of effects and mechanisms of drugs [55]. *Pharmaco-omics* is the study of effect of -omics (e.g., genomics, metabolomics, and transcriptomics) in drug response [31]. *Pharmacogenomics* is the study of the effect of genomics on the drug response. *Genomics* is the study of the structure, function, and evolution of the genome. *The genome* is the genetic make up (e.g., DNA and RNA) of an organism. There are about $23,000$ genes in the human genome, and each gene's DNA varies among people. Those biological variations in genes among individuals are referred to as *single-nucleotide polymorphism* (SNP), and are characterized by their location in the human genome (e.g., where they are in the 23 human chromosomes that comprise the human genome). The variations in the SNPs are called *genotypes*. Today, plasma blood samples for genome-wide genotyping can yield data for 7 million SNPs. For any given trait of clinical/biological interest (e.g., cancer vs. no cancer, disease vs. no disease, remission vs. no remission), a genome-wide association study (GWAS) can associate variations in SNPs (genotypes) with traits.

Metabolomics is a study of metabolites, which are products of metabolism (e.g., biochemical reactions in various parts of the body due to a drug or disease). *Pharmaco-metabolomics-informed-genomics* is the study of genome-wide associations with quantitative concentrations of metabolites that are associated with a trait of biological or clinical interest [27].

From the perspective of drug mechanisms, two concepts are relevant to understanding of the discussion of the CYP2C19 biomarker in Chapters 4 and 6. Pharmacokinetic markers allow for the study of an individual's ability to absorb, metabolize, and excrete a drug, and is often meant to guide the selection of a safe treatment option [55], i.e., selection of a drug or dosage that would not induce adverse effects (e.g., extreme blood pressure variations). Pharmacodynamic biomarkers, on the other hand, relate to drug concentration and treatment outcomes, including resulting side effects, if any [55]. In this dissertation, the pharmacogenomic SNPs used as predictors of antidepressant treatment outcomes in Chapter 6 are pharmacodynamic biomarkers.

2.3 The Most Compelling Challenge in Studying Depression? Heterogeneity

The following characteristics of MDD as a disease, and its subsequent response to antidepressant treatment, have thus far obstructed the achievement of high precision in treating MDD.

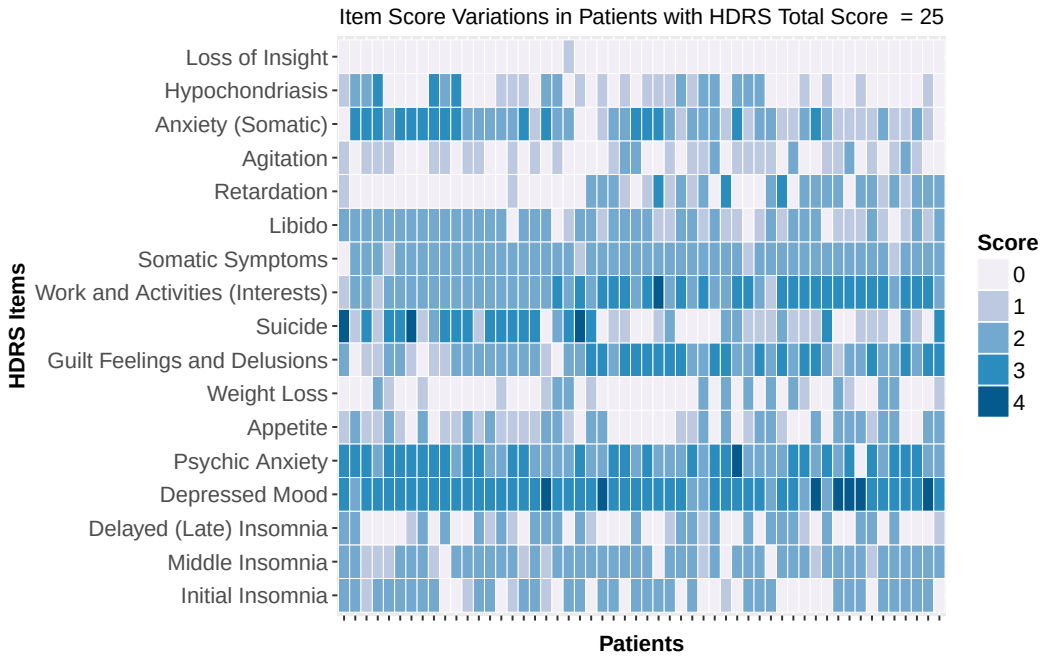
2.3.1 Heterogeneity in Disease Presentation

MDD is a disease comprising multiple individual symptoms [14]. The severity of these symptoms, as assessed by any of the rating scales, manifests differently in different patients, even if the patients have the same total score (as illustrated in Fig. 2.1(a)). Furthermore, among patients classified with the same outcome after 8 weeks of acute-phase treatment with antidepressants, there is considerable variation in symptom severity score profile between patients (as illustrated in Fig. 2.1(b)). To quantify the extent of the observed variability in depressive symptom presentation, we need five or more principal components to explain even 50% of the variability. This variability in manifestation of depressive symptoms is referred to as *heterogeneity in disease presentation*.

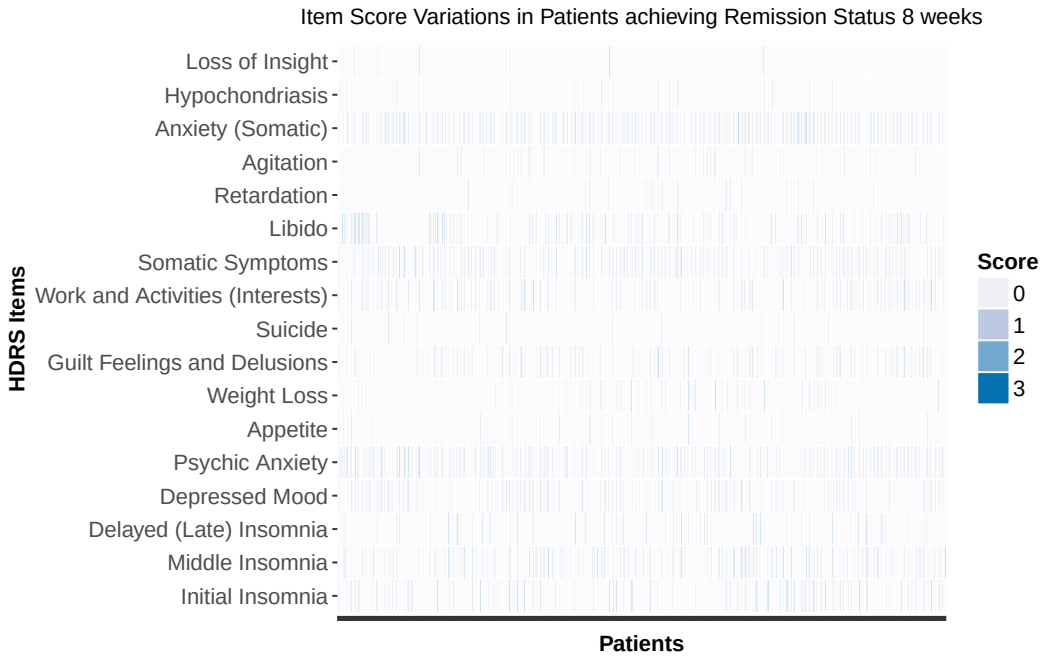
Treatment decisions are often based on a total score derived by summing the severity scores of symptoms, and a treatment's outcome is judged based on the improvement over the baseline total score at the end of the treatment's duration [30]. Treatment decisions made using a total score from a depression rating scale overlooks the sensitivity of individual item scores' changes to antidepressant treatment [30]. Thus, an opportunity to identify patients with more or less homogeneous presentations of depressive symptoms, albeit with similar total scores, is lost.

2.3.2 Heterogeneity in Antidepressant Response

Let us assume that each group of patients with the same total score are considered to be patients with the same disease state at each assessment during the antidepressant treatment. Then, if there are N states of depression, and k time-points of assessments in the clinic, the complexity in how depression states change is $O(N^k)$. That is a level of complexity that even experienced psychiatrists would not be able to assimilate in order to accurately forecast a treatment outcome, given a patient's total depression severity score at baseline. We illustrate the complexity in Fig. 2.2 for $1,400$ MDD patients who were treated with citalopram or escitalopram for 8 weeks in two large clinical trials. Figure 2.2 shows their depression



(a) Patients with HDRS total score = 25 at baseline



(b) Patients who achieved remission status with HDRS total scores ≤ 7 after 8 weeks of antidepressant treatment

Figure 2.1: Heterogeneity in symptom manifestation in patients with same the total score or clinical outcome after the acute phase of antidepressant treatment.

severity scores as assessed using the QIDS-C rating scale. To describe 100% variation of depression severity during the treatment, 986 unique paths between baseline and 8 weeks, with a follow-up at 4 weeks after treatment initiation were needed.

Furthermore, among patients who begin the treatment with the same depression severity score at baseline and who have the same sociodemographic characteristics, some go on to achieve remission, or response without remision status, after 8 weeks of acute-phase treatment, while some fail to respond to the same antidepressant treatment. That variability in how depression severity changes with time while resulting in an eventual outcome with antidepressant treatment is referred to as *heterogeneity in antidepressant response* [16–20].

2.3.3 Heterogeneity in Treatment Strategy

Let us assume that the sincerity of patients in reporting their depressive symptoms' severity reduces the effect of subjectivity"in diagnoses. Following the "artisanal medicine" approach in routine clinical practice to which Dr. Perlis refers, different psychiatrists choose different antidepressants. In many instances, they choose to combine multiple antidepressants (referred to as *co-medications*) to treat MDD patients. Indeed this end, there is no accepted consensus on which antidepressants are more effective than the others. Hence, it is easy to observe *heterogeneity in treatment strategy* even in the way medications are administered in treating MDD. It may not be entirely clear which combination of antidepressants, or which single antidepressant, was really the contributing factor to a particular treatment outcome.

2.3.4 How Does Heterogeneity Challenge Individualization of MDD Therapeutics?

To provide a context on how disease and antidepressant treatment response heterogeneity challenge individualization of antidepressant treatment, we will briefly look at the path to success in breast cancer therapeutics.

What Brought Success in Breast Cancer Therapeutics? Until breast tumors were uniquely characterized according to their molecular characteristics (e.g., hormone characteristics), breast cancer therapeutics was just as "artisanal" as MDD is treated today. Although no two individuals have identical genomes, high degrees of similarity in characterizations of tumors at the molecular level make the disease states very homogeneous. Thus a GWAS that treats unique disease states as traits (as opposed to breast cancer vs. no breast cancer as traits) yields very actionable pharmacogenomic markers (e.g., SNPs) as targets for drug design and choice, and that finding has been replicated in all clinical trials across the globe.

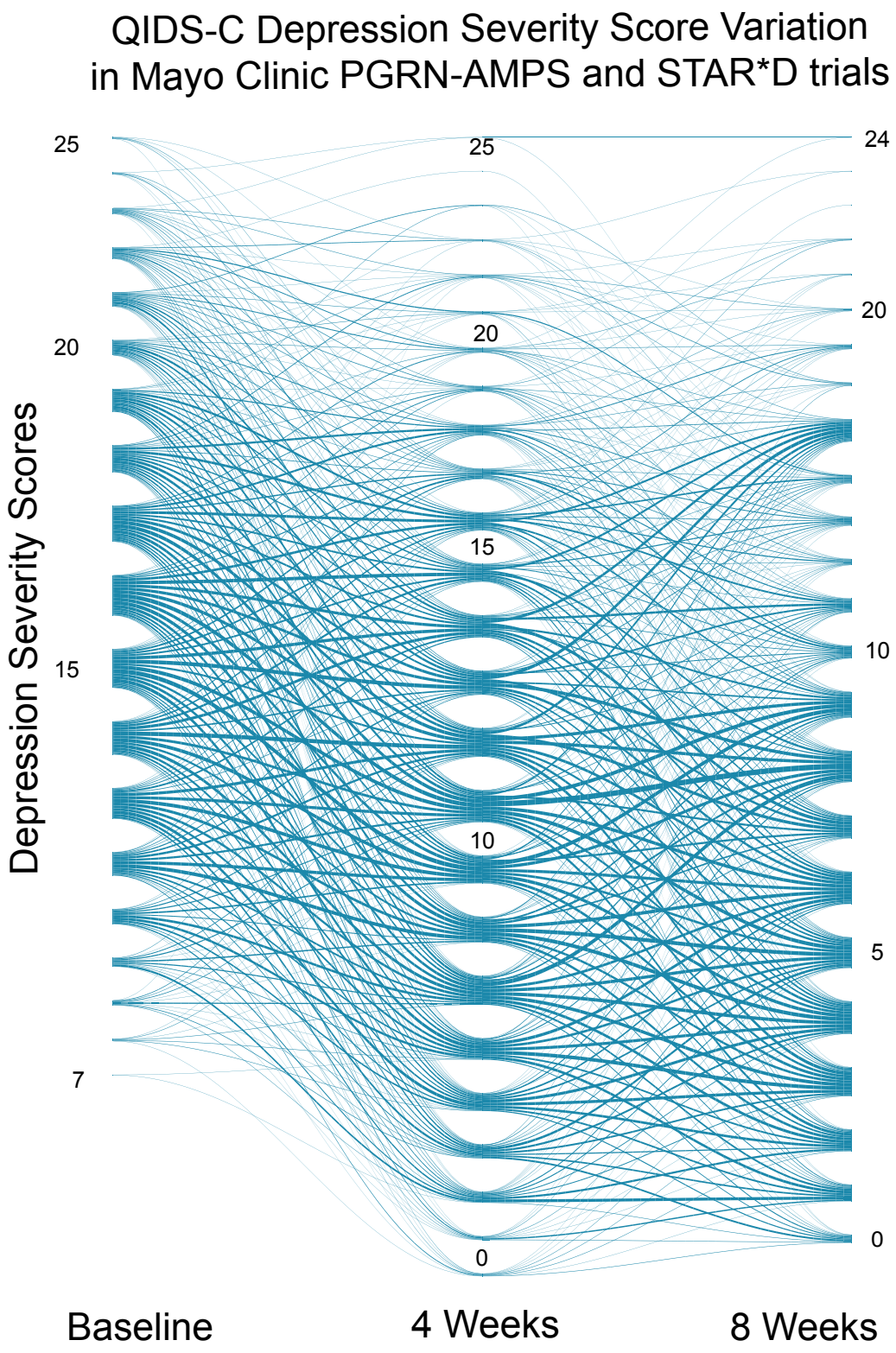


Figure 2.2: Variation in total depression severity scores of MDD patients treated with citalopram or escitalopram for 8 weeks, with a followup assessment of depression severity at 4 weeks after treatment initiation.

Hence, as long as a patient's tumor is correctly characterized and the treatment targeting the specific tumor subtype is chosen, the likelihood of an individual's therapeutic success is high [23].

On the Challenges in Individualizing MDD Therapeutics: GWAS in each large depression study have identified unique genomic markers as risk factors for developing depression (GWAS for disease vs. no disease as traits). Several large antidepressant studies have also established multiple mutually exclusive genomic markers of antidepressant response (GWAS for response vs. no response, remission vs. no remission as traits). However, none of these biomarkers replicated across trials [56, 57]. The key reason for the lack of replication in findings is the complexity of the disease in terms of differences between patients in how depressive symptoms manifest and respond to medication. Furthermore, even GWAS for the trait, "remission from medication vs. placebo treatment" has failed to find replicating sets of biomarkers, because even placebo-treated patients achieve remission status at nearly the same rate as patients treated with antidepressants.

2.4 Related Work

2.4.1 Prediction of Antidepressant Treatment Outcome

Transcriptome Variations Associated with Remission: Transcriptome variations have been observed in blood samples between non-remitting MDD patients after 12 weeks of citalopram with psychotherapy treatment and non-depressed control group patients [58]. The differentially expressed genes, used as predictor variables predicted eventual non-remission when the authors used support vector machines with linear kernels (SVM-Linear), with an accuracy of 76%. The work work showed the promise of studying transcriptome variations in response to antidepressant treatment. **Limitations.** Heterogeneity of the disease and response in the case of remitters was observed given that transcriptome variation was not significant between eventual remitters at baseline, and control subjects. Furthermore, the study called for additional analyses on the ability to improve predictions by using genotype and depression severity measures from full rating scales.

Sociodemographic Variables as Predictors of Remission: Elastic net regression has been used with recursive feature elimination to predict remission in citalopram-treated patients while also accounting for symptom clusters [28, 29, 59]. In all these approaches, baseline total depression severity was consistently identified as the top predictor of remission. However, among the lists of the other top five predictors, there was no significant overlap.

The lists of depressive items or sociodemographic variables that were strong predictors of remission varied among the prior efforts. **Limitations.** While the predictive accuracies of the prior efforts were better than chance and statistically significant (i.e., with a p-value ≤ 0.05), the choice of using only sociodemographic measures as predictors limits the ability to individualize treatment selection. The reason is that, individually or in subgroups, sociodemographic variables have been previously shown to be weak predictors of treatment outcomes [30, 60, 61]. Furthermore, these publications also showed that predictive models trained on treatment outcomes from one antidepressant (e.g., citalopram) could not predict treatment outcomes with comparable specificity and sensitivity in patients treated with a combination of antidepressants (e.g., venlafaxine plus smirtazapine) [29].

Pre-Treatment Severity of Symptom Dimensions as Predictors of Remission: Researchers have also attempted to predict treatment outcomes by using depressive "symptom dimensions" [61]. Clusters of symptoms defined as "symptom dimensions" are identified using factor analyses [62, 63]. It was shown that higher scores of interest-activity symptoms on the QIDS-C or HDRS scales at baseline were associated with lower chances of achieving remission after an acute phase of antidepressant treatment. **Limitations.** Further exploration of genomic associations between symptom dimensions' severity and potential patient subtypes are needed, in order to determine whether heterogeneity in treatment response and outcomes can be better understood.

Genomic SNPs as Predictors of Remission or Response: Instead of using GWAS as the approach for selecting genes, elastic net regression with variable selection was used to prioritize genomic predictors from about $500,000$ genes genotyped in the Genome-based Therapeutic Drugs for Depression (GENDEP) study [64]. The area under the receiver operator curve (AUC) in the work was 0.77; that agreed with prior work, which revealed that higher scores on interest-activity symptoms on the QIDS-C or HDRS scales at baseline were correlated with lower chances of achieving remission after acute phase of antidepressant treatment [61]. Some of the top predictors of treatment outcomes were associated with pathology of schizophrenia and bipolar disorder. In a Taiwanese study, remission following antidepressant treatment was predicted using top-hit SNPs (with a p-value $< 7.5 * 10^{-5}$) from GWAS and baseline clinical variables as predictor variables [65]. In that work, multilayer feedforward neural networks (MFNNs) that were used to predict the remission status achieved an AUC of 0.83. In another study, genes related to the hypothalamic-pituitary- adrenal (HPA) axis were used as predictors of remission to three commonly used antidepressants (escitalopram, sertraline, and venlafaxine) [66]. That study, which used data from the International Study to Predict Optimized Treatment in Depression (iSPOT-D) and Predictors of Remission in Depression to Individual and Combined Treatments (PReDICT), showed that

a particular SNP, h4, was associated with response to escitalopram and sertraline, and not venlafaxine. **Limitations.** In all these predictive approaches, the predictive performances when genomic predictors were used, were better than when sociodemographic factors were used as predictors, and that is encouraging. However, none of the identified genes had SNPs that were identified in pharmacogenomic studies of antidepressant response. Hence, it remains to be understood how these identified genes contribute to the mechanism of antidepressant response, which must be known in order to guide accurate treatment selection.

2.4.2 Longitudinal Analyses of Antidepressant Response

Early Improvements in Total Depression Severity Associated with Eventual Remission: Using an odds ratio metric, several studies suggest that early treatment response (improvement of at least 20% in the baseline total depression severity score) after 2 or 4 weeks of acute-phase antidepressant treatment suggests that there will be eventual response or remission at the end of the acute-phase treatment [67–70]. However, almost all of those studies also show that some patients (as many as 30%), early treatment response is not predictive of eventual remission. Further, the GENDEP study showed that about 40% of the patients treated with citalopram or nortriptyline who failed to achieve early treatment response at 2 weeks went on to achieve response status after 8 weeks of treatment [67]. One study used the least absolute shrinkage and selection operator (lasso) to predict eventual response to 12-week acute-phase treatment by using the improvement in total depression severity at 6 weeks with an AUC of 0.70 [71]. In that study, it was shown that predictive performance did not significantly improve when improvements in individual depressive items were also considered as predictors, in addition to improvement in the total depression severity score. **Limitations.** With the exception of de Vries et al. [71], all other studies call for methodological innovations that identify drug- and patient-specific changes in individual symptoms that extend beyond the overall improvement in total depression severity measured by either rating scales (e.g., QIDS-C or HDRS).

Trajectory Analyses for Modeling Antidepressant Response: Growth mixture models are a popular statistical approach for identifying longitudinal change in a variable of interest in unobserved subpopulations [72, 73]. Growth mixture models have been used to infer trajectories of improvement in total depression severity scores that lead to any of the categorical treatment outcomes [74–77]. A unique feature of the work of Kelley et al. [75], is their ability to associate genomic markers with trajectories of remission and non-response. Gueorguieva et al. [77] show that when trajectories of response are modeled, a third of the patients who are responders at the end of the acute-phase treatment will see worsening in

their depression symptom severity in the extended phase of the treatment. **Limitations.** In using growth mixture models, one must have sufficient domain expertise to define the number of latent classes and trajectories, and ensure appropriate model fit, and then interpret the results [78, 79]. The implicit assumption in defining the paths is that the population on each path remains homogeneous throughout the duration of the path (i.e., patients do not change trajectories during the treatment). Furthermore, growth mixture models are prone to overestimation of the number of trajectory classes [79]. Finally, growth mixture models do not find paths algorithmically by conditioning future improvements in depression severity upon improvements in depression severity at intermediate time-points, and they do not provide easy interpretation of the dynamics of symptom changes as a function of the percent improvement in total depression severity from baseline.

2.5 Summary: Research Gaps Addressed

In summarizing the related work in the context of predicting or modeling antidepressant response, we observe research gaps that call for analytical innovations to individualize antidepressant treatment selection. The prior work's demonstration of the predictive value of genomic measures shows the promise as a first step towards prediction of treatment outcomes based on patient's pharmacogenomic measures. The two limiting factors of the prior efforts to individualize antidepressant selection are (1) their lack of ways to parse the heterogeneity of MDD symptom manifestations and the antidepressant response, and (2) the lack of pharmacogenomic biomarkers as predictors in predicting antidepressant response. To move from artisanal medicine to individualized medicine, it will be important to present analytically inferred insights in such a way that psychiatrists or physicians can easily use them to augment the treatment strategies that they have refined through years of practice. To that end, this dissertation addresses the following research gaps.

- **Toward Pharmacogenomic-Guided Antidepressant Selection:** It is promising that several biomarkers of MDD risk and possible disease pathology have shown an ability to predict treatment outcomes. However, illustrated by the case of breast cancer therapeutics, drug selections consists of choosing the treatment that best suits the disease's biological characteristics of the patient [23]. To that end, our work demonstrates the predictability of antidepressant treatment outcomes based on pharmacogenomic markers as predictor variables, with cross-trial replications of the predictive performance (as shown in Chapter 6). Thus, as illustrated in Chapters 5 and 6, ALMOND comprises an analytical workflow to predict antidepressant response using pharmacogenomics

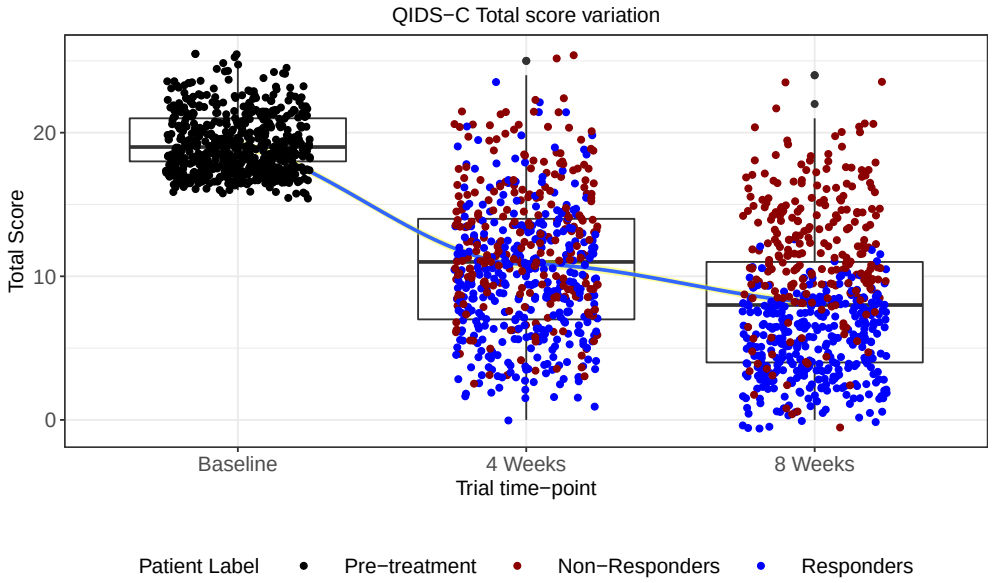


Figure 2.3: Challenges in interpreting aggregate trends of depression severity improvements by using regression-based approaches.

measures.

- **Addressing Homogeneity in MDD Presentation and Antidepressant Response:** As discussed in Chapter 4, ALMOND's analyses begin by inferring patient stratification as a strong foundation for studying biological differences in disease states. We then use the stratification as a basis for defining a probabilistic graphical model, wherein the strata serve as nodes of the graph, through which one can study the most likely longitudinal variation in depressive symptoms after treatment initiation. In standard care settings wherein biological measures are not routinely collected, ALMOND has a symptom-based analytical workflow using probabilistic graphs described in Chapter 7, to provide prognoses and predictions of antidepressant treatment outcomes by studying early changes in symptom severity after treatment initiation.

- **Providing Easily Interpretable Individualized Prognoses:** By using growth mixture models or regression techniques, we can infer the likely depressive severity variation during treatment while going on to achieve a certain categorical treatment outcome. For example, in Fig. 2.3, we show the trajectory of improvement in total depression severity of patients with baseline total QIDS-C scores in the range of 20 – 25. Just as all prior work in trajectory analyses has shown, an improvement in total depression severity at 4 weeks by at least 20% is prognostic of remission or

response at 8 weeks. However, when we look at the box-plot variation or even the scatter points colored by the 8-week response status (i.e., a $\geq 50\%$ improvement in total depression severity from baseline), we see that responders and non-responders are spread across both sides of the mean, which has a very strong estimate (which follows the thin and almost invisible 95% confidence interval band). To this end, ALMOND first uses probabilistic graphs that incorporates information about treatment outcome defined at intermediate time-points along with knowing the trajectories of improvement in severity scores. Then, we are able to provide additional degrees of specificity in individualized prognoses of treatment outcome by saying, *if a patient's total depression severity has changed from X to Y after 4 weeks of treatment, and if A or more core depressive symptoms have improved by B score points, this patient has a $Z\%$ chance of achieving remission.* This is of immense clinical value to psychiatrists in helping them identify what to focus on in individualizing treatment management plans. Today,they can measure the spectrum of depressive symptoms, but are often challenged by heterogeneity in MDD symptom manifestation and treatment response, as illustrated in Figs. 2.1 and 2.2.

CHAPTER 3

INNOVATIONS TO INDIVIDUALIZE ANTIDEPRESSANT TREATMENT MANAGEMENT

Clinicians' *deciding on the right patient-drug match* prior to treatment initiation is only the beginning of individualizing medicine. Prior to and during the course of treatment (e.g., at scheduled hospital visits), clinicians measure the severity of disease symptoms to assess therapeutic efficacy. Clinicians' *decisions to continue or alter current treatments* are based on their ability to forecast the sustained effect of the drugs by observing the patients' symptom variations during early stages of treatment. Those two important decisions that clinicians need to make in order to individualize treatment present opportunities for analytical innovations.

From an analytics perspective, a clinician's ability to (1) *match a patient with the right drug prior to treatment initiation* rests in the ability to analytically combine patient-derived pharmaco-omics measures to predict eventual treatment outcomes, and (2) *decide to continue or change treatment after the current treatment's initiation* rests in the ability to model longitudinal variations of symptom severity that are conditional upon patient's pre-treatment factors and subsequent symptom severity changes during the treatment. Available as inputs to the analyses are the routinely collected high-dimensional (sample sizes << number of variables), patient-level data (e.g., 7 million SNPs, and over 100 clinical measures). A clinician's treatment decision-making capabilities, augmented with analytically identified intelligence are successful when we can *identify strata of patients with homogeneous disease states and treatment response characteristics, such that a few associated clinical or biological measures are highly prognostic or predictive of eventual treatment outcomes.*

Contributions. ALMOND's analytical workflow (illustrated in Fig. 3.1) begins by stratifying MDD patients first by sex, and then by their depression severity, using unsupervised machine learning. Then, through an integration of multiple pharmaco-omics measures in patient strata, we use supervised machine learning to identify a few pharmacogenomic SNPs to augment psychiatrist's depression severity measures which can predict the antidepressant's efficacy prior to treatment initiation. Finally, using the patient strata as nodes of a probabilistic graph, ALMOND aids a psychiatrist's decision *to continue or change the current antidepressant treatment* by identifying depressive symptoms whose early changes (e.g., at 4 weeks) are highly prognostic and predictive of treatment outcomes at 8 weeks.

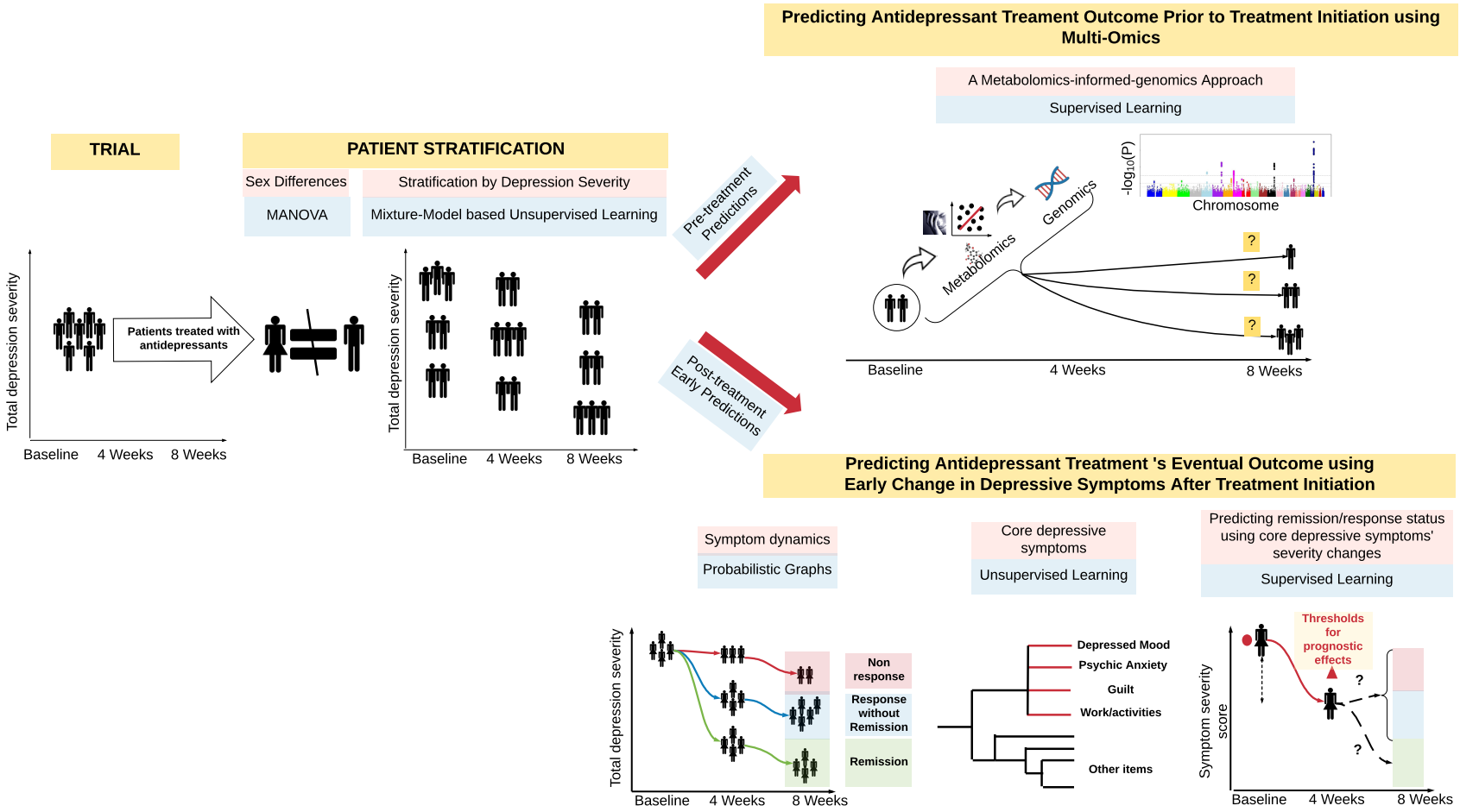


Figure 3.1: ALMOND's analyses workflow for predicting antidepressant treatment outcomes.

3.1 Problem Statements

Given that in current practice there is neither an ability to stratify MDD patients, nor a way to generate homogeneous response profiles to antidepressant treatment, prognoses or predictions of treatment outcomes to achieve individualized antidepressant treatment management have been challenging. Hence, ALMOND's analysis workflow was designed to answer the following questions sequentially; they begin with patient stratification and provide the flexibility of predicting treatment outcomes with and without biological measures.

1. Can MDD patients be stratified by the severity of their individual symptoms, or total depression severity, or both, at all time-points?

2. Are there biological measures (e.g., genomics) associated with patient strata at baseline (pre-treatment) that can predict eventual treatment outcomes?

3. Using patient strata as the disease states at each time-point of the antidepressant treatment, can we model the treatment's progression by jointly considering changes in depression severity and observations made by psychiatrists at time-points after treatment initiation?

4. Given a set of the progressions (e.g., paths of a graph) that patients in a given stratum at baseline are most likely to take to achieve categorical outcomes, is there a set of depressive symptoms (core symptoms) such that their early changes are highly prognostic of eventual (long-term) clinical outcomes?

5. For patients in a given stratum at baseline, can core depressive symptoms' pre-treatment severity and their early changes (e.g., at 4 weeks) predict eventual treatment outcomes at 8 weeks?

3.2 Patient Stratification

Rationale. While individualized treatment decisions for breast cancer are made based on well-defined criteria for patient stratification (e.g., tumor subtypes), there are no criteria for stratifying patients with MDD to guide antidepressant treatment selection. While the prevalence of MDD in women is double that in men, sex is often overlooked as a factor in guiding antidepressant selection [80–82]. Those reasons motivated us to find a way to stratify patients first by sex, and then by their total depression severity. Stratification by sex allows one to study biological and symptomatic variations in how men and women

respond to antidepressant treatment, instead of merely accounting for sex as a variable in predictive models. If the biological mechanisms of response is indeed different in men and women, the choice of an antidepressant will have to be sex-dependent. The stratification by total depression severity is justified even though individual depressive symptom severity manifests differently in different patients. There are two reasons. First, all existing machine learning prediction approaches have shown that pre-treatment depression severity is the highest-ranked predictor of remission from MDD symptoms in response to antidepressant treatment [28, 29, 59]. Second, an treatment's efficacy at each time-point of the treatment is measured by the extent of improvement in the total depression severity score [30].

3.2.1 First: Stratification by Sex

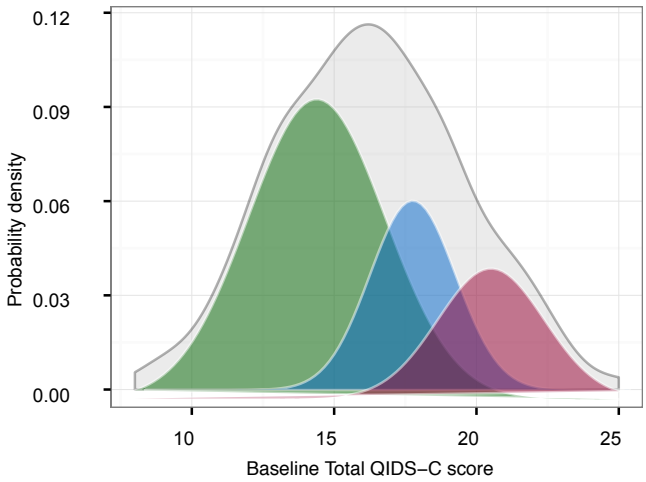
To anticipate sex differences in metabolomic profiles based on prior work, we used multivariate analysis of variance (MANOVA) to determine sex differences in metabolite concentrations of PGRN-AMPS data at baseline and after 4 and 8 weeks of treatment. The reason for choosing metabolomics data to identify sex differences at all time-points is that prior work has not consistently reported significant sex differences in clinical/demographic factors or MDD symptom manifestation of patients with MDD [28, 29, 59, 80–82].

3.2.2 Second: Stratification by Total Depression Severity Scores

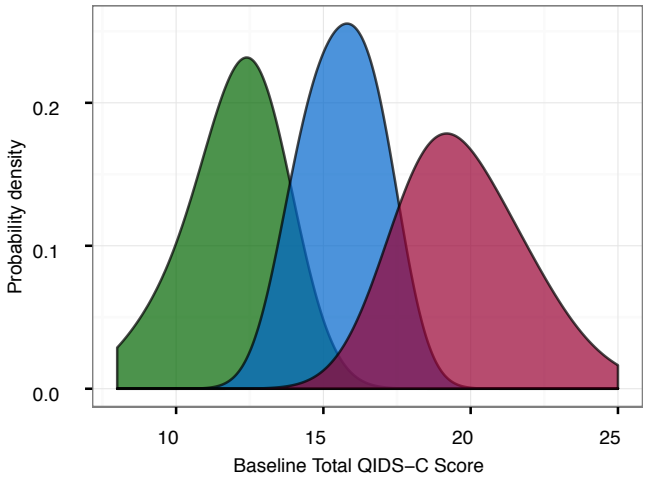
Unsupervised learning was used to identify clusters (stratification) of patients based on total QIDS-C and HDRS scores at baseline, 4 weeks, and 8 weeks.

Observation: The p-value from the Shapiro-Wilk test of the total score (e..g, in QIDS-C and HDRS) from all three time-points of the trial, and in both men and women, was less than the significance level ($\alpha = 0.05$). This meant that the symptom severity scores were not normally distributed, as we rejected the null hypothesis of the Shapiro-Wilk test (i.e., that the data are normally distributed).

Approach: The fact that symptom severity is not normally distributed meant that the k-means clustering algorithm would not be suitable as a clustering algorithm here. Without a loss in generality, under the assumption that the data (x: total QIDS-C/HDRS score) were distributed as a mixture of Gaussians (referred to as a *Gaussian mixture model*, or GMM), we developed the patient stratification workflow (Algorithm 1). Starting with an assumption that the data have at least two components in the GMM, we used the expectation maximization (EM) algorithm to estimate the sufficient statistics parameters of the Gaussian components (mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$) of the GMM as shown in Fig. 3.2(a). Using the function



(a) Estimating distributions



(b) Clusters from inferred GMM

Figure 3.2: Probability distributions of total depression severity scores and clustering using Gaussian mixture models. Fig. (a) illustrates the inference of mixtures comprising the distribution of symptom severity scores. Fig. (b) illustrates distribution of symptom severity within the clusters inferred using the sufficient statistics of the components that were inferred in Fig. (a).

generateSamples, $10,000$ samples were randomly drawn from the inferred distributions. Next, the Kolmogorov-Smirnov test (ks.test) was used to test whether the distribution of the generated data was statistically similar to that of the original data. If the p-value (p) was less than the significance level ($\alpha = 0.05$), then we rejected the null hypothesis that the two distributions were not similar. If that happened, the number of components was increased by one, and tested for similarity in the two distributions. Once we obtained the minimum number of components K in the GMM was obtained for which the generated and input data's distributions were similar, K clusters $\mathcal{C} = \{C^k; \forall k \in 1 : K\}$ ordered by the increasing mean (μ_k) of the components were the outputs of the workflow [83]. Patients were assigned to the component that maximizes the likelihood $\mathcal{L}(x)$ given the component's sufficient statistics (gmmCluster), as illustrated in Fig. 3.2(b) and described by Equation 3.1.

$$\underset{k \in [1:K]}{\operatorname{argmax}} \mathcal{L}_k(x) \text{ where } \mathcal{L}_k(x) = \mathcal{N}(x, \mu_k, \sigma_k^2) \tag{3.1}$$

Algorithm 1 Patient stratification

Input: $x \leftarrow$ Total QIDS-C Scores

$k \leftarrow 2$
$\mathcal{C} \leftarrow \emptyset$
$\alpha \leftarrow 0.05$
$p \leftarrow 0$
while $p \leq \alpha$ **do**
  $\{\boldsymbol{\mu}, \boldsymbol{\sigma^2}\} \leftarrow$ EM(x, k)
  $x' \leftarrow$ generateSamples $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$
  $p \leftarrow$ ks.test(x, x')
  if $p >$ significanceLevel **then**
    $\mathcal{C} \leftarrow$ gmmCluster$(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$
  end if
  $k \leftarrow k + 1$
end while

Output: $\mathcal{C}$

3.3 Multi-omics Integration to Predict Antidepressant Response

The formalism for integrating multiple biological measures in this case study is as follows and is illustrated in Fig. 3.3.

Just as tumor subtypes serve as a foundation for integrating biological measures in oncology, our formalism first established patient subtypes/stratification $\mathcal{C}$ by using mixture-model-based unsupervised learning techniques. In the first layer of overlaying of the biological measures, a set of metabolites $m \in \mathcal{M}$ were identified based on significant associations of their concentrations with symptom severity in previously inferred patient stratification. In the second layer of the overlay of biological measures, in what is referred to as a *metabolomics-informed-genomics* approach, we used GWAS to identify SNPs $g \in \mathcal{G}$ that are associated with concentrations of metabolites that comprise m.

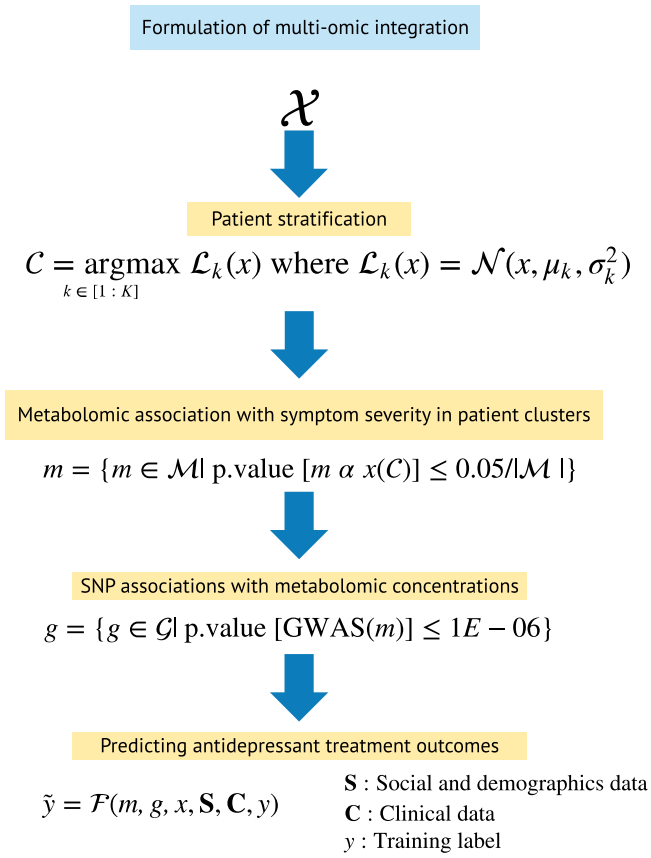


Figure 3.3: The proposed approach to integrating multiple omics (metabolomics and genomics) measures.

Through iterative overlaying of biological measures starting with metabolites (blood measures that reflect drug action) associated with depressive severity, and then addition of genes associated with metabolomic concentrations, the biological measures became more closely associated with the molecular mechanisms of antidepressant response. Finally, out of the more than 7 million possible predictor variables, the proposed approach identified about 65 predictor variables that comprised (1) SNPs (g) identified by the GWAS based on metabolomic concentrations, (2) metabolites (m) whose concentrations are significantly associated with depression severity in patient clusters, and (3) clinical measures (discussed in Chapter 5). Thus we made the size of the predictor data computationally tractable to predict clinical outcomes $\tilde{y}$ by using supervised learning methods $\mathcal{F}(m, g, S, C, y)$, where y is the treatment outcome labels of the training data.

3.4 Modeling the Symptom Dynamics of Antidepressant Response

Rationale. First, *symptom dynamics* in this work are defined as the likely changes in a patient's symptoms and associated clinical outcomes during the various stages of a trial (e.g., *response* at 4 weeks or 8 weeks) while he or she is being treated with antidepressants. To infer the symptom dynamics paths, we used the patient stratification as nodes of a probabilistic graph, wherein the graph's edges were the transitions patients make between clusters of consecutive time-points. Probabilistic graphs (e.g., Bayesian networks, hidden Markov models, and factor graphs) provide the mathematical foundation needed to model conditional dependencies that closely follow a clinician's treatment logic, i.e., accounting for improvement in total depression severity, conditioned upon both baseline depression severity and change in depression severity at intermediate time-points, in a purely data-driven manner, without a priori specification of trajectories. Our goal was to have a graphical model that not only models conditional dependencies of depressive symptom changes, but also incorporates any statistical functions that capture any annotations that clinicians routinely make during their practice (e.g., clinician defining an outcome at intermediate time-points). Hence, we used factor graphs [84, 85] to capture the relationships between clusters of patients at consecutive time-points of the trial and associated variables, such as clinical outcomes. The choice of factor graphs was driven by their ability to provide a compact, expressive representation of random variables and to subsume Bayesian networks, Markov random fields (MRFs), and hidden Markov models [84, 85]; further, they have been shown to be effective in modeling longitudinal electronic health record data of diabetic patients [13].

Probabilistic Graph from Patient Stratification: The factor graph is a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{F})$. We created separate factor graphs for men and women. Each graph has three layers at each time-point, as illustrated in Fig. 3.4; the *clinical observation layer*, where the clinician observes the clinical outcome based on symptom severity; the *patient symptom response layer* that keeps track of changes in symptoms; and the *patient stratification layer* to illustrate the cluster to which a patient's symptom score belongs. Each layer is associated one variable node $\in \mathcal{V}$ such as $\mathcal{O}$ (distribution of patients who demonstrate response (R) vs. no response (NR)), $\mathcal{X}$ (symptom measure at each time point), $\mathcal{C}$ (patient stratification at each time point) and one associated factor node $\in \mathcal{F}$ such as a decision rule to determine if a patient has demonstrated response (50% reduction in symptom from baseline) for random variable $\mathcal{O}$, a transition probability matrix for symptom severity between two time points for random variable $\mathcal{X}$, and what cluster $\mathcal{C}$ the patient belongs to based on his or her current symptom severity score $\mathcal{X}$. The graph can be evaluated at each time point of the trial $t \in T$ starting from baseline (t) to 4 weeks ($t + 1$) to 8 weeks ($t + 2$) and so on.

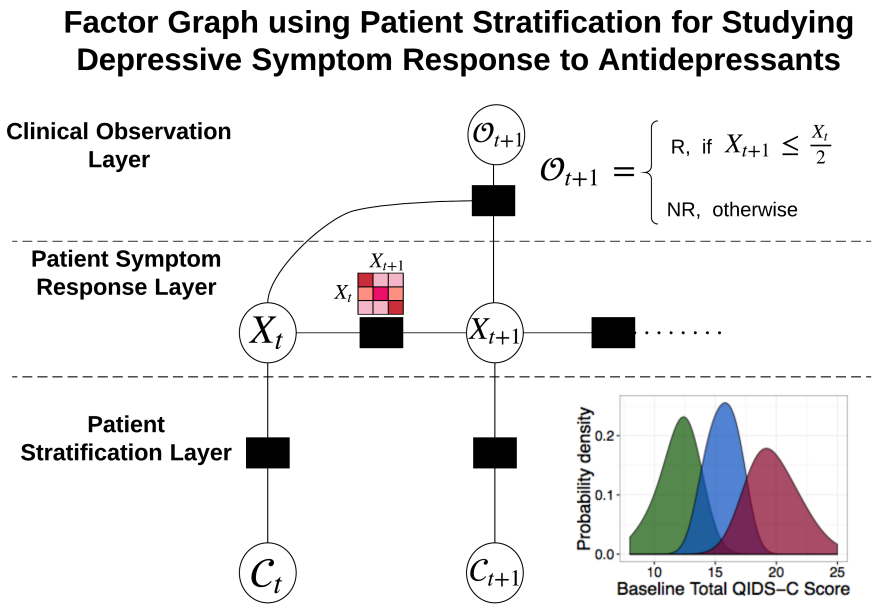
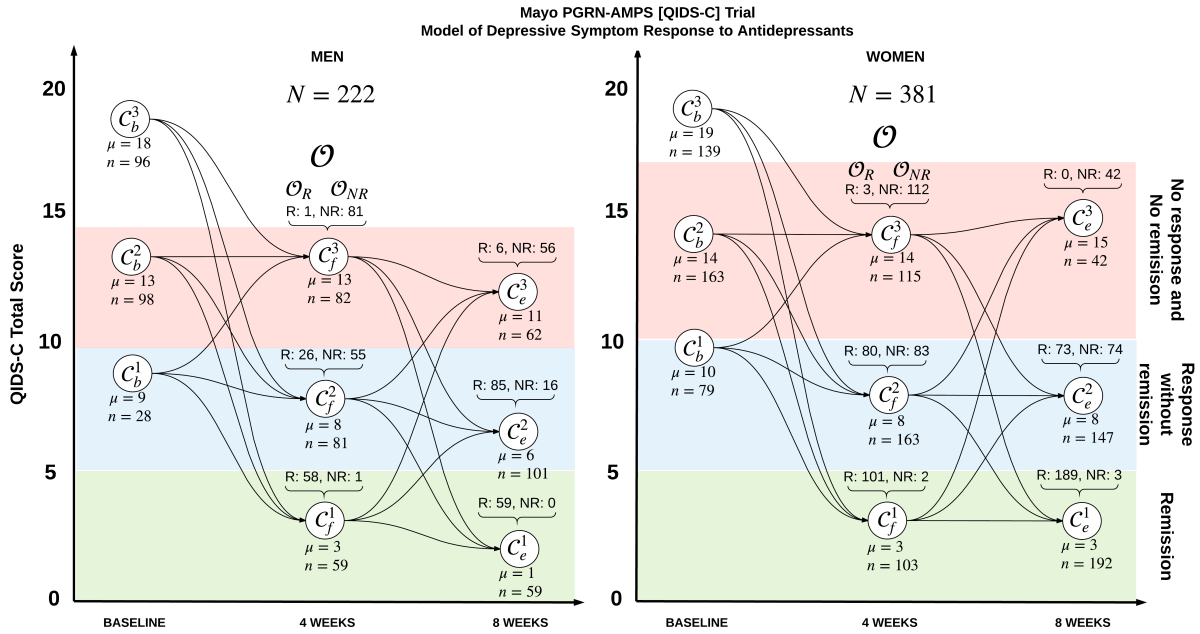


Figure 3.4: Patient stratification can be used to form the factor graph as illustrated. The graph is bipartite, with patient stratification (one set of nodes) and factor functions (the other set of nodes) that capture relationships between the symptom severity associated with the stratification and other data.

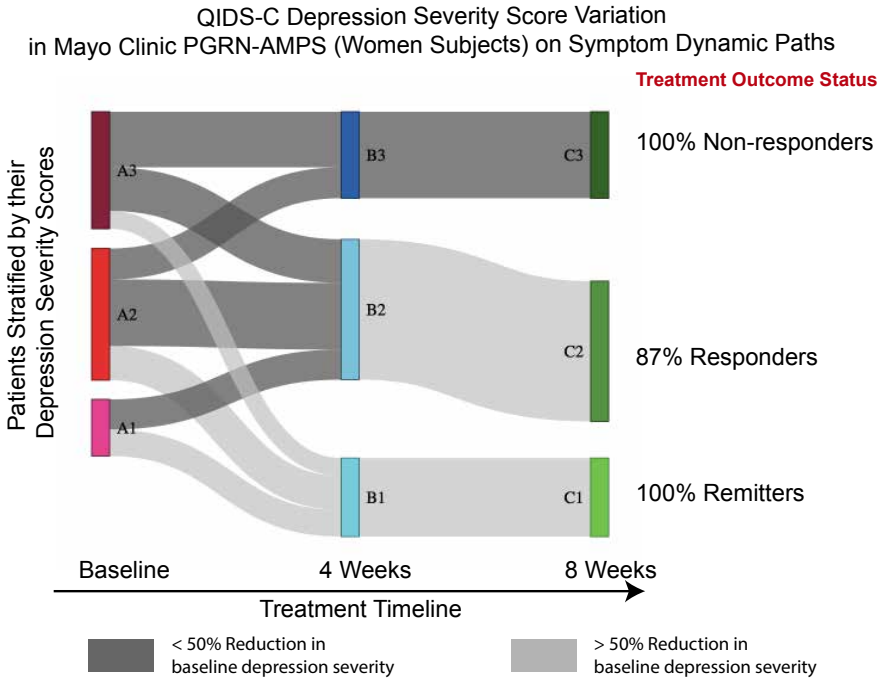
Computing Path Likelihoods: We use the forward algorithm [84, 85] to identify the most likely *forward* transitions a patient starting in any baseline cluster will make between clusters (*hidden states* –$\mathcal{C}$) of the trial, and also what the associated clinical outcomes will be during the transitions (*observed states* –$\mathcal{O}$). During transitions between the clusters, the clinician/psychiatrist assessing the patient observes the clinical outcome $\mathcal{O} = \{\mathcal{O}_R, \mathcal{O}_{NR}\}$, which is that the patient has demonstrated either *response* ($\mathcal{O}_R$), or *no-response* ($\mathcal{O}_{NR}$). For both men and women, the graphs with the number of patients (n), forward transitions, and observed outcomes $\mathcal{O} = \{\mathcal{O}_R, \mathcal{O}_{NR}\}$ in each cluster are shown in Fig. 3.5(a), which is similar in construct to a hidden Markov model (HMM). Now, the symptom dynamics for any patient starting in any of the clusters at baseline can be solved recursively by using the *forward algorithm*, which is described as,

$$P_{\mathcal{O}}(\mathcal{C}_t) = \sum_{t \in T} p(\mathcal{O}|\mathcal{C}_t) P_{\mathcal{O}}(\mathcal{C}_{t-1}) p(\mathcal{C}_{t-1} \rightarrow \mathcal{C}_t) \tag{3.2}$$

where $p(\mathcal{O}|\mathcal{C}_t)$ is the probability of the observation (response or no response) in a current state; $p(\mathcal{C}_{t-1} \rightarrow \mathcal{C}_t)$ is the probability of a transition from a state of a previous time-point to a state of the current time-point (e.g., $\mathcal{C}_b^1 \rightarrow \mathcal{C}_f^2$); and $P_{\mathcal{O}}(\mathcal{C}_{t-1})$ is the path probability for a



(a) Patient stratification and associated composition of clusters according to the associated clinical outcomes (R = *response*, NR = *no response*).



(b) Maximum likelihood symptom response paths that result from treatment with antidepressants.

Figure 3.5: Symptom dynamics using HMM. Fig. (a) illustrates the HMM of the symptom dynamics in men and women; Fig. (b) illustrates the inferred most likely symptom dynamics in women based on the cluster in which each patient starts in the trial using a Sankey diagram. Thicker lines indicate that a larger proportion of patients from the originating cluster are taking a particular path.

given set of observations $\mathcal{O}$ seen until $\mathcal{C}_{t-1}$.

Note that *the reduction from the factor graph to the HMM did not simplify the complexity of solving the forward algorithm, but rather allowed us to explain not only the symptom dynamics as a function of how symptoms change, but also, with the changes in symptoms, the associated clinical outcomes during various time-points of the trial.*

3.5 Identifying Core Depressive Symptoms and Associated Antidepressant Effects

3.5.1 Core Symptoms

To extract homogeneous patterns of antidepressant response, we defined "core depressive symptoms" based on three criteria: (1) similar response patterns at all time-points, (2) low inter-individual variability, and (3) patterns of change that were statistically distinct within each of the symptom dynamic paths (which we inferred in Stage 1 using total depression severity scores). First, unsupervised machine learning (i.e., hierarchical clustering with complete linkage) was used to identify individual QIDS-C and HDRS scale items with similar rating patterns (meaning that they were clustered together) within the patient clusters at baseline, 4 weeks, and 8 weeks. Second, we identified symptom clusters wherein clinician ratings for each of the scale items at baseline had a nonzero median and low inter-individual variability. A given item was defined as having low inter-individual variability if the chi-square test for the distribution of clinician ratings was significant after multiple comparisons, with the null hypothesis being that the distributions of ratings for that item were equal. Third, for each pair of symptom dynamic paths originating at a baseline cluster and transitioning to a cluster at 8 weeks, the Kolmogorov-Smirnov test was used to determine whether there were statistically significant differences between the associated distributions of core symptoms at 4 weeks. We used average smoothing kernels to visualize the variations in these core symptoms' scores within specific symptom dynamic paths.

3.5.2 Antidepressant Effects on Core Symptoms

The Mann-Whitney U-test was used to assess whether the severity of the core depressive symptoms (expressed as a rank order) changed significantly as a likely response to antidepressant treatment between two consecutive time-points on a given symptom dynamics path. By utilizing the replicating patient clusters across datasets, we satisfied the sample independence

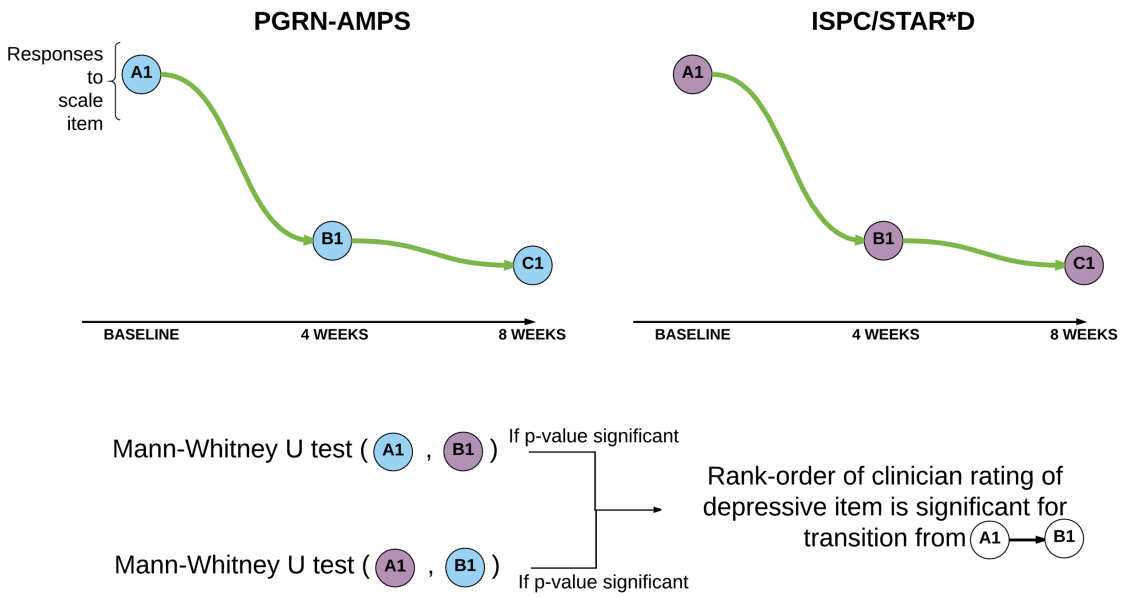


Figure 3.6: Illustrating the rank-order test construction.

requirement for this test by comparing patients in one cluster from one dataset with patients in the consecutive time-point cluster from another dataset (illustrated in Fig. 3.6).

An example of the test conducted in citalopram or escitalopram-treated patients is illustrated in Fig. 3.6. Consider a pair of clusters from consecutive time-points (A1 and B1) on a given path (A1 $\rightarrow$ B1 $\rightarrow$ C1), and a specific item from the QIDS-C scale (sad mood). We conducted two rank-order tests for the two clusters, A1 and B1. First, we tested whether changes in clinician ratings of severity for the QIDS-C sad mood item were significant between patients in the A1 cluster of PGRN-AMPS vs. patients who transitioned into the B1 cluster (4 weeks) from the A1 cluster (baseline) in STAR*D. Second, we tested whether changes in clinician ratings of severity of the same item were significantly different between patients in the A1 cluster of STAR*D vs. patients who transitioned into the B1 cluster from the A1 cluster in the PGRN-AMPS trial. If the p-value (with Bonferroni correction for multiple comparisons) was significant in both comparisons, then we conclude that the changes in clinician ratings of depressive items' observed severity were more likely due to antidepressants than to chance. Similar analyses were performed for all other pairs of clusters, including paths between 4 weeks and 8 weeks, and for other depression scales and antidepressants.

3.6 Prediction and Prognoses of Antidepressant Response Using Early Change in Core Depressive Symptoms

3.6.1 Prediction

We used random forests, a nonparametric supervised machine learning method, as a binary classifier to predict clinical outcomes at 8 weeks given a specific baseline cluster, using the associated baseline severity of the core depressive symptoms and their absolute changes at 4 weeks. Using five-repeat tenfold nested cross-validation, we trained the sex- and rating-scale-stratified classifiers with data from PGRN-AMPS subjects. Using R's randomForest library, we followed the recommended practice of grid search during training by setting the mTry parameter to one-half of the total number of predictor variables, and chose the number of trees from the range of 500 to $2,000$ with increments of 100. The trained prediction models were then externally validated using STAR*D subjects (for the QIDS-C scale) and ISPC subjects (for the HDRS scale). Prediction performance was reported using several metrics (AUC, PPV, NPV, sensitivity, and specificity), and the statistical significance of the classifier's accuracy was established using the null information rate (NIR, the prevalence of the class with the largest samples), which served as a proxy for chance.

3.6.2 Prognoses

This step defined the minimum number of core symptoms and levels of improvement in the core symptoms needed at 4 weeks (given a specific baseline cluster) to achieve specific outcomes at 8 weeks. First, the threshold of improvement vs. failure to improve was chosen based on changes in median scores on symptom dynamic paths between the baseline and 4-week clusters. Second, a chi-square test was conducted on a table comprising the number of core symptoms that exceeded (or failed to exceed) the threshold at 4 weeks, versus the outcome labels (e.g., remitters vs. non-remitters, or responders vs. non-responders). If the chi-square test's p-value was significant for remission or response/non-response, we computed the probability of the outcome based on how many symptoms had to exceed (or failed to exceed) the threshold. If the p-value was not significant, no conclusions about treatment outcome based on changes in any number of core symptoms were possible. We established standard deviations (SD) of the probabilities by creating ten sets of, five different random subsets (each of which maintained the same proportions of patients who achieved remission, response and non-response in the entire dataset).

3.7 Significance of Contributions

From a computer engineering perspective. ALMOND demonstrates that even when disease characteristics and drug response vary between patients, augmentation of a physician's assessment of disease severity with a few biological measures can significantly improve the predictability of the drug treatment outcomes. To achieve a high degree of predictability in treatment outcomes, we overcame two significant challenges. First, we developed the ability to integrate and analyze multiple biological measures that are high-dimensional in nature (i.e., the number of variables is significantly greater than the number of samples). Second, the we developed ability to derive a compact representation of trajectories of treatment response by reducing the state-space complexity of the longitudinal characteristics of treatment outcomes.

- **On Handling the Complexity of High-Dimensional Omics Data with Subjective Clinical Assessments:** Much as breast tumors can be stratified based on distinct genomic characteristics, other omics data (e.g., transcriptomics) can be associated with a few genes that differentiate disease strata. However, it has been challenging to find replicable patterns of genomic stratification in the context of MDD, since the disease manifestation and treatment outcomes are heterogeneous (as described in Chapter 2). Subsequent integration of multi-omics measures has therefore been both a clinical and a computational challenge. This dissertation shows that in the presence of heterogeneous disease or drug response traits, the data-driven unsupervised methods in ALMOND can infer a patient stratification in terms of disease severity measures that replicates in independent studies. Such stratification can serve as a foundation for the association of blood-based measures (e.g., metabolomics) with the severity of the disease or with drug response mechanisms. Finally, the metabolites associated with patient strata (e.g., serotonin) serve as quantitative biological traits that can be associated with genome-wide genomics data that comprise over 7 million SNPs. Hence, even when the patient sample size which was less than $3,000$ in our study is several orders of magnitude smaller than the genomic variations in genome-wide genomics (i.e., 7 million SNPs), ALMOND has shown biologically relevant and computationally tractable ways to integrate heterogeneous biological measures with disease severity measures obtained through clinical assessments. Such integration allowed us to find a few focused biological predictors of antidepressant response, which yielded significant improvement over use of sociodemographic factors with clinical assessments as the sole predictors. Those innovations taken together are novel, since several methodologies exist for integrating multiple biological measures when biological traits are well-defined, as in breast cancer [86], but their translation into clinical utility in treating depression

has been limited.

- **On Handling the State Space Complexity and Interpretability in Graphical Models:** To model any phenomenon (e.g., antidepressant response), an abstraction of the phenomenon is needed. The extent of abstraction should be realistic so that it meets the decision criteria of domain experts (psychiatrists, in this case), but also allow for ease of interpreting the results. Psychiatry, in treating mood disorders, is abundant with vast abilities to symptom severity, sociodemographic measures, and biological measures. However, there has not been a formal definition of "disease state" in the data. The ability of ALMOND to define the disease states allowed for vertical integration of biological measures to predict a treatment's outcome prior to its initiation, and also to explore longitudinal variations that best explained eventual treatment outcomes as a function of where patients began in the trial, and how they progressed during the trial from a clinician's perspective. Such interpretations have so far not been possible with existing machine learning approaches, which have attempted to predict treatment outcomes without accounting for patient subtypes either at baseline (pre-treatment), or during treatment. Hence, the ability to establish a patient stratification proved to be foundational in defining an abstraction for modeling and predicting antidepressant response.

In Psychiatry. The single greatest limiting factor in individualization of antidepressant treatment has been the heterogeneity of antidepressant response. ALMOND's workflow began with the intention of finding subgroups of patients at baseline who could potentially have longitudinal homogeneity in their response to antidepressant treatment. This approach has the following significance in augmenting clinicians' intelligence to individualize antidepressant treatment management.

- **On Reduced Complexity in Comprehending Antidepressant Response:** Even for a highly experienced psychiatrist, it would be impossible to memorize the 986 unique transitions ($\sim O(N^k)$ complexity) between depression severity scores ($N = [0:35]$ in QIDS-C) at three ($k = 3$) time-points in order to forecast a new patient's possible response to treatment. By using patient stratification and probabilistic graphs, ALMOND inferred nine unique paths originating at three baseline clusters traversing through three clusters at 4 weeks, and terminating in three clusters at 8 weeks. As the clusters at 8 weeks conformed to the three clinically accepted categorical outcomes, *these nine paths explained the most likely paths of antidepressant response in over* 80% *of patients while also explaining the most likely early clinical outcome at* 4 *weeks.* That

represents a $\sim 109x$ reduction in complexity, when not stratifying patients. To explain all 100% of the patient's response to treatment, 81 paths are needed, which is still a $\sim 12x$ reduction in complexity when not stratifying patients. In comparison to the response trajectories not stratifying patients (as previously shown in Fig. 2.2), the paths inferred after patient stratification are more comprehendable by clinicians as they begin to associate the dynamics of antidepressant response. This is possible by identifying which cluster the patient began at baseline, and clinical response needed at 4 weeks in order to most-likely observe an outcome at 8 weeks.

- **On Increased Specificity in Prognoses of Treatment Outcomes:** First, ALMOND describes the transition a patient makes from a baseline cluster to a cluster at 8 weeks, through a cluster at 4 weeks. Then, reflecting on the change in total depression severity, additional specificity was given to establish the prognoses of categorical treatment outcomes based on changes at 4 weeks in severity of a small set of depressive symptoms. ALMOND's prognoses are stated in this form: *"for total depression severity improvements between ranges a and b at baseline and 4 weeks, respectively, if x depressive symptoms have improved by $\geq$ y points, then the chance of remission at 8 weeks is z%."* Such specificity in prognoses of treatment outcomes has not been available in psychiatry for treating MDD thus far.

- **On the Promise for Predicting Treatment Outcomes When Pharmacogenomic Measures Are Used:** While there has been success of breast cancer therapeutics in matching patients with suitable treatments by using pharmacogenomic markers, previously there has been no exploration of whether pharmacogenomic SNPs could aid in antidepressant treatment selection. ALMOND provides the ability to algorithmically select an antidepressant prior to treatment initiation based on the highest likelihood of remission at 8 weeks, using patient-derived pharmacogenomic markers as predictors.

CHAPTER 4

PATIENT STRATIFICATION

Individualized medicine in practice optimizes therapeutic options based on patient "subtypes" (strata) based on specific biological/clinical characteristics. Close association of the strata's trait (e.g., tumor subtypes) with a set of biomarkers (e.g., genes) can be prognostic of treatment outcomes, as in the case with breast cancer therapeutics [87]. The need to stratify patients is at the heart of understanding heterogeneity in MDD disease states and antidepressant response characteristics. Currently, there are no established mechanisms by which patients with MDD are stratified at baseline (prior to treatment) or during the treatment's intermediate time-points. However, at the end-point of the trial, patients are triaged into remitters, responders without remission, and non-responders. Given the lack of replication in findings across trials, it has remained to be explored whether *patient stratification is possible that replicates across multiple trials and rating scales.*

Contribution. We first stratified patients by sex, based on observed differences in the metabolomic profiles of men and women prior to, during, and after citalopram/escitalopram treatment in the Mayo Clinic Pharmacogenomics Research Network Antidepressant Medical Pharmacogenomic Study (Mayo PGRN-AMPS) [24]. Then, using mixture-model based unsupervised learning, we inferred three clusters of men and women separately at all three time-points (baseline, 4 weeks and 8 weeks) of the Mayo PGRN-AMPS trial based on their total depression severity score derived separately from QIDS-C and HDRS. We validated our clustering approach in patients treated with the same drugs, in two independent datasets, Sequenced Treatment Alternatives to Relieve Depression (STAR*D) [25] for the QIDS-C scale, and International SSRI Pharmacogenomics Consortium (ISPC) [26] for the HDRS scale. In both men and women, clusters at 8 weeks had clinical validity based on outcome definitions outlined in Sec. 2.1.1. The first cluster included all patients who achieved remission; the 87% of those who achieved response but not remission comprised in second cluster; and the third cluster included all patients who achieved neither response nor remission. Our successful replication of patient stratification with clinical validity demonstrates that we have a strong framework for studying the integration of pharmaco-omics measures, and the longitudinal variations in depressive symptoms in response to antidepressant treatment.

4.1 Problem Statements

The following problem statements allow us first to stratify patients by sex, and then to stratify patients by their depression severity.

1. Are there sex-differences in biological profiles of patients with MDD before and after antidepressant treatment?

2. Can MDD patients be stratified by their depression severity?

3. Does patient stratification replicate across trials and rating scales?

4.2 Data

The Pharmacogenomics Research Network Antidepressant Medical Pharmacogenomics Study (PGRN-AMPS, NCT 00613470) was a single-arm, open trial designed to assess antidepressant effects of citalopram/escitalopram over 8 weeks in adults (aged 18 – 84 years) with MDD, and to examine metabolomic and genomic predictors of those outcomes [88]. Subjects were recruited from primary and specialty care clinics from March 2005 to May 2013. Psychiatric diagnoses were confirmed using modules A, B (screen-only version), and D of the Structured Clinical Interview for DSM-IV (SCID) [50].

Data from complete cases (baseline, 4-, and 8-week data) of step-1 of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (NCT 00021528) and International SSRI Pharmacogenomics Consortium (ISPC) were used to test the reproducibility of patterns of depressive symptom response inferred in the PGRN-AMPS study. Descriptions of the STAR*D and ISPC studies have been previously published [25, 26]. Briefly, phase 1 of STAR*D was a 12-week clinical trial of citalopram for adults (aged 18 –75 years) with MDD who were recruited from primary and specialty care settings in the United States between June 2001 and April 2004. Our analyses included data from 899 STAR*D subjects with complete step 1 clinical response data at 23 – 33 days (4 weeks) and 51 – 61 days (8 weeks). The ISPC dataset comprised pooled data from seven clinical trials of SSRIs for depression carried out in North America, Europe, and Asia, in order to examine genetic factors underlying variation in antidepressant responses [26]. Of the 998 ISPC subjects, we used data from 344 subjects who were treated with citalopram/escitalopram and had data at 4 and 8 weeks. Table A.1 summarizes the social and demographic characteristics of included subjects from each of the three datasets.

4.3 Approach Overview

We propose a two-stage analyses workflow to address the aforementioned problem statements.

4.3.1 Sage - 1: Stratification by Sex

Aim — Establish sex differences in metabolomics profiles. Anticipating sex differences in metabolomic profiles based on prior work [89], we used multivariate analysis of variance (MANOVA) to determine sex differences in metabolite concentrations of PGRN-AMPS data at baseline and after 4 and 8 weeks of treatment. By combining data from both sexes to predict antidepressant treatment outcomes, it is possible to investigate the predictive capability of sex as a variable. However, given a lack of consistent sex differences in antidepressant outcomes [28, 29, 80, 81], we stratified by sex to investigate potential sex differences in biological and clinical predictors of treatment outcomes.

4.3.2 Stage - 2: Stratification by Total Depression Severity

Aim — Identify depressive symptom severity clusters in PGRN-AMPS (Stage 2A), replicate the cluster patterns using STAR*D and ISPC data (Stage 2B), and identify sociodemographic factors associated with clusters (Stage 2C). As described in Sec. 3.2.2, we first observed that distribution of depression severity scores was a mixture of distributions. We then used mixture model-based unsupervised learning with Gaussian mixture models (GMM) was used to algorithmically identify the minimum number of Gaussians that best approximated the actual distribution of depressive symptom severity in Mayo PGRN-AMPS subjects. By using this approach, we assumed that the mixture comprised multiple Gaussians. GMM clustering has been applied in numerous fields requiring separation of data types characterized by unique distributions [11, 90]. Given the eventual goal of associating biological measures with depression severity during discrete treatment time-points when clinical assessments are performed, longitudinal clustering/trajectory techniques [91] were not suitable. This is because symptom improvement (trajectories) is conditioned upon baseline severity and subsequent improvement (i.e., not independent), and depression severity is assessed at discrete time-points (as opposed to continuous time measures). Therefore, using a GMM clustering algorithm in our approach, we assigned patients to clusters based on their total depression severity score at each time-point. To validate the clustering approach developed in Stage 2A, we used STAR*D (for QIDS-C) and ISPC (for HDRS) datasets in Stage 2B to investigate whether the distributions of depression severity using Kolmogorov-Smirnov

test, were the same between two independent datasets.

4.4 Results

4.4.1 Sex Differences in Metabolomic Profiles

Plasma concentrations of several metabolites differed significantly by sex at baseline and 4 and 8 weeks, regardless of response/remission status or the depression rating scale (QIDS-C/HDRS, defined in Table 2.1) used to define outcomes as tabulated in Table 4.1. The specific metabolite changes that occurred after citalopram/escitalopram initiation also differed by sex and by depression rating scale and outcomes as illustrated in Fig. 4.1. There were significant changes from baseline in the concentrations of 5HT in men and women who were classified as remitters (defined at 8 weeks), and as responders at 4 and 8 weeks, irrespective of the depression scale that was used (QIDS-C, HDRS). Significant changes from baseline in MHPG concentrations were also observed in men and women, for nearly all outcome types (remission, 4-week response, 8-week response) defined by the QIDS-C or HDRS. Based on these results, we proceeded through the remaining stages of the workflow using separate strata defined by sex and by depression rating scale.

4.4.2 Total Depression Severity Clusters

Our unsupervised learning approach algorithmically identified three distinct clusters of men and women ($p < 1.3E-09$) based on their total depressive symptom severity at baseline (A1, A2, A3) and after 4 weeks (B1, B2, B3) and 8 weeks (C1, C2, C3) of SSRI treatment in PGRN-AMPS as illustrated in Fig. 4.2. The nine depressive symptom clusters in men and women were labeled such that 3 represented the most severe symptom cluster, 1 represented the mildest symptom cluster, and 2 represented an intermediate symptom cluster. Importantly, in both men and women, C1 included all patients who achieved remission status, C2 included 87% of patients who achieved response but not remission status, and C3 included all patients who achieved neither response nor remission status at 8 weeks.

We also tested to see if by using responses to individual depressive symptom items on full rating scales, the clusters would replicate, and have the clinical validity at 8 weeks. Comparison of depressive symptom clustering behavior using hierarchical clustering when using individual depressive item scores are shown in Fig. 4.3. The ecological validity of the GMM clustering approach using univariate depression severity scores is represented by the

Table 4.1: Sex differences in metabolomic profile.

Time-point	Metabolite[a,b]	Men		Women	
		Mean[c]	Std. Dev	Mean[c]	Std. Dev.
Baseline	4HPLA***	116.85	38.52	96.26	34.37
	DTOCO**	84.31	35.04	69.52	43.18
	GTOCO1**	79.85	39.83	65.82	41.95
	GTOCO2**	112.94	114.14	87.30	45.34
	GUANOSINE*	122.92	38.88	112.24	32.50
	KYN*	108.39	27.40	100.28	32.59
	MET**	120.38	44.60	106.13	38.87
	TRP***	108.22	20.51	98.52	22.50
	URIC***	115.67	26.45	91.39	24.56
4 Weeks	4HPLA***	115.43	37.67	95.32	33.86
	GUANOSINE**	115.84	38.88	104.86	30.50
	KYN*	107.76	28.27	98.61	32.34
	PARAXAN***	123.98	100.71	87.50	71.27
	TRP***	107.30	22.12	96.32	20.30
	URIC***	120.58	26.23	89.414	26.25
	XAN*	114.10	145.68	83.55	68.82
8 Weeks	4HPLA***	124.90	44.17	96.49	32.32
	5HT**	41.78	89.21	23.94	20.33
	CYS**	100.78	40.08	85.29	36.70
	DTOCO*	84.70	43.23	72.63	38.11
	GTOCO1*	80.71	44.24	70.17	38.44
	GTOCO3*	70.52	45.43	84.37	56.55
	GUANOSINE***	118.93	37.04	104.39	32.71
	I3AA**	115.48	72.08	92.56	70.74
	KYN**	113.76	28.61	100.80	27.87
	TRP***	112.85	23.24	98.67	22.36
	TYR**	117.50	31.97	104.60	33.52
	URIC***	122.23	26.05	89.48	24.90
	XAN*	90.96	110.06	69.03	41.59

[a] See Table A.2 for definitions of abbreviated names for each metabolite.

[b] Between-group comparisons (men vs. women): *p<0.05, **p<0.01, ***p<0.001

[c] All mean concentration values are percent pools from the LCECA platform.

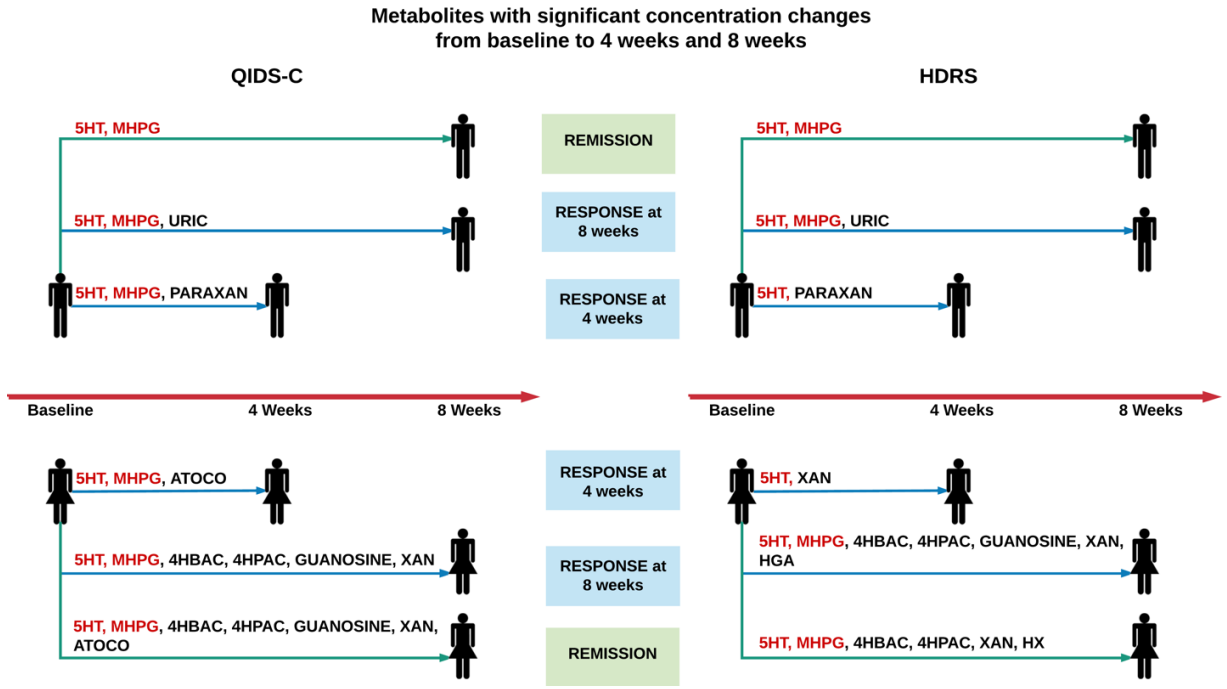


Figure 4.1: Metabolites exhibiting significant concentration changes between baseline and 4 weeks, and baseline and 8 weeks, in depressed patients, stratified by clinical outcome. Differences between men and women by outcome type were observed for the remaining metabolites (shown in black text).

fact that the C1 clusters for the QIDS-C and HDRS fall entirely within the range of scores defining remission. Neither of the two hierarchical clustering approaches yielded C1 clusters that fell entirely within the range of scores defining remission for either depression rating scale. Thus, we were unable to achieve the same clustering pattern using scores of individual scale items.

External validation. We also applied our unsupervised learning approach to STAR*D (QIDS-C rating scale) and ISPC (HDRS rating scale) datasets, and identified three clusters of men and women at all time-points. These clusters were not statistically different ($p > 0.1$) from the clusters of comparable depressive symptom severity inferred in PGRN-AMPS, providing external validation. At 8 weeks, the three clusters (C1, C2, C3) identified in STAR*D and ISPC also conformed to accepted clinical definitions of remission, response, and non-response, respectively, on both depression rating scales. Our externally validated clusters then allowed us to identify associations of depression severity with clinical, demographic, and biological factors.

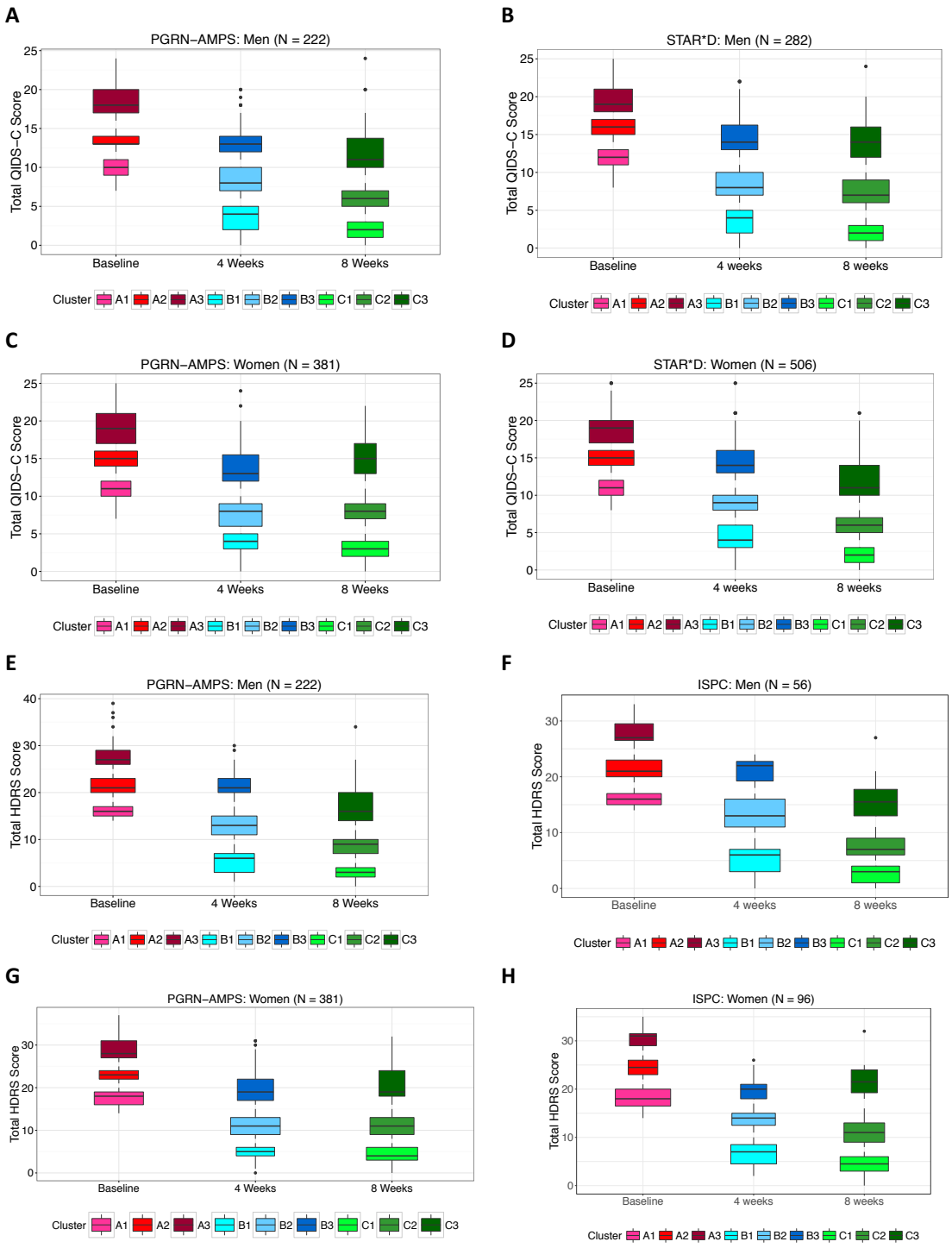


Figure 4.2: Depressive symptom-based clusters identified by data-driven unsupervised learning using Gaussian mixture models (GMM). The distribution of total depression severity scores in Caucasian subjects are represented by box plots; the width of the box is proportional to the number of patients comprising the cluster. The clusters are shown on both the QIDS-C (A-D) and HDRS (E-H) rating scales for the PGRN-AMPS (A, C, E, and G), STAR*D (B and D), and ISPC (F and H) for men and women.

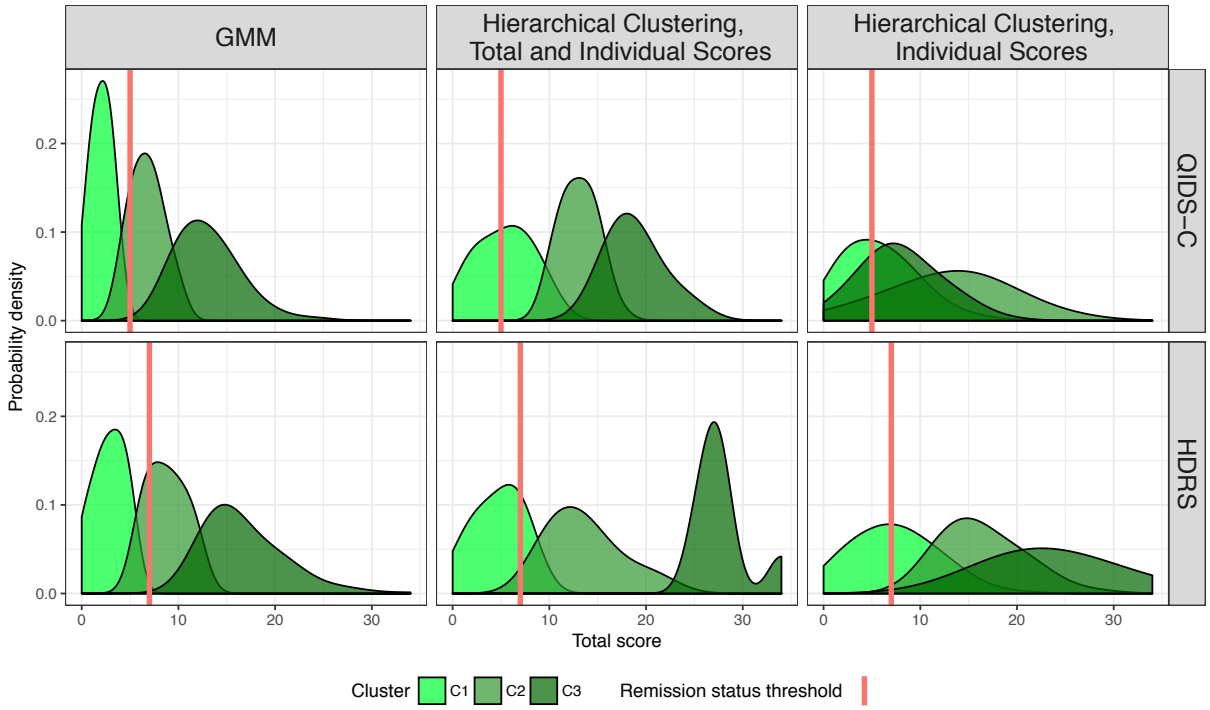


Figure 4.3: Comparing distribution of total depression severity scores in clusters inferred using GMM and hierarchical clustering approaches. Results with a Gaussian mixture model (GMM) are in the first column. The middle and third columns show clustering of patients done via hierarchical clustering approaches with multivariate data that comprise individual item responses at 8 weeks, both with (middle column) and without (third column) total depression severity scores at 8 weeks. The probability densities in each figure are represented on the y-axis, and depressive symptom scores using the QIDS-C (first row) and HDRS (second row) are represented on the x-axis. The threshold for remission for each depression rating scale is defined using the red vertical line in each plot.

4.4.3 Association of Clinical, Sociodemographic Factors with Patient Stratification

There were no significant differences in any of the clinical or soico-demographic factors (listed in Table A.1) among the three clusters in each sex ($p > 0.1$, body mass index variation is illustrated in Fig. 4.4(a)). For the 603 PGRN-AMPS patients who had complete response data, neither CYP2C19 metabolizer status nor plasma drug levels at 4 or 8 weeks were associated with depression severity in the respective clusters ($p > 0.4$; see Fig. 4.4(b)), or across different dosages of citalopram and escitalopram ($p > 0.3$).

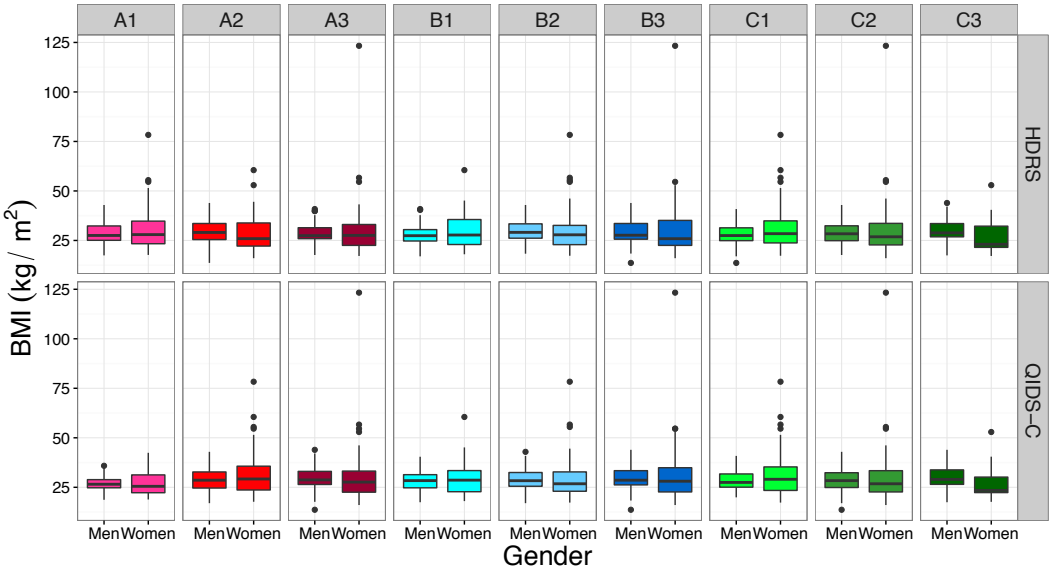
4.5 Discussion

The replication of patient clusters at all time-points that were identified using mixture-model-based unsupervised learning has the following clinical research implications.

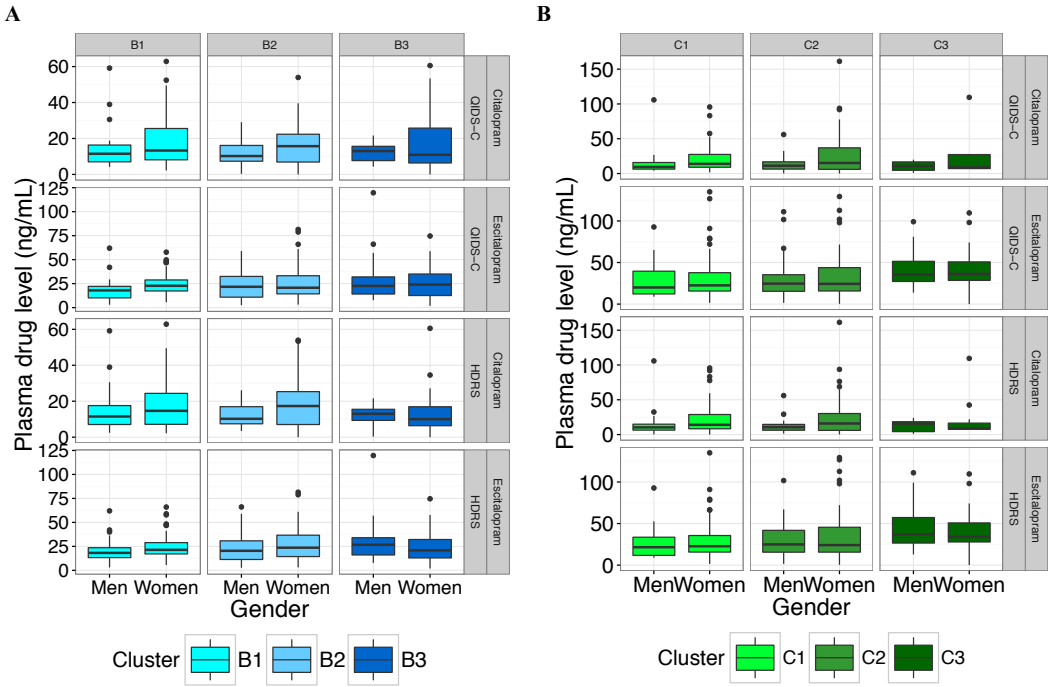
4.5.1 Clinical Implications of Patient Stratification

Toward Clinically Actionable Modeling of Longitudinal Effects of Antidepressants: It is known that eventual antidepressant treatment outcomes are conditioned upon baseline depression severity and subsequent changes in symptoms at intermediate time-points, before a therapeutic trial is complete. To study the longitudinal effects of antidepressants, the patient clusters inferred in this work can serve as nodes of a graph (e.g., a probabilistic graph) that enable capture of the longitudinal variation of depression symptoms over time, conditioned on baseline characteristics and changes in these characteristics at intermediate time-points — a process that we will refer to as "symptom dynamics" (discussed in Chapter 7). Here, the replicating cluster patterns at baseline and at 4 weeks are just as important as the clinically valid clusters at 8 weeks. Understanding the symptom dynamics within clusters of patients defined by depression severity and biological characteristics at baseline (predictive outcome markers) and at intermediate time-points (change markers) may lead to further improvement in our understanding of antidepressant response mechanisms. In addition, a detailed understanding of the symptom dynamics across multiple time-points may enable clinicians to change treatments if the predicted chances of response or remission are low.

Biological Associations with Depression Severity: Our clustering approach can be used to iteratively investigate the effects of multiple biological measures (e.g., metabolomics, genomics), individually and in groups, for predicting antidepressant response. Such systematic



(a) Body mass index (BMI) variation across clusters



(b) Plasma drug concentration levels after 4 and 8 weeks of treatment

Figure 4.4: Fig. (a) Illustrates the comparison of mean body mass indices (BMI, kg/m^2) for men and women in clusters with comparable symptom severity at baseline, 4 weeks, and 8 weeks. Fig. (b) Illustrates the comparison of citalopram and escitalopram plasma drug concentrations between men and women with each depressive symptom severity cluster at 4 weeks (Fig. a) and 8 weeks (Fig. b).

investigation, using a variety of biological measures and other antidepressants, may lead to improved understanding of the underlying neurobiology of antidepressant response, and an ability to match individual MDD patients with specific antidepressants based on their biological profiles.

Underlying Pathological Mechanisms Between Patient Clusters: The replicating patterns of patient clusters raise the possibility of investigating the variability of potential underlying pathological mechanisms of MDD between depression severity clusters — a possibility that can be tested by the application of "omics" analyses, such as GWAS that compare one cluster with another [92, 93].

4.5.2 CYP2C19 Metabolizer Status and Patient Stratification

Our observation that CYP2C19 metabolizer status was not significantly associated with eventual citalopram/escitalopram treatment outcomes or depression severity clusters is in line with findings from previous research. While functional CYP2C19 allele variants are associated with citalopram/escitalopram metabolism [94], the impact of P450 genotypes, including CYP2C19, on therapeutic outcomes has been less clear [95]. Some studies in depressed patients have found a significant association between CYP2C19 genotype and treatment response to citalopram or escitalopram [24, 96], while other studies have failed to demonstrate such an association [97, 98]. There are similar inconsistencies in results of studies attempting to link serum concentrations of antidepressants, including citalopram, with the likelihood of positive antidepressant response [99, 100]. The lack of improved predictability of treatment outcomes using CYP2C19 genotype does not suggest that pharmacokinetic mechanisms and the CYP2C19 genotype are not clinically relevant. This is particularly so with respect to the dose-dependent risk of QTc interval prolongation with citalopram [95], although evidence is mixed regarding the effect of the CYP2C19 genotype on the risk of citalopram-associated side effects [100, 101].

4.5.3 Methodological Considerations for Stratifying Patients

While it is relatively simple to establish and justify findings from a machine learning algorithm applied to a single dataset, it is much more difficult to replicate this behavior in independent datasets/trials because of between-trial differences in patient characteristics and other factors. Our multi-trial replication demonstrates the power of machine learning approaches to extract

consistent patterns of symptom severity in MDD, a disease that shows striking heterogeneity of symptoms and high inter-patient variability in response to antidepressants.

We focused on the use of total depression scale scores rather than individual depression scale items for three reasons. First, total depression scores at baseline were the most robust predictor of clinical outcomes in prior machine learning work [29]. Second, total depression scores are used to define response and remission in clinical trials [30]. And, finally, we showed in this work that multivariate clustering approaches using individual depressive item scores did not yield clustering patterns at 8 weeks that conformed to accepted definitions of response or remission. These observations provide strong evidence that our clustering approach is both computationally and clinically valid. The lack of associations between social/demographic factors and any of the depressive symptom severity clusters also agrees with prior work demonstrating that social/demographic factors individually or in aggregate cannot accurately predict antidepressant treatment outcomes [59, 102–104].

Others have attempted to cluster individual depression scale items to identify symptoms with similar responses [59]. A disadvantage of their approach is the inability to model the change in symptoms within potentially important patient subgroups defined by baseline characteristics and eventual treatment outcomes. Moreover, it is possible that the clustering behavior of each depression item is subject to variations conditional on baseline characteristics. This property also precludes the use of longitudinal clustering methods [91], which are modeled under the assumptions that sample movements between time-points are independent and identically distributed. Hence, our choice of method is justified given our eventual goal of associating biological measures with subgroups defined by depression severity during discrete treatment time-points.

4.6 Summary

The patient stratification presented in this work is significant at multiple levels. First, in a field of medicine in which findings of trials do not often replicate, our data-driven approach to stratifying patients demonstrated high degrees of replication in patient clusters across multiple clinical trials and rating scales. Second, the methodological approach to clustering of patients is further justified by the ecological validity of the clusters after 8 weeks of the acute phase of antidepressant treatment, in terms of accepted definitions of antidepressant treatment outcomes. When these two significant aspects are taken together, the stratification achieved in this work serves as a strong foundation for (1) multi-omic integration to predict long-term outcomes prior to treatment initiation, (2) probing of trajectories of depressive severity

that offer insights into homogeneity in antidepressant response, and (3) an opportunity to revisit GWAS on clusters of patients with similar depressive symptoms who are potentially having similar states of disease. Therefore, the replication of an ecologically valid patient stratification across multiple datasets, as presented in this chapter, will serve as the first step toward individualization of antidepressant treatment selection.

CHAPTER 5

ON THE PROMISE OF PHARMACO-OMICS MEASURES AS PREDICTORS OF ANTIDEPRESSANT TREATMENT OUTCOMES

In diseases that are characterized by complex phenotypes (traits), such as psychiatric disorders, inflammatory diseases, and migraines, therapeutic and treatment decisions are primarily based on the subject-reported and/or physician-rated severity of symptoms (which are an example of complex phenotypes/traits) in conjunction with standard social/demographic factors. In the context of antidepressant response in MDD, the ability of these measures to predict therapeutic success is slightly better than chance [28,29]. Genomic and transcriptomic biomarkers as predictor variables on the other hand, have demonstrated promise in improved predictive ability, relative to the use of using sociodemographic measures as predictors of antidepressant treatment outcomes [58,64–66]. However, it remains to be explored whether *pharmaco-omics* measures (such as metabolomics and genomics) that reflect the underlying molecular mechanisms of therapeutic agents (e.g., drugs) could also serve as stronger predictors of therapeutic outcomes.

Contributions. In this chapter, we propose a "*learning-augmented clinical assessment*" workflow to sequentially augment physicians' assessments of subject-specific ratings of symptoms with heterogeneous pharmaco-omics measures (such as metabolomics and genomics) to predict antidepressant treatment outcomes. The workflow was developed using data from the Mayo PGRN-AMPS, which is the largest single-center SSRI trial that has been conducted in the United States. Metabolomics and genomics data was derived from peripheral blood of study subjects in this trial. Through augmentation of those biological measures with psychiatric assessments and sociodemographic factors as predictor variables, the accuracy of predicted antidepressant treatment outcomes in MDD patients improved from 35% to 80% relative to the use of clinical measures alone as the predictor variables. This improvement in predictive accuracy of treatment outcomes motivates the need for developing antidepressant-specific prediction models, so that the choice of an antidepressant can be based on which one has the highest likelihood of causing remission of depressive symptoms.

5.1 Problem Statements

With access to metabolomics and genomics data for a smaller cohort of the Mayo PGRN-AMPS trial, we set out to answer the following questions (illustrated in Fig. 5.1):

1. Would augmenting social, demographic, and clinical data with *metabolomics* data improve the accuracies of treatment outcome predictions, relative to using only social, demographic, and clinical data as predictor variables?

2. Would augmenting social, demographic, and clinical data with *metabolomics and genomics* data improve the accuracies of treatment outcome predictions, relative to over using only social, demographic, and clinical data as predictor variables?

3. If the predictions improve as a result of augmenting existing clinical measures with biological measures, how many of the top predictors were biological measures?

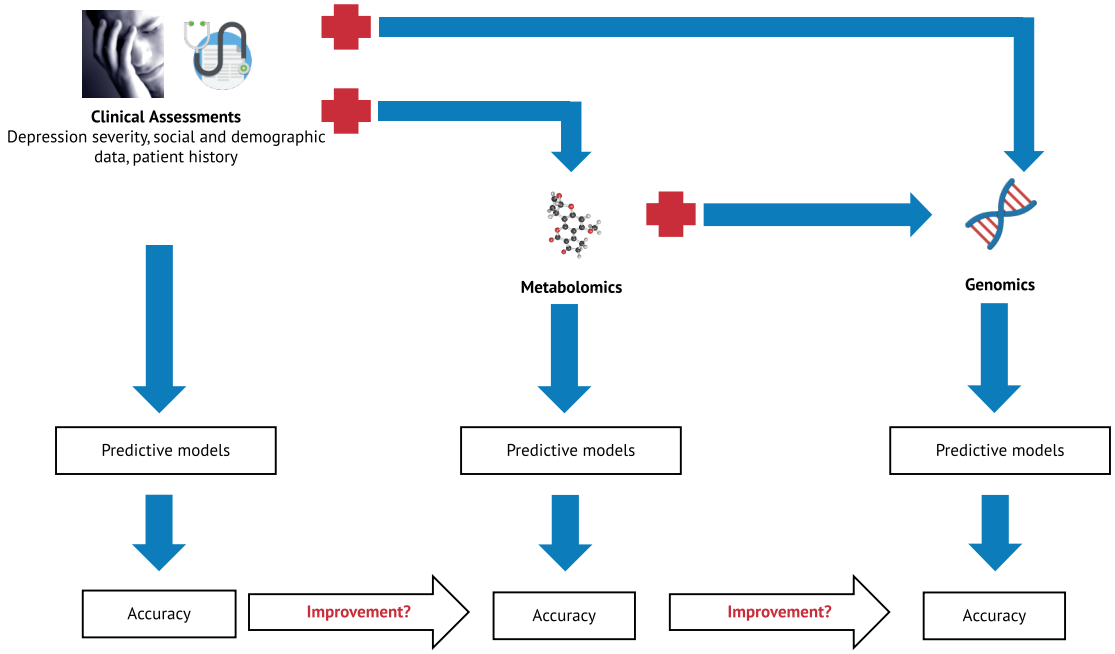


Figure 5.1: The proposed analyses to establish improved predictability of antidepressant treatment outcomes by augmenting clinicians' assessments with biological measures.

5.2 Data

The Mayo PGRN-AMPS trial (NCT 00613470) was designed to assess the clinical outcomes of adults (aged 18–84 years) with non-psychotic MDD after 4 and 8 weeks of open-label treatment with citalopram or escitalopram and to examine the metabolomic and genomic factors associated with those outcomes [24]. Subjects were recruited from primary and specialty care settings in and near Rochester, MN from March 2005 to May 2013. All psychiatric diagnoses were confirmed at the screening visit using modules of the Structured Clinical Interview for DSM-IV (SCID) administered by trained clinical research staff. The data $D = [S : C : B]$ analyzed in this work comprise social and demographic variables (S), clinical measures (C), and biological measures (B) and are listed in Table 5.1. The social and demographic data (S) were assessed only at baseline. The treatment outcomes were established using the 16-item, clinician-rated version of the Quick Inventory of Depressive Symptomatology (QIDS-C) at baseline, 4 weeks, and 8 weeks; the results comprise the clinical data C, which include the responses to the 16 QIDS-C questions and the total QIDS-C score of the symptom severity [52]. Biological measures for 290 of the 603 patients in this trial included GWAS genotype data that, after imputation, included approximately 7 million SNPs ($\mathcal{G}$), and plasma *metabolomic* concentrations (B in Table 5.1) for 31 metabolites ($\mathcal{M}$) taken from patients at three time-points of the trial (at baseline, 4 weeks, and 8 weeks). Samples were assayed on a high-performance liquid chromatography (HPLC) electrochemical coulometric array (LCECA) platform to obtain standardized measures of the concentrations of metabolites. Definitions of clinical outcomes are outlined in Sec. 2.1.1.

5.3 Approach Overview

To demonstrate the improved predictability in treatment outcomes, the workflow was developed using data from the Mayo PGRN-AMPS clinical trial [24]. This is the largest single-center selective serotonin reuptake inhibitor (SSRI) trial that has been conducted in the United States. There were 603 patients who completed the trial. They were administered citalopram or escitalopram (commonly prescribed SSRIs) for 8 weeks, and psychiatric assessments of depression severity at baseline (pre-treatment), 4 weeks, and 8 weeks were conducted by a clinician using the quick inventory of depressive symptom (QIDS-C). In this trial, biological measures for 290 of the 603 patients included genome-wide association study (GWAS) genotype data that, after imputation, included approximately 7 million single-nucleotide polymorphisms (SNPs) ($\mathcal{G}$), and plasma *metabolomic* concentrations (B in Table 5.1) for 31 metabolites ($\mathcal{M}$) from patients at three time-points of the trial (at baseline,

Table 5.1: Data ($D = [S : C : B]$).

Total patients: 603. Men: Total: 222. With omics: 99. Women: Total: 381. With omics: 191.
Social and demographic data (S) collected only at baseline:
Age (in years)
Body mass index (BMI in kg/m^2)
Depression in {parents, siblings, children}
Bipolar disorder in {parents, siblings, children}
Alcohol abuse by {parents, siblings, children}
Drug abuse by {parents, siblings, children}
Seasonal pattern in symptom occurrence
History of psychotherapy
Depressive severity assessment (C):
Clinician-rated Quick Inventory of Depressive Symptomatology (QIDS-C) questionnaire (16 questions)
QIDS-C total score
Biological data (B):
$\mathcal{M}$: 31 metabolites from the HPLC LCECA platform $\mathcal{G}$: 7 million single nucleotide polymorphism genotypes

4 weeks, and 8 weeks). Through augmentation of those biological measures with psychiatric assessments and sociodemographic factors as predictor variables, the prediction accuracy of antidepressant treatment outcomes in MDD patients improved from 35% to 80% relative to the use of clinical measures alone as the predictor variables.

The formalism for integrating multiple biological measures in this case study is as follows. Just as tumor subtypes serve as a foundation for integrating biological measures in oncology, our formalism first established patient subtypes/stratification $\mathcal{C}$ by using mixture-model-based unsupervised learning techniques. In the first layer of overlaying of the biological measures, a set of metabolites $m \in \mathcal{M}$ were identified based on significant associations of their concentrations with symptom severity in previously inferred patient stratification. In the second layer of the overlay of biological measures, in what is referred to as a *metabolomics-informed-genomics* approach, we used GWAS to identify SNPs $g \in \mathcal{G}$ that are associated with concentrations of metabolites comprising m. Through iterative overlaying of biological measures starting with metabolites (blood measures reflecting drug action) associated with depressive severity, and then adding in the genes associated with metabolomic concentrations, the biological measures became more closely associated with the molecular mechanisms of antidepressant response. Finally, out of the more than 7 million possible predictor variables, the proposed approach identified about 65 predictor variables that comprised (1) SNPs (g) identified by the GWAS based on metabolomic concentrations, (2) metabolites (m) whose concentrations are significantly associated with depression severity in patient clusters, and (3) clinical measures (as shown in Table 5.1). Thus we made the size of the predictor data computationally tractable to predict clinical outcomes $\tilde{y}$ by using supervised learning methods $\mathcal{F}(m, g, S, C, y)$, where y is the treatment outcome labels of the training data.

5.4 Identification of Pharmacogenomic Biomarkers

In operationalizing the described *metabolomics-informed-genomics* approach, we associated depression severity and metabolite concentrations of patients in baseline clusters (inferred in Chapter 4) with treatment outcomes at 8 weeks. Baseline metabolite concentrations of 5HT, KYN, 4HBAC, TRP, TYR, and PARAXAN, and baseline depression severity were significantly correlated ($p < 0.05$) with response and/or remission status at 8 weeks. These metabolites have been identified in previous studies of metabolomics and SSRI response using other datasets, supporting the biological relevance of these metabolite associations. Further, many of them are also related to the monoamine neurotransmitter pathways associated with MDD and its treatment response [32, 105, 106]. Those prior studies also identified SNPs in the

TSPAN5 (rs10516436), ERICH3 (rs696692), and DEFB1 (rs5743467, rs2741130, rs2702877) genes as top SNPs associated with plasma concentrations of 5HT or KYN [32, 33]. We used these SNPs as pharmacogenomic biomarkers for evaluating their predictive capabilities when combined with metabolomic and sociodemographic measures as predictors of remission or response to antidepressant treatment.

5.5 Predicting Antidepressant Response

Three classes of classifiers are used in this work, including kernel, linear, and ensemble methods. For predicting outcomes using baseline clinical, social, demographic, and metabolomic data, we used support vector machines with linear kernels (SVMLinear) and support vector machines that use radial-basis kernels (SVM-RBF) as kernel methods [107]; a generalized linear model (GLM) as a linear method [108]; and GBMs as an ensemble method [109]. As the creators of those methods have indicated, each of those broader types has its own merits, mathematical nuances, and complexities, and all of them have been used in other classification applications, such as in Kaggle [110]. To use all of the omics and clinical, social, and demographics data to predict outcomes, we used nonparametric classifiers such as SVM-RBF and random forests, as they are better suited to handling correlated features [111], and have been used in predicting treatment outcomes in other psychiatric diseases such as schizophrenia.

In addition to elastic-net regularization, recursive feature elimination (i.e., a wrapper method) was also used for the GLM and GBM classifiers; that made it possible to estimate the model performance not only by optimizing the parameters of the model, but also by searching for the right set of predictor variables. Based on our datasets, the prediction performance did not significantly vary with or without the use of any of the feature selection methods; the prediction accuracy remained within 4%. This observation could also be in part due to a reasonably small size of predictor variables.

To minimize the effects of overfit and information leak, nested cross-validation (nested-CV) with five repeats was used to train the classifiers. In each repeat, data were randomized, and the nested-CV comprised an outer loop and an inner loop. The outer loop had a fivefold cross-validation to split the data into training data (80% of the data) and testing data (the remaining 20%). The inner loop used the training data to train the classifier by using a tenfold cross-validation, and the trained classifier was tested on the testing data. To minimize the effects of class imbalance (i.e., unequal numbers of responders (60%) and non-responders (40%)) in the training data, we used the synthetic minority over-sampling (SMOTE) algorithm [112], which simulated patient profiles of the under-sampled class and

up-sampled the under-sampled class to ensure that the two classes had equal sizes. Prediction performance was reported using several metrics (AUC, sensitivity, and specificity), and the statistical significance of the classifier's accuracy was established using the null information rate (NIR, which is the prevalence of the class with the largest samples) that served as a proxy for chance.

5.5.1 Training with and without Biological Measures

In order to quantitatively assess the benefit of adding biological measures to predict outcomes $\tilde{y}$, we trained classifiers $\mathcal{F}(m, g, S, C, y)$ using (1) baseline clinical data that included only social and demographic data, $X = [S : C]$; (2) all baseline data (including metabolomics and genomics data), $X = [S : C : B]$, where $B = [m, g]$; and (3) training labels y for treatment outcomes. Metabolites (m) whose baseline concentrations were correlated with the symptom severity at 8 weeks, and SNPs (g) associated with their concentrations, were then normalized along with clinical data in order to train the chosen supervised learning methods. It is important to note that several other researchers have proposed the combination of other modalities of biological data [113–115], but it remains to be explored whether combination becomes less effective when patient-reported data are used, since there is considerable heterogeneity in subject-reported measures. Therefore, to the best of our knowledge, this is the first time that quantified biological measures comprising metabolomics and genomics measures have been integrated for analyses with the clinical measures of psychiatric assessments that comprise demographic data and patient-provided responses to symptom questionnaires (such as QIDS-C). For all the classifiers, we compared the AUC, in addition to the generalized prediction accuracies, to see whether the same model's predictive ability improved with the addition of metabolomics data. Further, if the predictability improved, we extracted the top five predictors of the model that provided the best balance of accuracy and AUC to see whether the top predictors were dominated by the metabolomics.

5.5.2 Prediction Performance

ASNPs shown in Table 5.2, for both men and women and for both the outcomes *response* and *remission*, there was a 30% improvement in the overall accuracy and corresponding AUC. The highlighted columns in Table 5.2 indicate the best-performing models with the metabolomics data included; four out of the top five predictors are metabolites, indicating that their addition to the prediction model likely explains the increase in the predictability of the outcomes. As shown in Table 5.3, there was a further improvement of at least 5% in the

Table 5.2: Clinical outcome prediction performance with metabolomics data in the Mayo Clinic PGRN-AMPS trial. Expansions of the abbreviations of the top predictors are listed in Table A.2.

Men									
RESPONSE									
Data	Clinical Data Only				Clinical and Metabolomics Data				Top Predictors
Model	SVM-RBF	SVM-Linear	GLM	GBM	SVM-RBF	SVM-Linear	GLM	GBM	ATOCO
Accuracy	28.2	32	52	40	48	48	64	48	URIC
Sensitivity	0	16.67	16.67	33.33	33.33	33.33	50	33.33	QIDS-1
Specificity	53.5	46.15	84.62	46.15	61.54	61.54	61.54	61.54	KYN
AUC	0.64	0.60	0.63	0.54	0.53	0.53	0.68	0.5	3OHKY
REMISSION									
Data	Clinical Data Only				Clinical and Metabolomics Data				Top Predictors
Model	SVM-RBF	SVM-Linear	GLM	GBM	SVM-RBF	SVM-Linear	GLM	GBM	AMTRP
Accuracy	28	44	44	48	64	68	64	45.65	I3PA
Sensitivity	38.46	38	53.85	46.15	76.52	76	76.92	65.22	Drug dosage
Specificity	16.67	50	33.33	50	50	50	50	26.09	GTOCO3
AUC	0.8	0.6	0.67	0.6	0.76	0.78	0.62	0.6	5HT
Women									
RESPONSE									
Data	Clinical Data Only				Clinical and Metabolomics Data				Top Predictors
Model	SVM-RBF	SVM-Linear	GLM	GBM	SVM-RBF	SVM-Linear	GLM	GBM	Seasonal Pattern
Accuracy	52.08	52.08	54.17	50	41.3	72.33	64.58	41.67	5HT
Sensitivity	18.18	18.18	27.27	18.18	34.78	18.18	36.36	0	MHPG
Specificity	80.72	80.76	76.92	76.9	47.83	92.83	88.46	76	MET
AUC	0.60	0.59	0.63	0.63	0.69	0.74	0.68	0.51	QIDS-13
REMISSION									
Data	Clinical Data Only				Clinical and Metabolomics Data				Top Predictors
Model	SVM-RBF	SVM-Linear	GLM	GBM	SVM-RBF	SVM-Linear	GLM	GBM	5HT
Accuracy	34.78	50	45.65	36.96	41.3	54.33	52.17	45.65	HGA
Sensitivity	26.09	65.22	56.52	47.83	34.78	56.52	76.92	65.22	3OHKY
Specificity	43.48	34.78	34.78	26.09	47.83	52.17	50	26.09	Seasonal pattern
AUC	0.64	0.52	0.58	0.58	0.56	0.53	0.53	0.53	PARAXAN

Table 5.3: Clinical outcome prediction performance when clinical measures were combined with metabolomics and genomics data from the Mayo Clinic PGRN-AMPS trial. Expansion of abbreviations of the metabolites are listed in Table A.2, QIDS-C13 is the 13^{th} item of the QIDS-C scale and DEFB1_1 is DEFB1 SNP rs5743467.

Gender	Men		Women	
Outcome	RESPONSE			
Model	SVM-RBF	Random Forest	SVM-RBF	Random Forest
Accuracy (%)	76	80	75	71
Sensitivity	0.66	0.83	0.5	0.6
Specificity	0.84	0.77	0.9	0.88
AUC	0.82	0.88	0.88	0.73
Top Predictors	URIC, ATOCO, 3OHKY, I3PA, DEFB1_1		MHPG, MET, 5HT, 4HPAC, XAN	
Gender	Men		Women	
Outcome	REMISSION			
Model	SVM-RBF	Random Forest	SVM-RBF	Random Forest
Accuracy (%)	68.75	80	86.7	73.7
Sensitivity	0.63	0.9	0.73	0.78
Specificity	0.74	0.7	0.8	0.68
AUC	0.74	0.87	0.9	0.88
Top Predictors	4HPLA, I3PA, 3OHKY, URIC, QIDSC-13		PARAXAN, 5HT, 4HBAC, CYS, 3OHKY	

AUC and corresponding accuracy when genomics data were integrated with the metabolomics, clinical, social, and demographic data. We have two observations about the inclusion of biological measures in all these predictions. First, the top predictors of outcomes when biological measures were used were different in men and women, likely pointing to different biological mechanisms determining how men and women respond to the same antidepressant. Second, except for the variable *seasonal pattern* and the *involvement* item in the QIDS-C scale, no other clinical/demographic measures were predictive of outcomes. Finally, it is biologically significant that many of the top predictor metabolites identified in this work are known to be correlated with mood in the behavioral sciences, which has additional promising implications, as discussed next.

5.6 Discussion

The improved predictive performance pharmaco-omics measures are augmented with routinely collected clinical and sociodemographic measures is of significance from two perspectives. First, from a pharmaco-omics perspective, that helps improve our understanding of MDD pathology and drug response. Second, the potential to begin choosing antidepressants based on predicted efficacy of the drug, a critical need given the high non-response rates to antidepressant treatment.

5.6.1 From a Pharmacogenomics Perspective

This work demonstrates its biological significance through its improvement of predictability by integrating metabolomics data with clinical measures, because metabolites, such as serotonin and kynurenine, are among the top predictors of outcomes. This development is important because for decades, the treatment of MDD has focused on biogenic amine neurotransmitter pathways, i.e., the synthesis and metabolism of catecholamines (such as norepinephrine) and indoleamines (such as serotonin) [116, 117]. Furthermore, the existing body of knowledge fits well with the findings of our study; note that the metabolites listed in Table 5.2 include serotonin (5HT) itself as well as two metabolites from the competing tryptophan metabolism pathway (KYN and 3OHKYN) and the major catecholamine metabolite (MHPG), which are known to play a role in behavior.

The addition of genomics data with metabolomics and clinical measures as predictor variables has further improved the predictability of antidepressant outcomes, as shown in Table 5.3. This result raises the question of whether the genes associated with serotonin

and kynurenine are also extending their effects in other metabolites used in this study through other mechanisms. The improvement in predictability should motivate researchers and clinicians to collect more biological measures for psychiatric diseases other than major depressive disorder (e.g., bipolar disorder, schizophrenia, and various dementias) that would not only help subtype or stratify patients by their symptom severity profiles, but also combine biological characteristics that would enable treatment strategies closer to the kinds used in breast cancer therapeutics.

5.6.2 The Need for Pharmaco-omics based Predictions in Psychiatry

The overarching significance of predictive capabilities of pharmaco-omics measures in psychiatry is suggested by the success of analogous "precision medicine" approaches in breast cancer therapeutics. Today, treatment strategies for each breast cancer patient are tailored to the tumor's specific molecular characteristics. That successful approach is facilitated by the close association of the phenotype (which is the molecular characteristics of the tumor, such as whether it is estrogen-receptor-positive (ER-positive), human-epidermal-growth-factor-receptor-2-amplified (HER2 amplified), and/or triple-negative) with a set of biomarkers, such as hormone receptors (e.g., ER), genes, and their SNPs, which, when taken together, can be prognostic of treatment outcomes [87]. However, in the study of treatment outcomes in patients with MDD (as for other diseases with complex phenotypes), some interesting key observations can be made. First, GWAS have often failed to associate SNPs with complex and non-binary phenotypes defined as, for example, "Did patients achieve a 50% reduction in baseline patient-reported/clinician-recorded symptoms?" As a consequence, it is acknowledged that methods of integrating widely heterogeneous biological measures without *a priori* biological knowledge become computationally intractable as the number of study variables increases to the order of millions [118, 119]. Second, the predictability of antidepressant treatment outcomes when clinical measures alone are used is at best slightly better than chance [28,29,102,103,120–122]. Third, antidepressant medications such as selective serotonin reuptake inhibitors (SSRIs) are the standard of care for drug therapy in adults with MDD, but less than half of patients have favorable outcomes from this treatment [120]. In light of these observations, if learning techniques could more accurately predict treatment outcomes in patients with MDD by integrating a few biological measures prognostic of antidepressant treatment's success with routine clinical measures, the impact would be far-reaching, because MDD affects over 350 million patients worldwide and is expected to be the leading cause of disabilities globally by 2030 [35, 36, 123].

5.7 Summary

Pharmaco-omics measures as potential predictors of citalopram/escitalopram outcomes are promising from the perspective of individualized antidepressant treatment selection. The improved predictive performance of these biological measures when augmented with routine clinical and sociodemographic measures highlights the need to study these measures for other antidepressants as well. Use of such a small set of measures (as opposed to millions of variables, as with genome-wide genotype data) makes inference of novel biology, or accurate prediction of clinical outcomes in diseases with complex phenotypes, computationally tractable. Of broader significance, these findings together motivate the use of our approach for other common diseases, such as rheumatoid arthritis or migraine headaches, for which a similar complexity in phenotype is seen, and will also motivate researchers and clinicians to collect additional biological measures for other psychiatric diseases for which the methods proposed in this work could identify novel mechanisms of therapeutic efficacy. Furthermore, the workflow could be further enhanced by considering other omics data, such as transcriptomics and/or proteomics, in addition to profiling of the microbiome. **Limitations.** While we demonstrated the capabilities of multiple pharmaco-omics measures as predictors of antidepressant treatment outcomes, their predictive value was not tested in independent citalopram/escitalopram trial datasets. It remains to be seen whether such multi-omic integration is possible for other antidepressants, to demonstrate the generizability of our approach. Such replication in findings would represent a strong foundation for investigating biological factors associated with pathophysiology of this disease with heterogeneous disease states, and additional mechanisms of drug action. In an attempt to addressing the limitation of external replications, we next explore if only pharmacogenomic biomarkers can reliably predict antidepressant treatment outcomes in multiple datasets.

CHAPTER 6

CROSS-TRIAL PREDICTIONS OF ANTIDEPRESSANT RESPONSE USING PHARMACOGENOMIC BIOMARKERS

A combination of clinician assessments of patients' depressive symptoms, individual sociodemographic characteristics, and functionally validated biomarkers could be used to individualize antidepressant selection [34,58,64,124,125]. However, the question of whether baseline depression severity assessments augmented with pharmacogenomic biomarkers (e.g., single nucleotide polymorphisms – SNPs) can predict antidepressant treatment outcomes remains insufficiently addressed. Predictions obtained using methods trained with clinical and sociodemographic factors alone have yielded predictive accuracies significantly better than chance [28,29,59,102]. The authors of these studies acknowledged the need to include biomarkers as predictors to further improve the predictability of treatment outcomes [28, 29, 59, 102]. Our prior work demonstrated that combining pharmaco-omics (metabolomics and genomics) measures with clinical and sociodemographic measures improved predictability of treatment outcomes as opposed to using only clinical, sociodemographic measures, using data from the Mayo PGRN-AMPS [88]) [124]. Improved predictability of treatment outcomes using pharmacogenomic biomarkers would be of immense clinical value by providing a quantitative basis for selecting an antidepressant with the highest likelihood of providing remission/response, as opposed to the current "try-and-wait" approach to treatment selection.

Contribution. In this chapter, we extend the two-stage patient stratification workflow proposed in Chapter 4 by adding a prediction stage, which uses pharmacogenomics biomarkers from PGRN-AMPS to train statistical/machine learning models for predicting remission/response to citalopram/escitalopram treatment. The pharmacogenomic biomarkers were identified using a "metabolomics-informed-genomics" approach described in Section 5.4. We then externally validated the models using data from the STAR*D [25] and ISPC [26] datasets. Supervised machine learning methods trained using pharmacogenomics SNPs and total baseline depression scores predicted sex-specific remission/response at 8 weeks with AUC > 0.7 in PGRN-AMPS, and with predictive accuracies $> 65\%$ ($p < 0.07$) and $> 76\%$ ($p < 0.07$) in STAR*D and ISPC, respectively.

6.1 Problem Statements

Chapter 5 demonstrated that the use of multiple pharmaco-omics measures could improve predictability of antidepressant treatment outcomes relative to use of sociodemographic factors alone. An important limitation of the work discussed in Chapter. 6 was its lack of external validation of prediction performance, and the reliance on metabolomic measures that are prone to variations based on activities prior to blood draw. Pharmacogenomic measures are stable measures that could be more reliable biomarkers in predicting treatment outcomes. To investigate their predictive capabilities, following questions are addressed in this chapter:

1. Can depression severity combined with pharmacogenomic biomarkers predict treatment outcomes with improved accuracies relative to the use of clinical and sociodemographic predictors alone?

2. Can improved predictive accuracies achieved with integration of pharmacogenomic biomarkers as predictors be replicated with multiple independent datasets and depression rating scales?

6.2 Data

For this study, we will restrict our analyses to Caucasian subjects from PGRN-AMPS (NCT 00613470), STAR*D (NCT 00021528) and ISPC trials. This is to avoid confounding effects due to linkage disequilibrium in the genomes of subjects with different races/ethnicity. For the present analyses, we utilized data from 603 Caucasian citalopram-treated PGRN-AMPS subjects, 788 Caucasian citalopram-treated STAR*D subjects, and 152 Caucasian citalopram/escitalopram-treated ISPC subjects who had complete clinical data at baseline and at 4 and 8 weeks. All 603 Caucasian citalopram-treated PGRN-AMPS subjects also had CY2C19 metabolizer genotype data at baseline, and plasma drug levels at 4 and 8 weeks. Details of genotyping and GWAS of the PGRN-AMPS, STAR*D, and ISPC subjects have been previously published [32, 33, 88]. Clinical outcome definitions are outlined in Sec. 2.1.1.

6.3 Approach Overview

We now extend the patient stratification workflow proposed in Chapter 4 with an additional stage for predicting treatment outcomes using pharmacogenomic markers identified in Sec. 5.4 as shown in Fig. 6.1. For the sake of simplicity, we have collapsed the two-stage workflow of

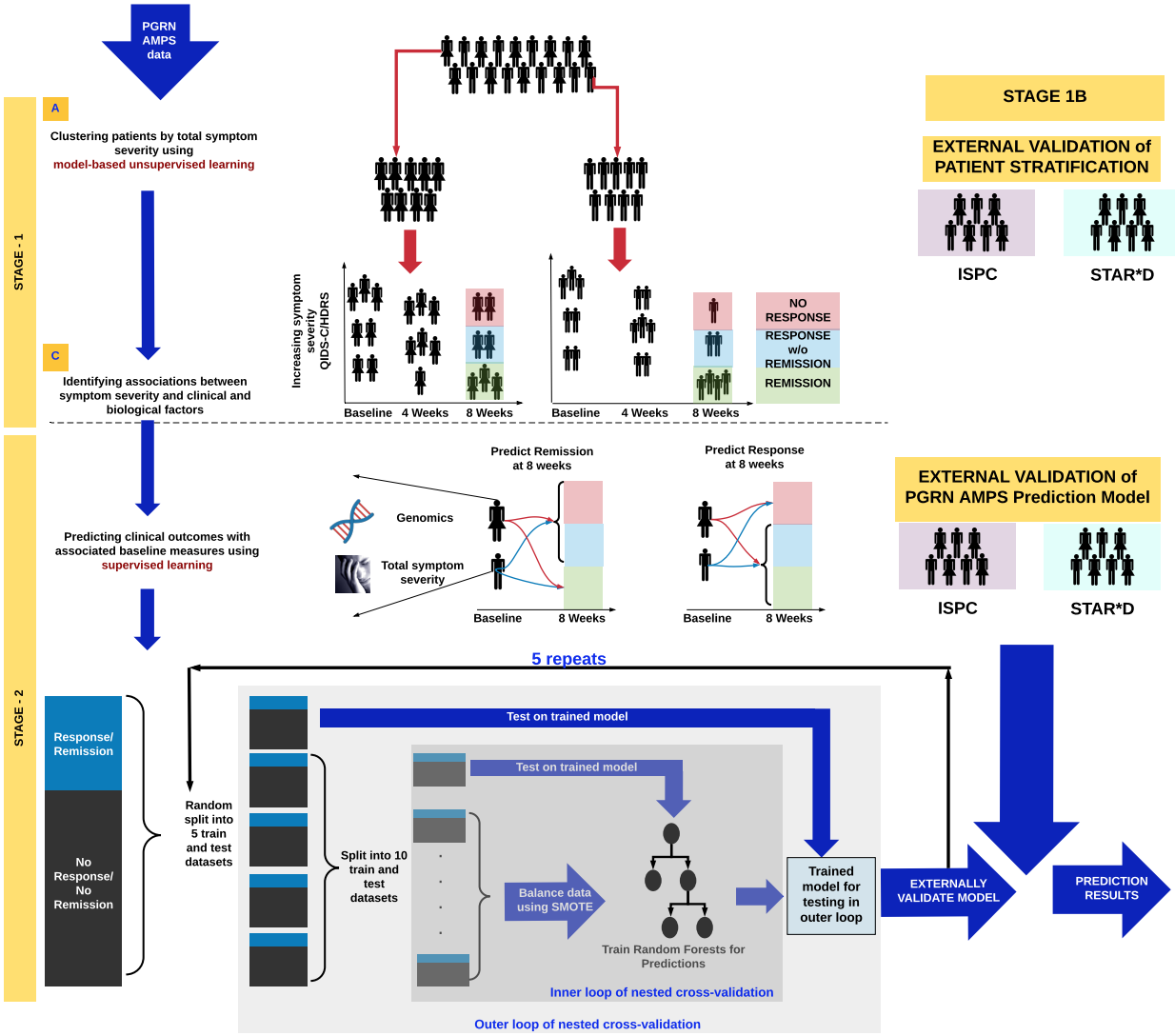


Figure 6.1: The proposed analyses to establish improved predictability in antidepressant treatment outcomes by augmenting the clinicians' assessments with biological measures.

Chapter 4 into one stage assuming we begin with sex-stratified analyses, and therefore we will still have a two-stage workflow in this chapter.

6.3.1 Stage - 1: Patient Stratification

Aim — Identify depressive symptom severity clusters in PGRN-AMPS (Stage 1A), replicate the cluster patterns using STAR*D and ISPC data (Stage 1B), and identify sociodemographic factors associated with clusters (Stage 1C). As described in Sec. 3.2.2, we first observed that distribution of depression severity scores was a mixture of distributions. We then used mixture model-based unsupervised learning with Gaussian mixture models (GMM) was used to algorithmically identify the minimum number of Gaussians that best approximated the actual distribution of depressive symptom severity in Mayo PGRN-AMPS subjects. By using this approach, we assumed that the mixture comprised multiple Gaussians. GMM clustering has been applied in numerous fields requiring separation of data types characterized by unique distributions [11, 90]. Given the eventual goal of associating biological measures with depression severity during discrete treatment time-points when clinical assessments are performed, longitudinal clustering/trajectory techniques [91] were not suitable. This is because symptom improvement (trajectories) is conditioned upon baseline severity and subsequent improvement (i.e., not independent), and depression severity is assessed at discrete time-points (as opposed to continuous time measures). Therefore, using a GMM clustering algorithm in our approach, we assigned patients to clusters based on their total depression severity score at each time-point.

To validate the clustering approach developed in Stage 1A, we used STAR*D (for QIDS-C) and ISPC (for HDRS) datasets in Stage 1B to investigate whether the distributions of depression severity using Kolmogorov-Smirnov test, were the same between two independent datasets.

In Stage 1C, Kolmogorov-Smirnov (continuous data) and two-way Chi-square (categorical data) tests were used to identify clinical and sociodemographic factors (listed in Table A.1) associated with the depression severity clusters at all time-points in all three datasets (i.e., Mayo PGRN-AMPS, STAR*D and ISPC). Any associated clinical/sociodemographic factors were then combined with pharmacogenomic SNPs to predict treatment outcomes in Stage 2.

6.3.2 Stage - 2: Predicting Antidepressant Response

Aim — Predict antidepressant remission or response using pharmacogenomic biomarkers and baseline depression severity. We trained random forests (randomForest

R library) [126] using PGRN-AMPS's baseline depression severity and pharmacogenomics data (represented as numerical genotypes [32,33,88]) to predict remission/response, and then externally validated the trained prediction model using STAR*D and ISPC data (see Fig. 6.1). Because clinical/sociodemographic factors, CY2C19 status and plasma drug levels were not associated with baseline or 4 week clusters, we only assessed the predictive capability of the pharmacogenomic biomarkers augmented by baseline depression severity, not stratified by the baseline depression severity cluster. Random forests were used because of their mathematical ability to handle discrete (e.g., numerical genotypes), correlated predictor variables, which has demonstrated robust predictive capabilities in several clinical applications [127], including psychiatric disorders [128].

To minimize the effects of overfit, information leak and use of all training data, nested cross-validation (nested-CV) with five repeats was used to train the classifiers, which maximized the area under the curve. In each repeat, data were randomized, and the nested-CV comprised an outer loop and an inner loop. The outer loop had a fivefold cross-validation to split the data into training data (80% of the data) and testing data (the remaining 20%). The inner loop used the training data to train the classifier using a tenfold cross-validation, and the trained classifier was tested on the testing data. To minimize the effects of class imbalance (i.e., unequal numbers of responders (60%) and non-responders (40%)) in the training data, we used the synthetic minority over-sampling (SMOTE) algorithm, which simulated patient profiles of the under-sampled class and up-sampled the under-sampled class to ensure that the two classes had equal sizes [112]. Following recommended practice of grid-search to find optimal number of trees that maximized AUC during training only, tuning parameter was set as square root of total number of variables, and number of trees selected was from 500 to 3,000 with increments of 100 [126]. The statistical significance of the classifier's accuracy was established using the null information rate (NIR, i.e., the prevalence of the class with the largest samples), which served as a proxy for chance. Only for the PGRN-AMPS data (in both scales), for which cross-validation was performed to train the prediction model, top predictors (using variable importance in R) and AUC will be reported in addition to PPV, NPV, sensitivity, specificity, and statistical significance of the classifier. For external validation of the trained model on the ISPC (for HDRS) and STAR*D (for QIDS-C) data, prediction performance will be reported using PPV, NPV, sensitivity, specificity, and statistical significance.

6.4 Results

In PGRN-AMPS (using nested cross-validation (CV) to train the prediction models), baseline depression severity combined with pharmacogenomic biomarkers predicted response and remission status with accuracies of 70 – 87.5% ($p < 0.01$, AUC 0.7 – 0.88) and 75 – 86% ($p < 0.04$, AUC 0.75 – 0.9), respectively, shown in Table 6.1. When CYP2C19 metabolizer status was included as a predictor variable, the prediction accuracies were reduced by at least 4% for both remission and response in both sexes and scales ($p > 0.3$). The classifier trained using PGRN-AMPS patients' total QIDS-C baseline depression severity and SNP data predicted response and remission status in STAR*D patients with accuracies of 64 – 70% ($p < 0.06$) and 65 – 75% ($p < 0.07$), respectively, shown in Table 6.1. The classifier trained using PGRN-AMPS patients' total HDRS baseline depression severity and SNP data predicted response and remission status in ISPC patients with accuracies of 76 – 77% ($p < 0.07$) and 80 – 83% ($p < 0.07$), respectively, shown in Table 6.1.

6.5 Discussion

6.5.1 Improved Predictions and Mechanistic Significance

We have shown that robust prediction of citalopram/escitalopram treatment outcomes can be achieved using machine learning approaches that integrate baseline depression severity with functionally validated pharmacogenomic biomarkers. The AUC of 0.70 or higher achieved in this work represents an advance over our prior work, in which we used sociodemographic and clinical factors as predictor variables in a machine learning algorithm applied to PGRN-AMPS data, which resulted in an AUC of 0.54 [124]. Importantly, the prediction model trained using PGRN-AMPS data predicted treatment outcomes in the STAR*D (QIDS-C scale) and ISPC (HDRS scale) trials with similar precision, thus providing cross-trial and multiple scale replication. Altogether, our findings represent an important step toward the goal of biologically guided selection of antidepressants for treating MDD patients.

The improvements in predictability of remission/response presented here are in line with the views expressed by others that including biomarkers as predictors could improve the ability of machine learning to predict remission with citalopram/escitalopram treatment as opposed to using only clinical and demographic variables [28, 29, 59, 102]. The SNPs that were included in our analysis were the "top hit" SNPs in GWAS of the plasma metabolites that were most highly associated with SSRI response (plasma serotonin) and baseline MDD symptom

Table 6.1: Prediction performance of pharmacogenomic biomarkers.

RESPONSE													
Rating scale	Trial	Training data	Gender	Accuracy (%)	95% CI in training CV	NIR	p-value	Sensitivity	Specificity	PPV	NPV	AUC in training CV	Top 3 predictors in CV
QIDS-C	PGRN-AMPS	10-fold CV	Men	73	(63,82)	0.66	0.01	0.7	0.78	0.86	0.57	0.85	DEFB1_2, Baseline severity, DEFB1_1
			Women	74	(65,83)	0.63	0.0003	0.71	0.8	0.85	0.62	0.7	TSPAN5, DEFB1_1, Baseline severity
	STAR*D	PGRN-AMPS	Men	69	NA	0.68	0.06	0.67	0.72	0.82	0.51	NA	NA
			Women	66	NA	0.68	0.0007	0.68	0.63	0.78	0.52	NA	NA
HDRS	PGRN-AMPS	10-fold CV	Men	86	(81,94)	0.68	5.70E-04	0.9	0.85	0.93	0.79	0.88	TSPAN5, DEFB1_1, DEFB1_2
			Women	88	(78,93)	0.7	7.30E-09	0.9	0.82	0.91	0.7	0.9	DEFB1_1, Baseline severity, DEFB1_2
	ISPC	PGRN-AMPS	Men	77	NA	0.69	0.05	0.8	0.71	0.85	0.62	NA	NA
			Women	75	NA	0.65	0.01	0.78	0.68	0.82	0.63	NA	NA
REMISSION													
Rating scale	Trial	Training data	Gender	Accuracy (%)	95% CI in training cross-validation	NIR	p-value	Sensitivity	Specificity	PPV	NPV	AUC in training cross-validation	Top 3 predictors in cross-validation
QIDS-C	PGRN-AMPS	10-fold CV	Men	78	(69,86)	0.63	5.40E-08	0.81	0.75	0.84	0.69	0.86	Baseline severity, DEFB1_1, DEFB1_2
			Women	69	(60,80)	0.62	0.0001	0.6	0.83	0.84	0.59	0.75	Baseline severity, DEFB1_2, DEFB1_1
	STAR*D	PGRN-AMPS	Men	75	NA	0.55	0.008	0.79	0.69	0.75	0.72	NA	NA
			Women	66	NA	0.5	0.001	0.59	0.72	0.67	0.63	NA	NA
HDRS	PGRN-AMPS	10-fold CV	Men	86	(75,90)	0.55	0.0001	0.9	0.84	0.87	0.87	0.84	Baseline severity, DEFB1_2, DEFB1_1
			Women	83	(75,90)	0.51	0.03	0.87	0.8	0.82	0.85	0.9	Baseline severity, DEFB1_2, DEFB1_1
	ISPC	PGRN-AMPS	Men	76	NA	0.56	0.04	0.8	0.71	0.77	0.73	NA	NA
			Women	74	NA	0.52	0.07	0.76	0.72	0.74	0.73	NA	NA
rs numbers of SNPs: AHR (rs17137566), TSPAN5 (rs10516436), ERICH3 (rs696692), DEFB1_1 (rs5743467), DEFB1_2 (rs2741130) and DEFB1_3 (rs2702877)													

severity (plasma kynurenine) in the PGRN-AMPS trial [32, 33]. The pharmacogenomic biomarkers for this study, which were SNPs in the DEFB1, AHR, TSPAN5, and ERICH3 genes, were chosen based on their important roles in serotonin or kynurenine biosynthesis, or in inflammation—mechanisms that are associated with MDD disease risk and/or antidepressant response [32, 33]. As noted earlier, knockdown of the expression of both TSPAN5 and ERICH3 in neuronally derived cell lines resulted in decreased serotonin release into the culture media [33]. Additionally, the DEFB1 gene encodes a protein expressed in gastrointestinal mucosa that can inactivate lipopolysaccharides and, as a result, inhibit both inflammation and the biosynthesis of kynurenine [32]. These observations are compatible with the rapidly evolving concept of a "gut-brain axis" [129, 130]. The identification of these SNPs during GWAS performed using quantitative biological traits (i.e., metabolite concentrations), rather than measures of MDD clinical symptom severity (i.e., HDRS or QIDS-C), as phenotypes represented a conscious attempt to move the analyses toward the biological underpinning SSRI response. With the goal of cross-trial replication, we focused on pharmacogenomic biomarkers in our predictive model because DNA data were more widely available than other "omics" data across datasets. Further, unlike metabolomics data, DNA sequences are stable and are less susceptible to variation related to environmental exposures (other medications, diet, etc.) or specimen handling and processing.

6.5.2 Sex Differences

When antidepressants are being chosen, potential sex differences in the underlying biology of antidepressant response are often overlooked. It is clear that sex represents an important risk factor for MDD [81]. Although sex has been reported to influence response to antidepressants in some studies [89, 131–133] prior machine learning approaches did not identify sex as a robust predictor of remission [28, 29, 59]. The sex-specific differences in some top predictors of treatment outcomes in our study suggest that sex-specific biological mechanisms may play an important role in antidepressant response.

Limitations. Our sample consisted of Caucasian subjects, limiting generalizability of the predictions. However, restricting our analyses to Caucasians may have reduced confounding by race. We had no direct measures of socioeconomic status and comorbid anxiety, factors associated with poorer response to antidepressants [134, 135]. With the use of complete cases, we cannot exclude the possibility of confounding by patients who dropped out. Although the improvement in outcome predictions replicated across clinical trials of citalopram/escitalopram, this work has not been replicated for other antidepressants. Finally, patients were not excluded

on the basis of BMI/comorbid general medical conditions that could have influenced the interactions between drug treatment and genomic profile.

6.6 Summary

By augmenting psychiatrists' total depression severity scores with pharmacogenomic biomarkers, we achieved superior prediction performances in comparison to those obtained when using sociodemographic factors as predictors of outcomes. More importantly, the predictive capabilities of the pharmacogenomic biomarkers replicated across two other large, independent citalopram/escitalopram studies. The overarching significance of the replicating predictability in treatment outcomes with pharmacogenomics biomarkers is suggested by the success of analogous "individualized medicine" approaches in breast cancer therapeutics. If pharmacogenomic biomarkers are available for other classes of antidepressants, antidepressant selection could potentially be guided by drug-specific predictive models, whereby an antidepressant is chosen that has the highest likelihood of causing a particular patient to achieve remission from his or her MDD symptoms. Then, there would be a true shift from "artisanal medicine" to "individualized medicine" in treating MDD.

CHAPTER 7

PROGNOSES AND PREDICTION OF ANTIDEPRESSANT RESPONSE BASED ON EARLY CHANGE IN DEPRESSION SYMPTOMS

During the pharmacological management of major depressive disorder (MDD), clinicians often make decisions about whether to continue or alter antidepressant treatment plans at intermediate treatment time-points, before a full therapeutic trial is complete. To operationalize the decision to continue or alter treatment plans at an intermediate treatment time-point, clinicians often focus on changes in individual depressive symptoms or total depression severity scores, measured using depression rating scales. However, clinicians are still unable to reliably predict eventual remission or non-response to current treatment, because of the heterogeneity in antidepressant response profiles [16–20] and the weak predictive effects of clinical and demographic variables for outcome prediction, with the exception of depression severity at baseline [28, 29, 59].

Prior studies using STAR*D and other large datasets have investigated whether early improvements in total depression rating scale scores can be used to predict eventual treatment non-response [67, 67–71, 74–77, 136, 137]. These studies relied on the use of growth mixture models and trajectory analyses [74–77], which cannot be used to individualize the prediction of eventual treatment outcomes by using specific improvements in the severity of total depression and individual depressive symptoms at an intermediate treatment time-point. Subscales (such as the Maier-6 [138], Bech-6 [139], HAMD7 [140], and VQIDS-C5 [141] subscales) have been derived from full-scale versions of the HDRS and QIDS-C to measure depressive symptoms that are more responsive to antidepressants. However, there is still a need to gain additional specificity by defining which specific set of depressive symptoms from the full rating scales must change at an intermediate time-point, by how much, and in which subgroup of patients, in order to individualize the predictions of eventual treatment outcomes, such as remission or non-response.

Contribution. Our main purpose in stratifying MDD patients was to parse, to the greatest possible extent, heterogeneity in antidepressant response. The replicating patient clusters at each time-point served as nodes in defining a probabilistic graph. That allowed us to model the probabilistic nature of the transitions patient make between clusters of consecutive time-points of the trial. Thus we allowed for eventual improvement in depression

severity to be conditioned upon, first, where patients began before the trial, and second, how they improved with time during the trial, while going on to achieve any of the clinical outcomes. In doing so, we established *symptom dynamic paths*, which are the most likely progressions through stages of depression severity that patients might experience, given their baseline cluster, while they go on to achieve any of the categorical treatment outcomes. Our inference of symptom dynamic paths then allowed us to identify a set of "core" depressive symptoms whose 4-week changes are enough to differentiate symptom dynamic paths that originate in the same baseline cluster, while leading to remission, response, and non-response at 8 weeks. Psychiatrists find that association statements that reflect the aggregate statistics (e.g., an $X\%$ improvement in total score has $Y\%$ chance of achieving eventual remission) clinically not actionable, because they overlook prognostic capabilities of changes in individual depressive symptoms and potential sex-differences. Instead, our approach says *if a patient's total depression severity has changed from X to Y after 4 weeks of treatment, and if A or more core depressive symptoms have improved by B score points, this patient has $Z\%$ chance to achieve remission.* Therefore, we now provide additional specificity at the individual patient level that can better augment the psychiatrists' judgments in changing treatments early if the early prognoses of treatment outcome are poor.

7.1 Problem Statements

Our goal is to extract longitudinal homogeneity in antidepressant response in our identified strata of patients. We addressed the following questions in this work:

1. Can we identify a specific set of (core) depressive symptoms, in patients with comparable levels of total depression severity, that exhibit high homogeneity in their longitudinal response to citalopram/escitalopram treatment?

2. Can baseline and early changes in core symptoms (at 4 weeks) accurately predict eventual outcomes after 8 weeks of citalopram/escitalopram treatment? Can predictive performance replicate across multiple independent trials and rating scales?

3. Can specific thresholds of change in core symptoms be identified at 4 weeks that are highly prognostic of eventual outcomes at 8 weeks?

7.2 Data

For this study, we used complete response data of subjects from PGRN-AMPS (NCT 00613470), STAR*D (NCT 00021528) and ISPC trials. For the present analyses, we utilized data from 603 Caucasian citalopram-treated PGRN-AMPS subjects, 899 Caucasian citalopram-treated STAR*D subjects, and 344 citalopram/escitalopram-treated ISPC subjects who had complete clinical data at baseline and at 4 and 8 weeks. Only complete cases were considered, as our explicit goal was to model longitudinal symptom responses to study drugs, conditioned on baseline depression severity and changes in depressive symptoms at intermediate time-points. Clinical outcome definitions are outlined in Sec. 2.1.1.

7.3 Approach Overview

A machine learning workflow comprising tasks (a) – (f) across five stages was developed to predict eventual treatment outcomes by using a set of individual depressive items with homogeneity in their longitudinal response to citalopram/escitalopram treatment. The background analyses needed for this workflow to be operational is the patient stratification workflow described in Chapter 4. We summarize the results from the chapter for ease in following material presented in this chapter.

Summary of patient stratification from Chapter 4: In each of the three datasets, the our clustering approach identified three clusters of patients based on total depressive symptom scores at baseline (labeled A1, A2, A3), 4 weeks (B1, B2, B3), and 8 weeks (C1, C2, C3). The ranges of depression severity scores for both scales were as follows. For the QIDS-C: A1 [7 — 12], A2 [13 — 16], A3 [17 — 25]; B1 [0 — 6], B2 [7 — 11], B3 [12 — 25]; C1 [0 — 5], C2 [6 — 11], C3 [12 — 24]. For the HDRS: A1 [14 — 18], A2 [19 — 24], A3 [25 — 39]; B1 [0 — 8], B2 [9 — 15], B3 [16 — 31]; C1 [0 — 7], C2 [8 — 15], C3 [16 — 34]. The clusters of comparable ranges of total depression severity between trials of a given rating scale (e.g., B1 of PGRN-AMPS and STAR*D for QIDS-C; B1 of PGRN-AMPS and ISPC for HDRS) were identically distributed ($p > 0.7$, from the Kolmogorov-Smirnov (KS) test). In clusters identified across all three datasets using our clustering approach, all patients in the C1 cluster achieved remission, and all patients in the C3 cluster were non-responders, i.e., failed to achieve remission or response. 87% of patients in the C2 cluster achieved response without remission (and the remaining 13% were non-responders). The clusters (A1,. . . ,C3) inferred at each time-point served as nodes of the probabilistic graph, which were then used to study the longitudinal effects of antidepressants.

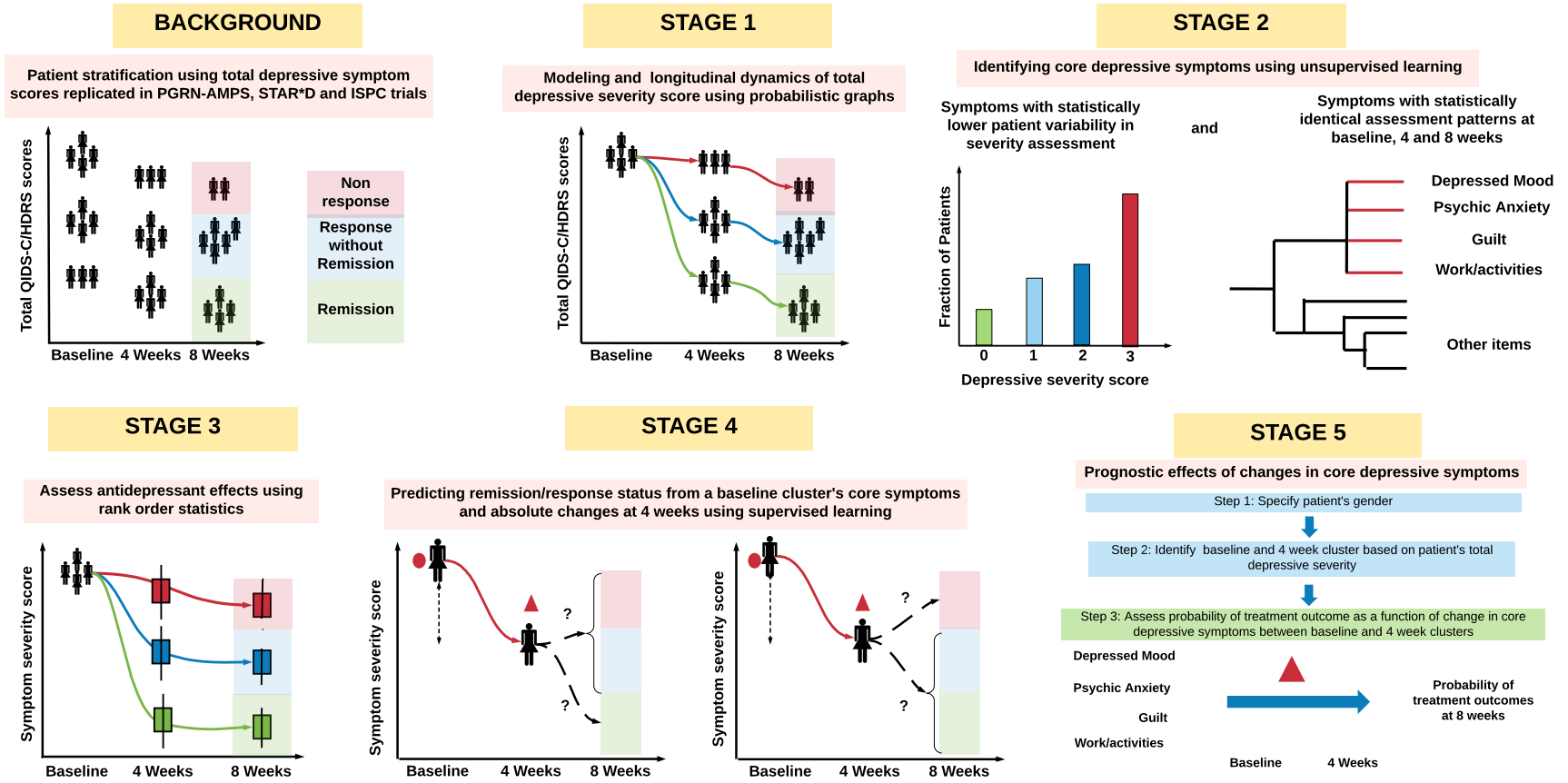


Figure 7.1: Using total depression severity scores, we first cluster patients at all three time-points (background stage) and then identify the symptom dynamic paths, i.e., the paths patients are most likely to traverse from baseline to 8 weeks through a cluster at 4 weeks (Stage 1). For the symptom dynamic paths originating at a given baseline cluster, we identify core depressive symptoms that demonstrate longitudinal homogeneity in their responses and prognostic effects based on early change (Stage 2). Antidepressant effects on these core symptoms are assessed due to the replication of results from Stages 1 and 2 in independent trials (Stage 3). Treatment outcomes at 8 weeks after starting from a baseline cluster are predicted using baseline severity and changes in core symptoms from baseline to 4 weeks (Stage 4). Finally, clinical utility of the prognostic effects of core symptoms is demonstrated in Stage 5.

7.3.1 Stage - 1: Modeling of Total Depressive Symptom Dynamics Paths

Aim — Identify paths from each cluster that explain the most likely trajectories to achieve remission, response and non-response. Probabilistic graphs with the forward algorithm were used to explore all possible paths connecting a given baseline cluster to patient clusters at 4 weeks, and then from 4 weeks to 8 weeks. We defined the graph as a hidden Markov model in order to use the recursive forward algorithm to compute the likelihood of each path, details of this process are explained in Sec. 3.4. For each of the nine pairs of baseline and 8-week clusters (e.g., A1, C1), we chose the "most likely" path, defined as the path between these two clusters through a given 4-week cluster (e.g., B1) that had the highest likelihood score, provided that the path was followed by more than 10% of the patients in the cluster in which the path originated. These unique "most likely" paths are subsequently referred to as symptom dynamics paths.

7.3.2 Stage - 2: Identification of Core Depressive Symptoms

Aim — Identify depressive symptoms that exhibit high homogeneity in their longitudinal response to citalopram/escitalopram treatment. To extract homogeneous patterns of antidepressant response, we defined "core depressive symptoms" based on three criteria: (1) similar response patterns at all time-points, (2) low inter-individual variability, and (3) patterns of change that were statistically distinct within each of the symptom dynamic paths (which we inferred in Stage-1 using total depression severity scores). First, unsupervised machine learning (i.e., hierarchical clustering with complete linkage) was used to identify individual QIDS-C and HDRS scale items with similar rating patterns (meaning that they were clustered together) within the patient clusters at baseline, 4 weeks, and 8 weeks. Second, we identified symptom clusters wherein clinician ratings for each of the scale items at baseline had a nonzero median and low inter-individual variability. A given item was defined as having low inter-individual variability if the chi-square test for the distribution of clinician ratings was significant after multiple comparisons, with the null hypothesis being that the distributions of ratings for that item were equal. Third, for each pair of symptom dynamic paths originating from a baseline cluster to a cluster at 8 weeks (e.g., A3→B3 → C3 and A3 → B2 → C2), the KS test was used to determine whether there were statistically significant differences between the associated distributions of core symptom at 4 weeks. We used average smoothing kernels to visualize the variations in these core symptoms' scores within specific symptom dynamic paths.

7.3.3 Stage - 3: Assessment of Antidepressant Effects on Core Depressive Symptoms

Aim — Check if symptom improvement is more likely due to antidepressants treatment than to chance. The Mann-Whitney U-test was used to assess whether the severity of the core depressive symptoms (expressed as a rank order) changed significantly as a likely response to antidepressant treatment between two consecutive time-points on a given symptom dynamics path. By utilizing the replicating patient clusters across datasets, we satisfied the sample independence requirement for this test by comparing patients in one cluster from one dataset with patients in the consecutive time-point cluster from another dataset. Details of constructing the test are discussed in Sec. 3.5.2 and illustrated in Fig. 3.6.

7.3.4 Stage - 4: Prediction of Clinical Outcomes from Early Changes in Core Depressive Symptoms

Aim — To test for capabilities of core symptom changes in patients stratified by baseline depression in predicting eventual treatment outcomes. We used random forests, a nonparametric supervised machine learning method, as a binary classifier to predict clinical outcomes at 8 weeks given a specific baseline cluster, using the associated baseline severity of the core depressive symptoms and their absolute changes at 4 weeks. Using five-repeat 10-fold nested cross-validation, we trained the sex- and rating-scale-stratified classifiers with data from PGRN-AMPS subjects. We followed the recommended practice of grid search during training by setting the mTry parameter to one-half of the total number of predictor variables, and chose the number of trees from the range of 500 to 2,000 with increments of 100. The trained prediction models were then externally validated using STAR*D subjects (for the QIDS-C scale), and ISPC subjects (for the HDRS scale). Prediction performance was reported using several metrics (AUC, PPV, NPV, sensitivity, and specificity), and the statistical significance of the classifier's accuracy was established using the null information rate (NIR, the prevalence of the class with the largest samples), which served as a proxy for chance. Because 98% of the baseline cluster A1 patients who achieved response were also classified as remitters, we trained prediction models for A1 cluster patients to predict remission at 8 weeks, which gave us additional samples for training.

7.3.5 Stage - 5: Establishment of the Prognostic Effects of Core Depressive Symptoms

Aim — Deriving prognoses of eventual treatment outcome using change in severity of core depressive symptoms. This step defined the minimum number of core symptoms and levels of improvement in the core symptoms needed at 4 weeks (given a specific baseline cluster) to achieve specific outcomes at 8 weeks. First, the threshold of improvement vs. failure to improve was chosen based on changes in median scores on symptom dynamic paths between baseline and 4-week clusters. Second, a chi-square test was conducted on a table comprising the number of core symptoms that exceeded (or failed to exceed) the threshold at 4 weeks, versus the outcome labels (e.g., remitters vs. non-remitters, or responders vs. non-responders). If the chi-square test's p-value was significant for remission or response/non-response, we computed the probability of the outcome based on how many symptoms had to exceed (or failed to exceed) the threshold. If the p-value was not significant, no conclusions about treatment outcome based on changes in any number of core symptoms were possible. Standard deviations (SD) of the probabilities were established by creating five random subsets (each subset maintained the same proportions of patients who achieved remission/response/non-response in the entire dataset), and with 10 repetitions (i.e., five different random subsets in each repeat).

Tasks (a) and (c) were in-situ (non-time-varying) inferential tasks, whereas task (b) was a longitudinal inferential task that required conditional dependencies (which motivated the use of probabilistic graphs), and task (e) was a predictive task that required supervised learning methods. Therefore, multiple statistical/machine learning methods (illustrated in Fig. 7.1) were needed to address these sequential tasks.

7.4 Results

7.4.1 Symptom Dynamics Based on Total Depression Scores

The symptom dynamic paths inferred using our machine learning workflow were identical for the PGRN-AMPS and STAR*D datasets (for QIDS-C), and for the PGRN-AMPS and ISPC datasets (for HDRS), based on the baseline and 4- and 8-week clusters on the path (Fig. 7.2). In clusters on the symptom dynamic paths in the PGRN-AMPS and STAR*D datasets (for QIDS-C), and in the PGRN-AMPS and ISPC datasets (for HDRS), the distributions of total depression scores and clinician ratings of individual scale items of QIDS-C or HDRS were statistically identical between associated trials ($p > 0.2$, as found using the KS-test for scores

of corresponding clusters between paths of two datasets).

7.4.2 Sex Differences in Symptom Dynamics

There were notable sex differences in the symptom dynamic paths. In both rating scales, only women demonstrated path A3 $\rightarrow$ B1 $\rightarrow$ C1, which linked patients with severe baseline depressive symptoms (cluster A3) to remission at 8 weeks, via cluster B1 at 4 weeks. In both rating scales, and in both women and men, the path from severe baseline depressive symptoms (cluster A3) to response (cluster C2) at 8 weeks was through cluster B2 at 4 weeks. On the A3 $\rightarrow$ B2 $\rightarrow$ C2 path when the HDRS rating scale was used, women were more likely to achieve an early response (i.e., $\geq 50\%$ reduction in baseline total depressive severity at 4 weeks) going from A3 to B2 (solid line in Fig. 7.2) than men (dashed line in Fig. 7.2). However, for the same path in the QIDS-C data, both sexes were likely to achieve an early response at 4 weeks, while achieving eventual response at 8 weeks.

7.4.3 Core Depressive Symptoms

In both the individual and combined datasets, five QIDS-C items (sad mood, concentration / decision-making, self-outlook, involvement, and energy/fatigability) and four HDRS items (sad mood, psychic anxiety, guilt feelings/delusions, and work/activities) met the core depressive symptom selection criteria outlined in Stage 2 of our analyses. These depressive symptoms clustered together at baseline and at 4 and 8 weeks. The five core depressive items of the QIDS-C and four core depressive items of the HDRS accounted for 62% and 43% of the variation in QIDS-C and HDRS total scores at baseline. The average contribution of the four core symptoms to the baseline HDRS total score was lower than the average contribution of the five core symptoms to the baseline QIDS-C total score. This is because 16% of patients had a score of zero for at least one of the core HDRS symptoms vs. only 11% for QIDS-C.

For the symptom dynamic paths (inferred using total depression severity scores) from a given baseline cluster (e.g., A3 in Figs. 7.3 and 7.4), there were significant differences in the distribution of core depressive item scores at 4 and 8 weeks for both the QIDS-C and HDRS ($p < 0.008$). Further signifying the degree of homogeneity in the scoring patterns of the core symptoms vs. non-core symptoms is visualized by the extent of overlap in bands of confidence interval of severity scores. Finally, on these paths, the changes in core depressive symptom severity in both rating scales were statistically significant from baseline to week 4, but were seldom significant from week 4 to week 8 (Table 7.1).

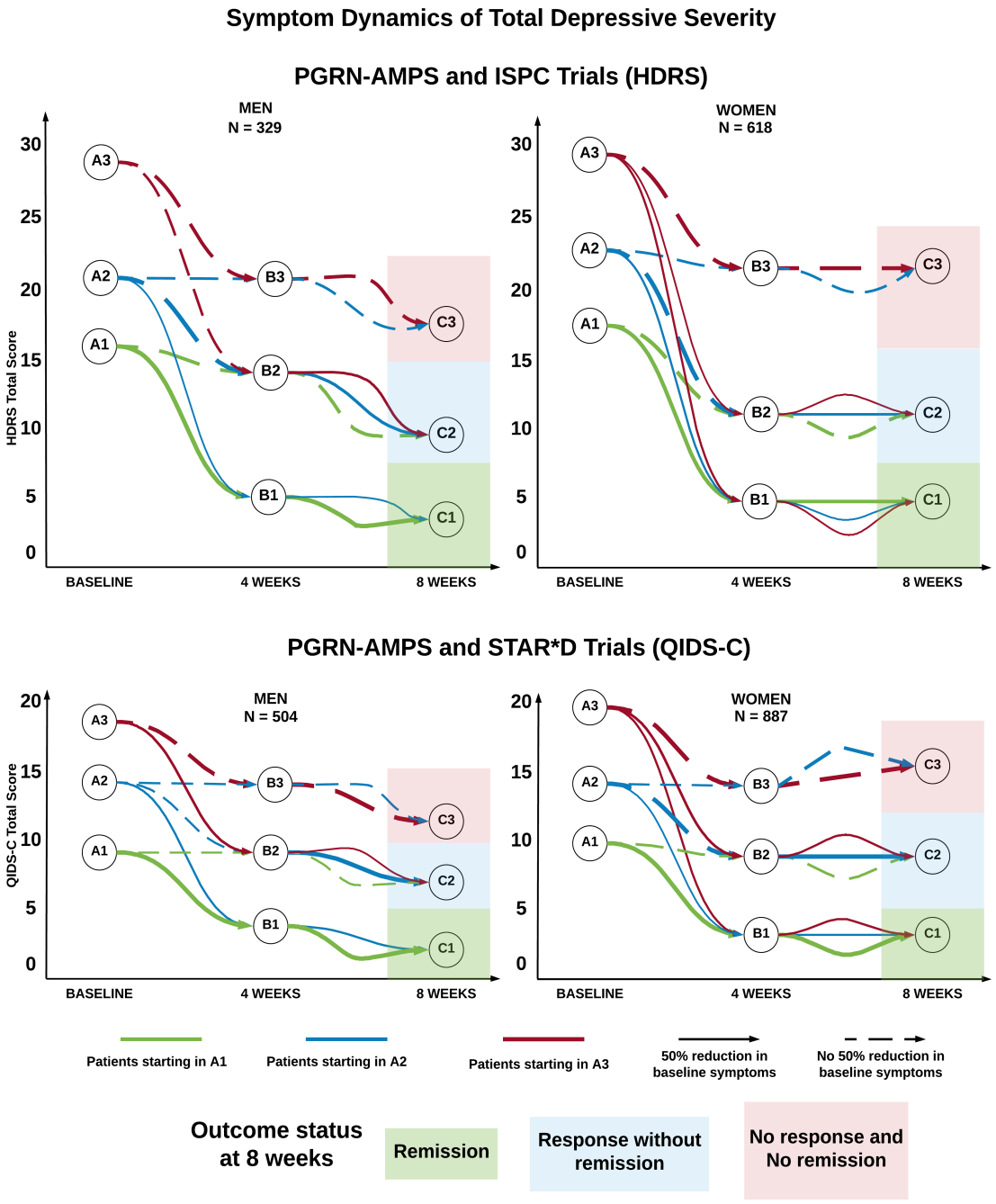


Figure 7.2: Patients are grouped into clusters at baseline, 4, and 8 weeks of citalopram/escitalopram using Gaussian mixture model clustering. These clusters allow for the use of probabilistic graphs to identify the most likely path of transition from a cluster at baseline to a cluster at 8 weeks, through an intermediary cluster at 4 weeks. These paths also account for the likely clinical outcomes observed at 4 and 8 weeks based on the improvement in the total depressive severity scores, relative to baseline total depressive severity. Thickness in lines indicates proportion of patients from the baseline cluster on the path, and the solid line indicates that between consecutive clusters, the total depressive symptom severity score is likely to be reduced by 50% relative to the baseline total depressive severity.

7.4.4 Visualizing the Benefits of Patient Stratification on Core Symptom Variations

Furthermore, elucidating the benefit of clustering patients and the subsequent finding of symptom dynamic paths is seen in the boxplots of scores of core symptoms. As shown in Fig. 7.3, although the 95% CI around the mean core symptom scores is narrow, the corresponding box plots at each time-point span the range of scores along the y-axis, indicating high inter-individual variation (heterogeneity) of core symptom scores. Adding the clustering framework (illustrated in Fig. 7.3) resulted in reduced heterogeneity of core symptom scores at all time-points, as illustrated by smaller box plots and confidence intervals, thus enhancing the capability of predicting 8 week treatment outcomes from patterns of core symptom responses at 4 weeks.

7.4.5 Prediction Performance of Core Depression Symptom Ratings

The performance characteristics of the machine learning algorithm that used baseline core depressive symptom severities and their absolute changes at 4 weeks to predict non-response or response and non-remission or emission at 8 weeks are summarized in Tables 7.2 and 7.3, respectively. In PGRN-AMPS (which was used to train the prediction models), the predictive accuracies for response and remission status were 65% – 77% ($p < 0.01$, AUC 0.66 – 0.8) and 66% – 79% ($p < 0.06$, AUC 0.67 – 0.87), respectively. The classifier trained using PGRN-AMPS data (with QIDS-C assessments) then predicted response and remission status in STAR*D patients with accuracies of 63% – 70%($p < 0.06$) and 69% – 73%($p < 0.07$), respectively. The classifier trained using PGRN-AMPS patients' data (with HDRS assessments) predicted response and remission status in ISPC patients with accuracies of 73% – 92%($p < 0.05$) and 75% – 91%($p < 0.04$), respectively.

In over 90% of patients for whom the prediction of clinical outcomes was incorrect, the core depressive symptoms accounted for $< 30\%$ of the total QIDS-C/HDRS scores at baseline. In these patients, no other set of symptoms met the criteria for core depressive symptoms. Furthermore, in all patients, the prediction performance of a classifier that used the remaining (non-core) depressive symptoms was significantly poorer (accuracy 56%, $p < 0.2$, AUC 0.59) than that of the classifier that used the core depressive symptoms. This result is illustrated by the significant overlap of the means of the confidence intervals of non-core symptom severities (illustrated in Fig. 7.4B). In a contrast, the core symptoms' variation on the symptom dynamic paths originating from a baseline cluster begin to show separation in confidence intervals at 4 weeks (illustrated in Fig. 7.4A).

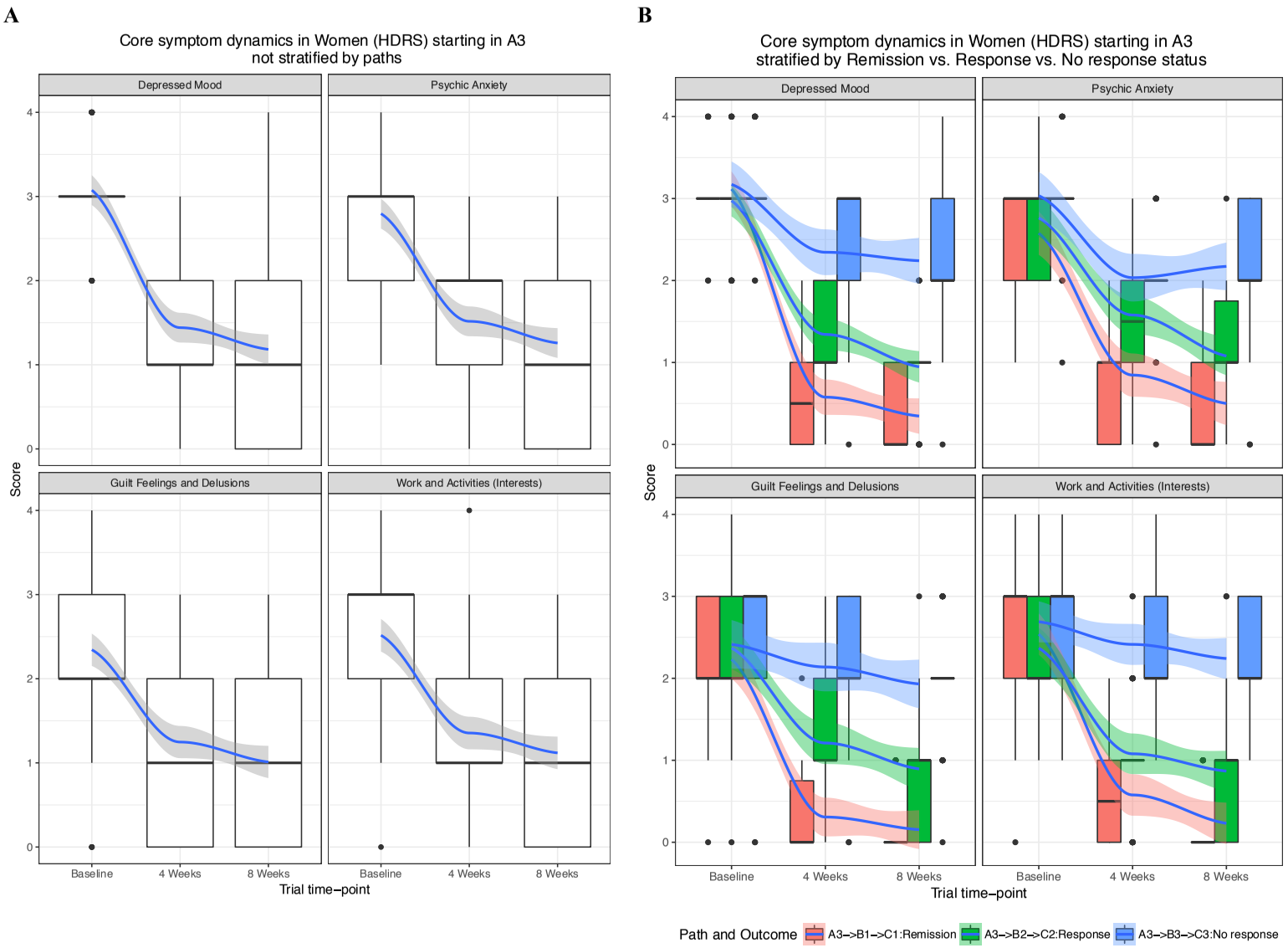


Figure 7.3: Variations in core symptom severity with and without patient stratification (and subsequently, symptom dynamic paths). The solid blue lines in each figure represent the variations (mean changes) in core symptom scores, and shaded regions around the mean illustrate their 95% confidence intervals (CIs).

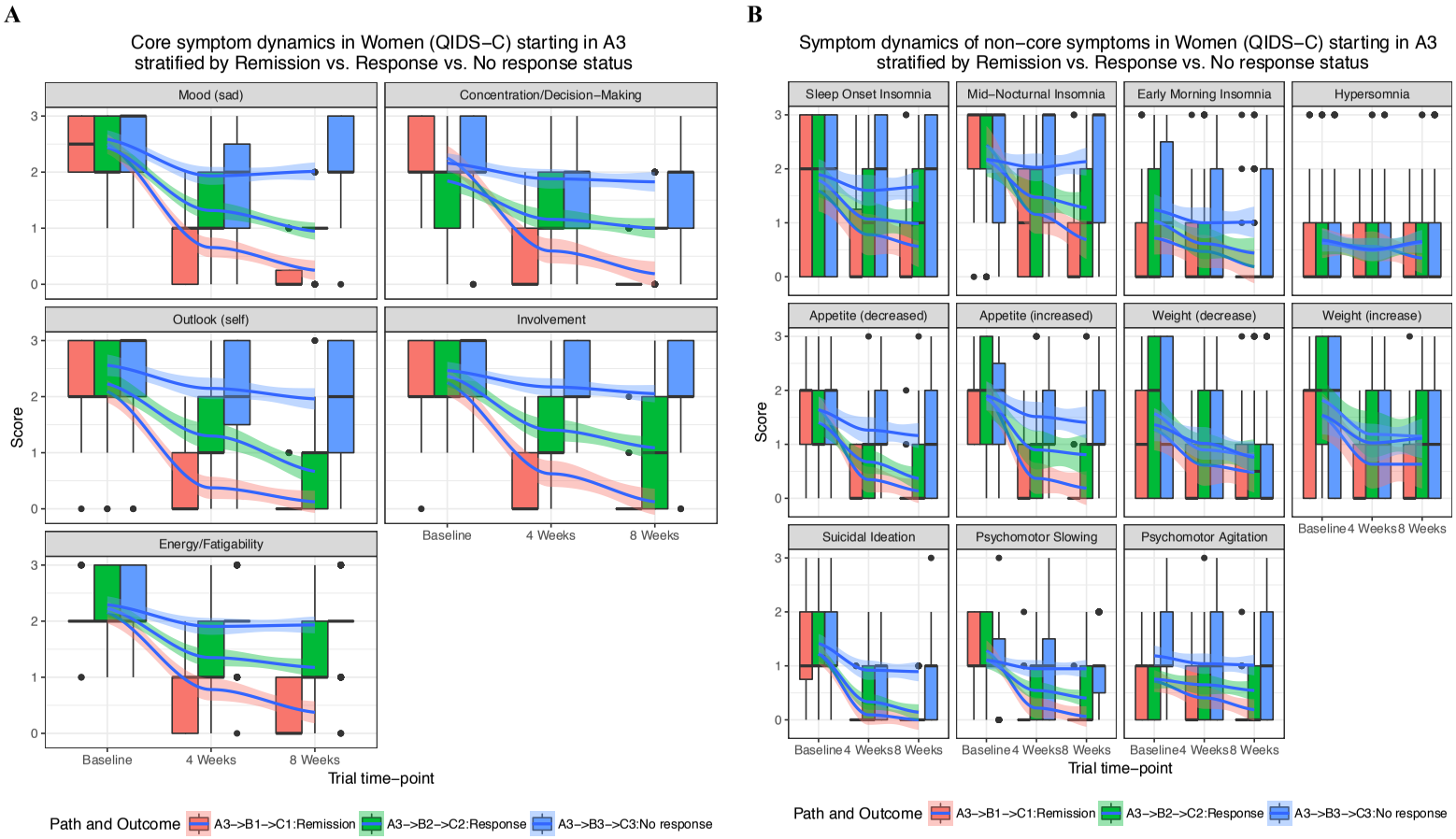


Figure 7.4: This figure illustrates the variation in scores of the QIDS-C's core symptoms (Fig. A) and non-core depressive symptoms' severity (Fig. B) on symptom dynamic paths leading to a categorical treatment outcome starting from a baseline cluster (e.g., A3 in this illustration).

Table 7.1: Median scores of core symptoms on the symptom dynamic paths: For each symptom dynamic path, severity measure, and core depressive symptom, we illustrate the median score on the paths. One table for both sexes is provided, as they had the same changes in median scores.

QIDS-C								
Symptom dynamics path	A1→B1→C1	A1→B2→C2	A2→B1→C1	A2->B2→C2	A2→B3→C3	A3→B1→C1	A3→B2→C2	A3→B2→C3
Outome	Remission	No response	Remission	Response	No response	Remission	Response	No response
Mood (sad)	2→0→0	2→1→1	2→1→0	2→1→1	2→2→2	3→1→0	3→2→1	3→2→2
Concentration/Decision Making	1→0→0	1→1→1	2→0→0	2→1→1	2→2→2	2→0→0	2→1→1	2→2→2
Outlook (self)	1→0→0	1→1→0	2→0→0	2→1→1	2→2→2	2→0→0	2→1→1	3→2→2
Involvement	1→0→0	1→1→1	2→0→0	2→1→1	2→2→2	2→0→0	2→1→1	2→2→2
Energy/Fatigability	1→0→0	1→1→1	2→1→0	2→1→1	2→2→2	2→1→0	2→1→1	2→2→2
HDRS								
Symptom dynamics path	A1→B1→C1	A1→B2→C2	A2→B1→C1	A2->B2→C2	A2→B3→C3	A3→B1→C1	A3→B2→C2	A3→B3→C3
Outome	Remission	No response	Remission	Response	No response	Remission	Response	No response
Depressed Mood	2→0→0	2→1→1	3→1→0	3→1→1	3→2→1	3→0→0	3→1→1	3→3→2
Psychic Anxiety	2→1→0	2→1→1	2→1→0	2→1→1	2→2→1	3→1→0	3→1→1	3→2→2
Guilt Feelings and Delusions	2→0→0	2→1→1	2→-0→0	2→1→1	2→2→1	2→0→0	2→1→1	2→2→2
Work and Activities (Interests)	2→0→0	2→2→1	2→-0→0	2→1→1	3>2→1	3→0→0	2→1→1	3→2→2

Table 7.2: Predicting response/non-response outcome at 8 weeks using change in severity of core symptoms at 4 weeks.

OUTCOME AT 8 WEEKS: RESPONSE														
Rating scale	Baseline cluster	Trial	Training data	Gender	Accuracy (%)	95% CI in training CV	NIR	p-value	Sensitivity	Specificity	PPV	NPV	AUC in training CV
QIDS-C	A3	PGRN-AMPS	10-fold CV	Men	77.32	(67.7,85.21)	0.67	0.01	0.62	0.84	0.67	0.82	0.8
				Women	66.75	(60.39,73.4)	0.65	0.03	0.57	0.72	0.53	0.76	0.66
		STAR*D	PGRN AMPS	Men	70.4	NA	0.6	0.04	0.59	0.78	0.64	0.74	NA
				Women	62.85	NA	0.57	0.06	0.6	0.65	0.56	0.68	NA
	A2	PGRN-AMPS	10-fold CV	Men	65.32	(56.23,73.64)	0.63	0.04	0.58	0.7	0.53	0.74	0.72
				Women	68.14	(60,73)	0.62	0.001	0.7	0.67	0.56	0.78	0.67
		STAR*D	PGRN AMPS	Men	63.2	NA	0.55	0.07	0.54	0.72	0.61	0.65	NA
				Women	66.7	NA	0.67	0.04	0.62	0.69	0.5	0.78	NA
HDRS	A3	PGRN-AMPS	10-fold CV	Men	89.3	(71.7,97.7)	0.75	0.05	0.71	0.95	0.83	0.9	0.91
				Women	73.08	(61.8,92.5)	0.6	0.01	0.77	0.7	0.63	0.82	0.79
		ISPC	PGRN AMPS	Men	75.2	NA	0.71	0.04	0.61	0.81	0.56	0.83	NA
				Women	68.05	NA	0.65	0.06	0.7	0.67	0.53	0.8	NA
	A2	PGRN-AMPS	10-fold CV	Men	89.5	(77.4,95.3)	0.77	0.02	0.72	0.94	0.8	0.92	0.92
				Women	92.9	(85.2,97.3)	0.67	1.20E-08	0.85	0.96	0.92	0.93	0.94
		ISPC	PGRN AMPS	Men	69.9	NA	0.73	0.04	0.67	0.71	0.47	0.85	NA
				Women	76.8	NA	0.7	0.001	0.68	0.8	0.61	0.85	NA

Table 7.3: Predicting response/non-response outcome at 8 weeks using change in severity of core symptoms at 4 weeks.

OUTCOME AT 8 WEEKS: REMISSION													
Rating scale	Baseline cluster	Trial	Training data	Gender	Accuracy (%)	95% CI in training CV	NIR	p-value	Sensitivity	Specificity	PPV	NPV	AUC in training CV
QIDS-C	A3	PGRN-AMPS	10-fold CV	Men	77.32	(67.7,85.2)	0.75	0.06	0.87	0.55	0.83	0.6	0.76
				Women	74.72	(64.5,76.46)	0.69	0.04	0.83	0.71	0.56	0.9	0.71
		STAR*D	PGRN-AMPS	Men	69.35	NA	0.71	0.04	0.8	0.65	0.5	0.87	NA
				Women	71.1	NA	0.65	0.07	0.75	0.69	0.57	0.84	NA
	A2	PGRN-AMPS	10-fold CV	Men	75.1	(63.9,80)	0.66	0.001	0.85	0.7	0.6	0.9	0.67
				Women	65.49	(58.9,71.67)	0.59	0.032	0.79	0.5	0.68	0.6	0.7
		STAR*D	PGRN-AMPS	Men	72.68	NA	0.61	0.04	0.8	0.68	0.61	0.84	NA
				Women	68.82	NA	0.51	0.06	0.78	0.6	0.65	0.73	NA
	A1	PGRN-AMPS	10-fold CV	Men	79.53	(63,84)	0.63	0.04	0.6	0.91	0.79	0.79	0.87
				Women	75.13	(60.9,83.24)	0.52	0.00038	0.75	0.71	0.7	0.75	0.82
		STAR*D	PGRN-AMPS	Men	71.6	NA	0.68	0.08	0.58	0.78	0.55	0.79	NA
				Women	70.56	NA	0.56	0.001	0.7	0.71	0.65	0.75	NA
HDRS	A3	PGRN-AMPS	10-fold CV	Men	85.71	(67.3,95.97)	0.65	0.011	0.83	0.93	0.93	0.75	0.88
				Women	89.4	(80.9,94.5)	0.75	0.0006	0.89	0.9	0.96	0.76	0.91
		ISPC	PGRN-AMPS	Men	83.9	NA	0.59	0.03	0.82	0.86	0.8	0.87	NA
				Women	75.8	NA	0.7	0.004	0.8	0.74	0.57	0.9	NA
	A2	PGRN-AMPS	10-fold CV	Men	87.5	(74.5,94.0)	0.54	8.80E-07	0.9	0.84	0.83	0.91	0.93
				Women	83.3	(73.9,90.6)	0.6	2.20E-06	0.95	0.63	0.8	0.93	0.9
		ISPC	PGRN AMPS	Men	82	NA	0.5	0.0008	0.88	0.76	0.78	0.86	NA
				Women	81.74	NA	0.71	0.04	0.86	0.8	0.65	0.9	NA
	A1	PGRN-AMPS	10-fold CV	Men	96.67	(83.3,98.2)	0.58	1.10E-06	0.92	0.9	0.95	0.4	0.95
				Women	87.84	(78.16,94.3)	0.57	7.20E-06	0.75	0.96	0.95	0.84	0.93
		ISPC	PGRN AMPS	Men	91.1	NA	0.65	0.0001	0.9	0.92	0.9	0.91	NA
				Women	82.6	NA	0.62	0.05	0.87	0.8	0.73	0.9	NA

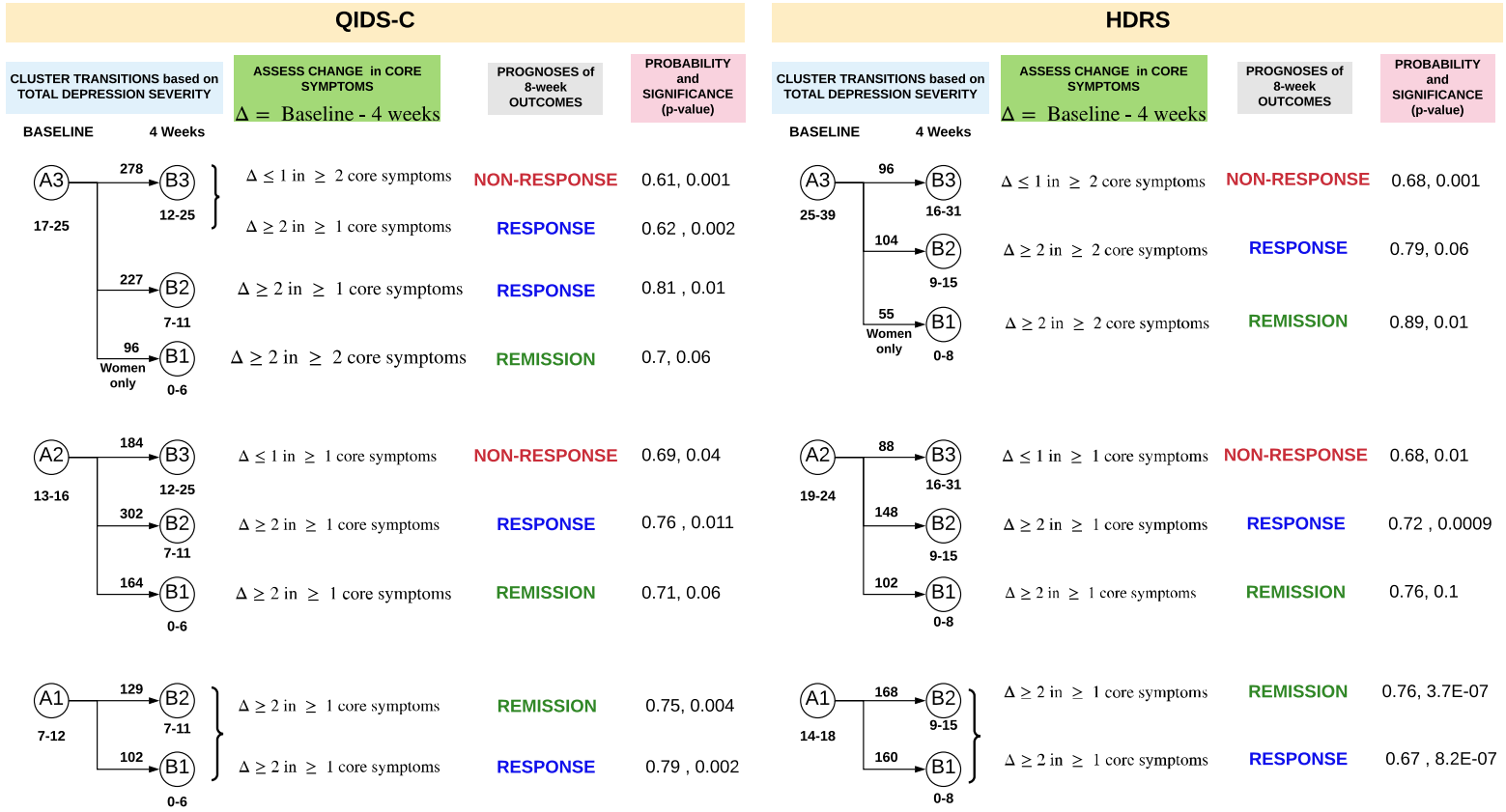


Figure 7.5: Treatment prognoses based on transition between clusters and degrees of improvements in core symptoms.

7.4.6 Prognostic Effects of Change in Core Depressive Symptoms

We identified thresholds of change needed in the core symptoms at 4 weeks to achieve the most likely treatment outcome at 8 weeks on the symptom dynamic paths (Fig. 7.5). The thresholds were derived using the change in median scores on symptom dynamic paths, as illustrated in Table 7.1. For example, as shown in Table 7.1, failure to improve by ≥ 1 point in each of the core symptoms was highly predictive of eventual non-response, whereas improvement by ≥ 2 points in each of the core symptoms was highly predictive of eventual remission/response. This result was observed in both men and women, and in both rating scales. Although sex differences were observed in symptom dynamic paths with respect to the total depression severity scores, the distributions of core symptom severity scores on the same path were statistically identical ($p > 0.6$) for men and women (with the exception that the A3→B1→C1 path was observed in women only). Furthermore, as changes in median scores were not sex-dependent (again with the exception of the A3→B1→C1 path that was observed only in women), we computed prognostic effects by combining data from both sexes with sample sizes of at least 88 patients in each path.

The analyses conducted to identify thresholds of core symptom change at 4 weeks that predicted non-response at 8 weeks were focused on patients with moderate (A2) or severe (A3) baseline total depressive severity. As shown in Table 7.1, eventual non-responders with moderate or severe baseline depression generally had absent to minimal improvement (reduction by 1 point) in core depressive symptom scores at 4 weeks (B3). On the A2→B3 path, 69% (SD 1.9) and 68% (SD 2.0) ($p < 0.03$) of patients achieved non-response status on the QIDS-C and HDRS scales, respectively, by reaching cluster C3 at 8 weeks (A2→B3→C3), if at least one of the core depressive items reduced by 2 or fewer points at 4 weeks. On the A3→B3 path, 61% (SD 1.3) and 68% (SD 1.7) ($p < 0.04$) of patients achieved non-response status on the QIDS-C and HDRS scales, respectively, by reaching cluster C3 at 8 weeks (A3→B3→C3), if at least two core depressive symptoms reduced by 2 or fewer points at 4 weeks.

The thresholds of change in core symptoms at 4 weeks needed to achieve response and remission at 8 weeks were also investigated. As illustrated in Fig. 7.5, across all transitions from clusters at baseline to the B1/B2 clusters at 4 weeks, the probabilities of response at 8 weeks were 62% – 79% (SD 3, $p < 0.01$) and 72% – 79% (SD 3.4, $p < 0.06$) if at least one core depressive item's score improved by ≥ 2 on the QIDS-C and HDRS scales, respectively. Starting from the A3 cluster (the highest depression severity at baseline), the probabilities of remission (which occurred for women only) at 8 weeks were 70% and 89% (SD 2.4, $p < 0.06$) if at least two core depressive items improved by ≥ 2 points on the QIDS-C and HDRS scales at 4 weeks, respectively.

7.4.7 Prognostic Effects of Core Depressive Symptoms without Patient Stratification

We contrasted the prognostic effects of core depressive symptoms with and without the use of our clustering algorithm, as shown in Fig. 7.3 (women starting from the baseline cluster A3). In Fig. 7.3, the precision of the mean values of individual symptom severity scores at 4 and at 8 weeks are represented by the 95% confidence intervals, whereas the variance in the distribution of depression severity scores is represented by box plots with whiskers. In the absence of symptom dynamic paths that are dependent on clusters inferred in this work, the confidence interval around the mean is narrow but the box plots of symptom severity score variation at 4 and 8 weeks in Fig. 7.3 are large. The large variation in scores illustrates the heterogeneity in responses, with the upper quartile of the box-plot revealing the very limited improvement in symptom severity. Furthermore, in the absence of patient clustering, the probability of any one of the 8-week outcomes, based on changes in the severity of core depressive items on either scale, was between 49% and 53%, and not statistically significant ($p > 0.7$). By contrast (and as shown in Fig. 7.3B), clustering patients using our algorithm minimized the previously observed heterogeneity in scores at 8 weeks from the patterns of response at 4 weeks, as indicated by the smaller variation in variation of symptom scores, as seen in the box plots for the paths that originated in A3 cluster (illustrated as an example).

7.5 Discussion

7.5.1 Clinical Treatment Implications

The initial evaluation of the clinical effects of antidepressants at 4 weeks following the initiation of treatment is crucial for treatment planning in patients with MDD who are managed in outpatient settings. In clinical trials, improvement in depressive symptoms with antidepressant treatment can be observed as early as within the first 2 weeks of treatment [142] and lack of early improvement during the first 2 weeks of antidepressant treatment may accurately predict eventual non-response defined at 6 weeks (NPV 89%), although the predictive value for stable response and remission for patients who achieve early improvement is much more limited (PPV 53%) [143], and not all studies have demonstrated a predictive effect of early improvement (within the first 2–3 weeks of antidepressant treatment) for eventual response [92, 144]. It is known that certain levels of improvement in total depressive symptoms at 4 weeks indicate the need to alter treatment [145–148], and the reevaluation of antidepressant effects at 4 weeks has been broadly advocated for routine practice [149–153]. Moreover, in a comprehensive review

of 4 meta-analyses of randomized trials and 10 individual open or naturalistic trials, the NPVs of early improvement at 3 – 4 weeks (73% – 81%) for predicting eventual response at 5 – 12 weeks were less subject to variation than the NPVs for early improvement at 2 weeks (35% – 92%) [154]. In the reviewed studies, early improvement was defined as a$\geq$ 20% reduction from baseline in total depression scale scores. Our results provide additional specificity by defining which depressive symptoms must improve at 4 weeks, by how much, and in which subgroup of patients, in order to individualize the predictions of eventual treatment outcomes. For instance, citalopram-treated patients who show improvement in total depression scores after 4 weeks, but do not achieve sufficient improvement in the core symptoms that predict response (or remission) at that time-point, could require a change in antidepressant treatment based on the high likelihood of non-response at 8 weeks. This work therefore provides a quantitative framework for early "triaging" of patients based on specific thresholds of a few symptoms.

7.5.2 On the Necessity of Patient Stratification for Homogeneity in Antidepressant Response

The predictive accuracies achieved in this work stemmed, in part, from the novel use of unsupervised learning and probabilistic graphs to stratify patients in an unbiased manner into unique clusters based on depression severity. The goal of this approach was to minimize the effects of heterogeneity in depressive symptom responses — a major barrier to the accurate prediction of categorical treatment outcomes in clinical antidepressant trials and real-world practice. We ran our analyses separately by stratifying patients to clusters, and without the cluster assignment. In the absence of patient stratification, the prognoses (probability) of sex-specific remission/response at 8 weeks is relatively weak ($49\% - 53\%$) when using core symptoms' baseline severity and associated changes at 4 weeks. However, when the clustering algorithm was applied, the same set of core symptoms' baseline severity and associated changes at 4 weeks accurately predicted sex-specific remission/response at 8 weeks with a minimum accuracy of 66% and going as high as 96% depending on the cluster patients began at baseline.

7.5.3 Sex-differences in Antidepressant Response

We modeled antidepressant effects separately in women and men even though individual sociodemographic factors, including sex, have not demonstrated strong and consistent predictive properties in previous machine learning studies of antidepressant effects. Of the patients

with the highest depression severity at baseline (cluster A3), only women achieved early (at 4 weeks) and eventual (at 8 weeks) remission in our study, under both rating scales. Further, women were more likely than men to achieve early response at 4 weeks on the way to eventual response at 8 weeks on the HDRS scale. These findings are consistent with previously published results that show an association between female sex and greater response to citalopram [25, 82, 132], which was one of our reasons for stratifying analyses by sex.

7.5.4 Clinical Research Implications

The criteria used to define "core" depressive symptoms in this work do not imply that these symptoms are "core" to the syndrome of MDD. However, the core symptoms identified in this report substantially overlap with symptoms derived from HDRS and QIDS-C that, according to prior work [138–141], are most strongly linked to the effects of treatment with antidepressant medications. The overlap of our core symptoms with existing subscales' symptoms further validates our machine learning-based approach, which establishes the prognostic capabilities of these symptoms. We do not suggest that the full versions of depression rating scales should be replaced with shorter versions based on core symptoms only, which would fail to consider all of the important elements of MDD severity for individual patients [155]. Suicidal ideation and sleep disturbances, for example, were not core depressive symptoms in this work, but are important clinical symptoms for many MDD patients [28, 59, 156–158]. However, the core symptoms identified here could inform future depression rating scale development, for instance, by assigning greater weight to the core symptoms relative to other symptoms in determining an eventual total score, or by providing finer gradations of the scale response items for the core symptoms. Our results suggest that focusing on early changes in core symptoms may increase the prognostic value of full-scale depression measures, which were designed to measure disease severity but not necessarily to predict outcomes. By applying our machine learning workflow to other large antidepressant datasets, one could potentially derive drug-specific core depressive symptoms, as well as a "superset" of core depressive symptoms whose early changes may be highly prognostic of eventual response across numerous antidepressants.

From the perspective of finding homogeneity in the responses of groups of symptoms, the integration of specific in-situ inferential tasks (clustering), longitudinal inferential tasks (probabilistic graphs), and predictive tasks (supervised learning) for predicting outcomes of antidepressant treatment is new. One might argue that homogeneity in depressive symptoms could be inferred without requiring patient stratification or symptom dynamic paths. However, we demonstrated that changes in the core symptoms' severity do not have meaningful

prognostic value in the absence of patient stratification and information on the paths the patients traversed during treatment. These insights into homogeneity based on a few prognostic depressive symptoms may translate to clinical practice and to future research efforts to develop genomic (or other) biomarkers of antidepressant mechanisms of action and response, wherein the antidepressants are used as "molecular probes."

7.5.5 On Clinical Interpretability with Probabilistic Graphs

Latent variable analyses with growth mixture models have been used previously to study trajectories of depression severity with antidepressant treatment [62, 63, 74–77, 136, 137]. We used probabilistic graphs in this work because growth mixture models (1) do not find paths algorithmically by conditioning upon improvements in symptoms at intermediate time-points, (2) offer very limited interpretability of dynamics of symptom changes, and (3) often need some domain expertise to reconcile results that might be perturbed by differences in model specifications [78, 79, 159, 160]. Growth mixture models and trajectory analyses are useful (and often computationally simpler) when aggregate statistics of longitudinal responses are sought, provided that the analysis is not exploratory (i.e., the numbers of paths and trajectories are already known), path parameters are stable, and the number of unique paths is small. However, the computational cost incurred in the construction of probabilistic graphs and their associated inference/optimization methods is worthwhile given the rich insights gained.

Limitations. There are limitations to our approach that must be considered. We were unable to investigate whether changes in core depressive symptoms at time-points earlier than 4 weeks can accurately predict clinical outcomes at 8 weeks—an important consideration, given evidence that eventual response may sometimes be predictable as early as 2 weeks [143]. None of the datasets used here included a placebo arm; however, the replication of depressive symptom clusters at all time-points in our datasets provides a basis for concluding that observed changes in clinicians' individual symptom ratings are more likely to reflect antidepressant effects than chance occurrences. There was no dose standardization across datasets, but this is less concerning given that drug dosage was not associated with clinical outcomes here or in previous studies [59, 161]. Despite replication across three independent datasets, to bring this methodology into practice, future studies are needed to establish the generalizability of our approach to other medications, other treatment approaches (including evidence-supported psychotherapies), and longer follow-up durations. Further, we were unable to address which treatments should be considered after failure to respond to citalopram/escitalopram, based on their comparative likelihoods of achieving response or remission. Finally, our analyses focused

on trial completers, so the impact of our findings on those who dropped out of treatment prior to 8 weeks is unknown.

7.6 Summary

We used statistical and machine learning methods to identify a subset of core depressive symptoms that are highly homogeneous in their longitudinal response to citalopram/escitalopram, and investigated their utility for predicting clinically relevant outcomes after 8 weeks of treatment. Algorithms that utilized core depressive symptom severity at baseline and associated changes at 4 weeks accurately predicted remission, response, and non-response at 8 weeks (AUC > 0.66; accuracy $> 63\%$). We established thresholds of change in the core depressive symptoms at 4 weeks that were highly prognostic of eventual non-response and remission with citalopram/escitalopram treatment at 8 weeks. We replicated these results across three independent clinical trials' datasets and two separate depression rating scales. The results presented in this chapter using novel application of machine learning methods represent significant advance because, (1) probabilistic graphical models allowed for a compact representation of longitudinal response to antidepressants (as opposed to 986 paths illustrated in Fig. 2.2), (2) we transformed diagnostic instruments (e.g., QIDS-C, HDRS) into prognostic instruments using changes in individual depression severity scores, and (3) we replicated predictive performance across multiple datasets and rating scales. Replications in symptom dynamics and predictions are particularly significant in the context of psychiatric research, considering that heterogeneity in disease manifestation and treatment response (as described in Chapter 2) has been main challenge in achieving replications of findings across independent studies.

CHAPTER 8

CONCLUSION

This dissertation introduced, *ALMOND*, the Analytics and Machine Learning Framework for Actionable Intelligence from Clinical and Omics Data. ALMOND has the capability to advance both basic and translational health science research, as illustrated in this dissertation with examples from many fields of medicine. ALMOND augments a clinician's disease knowledge and treatment selection logic with analytically, inferred patient-specific biological measures to individualize treatment selection. A clinician's decision to individualize treatment are made prior to treatment initiation, and again soon after treatment initiation to ensure that the chosen treatment option is as effective as desired. To that end, ALMOND demonstrated the capability to aid in treatment management prior to and after treatment initiation in treating major depressive disorder as a case study.

Major depressive disorder (MDD) is a globally prevalent disease, and individualized medicine practices to treat its symptoms do not exist. ALMOND's analytical workflow for individualization of antidepressant treatment systematically addresses the challenge of heterogeneity in major depressive disorder symptoms and antidepressant response, with the goal being to achieve a "right patient, right drug, right time" approach to treatment management. First, we identified the "right patients" by using unsupervised learning to stratify patients. Patient stratification served as a foundation for associating disease states with multiple pharmacological (drug-associated) measures. Second, "right drug" selection in the context of antidepressants was shown to be possible, as psychiatrists' depression severity assessments augmented with pharmacogenomic measures robustly predicted treatment outcomes, with replications across multiple independent trials. Finally, probabilistic graphs provided early and easily interpretable prognoses at the "right time" for psychiatrists by accounting for changes in routinely assessed depressive symptoms' severity.

Important outcomes of this work include (1) discovery of compelling sex differences in the dynamics of symptom changes due to antidepressant treatment — an aspect often overlooked in antidepressant treatment management; (2) recognition of the ability of pharmacogenomic biomarkers to predict antidepressant treatment outcomes, an important measure of success in individualizing medicine; and (3) development of a clinician-friendly interface with which

a psychiatrist can input depression severity measurement, and obtain easily interpretable prognoses of eventual treatment outcomes. Our usable interface will enable deployment of ALMOND in primary care settings, which is where the majority of early depressive episodes are treated. Hence, ALMOND's early poor prognoses will allow primary care physicians to immediately refer the patients with poor prognoses to a mental healthcare specialists.

The broader significance of ALMOND in potentially individualizing the treatment of MDD is twofold. First, helping clinicians choose antidepressants that have the highest likelihoods of enabling patients to achieve remission, or by helping them decide to change treatments based on poor early prognoses. ALMOND will help patients avoid ongoing suffering and the burden of possible side effects of suboptimal treatments. The sooner an antidepressant treatment can control depressive symptoms, the lower the likelihood that the patient will develop other chronic diseases. Second, treatment approaches for many other diseases that have similar heterogeneity in disease presentation such as rheumatoid arthritis or migraine headaches could benefit by using from the approach of ALMOND's workflow design to individualize treatment management.

APPENDIX A

ADDITIONAL MATERIALS

A.1 Additional Tables

Table A.1: Socio-demographic variables studied.

DATA	DESCRIPTION
Age at study enrollment	[Continuous, age in years]
Body mass index at enrollment	[Continuous, kg/m^2]
Smoking status	Current smoker
	Former smoker
	Non (never)-smoker
History of major depression in first-degree relative	
Parent	Yes/No
Sibling	Yes/No
Child	Yes/No
History of bipolar spectrum disorder in first-degree relative	
Parent	Yes/No
Sibling	Yes/No
Child	Yes/No
History of alcohol abuse in first-degree relative	
Parent	Yes/No
Sibling	Yes/No
Child	Yes/No
Pregnant (women only)	Yes/No/Did not answer
Seasonal pattern to depressive episode occurrence	Yes/No/Unknown

Table A.1: Socio-demographic variables studied (*Continued*).

DATA	DESCRIPTION
History of any other substance abuse in first-degree relative	
Parent	Yes/No
Sibling	Yes/No
Child	Yes/No
Transplantation or transfusion	Yes/No
History of liver or bone marrow transplant, or blood transfusion within 6 weeks of study enrollment	Yes/No
Marital status	Never married
	Cohabitating/life partner
	Married
	Separated
	Divorced
	Widowed
Education level (highest degree received)	No degree received
	High school diploma
	Passed the General Educational Development Test
	Some college
	Associate degree/Technical degree
	College diploma
	Master's degree
	Doctorate or professional degree (e.g., MD, PhD, JD)
Cohabitation	Spouse or partner lives in same home as patient
	Spouse or partner does not live in same home as patient
	Not applicable

Table A.1: Socio-demographic variables studied (*Continued*).

DATA	DESCRIPTION
Employment status	Unemployed, not looking for employment
	Unemployed, looking for employment
	Full-time employed
	Part-time employed
	Self-employed
	Retired, not working
Student status, current	Not a student
	Full-time student
	Part-time student
Years of education	[Continuous, total number of years of formal education]
Drug dosage	[Continuous, milligrams per day]
Plasma drug levels	[Continuous]

Table A.2: Metabolite abbreviations and pathways.

Metabolite	Metabolite Abbreviation	Pathway
(+)-alpha-Tocopherol	ATOCO	Antioxidants
(+)-delta-Tocopherol	DTOCO	Antioxidants
(+)-gamma-Tocopherol (redox state #1)	GTOCO1	Antioxidants
(+)-gamma-Tocopherol (redox state #2)	GTOCO2	Antioxidants
(+)-gamma-Tocopherol (redox state #3)	GTOCO3	Antioxidants
Cysteine	CYS	Cysteine/Methionine
Methionine	MET	Cysteine/Methionine
4-Hydroxybenzoic acid	4HBAC	Phenylalanine
4-Hydroxyphenyllactic acid	4HPLA	Phenylalanine
Salicylic Acid	SA	Phenylalanine
1,3-diMethylxanthine	THEOPHYLINE	Purine
1,7-diMethylxanthine	PARAXAN	Purine
Guanine	GUANINE	Purine
Guanosine	GUANOSINE	Purine
Hypoxanthine	HX	Purine
Uric acid	URIC	Purine
Xanthine	XAN	Purine
Xanthosine	XANTH	Purine
3-Hydroxykynurenine	3OHKY	Tryptophan
5-Hydroxyindoleacetic acid	5HIAA	Tryptophan
5-Hydroxytrptophan	5HTP	Tryptophan
Alpha-methyltryptophan	AMTRP	Tryptophan
Indole-3-acetic acid	I3AA	Tryptophan
Indole-3-propionic acid	I3PA	Tryptophan
Kynurenine	KYN	Tryptophan
Serotonin	5HT	Tryptophan
Tryptophan	TRP	Tryptophan
4-Hydroxyphenylacetic acid	4HPAC	Tyrosine
Homogentisic Acid	HGA	Tyrosine
Homovanillic Acid	HVA	Tyrosine
Methoxy-Hydroxyphenly Glycol	MHPG	Tyrosine
Tyrosine	TYR	Tyrosine
Vanillylmandelic Acid	VMA	Tyrosine

APPENDIX B

MIMOSA: MIXTURE-MODEL BASED SINGLE-CELL ANALYSIS

B.1 Introduction

Population studies have shown that an anti-diabetic drug called metformin inhibits cancer growth in various types of cancer, including triple-negative breast cancer [162, 163]. Triple-negative breast cancer (TNBC) is a molecular subtype of breast cancer that does not have any standard targeted therapies [164, 165]. Pharmacogenomics research focuses on understanding the interplay between drug effects and functions of the genome (i.e., human DNA). Using the example of metformin response in TNBC, this work shows that when biomarkers of drug response are not known *a priori*, it is possible for unsupervised learning methods to augment pharmacogenomics experts' knowledge by identifying a few genes, out of the entire human genome (i.e., 23, 398 genes), as candidates for laboratory experiments to establish novel biological mechanisms of a drug.

The purpose of the overall project driving this paper is to use machine learning methods to help infer the molecular mechanism by which metformin inhibits cancer growth in TNBC. The workflow of our analysis is illustrated in Fig. B.1. In order to identify metformin's impact on the TNBC cells, we used two identical MDA-MB-231 TNBC cell populations, including 192 cells (two assays, each comprising 96 cells) not treated with metformin (referred to as *baseline cells*), and an equal number of the same cells treated with metformin (referred to as *metformin-treated cells*). We sequenced the cells by using single-cell RNA sequencing (scRNA-seq) technology, and the resulting data comprise the expression measure for each gene of the sequenced genome contained in each of the cells under study [9]. The data reflect 23, 398 genes and their associated gene expressions for baseline and metformin-treated cells. Thus, the overall data consist of 9M[1] gene expression values. The goal of the analytics in this work was to infer clusters of metformin-treated cells by using unsupervised learning and then identify a small group of differentially expressed genes across clusters. Those genes can then be used to identify associated diseases and pathways (where a pathway is a "series of

[1] [192*2 cells]*23, 398 $\approx$ 9M.

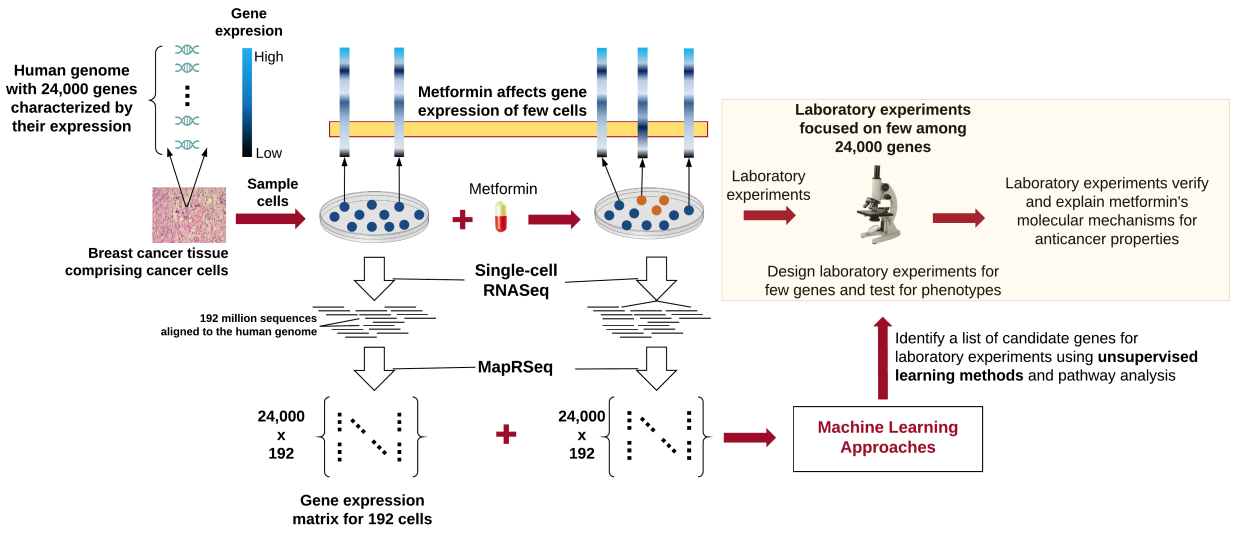


Figure B.1: The single-cell RNAseq analysis workflow. It starts with sequencing of cancer cells and goes to the use of unsupervised learning methods, combined with pathway analysis, to generate a list of a few genes. Those genes become candidates for informing the design of focused laboratory experiments for learning novel biology of drug action.

actions among molecules in a cell that leads to a certain product or a change in the cell") [2] by performing pathway analysis. By combining differentially expressed genes that overlap with relevant pathways (as found by our pathway analysis) and available data on these genes from the existing literature in the context of metformin and anticancer mechanisms, we choose genes that have been implicated as having anticancer functions as candidates for laboratory experiments.

In this work, we used mixture-model based clustering, hierarchical clustering, and k-means clustering as unsupervised learning methods. Based on the observations of multiple normal distributions in gene expressions of single cells (see Sec. B.4 and Fig. B.2), we first used a mixture-model based unsupervised learning approach embodied in a tool we created called MiMoSA, for "mixture-model-based single-cell analyses" (described in Sec. B.5.1) [10]. MiMoSA found 310 of the $23,398$ genes to be significantly differentially expressed in six metformin-treated cells. As a validating step for MiMoSA's findings, hierarchical clustering approaches (divisive and agglomerative clustering) also identified the same set of 310 differentially expressed genes identified by MiMoSA (see Sec. B.5.2). However, the unsupervised and lightly supervised approaches using k-means clustering for a range of $k = [2 : 9]$ were unable to find a set of clusters that could either identify the same set of 310 differentially expressed genes found by MiMoSA and hierarchical clustering, or identify a different set of

[2] `https://www.genome.gov/27530687/biological-pathways-fact-sheet/`

differentially expressed genes (see Sec. B.5.4).

That set of 310 genes is small enough to be handled by well-understood bioinformatics approaches, such as pathway analysis. As a substantiation of our learning approach, pathway analysis of the downregulation of these 210 genes showed strong correlations with three pathways: (i) oxidative phosphorylation (p-value $3.81E-21$), (ii) the citric acid (TCA) cycle, and the respiratory electron transport (p-value $2.10E-19$), and (iii) mitochondrial translation (p-value $1.41E-07$). All of those pathways were recently found to have anticancer properties, via both in-vivo and in-vitro experiments [166–169]. Further, among the differentially expressed genes that overlap with those pathways, we have identified the NDUFB9, COX5B, MRPS7, and CDC42, which have been implicated in other anticancer mechanisms for other cancers not driven by metformin; these genes are now candidates for laboratory experiments. In Sec. B.6, we present a summary of laboratory experiments on CDC42's downregulation by metformin that explain the inhibition of cell migration and cell proliferation in triple-negative breast cancer [11]. Results from the laboratory experiments demonstrate the power of unsupervised learning that can not only identify candidate genes for laboratory experiments, but also identify genes that could lead to the establishment of novel mechanisms of drug action.

Traditional bulk sequencing enabled the study of aggregate gene expressions in a tumor sample. However, with scRNA-seq, the amount of data is significantly larger, and we have gained finer differentiation of cells by using distributions of gene expression in the cells, as opposed to the single aggregate value of gene expression provided by bulk sequencing. For example, scRNA-seq generates about 1 million RNA sequences per cell comprising roughly $24,000$ genes. When two sequencing panels are analyzed, where each panel consists of 96 cells, 192M sequences are generated (see Fig. B.1). Several prior efforts (discussed in Sec. B.3) have analyzed single-cell data, but our work is unique in that it demonstrates the ability of data-driven unsupervised learning analytics to help establish novel biological mechanisms.

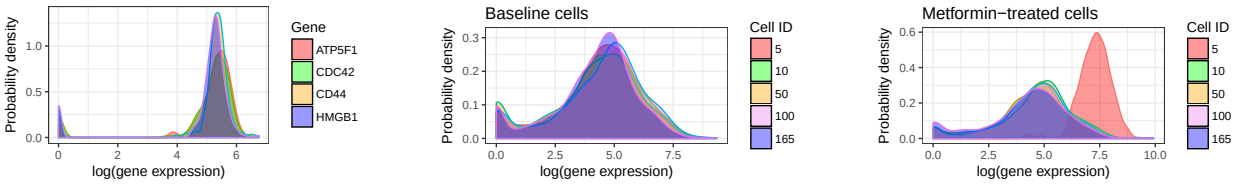
B.2 Contribution

Key additional contributions of this work are as follows.

1. **We demonstrate the feasibility of using learning methods to inform novel biology:** This work demonstrates the complete workflow of analyses that proceed from data generation, to machine learning analyses, to laboratory experiments, and finally to identifying novel mechanisms of drug action in triple negative breast cancer (see Sec. B.6). This work represents a significant advancement considering that the

molecular mechanisms of metformin's response in TNBC are not yet known.

2. **Test dataset and tool access:** We provide access to a test dataset and MiMoSA, which is compatible with multiple operating systems and computation architectures.



(a) Gene (b) Baseline cells (c) Metformin-treated cells

Figure B.2: The existence of mixtures in the feature space and samples is illustrated by the probability density functions (PDF) of gene expressions in a set of genes across (a) all baseline cells, (b) a set of baseline cells and (c) a set of metformin-treated cells.

B.3 Related Work

The recently proposed methods for analyzing single-cell data have largely focused on finding subpopulations of cells in a population of cells [170, 171]. All of the proposed methods include two steps of processing, first reducing the number of genes being used to cluster cells, and then using a clustering method to find subpopulations of cells. Further, all these methods have found that only a few thousand genes are significantly differentially expressed in cell samples [170–172]. For the second step of these analyses, supervised, unsupervised, or graphical model approaches [172] are used.

The first step of the analysis tries to retain the genes that show variation in their expression levels across the samples. For example, in the transcriptome analysis of lung adenocarcinoma [170], the method to reduce the dimensionality of the data (for genes in this context) was to start by looking at genes (expressed across all the samples) whose gene expressions were measured as greater than 0, thereby reducing the gene list to about $9,000$ genes. Then, the authors of [170] studied the correlation of gene expression among these genes by using Pearson's correlation analysis, and reduced the gene list to about $5,500$ genes by choosing genes with correlation coefficients greater than 0.9. In an analysis of cell-to-cell heterogeneity that revealed subpopulations [171], prior knowledge of cell-cycle genes was used, and only genes that showed significant correlation were chosen, bringing the gene list down from $23,398$ to $2,881$.

To reduce the number of genes needed to infer cell types (assuming that the data are normally distributed), shared-nearest neighbors (SNN) or k-means clustering is used [172]. In particular, when SNN is used on simulated and real cell data, it has been shown that it performs better than k-means clustering when no biological priors are used [172]. However, there remains an open problem on how to choose the optimal number of neighbors and k values; currently, we must perform an extensive search of values or use heuristics to estimate the best values. One particular approach that is different from the normal two-step process is the use of diffusion maps [173], a supervised approach. The authors of [173] assume the existence of known types of cells and then use a transition matrix for classifying cells based on the state they best match their signature to. To do so, the authors needed to define the Gaussian kernel and further approximate the transition probabilities, and those steps are hinged on the assumption that the cell types are known.

A significant number of single-cell analyses have used hierarchical clustering [174, 175] to infer cell heterogeneity and then compare if the inferred clusters matched with known cell-types inferred using human observation (e.g., pathology) [176–178]. While making no implicit assumptions on the data's distribution, these works have identified several novel mechanisms in the context of circulating tumor cells, preimplantation embryos and embryonic stem cells and phospho-protein networks in cancer cells among other studies.

All the aforementioned methods for single-cell analysis are driven either by implicit assumptions in normality, known correlations between genes and biological mechanisms, or by supervised methods that use cell signatures. However, in problems where either there are many mechanisms related to drug response or we do not know the mechanism by which the drug impacts the cells, we have to turn to data-driven methods that first study characteristics of the data and then choose a method/algorithm to apply on the data. Further interpretation of the chosen method's results requires interactions with domain experts to address the primary goal of the analysis. To the best of our knowledge, no methods exist that can use mixture-model distributions to infer clusters of cells, despite the observation that gene expression of cells is best described using mixture models.

Through (1) prior methodological innovations presented in describing MiMoSA [10] to cluster cells whose gene expressions are characterized as a mixture of Gaussians, (2) novel biological mechanisms of metformin in TNBC established using candidate genes identified by MiMoSA [11], and (3) additional methodological investigations presented in this work to identify novel candidates for laboratory experiments, multiple research gaps are addressed in our work that have been overlooked by previous analysis methods.

1. Our case study demonstrates a consistent method to go from data generation from drug intervention, to identifying major candidates for focused laboratory experimentation to

establish a drug's molecular mechanisms.

2. Our work does not make prior assumptions on gene correlations with drug response to reduce the number of genes for analysis.

3. Our choice of method was driven by observations made in our preliminary analysis, which revealed that distributions of gene expression in cells were best described by mixtures of Gaussians; this observation was aided by the fine resolution of data provided by scRNA-seq.

B.4 Data and Data Characteristics

B.4.1 Data

The MDA-MB-231 breast cancer cell line (ATCC HTB-26) was cultured in Leibovitz's L-15 medium with 10% fetal bovine serum for 5 days with and without metformin. Duplicate cultures were processed for single-cell analysis. Single cells with and without metformin were captured on a large-sized ($17 - 25\,\mu m$ cell diameter) microfluidic mRNA-seq chip known as the C1 Single-Cell Auto Prep IFC, using the C^{TM} Single-Cell Auto Prep System (Fluidigm Corporation, South San Francisco, CA). Cells were loaded onto the chips at a concentration of 300 cells/ μl, stained for viability with a LIVE/DEAD cell viability assay kit (Life Technologies, Carlsbad, CA), and imaged by phase contrast and fluorescence microscopy to assess the number and viability of cells per capture site. Only single, live cells were included in the analysis. cDNAs were prepared on-chip using the SMARTer Ultra Low RNA kit for Illumina (Clontech Laboratories, Mountain View, CA). Single-cell cDNA size distribution and concentration were measured with a Quant-iT Pico green dsDNA assay kit (Life Technologies). Illumina libraries were constructed in 96-well plates using the Illumina Nextera XT DNA Sample Preparation kit using the protocol supplied by Fluidigm. Libraries were quantified by Agilent BioAnalyzer, using a high-sensitivity DNA analysis kit. Single-cell Nextera XT (Illumina) libraries of one experiment were pooled and sequenced at 100 bp paired-end on Illumina Hiseq to a depth of about 1 million reads. Single-cell mRNA-Seq data were processed using MAP-RSeq pipeline [179].

B.4.2 Data Characteristics and Pre-processing

The data characteristics are as follows.

1. The data comprises 192 baseline cells, and an equal number of cells treated with metformin. In each cell, $23,398$ genes were sequenced, and MAP-RSeq [179] was used obtain gene expression from sequencing data.

2. The expression value for each gene was normalized by accounting for the sequencing depth (number of short DNA sequence strings from the sequencing platform aligned to a gene) and length of the gene. The measure of gene expression is Reads Per Kilo Million (RPKM). The range of the RPKM is between 0 and $2,000$.

3. Only 20% of the $23,398$ genes show expression levels greater than 32 on the RPKM scale, which is a heuristic that can be used to decide whether a gene is expressed or not as recommended by MAP-Rseq [179].

4. Roughly 10% of the baseline and metformin-treated cells had low sequencing coverage ($<$ 1M reads per cell), so we excluded those cells from our analysis.

5. The density functions of gene expression of a gene across all cells, and for all genes in a cell comprise mixtures as shown in Figures B.2(a) and B.2(b) respectively. We used the mclust package to fit a distribution for each gene and the best fit was a Gaussian mixture model, where each component of the mixture was a Gaussian. Using the fitted distribution, we performed non-parametric tests (Kolmogorov-Smirnov and Wilcoxon-rank tests) against the distribution derived from the data. The null hypothesis in this test is that the two distributions have equal means. The p-value in these tests were greater than the significance level of 0.05, thereby failing to reject the null hypothesis, meaning that the model fit was statistically close to the actual data's distribution.

6. In some metformin-treated cells, at least one component of the mixture had phase-shifted significantly. Therefore, metformin was affecting these cells in a way differently from other cells, thereby making the drug's effect on the cells non-uniform as shown in Fig. B.2(c).

We observed that 80% of the genes were considered inactive in the data and as our focus in this analysis was on the impact of metformin measured by changes in gene expression, only genes in the top 5% of variance across cells were considered. The dataset for available for analyses reduced the number of genes (features) for our analysis from $23,398$ to $1,170$. We used the reduced set of genes and the samples as inputs to the unsupervised learning methods. This way of reducing feature space is common in bioinformatics practices, although there is no standard threshold for the amount of variance to consider in gene expression profiles. The general assumption is that only a few biological pathways, comprising $100 - 400$ genes, are

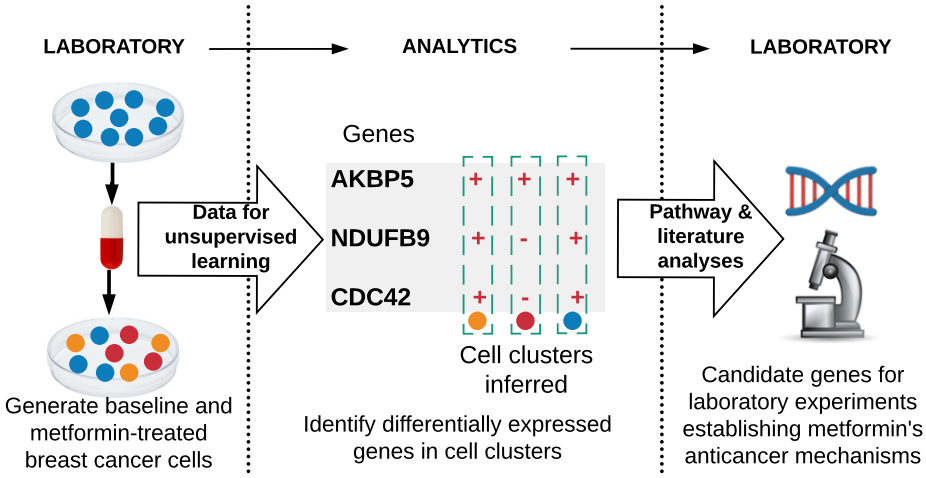


Figure B.3: The unsupervised learning view of single-cell analysis is explained as follows. Cells exposed to a drug may exhibit differences in their gene expression behavior due to the molecular interactions with the drug, and these differences are not known. The computational problem is to cluster these cells (samples), and extract the genes (features) that are behaving differently compared to other clusters (referred to as differentially expressed genes). These differentially expressed genes are analyzed to study their biological significance in all known disease and molecular pathways.

affected by a treatment, and thus these genes would highly likely be present within the top 5% most variable genes.

B.5 Methods and Results

We describe the approaches used to analyze single cells with an unsupervised learning approach as shown in Fig. B.3, as cell-types induced by metformin are not yet known.

Analytical Approach and Visualization

Our overarching analytical approach was to begin with a method that best suited the data characteristics and the method's assumptions. Hence, we first began with mixture-model based unsupervised learning embodied in MiMoSA, given the observation that distribution of gene expressions in a single-cell population is a mixture of multiple distributions. Then, assuming that cell distributions are characterized by a few differentially expressed genes, we used hierarchical clustering approaches to verify the clusters. We chose that approach because the clusters are formed based on similarity in Euclidean distances; hence, if the expression

of a few genes is sufficient to cluster cells into groups, then hierarchical clustering could, in theory, capture cell clusters that are highly concordant with those found by MiMoSA. Finally, for a range of k, we chose k-means clustering to assess whether the inferred clusters captured the same differentially expressed genes identified by MiMoSA.

Since scRNA-seq allows one to study the expression of the genome in each cell, it would also be of interest to visualize the subpopulations (i.e., inter-cluster separations) inferred using single-cell analyses. Earlier efforts have used either linear methods, such as principal component analysis (PCA) [180], or nonlinear methods such as t-distributed stochastic neighbor embedding (t-SNE) [181] to reduce the dimensionality of the data to two or three dimensions, and then projected the clusters onto these lower dimensions [172, 182]. Our prior work demonstrating the clustering capabilities of MiMoSA showed that projection of the cluster labels onto first two principal components computed using PCA helped visualize cluster separations better than t-SNE did [10]. Hence, we demonstrate the inter-cluster separation using only PCA in this work.

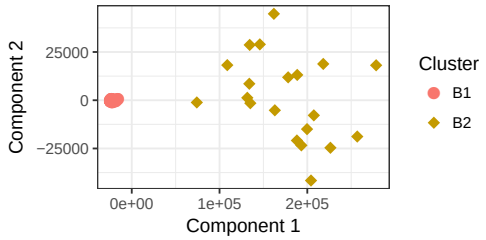
B.5.1 Inferring Cell Subpopulations

MiMoSA

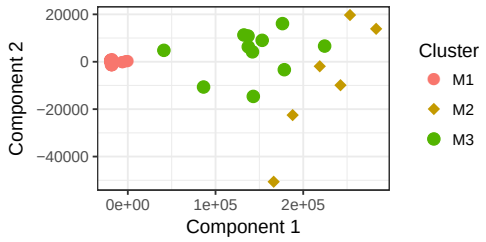
Based on our knowledge of the presence of multiple distributions in our data, as illustrated in Fig. B.2, we chose to use probability distribution models to cluster the cells. Probability models cluster data according to a model (distribution) that best defines the data (such as Gaussian mixture models, in this work), and treats each sample (cell) as being independent from other samples. The assumption of independence is acceptable, since each cell can behave differently and independently of other cells in response to metformin. We define the likelihood $\mathcal{L}_i$ of the mixture model for each cell y_i of N cells as shown in Equation B.1, where there are K mixtures/components, f_k is the distribution of the component k, and θ_k is the distribution's parameters.

$$\mathcal{L}_i(\theta_1, ..., \theta_N; \tau_1, ...\tau_K|\mathbf{y}) = \prod_{i=1}^{N} \sum_{k=1}^{k=K} \tau_k . f_k(y_i, \theta_k) \tag{B.1}$$

We chose to fit the data with Gaussian mixture models (GMM) with varying volumes and finite mixtures, as this provided the best likelihood score for fitting the data. The multivariate Gaussian distribution function is as defined in Equation B.2, where $(\mu_k, \sum_k)$ is the mean and covariance of the component k. Because our model has mixtures of Gaussians, we need to compute the model parameters using the maximum likelihood (ML). For model estimation,



(a) Baseline cell clusters projected using PCA



(b) Metformin-treated cell clusters projected using PCA

Figure B.4: In (a) and (b), we project respectively, the baseline and metformin-treated cell clusters found by MiMoSA onto the first two principal components derived from PCA.

there are two popular algorithms to choose from: the expectation maximization algorithm (EM), and the variational Bayesian EM algorithm (VBEM). Both algorithms are iterative and are known to have similar time complexities; VBEM can be used to perform automatic model selection and is less prone to over-fitting than is EM. However, implementations of VBEM required binning of the gene expression measures, and with small sample space and large variation in the range of gene expression, the binning proved to be a challenge. Hence, we used the EM algorithm for ML to learn the model parameters. The EM algorithm is a two-step process. First, the E-step computes the conditional expectation of the observable data and current parameter estimate. Then, the M-step maximizes the log-likelihood of the parameter estimates learned in the E-step.

$$\phi_k(y_i|\mu_k, \sum_k) = \frac{0.5exp\{(y_i - \mu_k)^T \sum_k^{-1}(y_i - \mu_k)\}}{\sqrt{det(2\pi \sum_k)}} \tag{B.2}$$

Once we have learned the model parameters, we then need to decide on an optimal number of clusters. One method that has historically provided a consistent estimator of the number of clusters is Bayesian information criteria (BIC) [183], where the value of K at which the BIC value asymptotically converges is chosen. We implemented the model-based clustering from the "mclust" package in R [184]. MiMoSA identified two clusters ($B1$, $B2$) in baseline cells,

Table B.1: Subpopulations inferred in metformin-treated cells.

Method	Cluster 1	Cluster 2	Cluster 3
MiMoSA	160	6	12
Agglomerative	163	5	10
Divisive	162	5	11

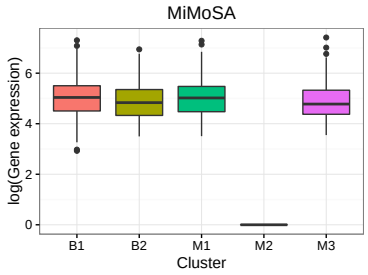
and three clusters ($M1$, $M2$, $M3$) in metformin-treated cells. The clustering visualizations obtained using PCA for baseline and metformin-treated cells are shown in Figs. B.4(a)-B.4(b). We observe that metformin treatment induced very little variability in the gene expression across cells. Hence, the majority of the cells are tightly clustered together in $M1$.

B.5.2 Cluster Validation: Hierarchical Clustering

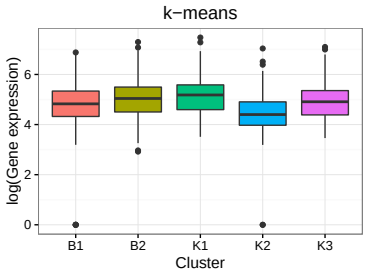
We validated the clusters inferred by MiMoSA by using hierarchical clustering methods. We performed both agglomerative and divisive hierarchical clustering on metformin-treated cells. Hierarchical clustering was chosen because it is based on comparison of pairwise similarity of features. Principal component analysis of metformin-treated cells showed that 97% of the variability was captured within the first two components. Thus, it is likely that a few genes are probably significantly altered by metformin, while the rest of the genes show little change in their expression. Hence, pairwise comparison of the cells would tend to cluster cells with similar changes in expression of the genome.

Hierarchical clustering treated each of the N cells as a data point described by a set of M feature coordinates, where each individual feature coordinate is the expression of a gene. We computed the relative measure of proximity between the cells that encompassed all of the M gene expression levels. We performed agglomerative clustering through successive merging of N total clusters, based on proximity, into a single, global cluster. We performed divisive clustering in the inverse direction, which is to begin with a single cluster (comprising all N cells) and successively splitting it into N remaining clusters. We performed agglomerative and divisive hierarchical clustering on the metformin-treated cells using the proximity matrix and complete linkage; the details have been previously published [10]. We began pruning the clustering hierarchy from both approaches by starting from the root until we arrived at three sub-trees (clusters). We note the similarity among the number of cells present the three main clusters, shown in Table B.1. Further, we find that all the cells in cluster 2 of both hierarchical methods overlapped with $M2$ from MiMoSA, with $M2$ is having one additional cell. Thereby, we found that clusters inferred by MiMoSA were validated by at

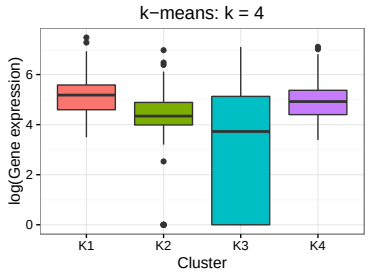
least one other clustering approach with a different mathematical formalism such as one of those used in hierarchical clustering methods. Given the evidence of cluster replication in these independent methods, we proceeded to analyze the clusters next to identify a set of genes that are differentially expressed across these clusters.



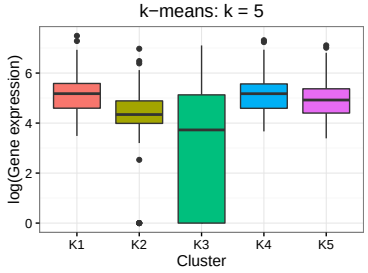
(a) MiMoSA



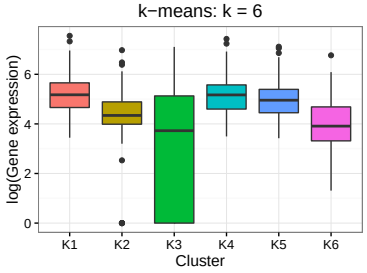
(b) k-means: $k = 2$ in baseline cells, $k = 3$ in metformin-treated cells



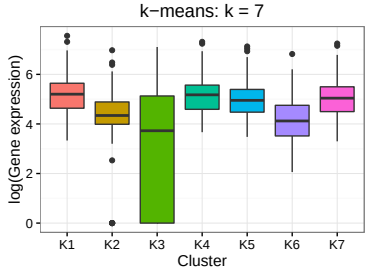
(c) k-means: $k = 4$ in metformin-treated cells



(d) k-means: $k = 5$ in metformin-treated cells



(e) k-means: $k = 6$ in metformin-treated cells

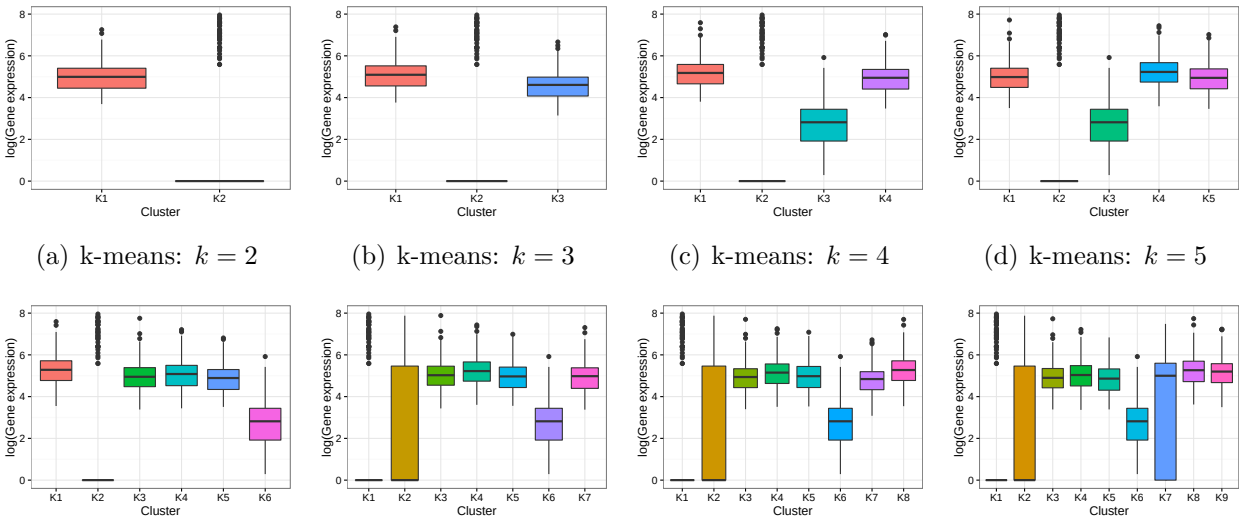


(f) k-means: $k = 7$ in metformin-treated cells

Figure B.5: The average gene expression of downregulated genes identified by MiMoSA and k-means clustering using various values of k are illustrated. Specifically, (a) shows the average gene expression levels in clusters found by MiMoSA, while (b) shows the same number of clusters identified by k-means clustering in baseline and metformin-treated cells. For a wide-range of k, it can be seen in Figs. (b)–(f) that k-means clustering was unable to establish the same cluster of cells that helped identify the significantly downregulated genes.

B.5.3 Cluster Analysis

The clusters inferred by MiMoSA in metformin-treated cells are characterized by 310 differentially expressed genes of which about 200 are downregulated and about 100 are upregulated in cluster $M2$, compared to $M1$ and $M3$. Clusters $M1$ and $M3$ showed little variation in gene expression, and the cells comprising cluster $M2$ were the most affected by metformin that saw a striking downregulation of over 200 genes, as shown in Fig. B.5(a). We use



(a) k-means: $k = 2$ (b) k-means: $k = 3$ (c) k-means: $k = 4$ (d) k-means: $k = 5$

(e) k-means: $k = 6$ (f) k-means: $k = 7$ (g) k-means: $k = 8$ (h) k-means: $k = 9$

Figure B.6: These box plots of the distribution of average gene expression of respective clusters illustrate the inability of k-means clustering to capture the cells of $M2$ in one cluster even if the features comprise only the downregulated genes identified by MiMoSA. In Figs. (a) – (e), although one could observe downregulation of the genes, they are statistically not significant because cluster $K2$ comprised only one cell, and in Figs. (f)–(h), clusters $K1$ and $K2$ comprised only one cell.

these differentially expressed genes in studying their biological relevance, which is discussed in Sec. B.6.

B.5.4 Demonstrating the Unsuitability of k-Means Clustering

We set out to demonstrate that the data-driven clustering approach is better suited for observing mixtures in the distribution of gene expressions. We first normalized (centering followed by scaling) the baseline and the metformin-treated cells, which still keeps the original variability in the data. Normalizing the data rendered the data to be normally distributed with zero mean and unit variance. We looked to see if we could identify the downregulated genes in metformin-treated cell clusters obtained using the k-means clustering algorithm. The k-means clustering algorithm is an iterative algorithm that assigns a data point to a cluster that minimizes the distance from the point to the cluster's mean [185]. After performing the clustering, we mapped the cluster labels with the cells and their associated expression levels of before normalizing. Figure B.5(a) shows the baseline and metformin-treated cell clusters with the expression of genes that are downregulated in $M2$, but upregulated in all other clusters; the downregulation is not visually observable, but is statistically significant (p-value<0.05). However, Fig. B.5(b) shows that for the same number of baseline and metformin-treated cell clusters, the clusters comprising the cells were different enough that we do not observe any significant variations in the average gene expressions of the downregulated genes identified by MiMoSA. We then increased the number of clusters (k) in metformin-treated cells from 3 to 7, and in all of these cases, we did not observe any clusters that could capture the same 6 cells of $M2$ found by MiMoSA. Therefore the downregulation was not observed in any of the clusters across the different values of k, as can be seen in Figs. B.5(c)–B.5(f).

k-Means Clustering of Cells Using Only Downregulated Genes

Instead of clustering cells using genes among those with the highest variance in their expression, we attempted to cluster the metformin-treated cells using only the downregulated genes identified by MiMoSA. When we started with $k = 2$ as the initial value, we observed that cluster $K2$ was created with only one cell in it with some evidence of downregulated genes as observed in $M2$. Since the cluster was made up of only one cell, making any further analysis statistically insignificant. To find out whether the six cells of $M2$ would be captured together, we increased k from 2 to 9. As illustrated in Fig. B.6, none of the clusters captured the behavior observed in $M2$ that was identified by MiMoSA, while the same singe-cell (in $K2$ in $k = 2$) continued to be clustered by itself in $K2$ ($k = 2 : 6$) and in $K1$ ($k = 7 : 9$).

A Semi-Supervised Approach to k-Means Clustering

Unlike unsupervised learning where no labeled data is used to infer clusters, semi-supervised learning uses a small fraction of the overall data to guide the clustering behavior [186]. We chose three cells (50% of $M2$) from each of the metformin-treated cell clusters identified by MiMoSA. Using a semi-supervised k-means clustering approach proposed by Jain [185], we obtained three clusters of cells. The semi-supervised approach also failed to cluster the six cells of $M2$ together, which meant that we could still not observe the drastic downregulation observed using MiMoSA.

All these results show that if single-cell subpopulations are identified based on subtle variations in their gene expressions, a data-driven model-based unsupervised learning methods could be better suited than k-means clustering algorithm.

B.6 Unsupervised Learning Informing Biology

Pathway analysis was performed using the differentially expressed genes to further understand the biological relevance of the differentially expressed genes. The top pathways (and the associated p-values) were oxidative phosphorylation ($3.8E-21$), the citric acid (TCA) cycle, and respiratory electron transport ($2.1E{-}19$) and mitochondrial translation ($1.4E{-}07$). These pathways are relevant in the context of metformin in several ways. (1) The possibility of chemoprevention with metformin is being investigated by targeting the oxidaptive phosphorylation pathway [167], (2) it has been shown that metformin inhibits cancer cell proliferation by regulating the TCA cycle [166], and (3) metformin has been shown to target mitochondrial metabolism in cancer therapies [168, 169].

It is clear that the differentially expressed genes inferred from MiMoSA's clusters are on pathways known to have anticancer effects driven by metformin. Among these differentially expressed genes in the above listed pathways, we have identified the following genes which are implicated in anticancer mechanisms. (1) NDUFB9: an accessory subunit of the mitochondrial membrane respiratory chain NADH dehydrogenase (complex I), and loss of NDUFB9 promotes MDA-MB-231 cells proliferation, migration, and invasion; because of elevated levels of reactive oxygen species (ROS) [187], (2) COX5B is a peripheral nuclear-encoded sub-unit of CcO (cytochrome c oxidase), and loss of COZ5B induces mitochondrial dysfunction and subsequently leads to suppression of cell growth and cell senescence [188], (3) MRPS7 is a mitochondrial ribosomal protein, involved in mitochondrial translation, that is significantly elevated in human breast cancer cells, leading to amplified mitochondrial biogenesis and/or mitochondrial translation in epithelial breast cancer cells [189]. Therefore, mitochondrial

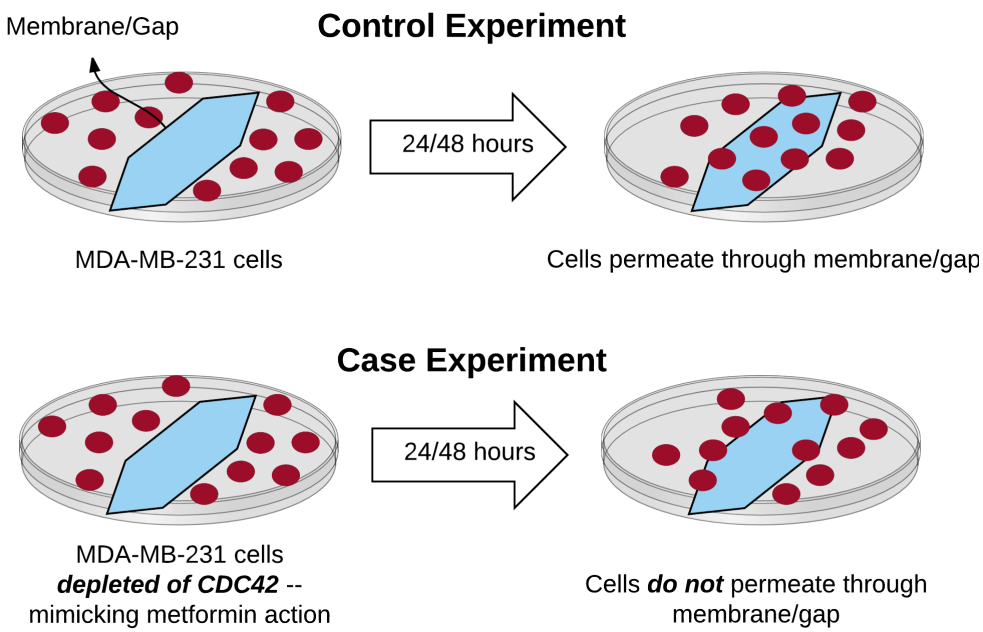


Figure B.7: Illustration of the laboratory experiments performed to establish CDC42's role in meformin's anticancer mechanisms in triple-negative breast cancer.

biogenesis could be a potential target for anticancer agents and therefore could explain the retrospective success of metformin, which prevents the onset of nearly all types of cancer in diabetic patients, likely because it functions as a "weak" mitochondrial poison, and (4) CDC42: known to play a role in cell-migration and cell-proliferation.

NDUFB9, COX5B and MRPS7 are known to play role in anti-cancer mechanisms in breast cancer. CDC42 has shown to be downregulated in breast cancer patients treated with metformin and is also downregulated in our study.

B.6.1 Cell Migration and Cell Proliferation Experimental Study

Existence of prior knowledge of CDC42's downregulation in triple-negative breast cancer patients treated with metformin led us to perform an elaborate set of laboratory experiments to study whether downregulation of CDC42 by metformin could demonstrate any anticancer properties in triple-negative breast cancer [11]. We next give an overview of the experiments as shown in Fig. B.7 and the summarize the findings that are detailed in [11].

Control experiment: Baseline MDA-MB-231 cells were cultured in a transwell dish and separated by a membrane in one setting (for cell proliferation), and by a gap made by scratching in another setting (for cell migration). If we let the dish sit for 24 hours (for cell

proliferation) or 48 hours (for cell migration) hours, the cells invaded the membrane or the gap.

Case experiment: Another set of MDA-MB-231 cells were cultured in another transwell dish, but this time the CDC42 was depleted in the cells (mimicking downregulation by metformin). After we let the dish sit for 24 hours (for cell proliferation) or 48 hours (for cell migration) hours, it was observed that only a few cells whose CDC42 was depleted were able to permeate through the membrane or the gap.

These findings established that downregulation of CDC42 induced by metformin inhibited cell migration and cell proliferation. Therefore, at least one new mechanism of metformin's anticancer property has been is established, via the use of mixture-model based unsupervised learning's ability to identify candidate genes.

Future work based on the current findings will include the following: (1) We will investigate what makes the six metformin-treated cells "diagnostic" in terms of inferring metformin's response, as these cells seem to be more sensitive to metformin than the other cells, and (2) We will conduct laboratory experiments for the remaining candidate genes identified in this work based on their differential expression after metformin treatment and their biological relevance as shown by pathway analysis.

B.7 Summary

Using metformin and TNBC as a case study, this work demonstrates an end-to-end workflow whereby learning methods can augment the drug and disease knowledge of pharmacogenomics experts by identifying biomarkers of novel drug actions. Considering that TNBC currently has no targeted treatments, this work represents an important step toward the design of targeted therapies for treating this aggressive cancer in women. Identification of a few novel and biologically significant candidates for laboratory experiments in the absence of *a priori* knowledge is important given the large size of the human genome and limitations in costs of laboratory experiments. The broader impact of this work in identifying a small set of differentially expressed genes after drug treatment lies in its potential to augment the drug and disease knowledge of pharmacogenomics expert, to support laboratory investigations, and thereby help establish novel biological mechanisms associated with drug response in diseases beyond triple-negative breast cancer.

APPENDIX C

GITA: GAME-THEORETIC TRANSCRIPTOME ANALYSIS

C.1 Introduction

It is known that gene expression is modifiable by both the environmental exposures and the genetic background of a cell [190]. It is also known that oncogene activation (increased gene expression level) is a driver of cancer progression and that tumor suppressor gene inactivation (decreased gene expression level) can also drive cancer progression, as shown in Fig. C.1 (in which the solid arrows indicate shifts in modes of distribution) [191]. Furthermore, in the two largest publicly available transcriptome datasets for adenocarcinoma [192, 193], in more than 50% of the matched samples (cancerous and histologically non-cancerous tissue from the same lung), (1) the sum of normalized gene expressions of tumor-suppressor genes was greater than that of the same number of oncogenes for histologically labeled healthy samples, and (2) the sum of normalized gene expressions of tumor suppressor genes was less than the sum of normalized gene expressions of oncogenes in tumor samples. The exact causes of activation or inactivation of genes are multifactorial, but tumorogenesis is correlated to increased oncogene expression and tumor suppressor gene suppression [194].

Conversely, if sustained tumor suppressor gene expression overwhelms that of the oncogenes, the lung will continue to stay healthy (non-cancerous), but not without a risk of developing tumors in the future. Therefore, we expect that in reaction to environmental insults such as tobacco smoke intake in the lungs, there is a competitive relationship between the oncogenes and the tumor suppressor genes that dictates whether the lungs will remain healthy, be at risk of cancer proliferation, or have cancerous tumors. We use that competitive relationship between oncogenes and tumor suppressor genes to develop a novel game-theoretic model to predict the proliferation of adenocarcinoma. Key contributions in our game-theoretic model are, (1) **a data-driven payoff function** incorporating the expression of several tumor-suppressor and oncogenes, as no one gene's expression is sufficient to describe cancer proliferation in every individual, and (2) **solutions in Nash equilibrium** from a 2-player game with tumor-suppressor and oncogenes as players to predict the health of the lungs, thereby reducing the complexity of the game, which would otherwise be played by N genes

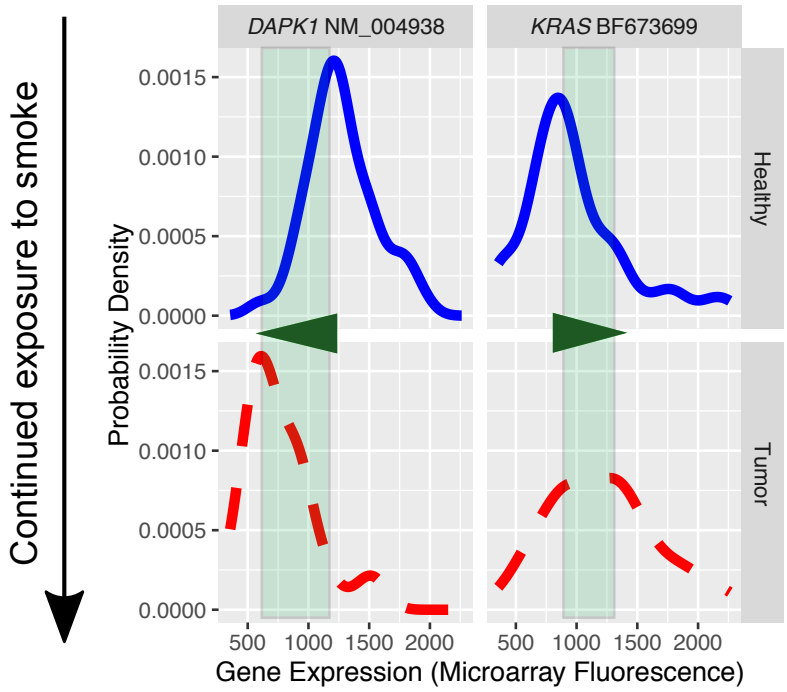


Figure C.1: Illustration of the oncogene's (*KRAS*) activation and tumor suppressor gene's (*DPAK1*) inactivation driving cancer progression, using probability density functions of gene expression. The shaded regions of the PDFs indicate the overlap in their distributions.

whose molecular interactions are not yet understood.

Lung cancer was chosen for this case study as it is the number one cause of cancer death worldwide [195]. The five-year survival rate for lung cancer is 17.8%, which is much lower than that for other cancers, as it is most often diagnosed after the cancer has grown to the point that it is difficult to treat effectively. However, the five-year survival rate for the disease when it is diagnosed early is 54%. Therefore, there is an urgent need to develop techniques to detect lung cancer in its earliest stages, which is why we have chosen early-stage adenocarcinoma data for this study.

C.2 Contribution

Using changes in gene expression as a basis, this work proposes a data-driven 2-player game-theoretic model to predict the risk of adenocarcinoma based on Nash equilibrium. A key innovation in this work is the pay-off function which is a weighted composite of the expression of a cohort of tumor-suppressor genes (as one player) and an analogous cohort of oncogenes

(as the other player). Another novelty of the model is its ability to predict the risk that a healthy sample will develop adenocarcinoma, if its associated gene expression is comparable to that of early-stage tumor samples. The model is validated using two of the largest publicly available adenocarcinoma datasets. The results show that (1) the model is able to distinguish between healthy and cancerous samples with an accuracy of 93%, and (2) 95% of the healthy samples said to be at risk had gene expressions comparable to those of samples with stage I or stage II tumors, thereby predicting the imminent onset of adenocarcinoma.

C.3 Related Work

Conflicts in biology have been studied since the pioneering work of John Maynard Smith in the context of species' survival in a population [196]. Smith's work has been extended to study the evolution of diseases, including cancers [197, 198]. From a biological perspective, to the best of our knowledge, previous studies have not modeled predictability in adenocarcinoma using gene expression. Work on evolutionary theory of disease development generally assumes that an evolutionary process has led either to mutations or to the development of an unhealthy cell (such as cancer cells), and then develop stochastic models (such as the Moran process) to derive the probability that the mutation or unhealthy cell will take over the entire population of cells. The analytical solutions of such models are simple if the difference in player's fitness is assumed to be constant. The practical limitations of these models are (1) the impossibility of learning the fitness of cell types for each patient while the cancer is developing, because generating such data from biopsies is not tractable; and (2) the high complexity of the Moran processes that models cell types based on multiple genes and their associated expressions. We address these limitations in Sec. C.4 by formulating a payoff function that combines the expression of many genes by using distributions learned from a population, thereby allowing for a compact model that captures the dynamics of cancer proliferation. While miRNA of specific biomarkers has been used to predict adenocarcinoma in sputum samples [199], to the best of our knowledge, the proposed approach is unique in using gene expression to predict adenocarcinoma.

C.4 The Data-Driven Game

Players: In a game, players fight/play for some utility/payoff [200]. The players are two types of genes, namely G_1 (tumor suppressor genes) and G_2 (oncogenes). The rationale for that grouping is that several oncogenes have been identified as drivers of lung adenocarcinomas,

Figure C.2: Illustration of risk distribution derived from population data.

while many tumor suppressor genes are known to be inactivated in individuals with this type of cancer. Genes are chosen if (1) they have shown higher scores of Kullback-Liebler divergence and significant p-values ($p < 0.05$), relative to other tumor suppressor genes and oncogenes, in nonparametric statistical tests such as the Komogorov-Smirnov test (which helps differentiate the PDFs for genes derived from healthy and from tumor samples); and (2) they have been found to be prevalent in most adenocarcinoma cases, and are regulated by epigenetic pathways [201, 202]. The model does not limit how many genes compose each group.

Strategies: To differentiate between healthy and tumor samples, we let the strategy space $S = \{healthy(H), tumor(T)\}$ be defined based on the biological functions of the genes in the particular context of adenocarcinoma. However, to assess whether a healthy sample is at risk of becoming cancerous, we let $S = \{healthy(H), risk(S), tumor(T)\}$.

Payoff: For a new patient, the player's payoff is a weighted sum of the gene expressions of genes comprising G_1 and G_2. Suppose we have M strategies, and each strategy is denoted by m. For each gene g_k in a type j, the corresponding payoff $U_j^{m,m'}$ is computed as shown in Equation C.1. $p(m|e(g_k))$ is the likelihood, derived from population data that given the gene g_k's expression $e(g_k)$ and explains which of the m strategies the gene is likely supporting. E.g., $p(sample \text{ is healthy}|e(DAPK1) = 1,250) > p(sample \text{ has tumor}|e(DAPK1) = 1,250)$ in Fig. C.1; similar to the formulation of a naive Bayesian classifier. $p(m|m')$ is the probability that one player's strategy is m if the other player plays m'. This probability is derived from data as the ratio of samples for which the normalized sum of the expression with label m

Table C.1: Adenocarcinoma prediction performance using game theory. ("NA" means ground truth labels are not available in dataset.)

				Strategies				
Dataset	Num. Strategies	Histology labels	Num. Samples	H	R	T	Sensitivity/Specificity	Accuracy(p-value)
Microarray	Two	Healthy	49	48	NA	1	88.89/98.91	93.45(2.1E-16)
		Tumor	58	6	NA	52		
	Three	Healthy	49	23	25	1	NA/NA	NA
		Tumor	58	6	0	52		
NGS	Two	Healthy	37	36	NA	1	76.5/99.10	90.7/(5.466E-6)
		Tumor	125	14	NA	111		
	Threee	Healthy	37	26	11	0	NA/NA	NA
		Tumor	125	14	0	111		

was greater than m'. Finally, the game is defined in the normal form with a bimatrix U, where player G_1 is the row player; player G_2 is the column player; and their payoffs have the corresponding subscripts.

$$U_j^{m.m'} = p(m|m') \sum_{k=1}^{|G_j|} p(m|e(g_k))e(g_k) \tag{C.1}$$

$$U = \begin{array}{c} \\ Healthy \\ Tumor \end{array} \begin{array}{c} \begin{array}{cc} Healthy & Tumor \end{array} \\ \begin{pmatrix} (U_1^{1,1}, U_2^{1,1}) & (U_1^{1,2}, U_2^{2,1}) \\ (U_1^{2,1}, U_2^{1,2}) & (U_1^{2,2}, U_2^{2,2}) \end{pmatrix} \end{array}$$

Notion of risk: We focus on healthy samples that have gene expression in the overlap regions of the distribution of a gene's expression in tumor and healthy samples of a population as shown in Fig. C.2. For a histologically healthy sample, the fact that its expression lies in the overlap region might suggest that it is more likely to develop a tumor in the future. Therefore, we call this distribution from these gene expressions the distribution of *risk*. The 95^{th} percentile of the gene expression that comprises the PDF is such that the $(p(e(g_i)|risk) > p(e(g_i)|healthy))$ and $(p(e(g_i)|risk) > p(e(g_i)|tumor))$.

Solution concept: With a payoff bimatrix U, the game is analyzed using solution concepts such as the Nash equilibrium. In an N-player game, the strategies are said to be *in equilibrium* if one player cannot unilaterally change its strategy to increase its own payoff. Suppose there exist actions in S in a 2-player game. The pair of actions/strategies for player G_1 and player G_2 are said to be in Nash equilibrium (NE) if $(U_j^{m_1^*,m_2^*}, U_j^{m_2^*,m_1^*}) \geq (U_j^{m_1,m_2^*}, U_j^{m_2,m_1^*})$ $\forall m \neq m_1^*$, i.e., the utilities for the players with strategy pair cannot be improved by one player's changing to another strategy, $\forall m \neq m_1^*$. A pair of actions/strategies (m_2^*, m_1^*) is said to be in *strict Nash equilibrium* (SNE) if $(U_j^{m_1^*,m_2^*}, U_j^{m_2^*,m_1^*}) > (U_j^{m_1,m_2^*}, U_j^{m_2,m_1^*})$ $\forall m \neq m_1^*$. A mixed strategy is one in which a player chooses an action from a distribution over all

actions. A game in normal form has at least one mixed-strategy Nash equilibrium [203]. Pure strategies are a degenerate case of mixed strategies, in which one action is chosen with probability $p = 1$ and all others with $p = 0$ [203].

C.5 Results

Data: The effectiveness of the model is demonstrated using two of the largest publicly available transcriptome datasets of both solid adenocarcinoma tumors and histologically normal lung tissue from patients with lung adenocarcinoma. The tissues for both of these datasets were diagnosed and staged by pathologists using histology. The transcriptomes in these studies were measured by two different techniques: for the first, we used Affymetrix HU133A microarrays [192]; for the second, we used next-generation sequencing (NGS) [193] of RNA. We used two distinct methods of measuring expression to demonstrate that our model is not unduly influenced by the underlying technology and to show that more cost-effective methods of assaying expression, such as qPCR, could also be employed. Finally, as each of these studies includes adjacent, histologically normal samples, our model is unlikely to be affected by systemic confounders and stratifiers, such as environmental exposure, genetic background, pharmaceutical use, sex, and age. Two histologically defined states of lung health are defined in both the datasets: the lungs either have tumors, or are healthy (are histologically noncancerous). We will use these labels to differentiate the samples in this work.

Genes: The expressions in 237 tumor suppressor genes and 248 oncogenes were examined. Ultimately, three tumor suppressor genes (*DAPK1, APC*, *RASSF1*), which constitute G_1, and three oncogenes (*KRAS, BRAF*, *CCNE2*), which constitute G_2 were chosen using criteria discussed in Sec. C.4.

Cross-validation and equilibrium selection: Using leave-one-out and k-fold cross-validation, we trained our model for all but one sample (or $k-1$ sets of samples), and learned the distributions for each gene; each gene's distribution is one of the two (or three) strategies in this game. Then we input the test sample's gene expression into the model and used Equation C.1 to compute the utilities, using the distributions learned from the population data. We then subjected the game to a test sample; the prediction of the health of the sample's lung was based on the strategies either in strict Nash equilibrium (SNE) or in the pure strategy Nash equilibrium (PSNE) that maximized the sum of the player's payoffs.

Performance: First, two strategies (H, T) were used to predict whether the samples were healthy or had tumor. Predictions were done for about 80% of the samples using solutions in

SNE, and the model discriminated between healthy samples and the samples with tumor with a specificity of 91.29% and a sensitivity of 100%. All the samples for which predictions were made using SNE had expressions that did not fall in the overlap regions of their respective gene expression distributions drawn from healthy and tumor samples. For all the samples (i.e., in totality), the model discriminated between healthy samples and the samples with tumor with a specificity of 82.7% and a sensitivity of 98.61% (see Table C.1). Prediction of tumor samples saw higher false-negative rates because the expression profiles of some of the healthy samples overlapped with those of the tumor samples, as seen in Fig. C.1. Other efforts to discriminate among stages of cancer and between healthy and tumor samples have made similar observations on the overlap of gene expression profiles in healthy and tumor samples [190].

We define a quality metric for the PSNE used for prediction (among other possible PSNEs) as the ratio of the sum of utilities from a strategy with maximum utility (such as that of a strict Nash) and the sum of all utilities from strategies in NE. This metric provides the proportion of dominance of the strategy in the Nash equilibrium chosen for prediction, relative to other strategy pairs in NE. Figure C.3 illustrates the quality of the chosen PSNE for prediction; we observed that the average quality score of PSNE in correct predictions (box-plots along principal diagonal) was significantly higher than and differently distributed from those found when the predictions were wrong ($p < 0.0045$).

Game versus standard classifiers: For the two-strategy case, since the truth labels were known from the datasets, we used support vector machines (SVM) with linear and radial-basis function kernels and random forests with leave-one-out and 10-fold cross-validations for training the classifiers with an extensive grid-search to choose model parameters that maximize accuracy and area under curve (AUC). The proposed model's prediction accuracy was better than these two standard classifiers by at least 8.4%. The misclassification was for samples in the overlap regions of gene expression distributions (e.g., Figs. C.1 and C.2), thereby establishing robustness of the data-driven game-theoretic model.

Risk evaluation: Given that our model discriminates between the histologically healthy and tumor samples, we now focus on the healthy samples that have gene expression in the overlap regions of the two distributions. We subjected the same datasets to the game, with three strategies for each type of genes. In Table C.1, we show that 51% (25 out of 49 samples) and 27% (11 out of 37 samples) of the healthy samples from the microarray and NGS datasets, respectively, are at risk of developing tumors. Further, the samples identified as being at risk were found to have quality scores comparable to those of samples whose health was correctly predicted, indicating the confidence in risk prediction.

In recent research, a risk model for survivability using gene expression has been proposed

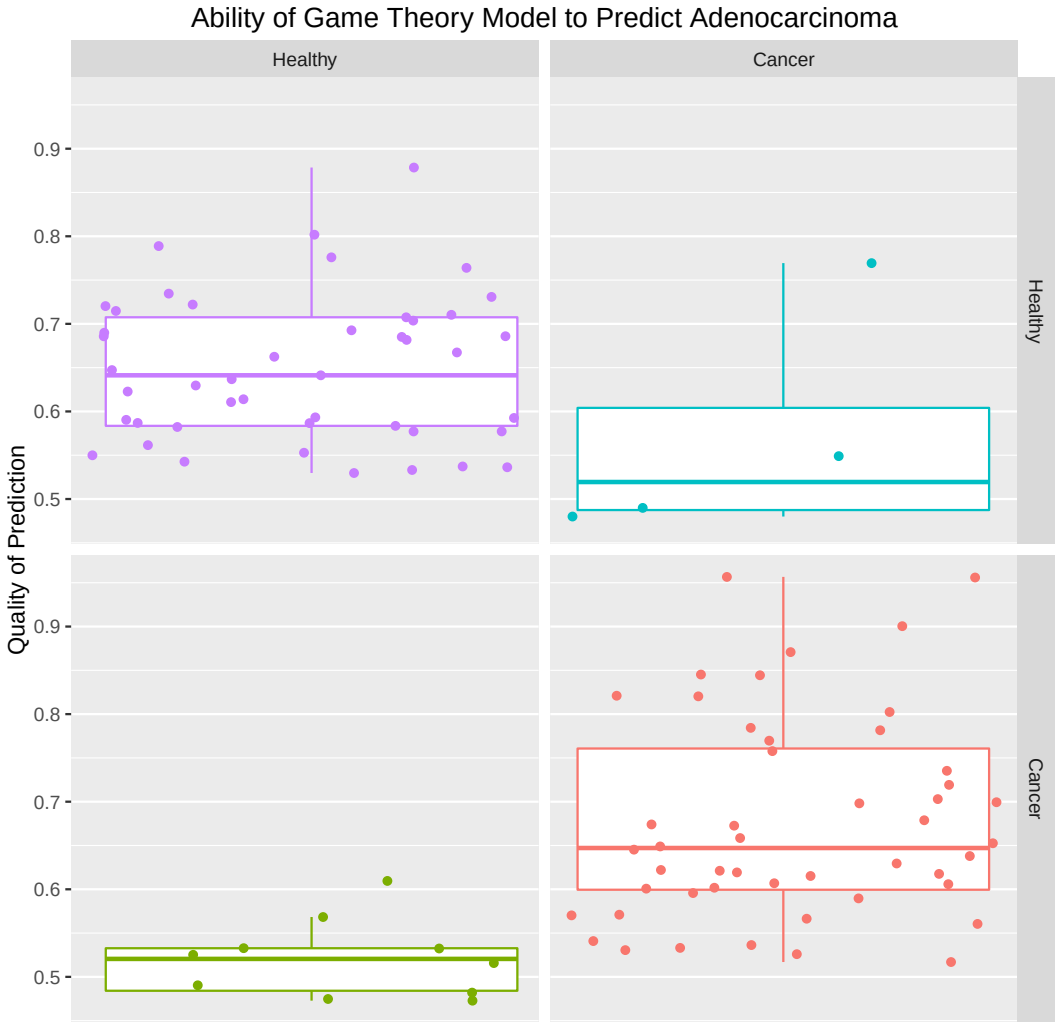


Figure C.3: Game's prediction results for the Microarray dataset. The x-axis are the predicted labels and y-axis are the labels from the data. The box-plots show the variability in quality of prediction.

in the context of clinical outcomes [204]. The study of risk was beneficial for early-stage adenocarcinoma (stage I) from an intervention perspective. In our work, the game theory model showed that an average of 95% of the healthy samples that were predicted to be at risk of developing tumor in the future (including 92% of, or 23 out of 25 samples in, the microarray data, and 100% of the 11 samples in the NGS data) were very similar in expression to the lung samples diagnosed with early-stage cancers (stages IA, IB, IIA, and IIB) in both the microarray fluorescence and NGS data, in that there was a minimal L2-norm between the gene expression vectors of healthy samples predicted to be at risk and those of tumor samples. This result from our risk evaluation is consistent with the findings of the risk model using survivability [204], suggesting a high likelihood of adenocarcinomas onset in these histologically healthy samples.

C.6 Summary

By combining the analytics of learning and game theory with measures of gene expression, we are able to accurately distinguish tumor samples from healthy samples. The additional value of the model is its ability to predict, with a high degree of confidence, the risk that a healthy lung sample will turn tumorous. Moreover, our game-theoretic approach can be generalized to other types of disease and diverse phenotypes as we develop more complex and repeated games with intricate payoff functions, as we continue to improve our understanding of the underlying biological mechanisms and interactions. When combined with an easy way of collecting samples from biofluids such as blood and sputum, the model, in focusing on molecular interactions within and among cells, can augment the use of current standard histological technologies, in which image patterns alone cannot identify risk of cancer progression. This model can be extended to enable visualization of environmental impact on adenocarcinoma development.

APPENDIX D

SINGA-DRAGN: SINGAPORE DIABETIC READMISSION GRAPHICAL MODEL

D.1 Introduction

This work uses a factor-graph-based probabilistic graphical model to analyze longitudinal data presented by electronic health records (EHR) to forecast a series of future health complications that might warrant hospital readmission. The choice of factor graphs is driven by their ability to provide a compact expressive representation of random variables and can subsume both Bayesian networks and Markov random fields (MRFs) [84, 85]. Furthermore, factor functions learned from the data facilitate efficient mechanisms to forecast future events. Although factor graphs have been pursued in information-theoretical settings, recent work has shown that factor graphs can also be used in continuous monitoring of cyber-physical systems [205].

The EHR comprise details pertaining to a patient's visit to a healthcare provider [206]. The primary contents of the EHR include demographic information (e.g., age, gender, race, marital status), epidemiological information (e.g., disease exposure), diagnosis history, laboratory tests and results, drug prescriptions, and clinicians' notes. The nature of the data in EHR can be structured or unstructured. For example, structured data might include age, gender, drug name and drug dosages; and unstructured data might include radiology, microbiology, and histology reports as well as a clinician's text inputs.

This work is motivated by the need to predict/forecast a diabetic patient's short-term post-surgical health complications. Type II Diabetes (T2DM) is a major chronic disease globally, but especially in Asia. T2DM patients have increased risk of post-operative complications due to pre-existing chronic diseases and the immunosuppressive effects of diabetes. While doctors are able to provide value judgments on a patient's ability to recover from surgery and can implement preemptive intervention such as prophylactic antibiotics, they are unable to accurately predict which patients are likely to suffer short-term complications (within 30 days) due to the interaction of preexisting chronic diseases and surgical factors. The ability to accurately predict outcomes of surgery (even if performed using surgical robots) based on multiple features of patients and details of operations to optimize perioperative care in diabetic patients would represent a significant advance in the care of these patients.

In particular, the prevention of readmissions secondary to post-operative complications would represent a significant reduction in patient morbidity as well as cost savings to the hospital. Furthermore, currently it is not possible either to make long-term forecast of health conditions that will warrant readmissions and surgical interventions, or to query population-wide comorbids (simultaneously presented health conditions) that contribute to readmissions (including readmissions within 30 days).

D.2 Contribution

Toward that end, by demonstrating the use of factor-graphs embodied in a tool, SINGADRAGN (Singapore Diabetes Readmission Graphical Network), this work makes the following key contributions.

1. It demonstrates our ability to forecast ten test patients' future health complications and their expected times to hospital readmission given their current comorbids. The forecast uses the factor functions inferred from EHR spanning 10 years of 100 diabetic patients who have undergone surgeries at the National University Hospital, Singapore. As an example, for *diverticulosis* as current diagnosis in test patients, we show that we are able to forecast accurately their future complications.

2. We provide a technique that can use the most highly weighted factor functions to facilitate the identification of common comorbids warranting readmission to the hospital within 30 days.

D.3 Related Work and Analysis Challenges

Current EHR analyses have largely focused on (1) inferring comorbids (simultaneous presence of multiple conditions) associated with specific background health conditions/diagnoses [207], (2) early detection of specific events (for example, heart failure, atrial fibrillation and/or atrial flutter, tumor relapse) [208–211], (3) recommending therapeutic options [212], and (4) predicting adverse drug events (ADE) [213]. All these existing analyses use diagnoses codes (ICD-9, ICD-10) or, diagnoses descriptions, or discharge codes associated with events/diagnoses prior to specific events.

That leads to two key observations both of which reveal shortcomings in the context of this work.

1. If the analyses are customized for a single class of health problems, they alone might not be sufficient in a large multi-speciality hospital setting. A physician might be interested in information beyond prediction of specific health condition such as possible downstream health effects as patients continue to age.

2. To predict specific health events/conditions, the analyses first identify patterns or trajectories of diagnoses that lead to the event of interest. Then they train classifiers, such as neural networks or random forests, which help identify important features in addition to making predictions. Prediction based on trained patterns from high-frequency events implicitly assumes causality of observed patterns of diagnoses. However, this approach will overlook patterns of rare but important events if their occurrence is very scarce in the training data, potentially leading to false or missed predictions.

Key challenges in analyzing the EHR data are as follows:

1. A current health complication in an individual could be a manifestation of several other current and past complications. For example, a current complication such as chronic renal failure might have resulted from early-stage renal failure (ESRF) in the past and may be an outcome of type II diabetes as a background disease. On the other hand, renal failure could be caused by other antecedent conditions, such as hypertension, and might not progress to chronic renal failure. Hence a robust probabilistic model is required in order to estimate the likelihood that any given individual will develop a disease, given his/her medical history relative to a particular population.

2. With longitudinal clinical data alone, identified disease associations do not imply causality. Through the availability of population-level clinical data, it is hoped that such associations will capture trends that warrant investigation through a clinical trial or from additional data, such as genomic data. For example, studies have shown a higher incidence of cancer in type I and type II diabetics [214, 215]. However, that does not imply that cancer observed in diabetic patients is caused by diabetes. Indeed, there could be other genetic predispositions for cancer in such patients, which may be elucidated only if other data, such as genetic information, is made available.

3. A predictive model must be able to distinguish repeat diagnoses and complications to avoid spurious outcomes as a result of administrative or syntax-related repetitions. For example, a patient might appear to have multiple admission events for the same diagnosis because of documentation requirements, but the model should recognize them as a single episode of that diagnosis. For another example, an individual might be admitted to the hospital for fever several times in their lifetime, but the causes and

contexts of the fevers may be different. Furthermore, several other complications might be driving the fevers and each combination of such complications in the context of fever must be learned from the population's EHR data.

Our work addresses the shortcomings of existing EHR analyses in the following ways:

1. To the best of our knowledge, our factor-graph based graphical model-based tool is the first of its kind that can be trained on all combinations of diagnoses observed in a population. By design, we are able to provide a global view of an individual's health by forecasting future health complications with current comorbids (diagnoses) as inputs.

2. We track every combination of comorbids associated with a current diagnosis that lead to different sets of comorbids of the next diagnosis, embodied in what we call factor functions. We then rank the likelihood that these factor functions will be associated with specific combinations of current comorbids to determine the most common population-wide combinations of comorbids.

3. We identify every combination of comorbids associated with current diagnoses that act as precursors to subsequent diagnoses that warranted hospital readmission within 30 days.

4. Because we rank every combination of comorbids observed in a population, we are now able to provide population-level statistics of all prevailing health conditions and common complications associated with hospital readmissions.

D.4 Data

The data were derived from a longitudinal inpatient dataset comprising approximately $500,000$ medical records, including lab and radiology reports, emergency department notes, prescribed and dispensed medications, surgical notes, and discharge summaries. It is a National Healthcare Group (NHG) Domain Specific Review Board (DSRB) approved database and resides in NUH servers and workstations governed by institutional data policies.

Records of $11,000$ unique T2DM patients who underwent surgery at NUH over a period of 10 years were extracted. Diabetic surgical patients were identified according to the multiple text permutations of diabetes diagnoses and further stratified according the diabetic subtype (e.g., gestational diabetes). Each record contains primary (raw) data such as anonymized demographic information (nationality, race, age, gender, blood type), the condition in which

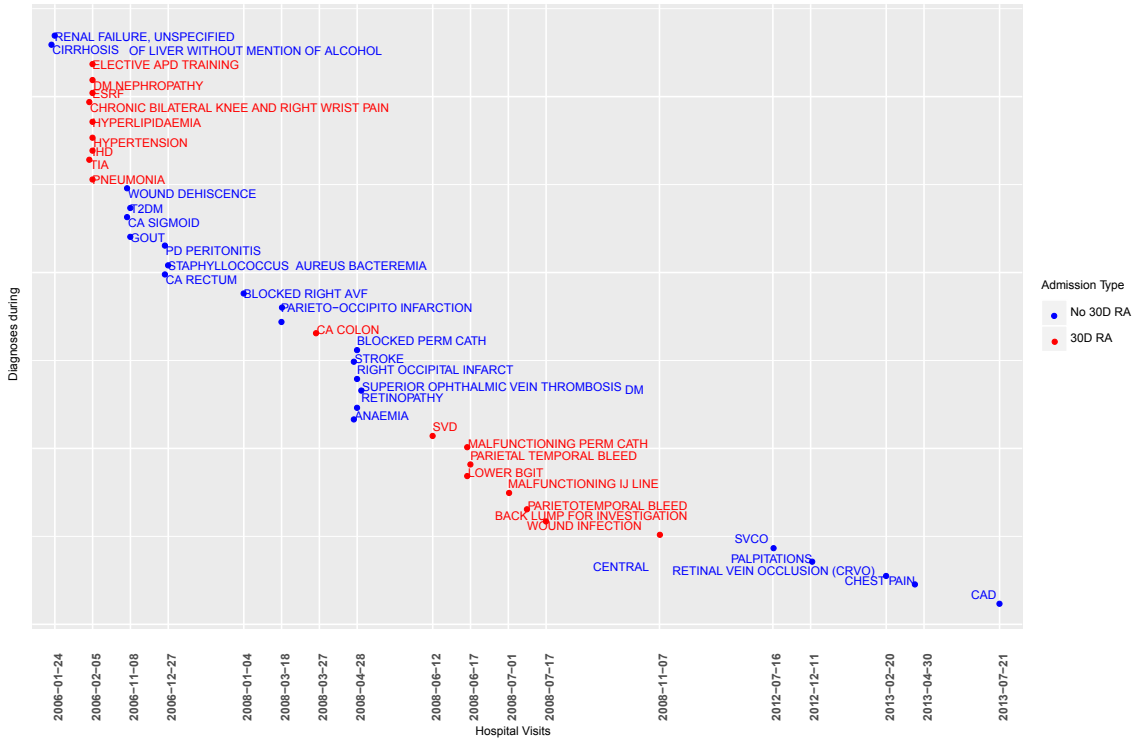


Figure D.1: Course of one patient's health over 10 years derived from the person's electronic health records. Complications in blue are those for which the patient was readmitted to the hospital, but not within 30 days of the previous discharge. Complications in red are those for which the patient was readmitted to the hospital within 30 days of the previous discharge.

the patient was admitted (heart rate, sugar levels, weight, etc.), emergency admission notes, lab report information (blood tests, urine analysis, etc.), surgical notes (type of surgery), patient discharge summaries, and medications prescribed and dispensed. Secondary (processed) data include patient conformance (whether the patient conforms to the treatment prescribed and manages sugar levels), time sequence of admission diagnoses, and so on. If we were to treat each of these labels in the data as a feature, the dataset would have about 200 features in total.

D.4.1 Data Format

The data are presented in an XML file format provided by the database software engineered by Oracle. This is the native enterprise data storage format, and significant processing is required to transform the data into analyzable data. The dataset is presented as a compressed dump file approximately 2.5 Tb in size divided into nine semantic groups in separate databases.

D.4.2 Data Transformation

Each attribute in the medical record is a container in the XML file. Using a standard XML to .CSV conversion software, we extracted and flattened the files. In this initial study, 100 randomly selected individual patient records, including all semantic groups, were manually reviewed by doctors to check for systemic errors and to identify inaccuracies. This process identified major errors in data transformation that resulted in omissions and were subsequently fixed through alterations to the data-flattening program to account for idiosyncratic variations of the source index files through the years.

After the data-flattening program was altered, the extraction software was unable to fully convert all the files because of the size and complexity of the database. The flattening software had to be specifically engineered to reduce the time needed to extract the 100 patient's data to just under an hour for the same batch size.

Significant effort was employed to ensure data veracity at every step. After data transformation, another error-checking step using one hundred randomly selected patients was performed to ensure that no packet losses or frame-short errors occurred during transformation. The completed data package was presented as a MS-SQL database for analysis.

D.4.3 Data Exploration and Curation

The nine semantic groups in the database contain many features required for routine clinical operations, such as ward transfer locations and duplicate demographic information. We indexed the database according to diagnoses and relevant fields we selected to optimize the size of the dataset for analysis. The feature selection strategy is inclusive to incorporate known as well as potentially unknown variables in the T2DM and surgical readmission literature while reducing the dataset size through elimination of duplicate, redundant, or unfilled features.

In addition, there were many sparse features because of changes in the data capture methodologies or creation of new fields over the 10-year period. In situations where a sparse data variable was critical to the analysis, statistical imputation techniques were employed to enable the representation of the feature.

Next, medication lists were consolidated, and variations in medication dictionaries were regularized according to the hospital's current pharmacopeia. To compare drug doses in the analysis, a "standard dose equivalence" (SDE) list was established, against which the various doses, frequencies and duration of drugs used were calibrated.

To address the issue of changes in classification standards (such as ICD-9 to ICD-10

transitions through the years [216], or the absence of such coding in the data, a separate program was developed by the NUS team to assign codes to diagnoses. Using the UMLS metathesaurus and a text-mining engine, the program was able to assign ICD-10 codes to the Concept Unit Identifier (CUI) level for analysis. There is an ongoing effort to complete ICD code assignment to term (LUI) and even-string level (SUI) concept identifiers, which would greatly improve the granularity of the data field. That process is eliminating incorrectly assigned diagnosis codes due to spelling errors and semantic duplications (e.g., heart failure and congestive cardiac failure) and regularizing ICD coding standards. An example of a patient's record is shown in Fig. D.1.

Anonymization of data is carried out at the data administrator level and governed according to institutional data privacy polices. Structured identifiers (e.g., identity numbers, names) are assigned random numbers and with a re-identification key is kept by the administrator. Any re-identification needs are subject to review by the project IRB and data committee. For unstructured identifiers, another program developed by NUS researchers is used to remove patient identifiers in local medical text data (such as discharge summaries and notes). The program is able to remove 99.8% of identifiers and has been vigorously tested on a local medical text lexicon to ensure complete removal of identifiers without eliminating matched terms that are non-identifiers.

D.5 Longitudinal Analysis Using Factor Graphs

Factor graphs provide an expressive representation of random variables. Factor graphs can subsume both Bayesian networks and Markov random fields (MRFs). While Bayesian networks have been quite extensively used in probabilistic methods, their application in this domain is limited by their implicit assumption of causality in observed events, which might not be biologically substantiated.

A factor graph is a bipartite, undirected graph $G = (V; E)$ that represents the relations among random variables, which can be causal or non-causal relations. A vertex (node) $v \in V$ corresponds to a random variable or a factor function. An undirected edge $e \in E$ connects a factor function to a random variable. In a factor graph representation, the relations among the variables are explicitly specified by factor functions $f(X_i)$ that describe the relation among variables in the set X_i. The undirected nature of the graph does not assume causality in the observed events. The variable in this work correspond to comorbids such as hypertension, chronic renal failure (CRF), heart failure and anemia or any combination of them observed in hospital visits and are inferred from the EHR of a population. A factor function in a factor

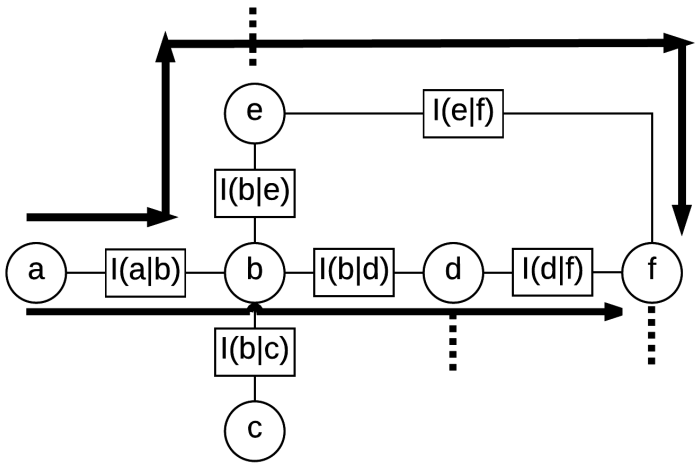


Figure D.2: For pairwise relationships between health complications expressed by factor functions established from population data, we can trace paths from health complications in node "a" to health complications in node "f" in two possible ways.

graph can be any function, e.g., a probability mass function or any real-valued function. In this work, we define the factor function as an imply function comprising the current set of complications and possible future complications. The imply function $I(a|b)$ finds the occurrences (and hence the probability) of health complication(s) "b", given the current health complication(s) "a", where $[a, b] \in X$. Conversely, $I(b|a)$ finds the occurrences (and hence the probability) of health complication(s) "a", given the current health complication(s) "b", where $[b, a] \in X$.

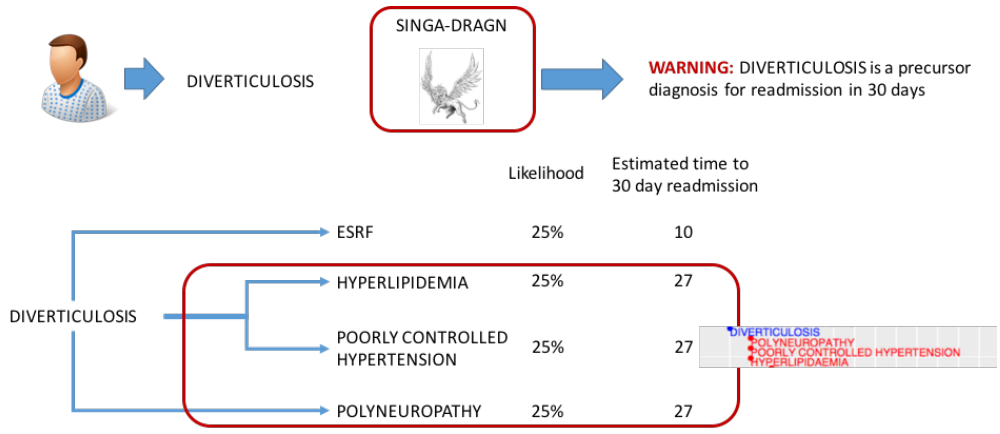
If every patients' course of health over 10 years is treated as a graph, using our approach, the factor functions provide an understanding of all pairwise relationships between pairs of comorbids (diagnoses during visits) in the population as shown in Fig. D.2. For a given starting health complication as "a", we can find all factor functions with current health complication as "a" that was observed in the population. Using a set of rules (for e.g., most occurring transitions, transition more likely in a specific race etc.), we can choose a transition that is most relevant. We can build the future health complications by recursively looking up factor functions with likely current health complications. For a future health complication"f", if a patient is starting with diagnoses "a", two possible paths are $a \rightarrow b \rightarrow d \rightarrow f$ and $a \rightarrow e \rightarrow f$ as shown in Fig. D.2.

Table D.1: An example factor function table for a patient.

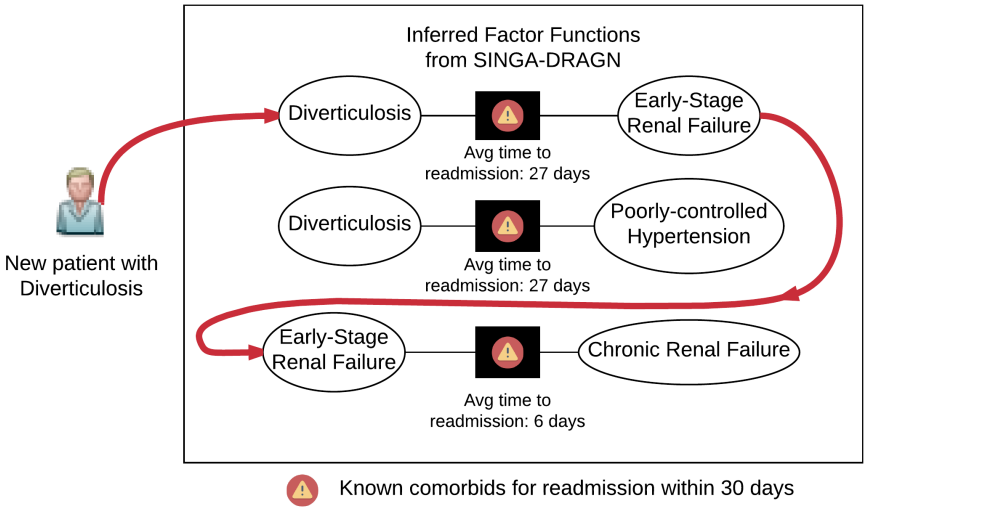
From complications	To complications	Time between complications (days)
Postural hypotension	DVT	380
Anemia, ESRF	Herpes zoster, hypertension	100
Diverticulosis	ESRF	27
Diverticulosis	Hyperlipidemia	25
Diverticulosis	Poorly controlled hypertension	24
Diverticulosis	Polyneuropathy	29

Table D.2: An example of factor functions across patients.

From complications	To complications	Number of occurrences	Time between complications (days)	30-day readmission flag
Postural hypotension	DVT	1	380	Yes
Anaemia, ESRF	Herpes zoster, hypertension	2	100,56	No
Diverticulosis	ESRF	3	3,11,7	Yes
Diverticulosis	Hyperlipidemia	3	25, 28, 26	Yes
Diverticulosis	Poorly controlled hypertension	3	24, 30, 25	Yes
Diverticulosis	Polyneuropathy	3	29, 25, 27	Yes
ESRF	Chronic renal failure	4	3, 9, 4, 8	Yes



(a) An example test patient



(b) Establishing course of health using factor functions

Figure D.3: Fig. (a) illustrates an example of a patient with a current diagnosis of *diverticulosis* as input to the SINGA-DRAGN tool, which has precomputed the factor functions. Fig. (b) shows how one plausible course of health for the next two visits to the hospital are computed based on the functions derived from Table D.2. The red line in Fig. (b) traverses the functions starting with the patient's current diagnoses.

D.5.1 Computation Model: Training on Population Data

The entire development of the tool was done in R, version 3.2.2. The tool was first trained on the population data using the training module of SINGA-DRAGN and then tested with a patient's current complication t forecast the individual's health. We trained SINGA-DRAGN with 12,000 EHR of 100 T2DM patients (all of whom were older than 40) who underwent

surgery at NUH during a 10-year period. We used EHR of 10 other T2DM patients who underwent surgery at NUH during the same 10-year period for testing. For the initial development of the model, we chose the EHR of those 100 patients because those EHR had been manually verified by physicians at NUH.

First, each patient's record was individually processed. A data structure describing the factor functions in terms of the relationship of the imply function and the time between observed complications was output for each patient. We can output of each those data structures as a table in a .CSV file. As an illustration, an example factor function table with a few descriptive diagnoses, derived from a patient's record is shown in Table D.1. (in the tool, the human-readable diagnoses are replaced by in ICD-9/ICD-10 codes). There are multiple diagnoses of complications in some functions, as these diagnoses were all made during the same visit to the hospital. For example, the stacked diagnoses in each hospital visit shown in Fig. D.3(a).

Next, once the factor function tables have been computed for all patients, a script looks for identical factor functions. Identical factor functions are those that have the same diagnoses in the *from complications* and the same diagnoses in the *to complications*. We also count the occurrences of identical factor functions and obtain the distribution of *time between complications* during each occurrence. The training module of SINGA-DRAGN then outputs the data-structure for all the factor functions learned. Table D.2 shows few of the factor functions. For the 100 patients that were used to train the model, a total of $603,475$ factor functions were computed. (We discuss computational performance issues in Sec. D.6.)

D.5.2 Patient-specific Forecast

We now describe the order in which we process a test patient's current comorbid being *Diverticulosis*, using the factor functions inferred from the training cohort and tabulated in Table D.2.

1. From Table D.2, we extract all functions that have the current comorbid of the patient. We subset Table.D.2 with *from complications* having *Diverticulosis*.

2. In Table D.2, there are four entries with *Diverticulosis* in the from *from complications* column and each of them have occurred three times. Therefore, the likelihood of each of these complications in the future is equal. We believe that this particular observation is an artifact of a sampling bias in a very small cohort. However, we are unlikely to observe this uniform distribution of likelihoods when our model is trained on the larger cohort.

3. Since we have recorded the readmission intervals associating current comorbids to future diagnoses in all their occurrences in the training cohort, we can compute their statistical average. For example, for the transition from *Diverticulosis* to *Hyperlipidemia*, the average time to readmission is 27 days ($\{27 + 26 + 28\}/3$). These likelihoods will be different across different demographic factors when a larger cohort is processed.

4. Next we establish plausible courses of health using the computed factor functions from the Table D.2, as shown in Fig. D.3(b). As an example, this test patient currently diagnosed with *diverticulosis* could later be diagnosed with *early-stage renal failure*, and next be diagnosed with *chronic renal failure*. The course of health forecast is *diverticulosis* $\rightarrow$ *early-stage renal failure* $\rightarrow$ *chronic renal failure*. We can not only provide the average time to readmissions for every pair of events in this patient's course of health, but also raise warnings if at least one patient during the training phase was readmitted within 30 days with this pair of comorbids (*diverticulosis*, *early-stage renal failure*) using the 30-day readmission flag is set to "Yes" in Table D.2.

Although we learned over half a million factor functions from the EHR associated with the 100 patients, for *diverticulosis*, we needed only a few factor functions to find the next possible health conditions of this patient. From this patient's actual medical record, we learned that the model predicted all three actual complications correctly as shown in Fig. D.3(a). However, a new possible diagnosis was found, which is *early-stage renal failure* (ESRF), which might imply that the cohort of patients with similar characteristics might suffer this complication in the future. We believe that use of our approach would change the way physicians screen patients who present with certain diseases and bundle interventions that are common to patients with certain complications. Currently in our tool, for each of the diagnoses in the forecast, we recursively query the factor function obtained from the training module and forecast complications in upto five hospital visits in the future, as shown in Fig. D.3(b).

For testing our model, we used ten new test patients (not in the training cohort) to predict their next potential health complication (diagnosis) given that they had *Diverticulosis* as the current health condition. Only five of the ten patients had *Diverticulosis* in their EHR as a diagnosis. In all these five patients, the future diagnoses that were learned from the training data was present in all of their diagnoses when they visited the hospital after having *Diverticulosis* diagnosis in their previous visit. Furthermore, in three among the five test patients, their time to readmission was on average two weeks more than the estimated time to readmission from the factor functions and in the remaining two patients, the time readmission was within a week of previous discharge. While the accuracy of these forecast are promising and take this feasibility study a step in the right direction, we are aware of

several other variables that were not considered while we were training our model as well as biases introduced by a small training cohort. We will discuss these factors in Sec. D.6.

D.5.3 Population-specific Statistics

From an epidemiological perspective, it is interesting to ask questions such as, "what complications are most prevalent diabetic patients in Singapore, which increases healthcare costs and adversely affect the population's health?" Our model keeps count of the occurrences of complications (which are weights of the factor functions in this work) as shown in Table D.2. Further, we can query the model about the health complications that can reveal potential precursors and future complications based on population data.

The model is being developed to accommodate more training data, and eventually will scale to health-system level populations. The validity of the model can be further tested in other hospitals in Singapore. This will work better inform subgroups of patients about their future health complications, and will provide more personalized information to allow patients and physicians to make better decisions on early intervention.

D.6 Discussion

The goal of this work was to demonstrate the feasibility of a factor graph-based approach to analyzing longitudinal data from electronic health records. Since the training dataset used was very small compared to the actual diabetic cohort in Singapore, our model has several limitations which we discuss next and will address in our future work.

D.6.1 Performance and Scalability

The training module of SINGA-DRAGN was designed to allow for processing of multiple patients' data in parallel based on the threads available in the computing environment. On a 2.7 GHz Intel i7 processor with Mac OS X and an IBM POWER8 machine with Linux, patient data were processed eight patients at a time and each patient's analysis took on an average of 40 seconds with a standard deviation of 13 seconds. The greater the number of visits, the larger the time to compute the factor function table for that particular patient. The script that computed the factor function table took roughly three minutes to coalesce the factor functions from all patients. The total number of factor functions was a little more

than half a million. We anticipate that the number of factor functions will grow when we incorporate the entire cohort's records.

In our future work, we intend to make SINGA-DRAGN compatible with a MapReduce framework that can process the patient's data in parallel in the Map() procedure and then combine the factor function tables into a composite one with a Reduce() procedure. That will allow us to use high-performance computing facilities, such as the Blue Waters supercomputer at the University of Illinois at Urbana-Champaign, or the cluster facilities at the National Supercomputing Center, Singapore for executing the training module.

D.6.2 Demographic Integration and Forecast Accuracy

For the 100 patients in this trial phase we did not incorporate any demographic features. However, we plan to incorporate demographic information such as age, gender, and race as priors in our future work to improve prediction and make the model very expressive. One challenge in EHR analysis we mentioned in Sec. D.3 was the need to manage repeated diagnoses. Let us suppose a patient who is currently diagnosed with *hypertension* gets treated with medications and the patient conforms to the same. Because of medication, let us assume that in a few subsequent hospital admissions, *hypertension* is not listed among other diagnoses. It is highly likely that this same patient has other conditions along with *hypertension* in the future, since aging introduces tends to compound health complications. Then, a different factor function that has other comorbids as part of the patients' health will be used to forecast future health complications. If *hypertension* is the only diagnosis in this patient's health after many years, then, the same factor function that was used to forecast this patient's health with this diagnosis as the only input will be used, and therefore might be prone to errors in forecast. We believe that our approach has the ability to capture as many possible health conditions individuals can transition into, based on training data. At the same time, we are also aware that we will not be able to learn every possible transition between combination of comorbids if they are not observed in the training cohort.

To forecast possible courses of health, we currently generate forecasts for up to three potential hospital visits in the future. However, we intend to generate up to ten hospital visits in the future and rank the plausible forecasts by their likelihoods, using a combination of the occurrence of the factor functions along with the associated estimated times to readmission and information on whether the comorbids are associated with 30-day readmissions.

D.6.3 Cost-Benefit Analysis for Early Intervention

The current version of SINGA-DRAGN has provided physicians with the first tool that quantitatively assesses common health complications that cause recurring hospital re-admissions. Further, insights on which complications warrant surgeries and how the aftereffects of surgeries affect the patients' health are being gained with this feasibility study. Currently, physicians in collaboration with hospital administration are assessing the downstream cost of care for these common complications, and as well as the degradation in quality of life resulting from associated surgeries. When our future analyses encompass the entire cohort, we will be able to identify a tipping point in an individual's predicted health, beyond which the patient's aging can be improved through preemptive clinical/surgical intervention.

D.7 Summary

This work describes the success of a feasibility study in a factor graph-based approach that was used to analyzing data from electronic health records (EHR) to predict the future health complications and patients' expected time to hospital readmission. Factor functions were learned from over 10 years of EHR data for 100 diabetic patients who have undergone surgeries at the National University Hospital, Singapore. Furthermore, we used the most frequently occurring factor functions to identify comorbids that warrant hospital readmissions. Such information can inform the physician/clinician about when to intervene in order to maximize patients' quality of life and minimize the cost of their care.

REFERENCES

[1] R. Weinshilboum and L. Wang, "Pharmacogenomics: Bench to bedside," *Nature Reviews Drug Discovery*, vol. 3, no. 9, pp. 739–748, September 2004.

[2] W. E. Evans and M. V. Relling, "Moving towards individualized medicine with pharmacogenomics," *Nature*, vol. 429, no. 6990, p. 464, 2004.

[3] K. N. Lazaridis, T. M. Mcallister, D. Babovic-Vuksanovic, S. A. Beck, M. J. Borad, A. H. Bryce, A. A. Chanan-Khan, M. J. Ferber, R. Fonseca, K. J. Johnson et al., "Implementing individualized medicine into the medical practice," in *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, vol. 166, no. 1. Wiley Online Library, 2014, pp. 15–23.

[4] F. S. Collins and H. Varmus, "A new initiative on precision medicine," *New England Journal of Medicine*, vol. 372, no. 9, pp. 793–795, 2015.

[5] E. J. Topol, "Individualized medicine from prewomb to tomb," *Cell*, vol. 157, no. 1, pp. 241–253, 2014.

[6] M. D. Ritchie, "The success of pharmacogenomics in moving genetic association studies from bench to bedside: Study design and implementation of precision medicine in the post-gwas era," *Human Genetics*, vol. 131, no. 10, pp. 1615–1626, 2012.

[7] B. Palsson and K. Zengler, "The challenges of integrating multi-omic data sets," *Nature Chemical Biology*, vol. 6, no. 11, p. 787, 2010.

[8] Z. Wang, M. Gerstein, and M. Snyder, "RNA-Seq: A revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, p. 57, 2009.

[9] C. Trapnell, "Defining cell types and states with single-cell genomics," *Genome Research*, vol. 25, no. 10, pp. 1491–1498, 2015.

[10] A. P. Athreya, A. J. Gaglio, Z. T. Kalbarczyk, R. K. Iyer, J. Cairns, K. R. Kalari, R. M. Weinshilboum, and L. Wang, "Unsupervised single-cell analysis in triple-negative breast cancer: A case study," in *Proceedings of The 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM),*. IEEE, 2016, pp. 556–563.

[11] A. P. Athreya, K. R. Kalari, J. Cairns, A. J. Gaglio, Q. F. Wills, N. Niu, R. Weinshilboum, R. K. Iyer, and L. Wang, "Model-based unsupervised learning informs metformin-induced cell-migration inhibition through an ampk-independent mechanism in breast cancer," *Oncotarget*, vol. 8, no. 16, p. 27199, 2017.

[12] A. P. Athreya, D. Armstrong, W. Gundling, D. Wildman, Z. T. Kalbarczyk, and R. K. Iyer, "Prediction of adenocarcinoma development using game theory," in *Proceedings of The 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 1668–1671.

[13] A. P. Athreya, K. Y. Ngiam, Z. Luo, E. S. Tai, Z. Kalbarczyk, and R. K. Iyer, "Towards longitudinal analysis of a population's electronic health records using factor graphs," in *Proceedings of The 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*. ACM, 2016, pp. 79–86.

[14] C. Otte, S. M. Gold, B. W. Penninx, C. M. Pariante, A. Etkin, M. Fava, D. C. Mohr, and A. F. Schatzberg, "Major depressive disorder," *Nature Reviews Disease Primers*, vol. 2, p. 16065, 2016.

[15] C. D. Mathers and D. Loncar, "Projections of global mortality and burden of disease from 2002 to 2030," *PLoS Medicine*, vol. 3, no. 11, p. e442, 2006.

[16] B. A. Arnow, C. Blasey, L. M. Williams, D. M. Palmer, W. Rekshan, A. F. Schatzberg, A. Etkin, J. Kulkarni, J. F. Luther, and A. J. Rush, "Depression subtypes in predicting antidepressant response: A report from the iSPOT-D trial," *American Journal of Psychiatry*, vol. 172, no. 8, pp. 743–750, 2015.

[17] E. Fried, "Moving forward: How depression heterogeneity hinders progress in treatment and research," *Expert Review of Neurotherapeutics*, vol. 77, pp. 423–425, 2017.

[18] R. Kessler, H. Van Loo, K. Wardenaar, R. Bossarte, L. Brenner, D. Ebert, P. de Jonge, A. Nierenberg, A. Rosellini, N. Sampson et al., "Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder," *Epidemiology and Psychiatric Sciences*, vol. 26, no. 1, pp. 22–36, 2017.

[19] K. L. Musliner, T. Munk-Olsen, W. W. Eaton, and P. P. Zandi, "Heterogeneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes," *Journal of Affective Disorders*, vol. 192, pp. 199–211, 2016.

[20] S. Senn, "Mastering variation: Variance components and personalised medicine," *Statistics in Medicine*, vol. 35, no. 7, pp. 966–977, 2016.

[21] R. H. Perlis, "Abandoning personalization to get to precision in the pharmacotherapy of depression," *World Psychiatry*, vol. 15, no. 3, pp. 228–235, 2016.

[22] A. J. Rush, M. H. Trivedi, S. R. Wisniewski, A. A. Nierenberg, J. W. Stewart, D. Warden, G. Niederehe, M. E. Thase, P. W. Lavori, B. D. Lebowitz et al., "Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR* D report," *American Journal of Psychiatry*, vol. 163, no. 11, pp. 1905–1917, 2006.

[23] L. Wang, J. Ingle, and R. Weinshilboum, "Pharmacogenomic discovery to function and mechanism: Breast cancer as a case study," *Clinical Pharmacology & Therapeutics*, vol. 103, no. 2, pp. 243–252, 2018.

[24] D. A. Mrazek, J. M. Biernacka, D. J. O'kane, J. L. Black, J. M. Cunningham, M. S. Drews, K. A. Snyder, S. R. Stevens, A. J. Rush, and R. M. Weinshilboum, "CYP2C19 variation and citalopram response," *Pharmacogenetics and Genomics*, vol. 21, no. 1, p. 1, January 2011.

[25] M. H. Trivedi, A. J. Rush, S. R. Wisniewski, A. A. Nierenberg, D. Warden, L. Ritz, G. Norquist, R. H. Howland, B. Lebowitz, P. J. McGrath et al., "Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice," *American Journal of Psychiatry*, vol. 163, no. 1, pp. 28–40, January 2006.

[26] J. Biernacka, K. Sangkuhl, G. Jenkins, R. Whaley, P. Barman, A. Batzler, R. Altman, V. Arolt, J. Brockmöller, C. Chen et al., "The international SSRI pharmacogenomics consortium (ISPC): A genome-wide association study of antidepressant treatment response," *Translational Psychiatry*, vol. 5, no. 4, p. e553, 2015.

[27] R. Kaddurah-Daouk, B. S. Kristal, and R. M. Weinshilboum, "Metabolomics: A global biochemical approach to drug response and disease," *Annu. Rev. Pharmacol. Toxicol.*, vol. 48, pp. 653–683, 2008.

[28] R. Iniesta, K. Malki, W. Maier, M. Rietschel, O. Mors, J. Hauser, N. Henigsberg, M. Z. Dernovsek, D. Souery, D. Stahl et al., "Combining clinical variables to optimize prediction of antidepressant treatment outcomes," *Journal of Psychiatric Research*, vol. 78, pp. 94–102, 2016.

[29] A. M. Chekroud, R. J. Zotti, Z. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett, "Cross-trial prediction of treatment outcome in depression: A machine learning approach," *The Lancet Psychiatry*, vol. 3, no. 3, pp. 243–250, April 2016.

[30] A. J. Rush, H. C. Kraemer, H. A. Sackeim, M. Fava, M. H. Trivedi, E. Frank, P. T. Ninan, M. E. Thase, A. J. Gelenberg, D. J. Kupfer et al., "Report by the ACNP task force on response and remission in major depressive disorder," *Neuropsychopharmacology*, vol. 31, no. 9, p. 1841, 2006.

[31] R. Weinshilboum, "Pharmacogenomics to pharmaco-omics: Precision medicine and drug response," *Drug Metabolism and Pharmacokinetics*, vol. 33, no. 1, p. S14, 2018.

[32] D. Liu, B. Ray, D. R. Neavin, J. Zhang, A. P. Athreya, J. M. Biernacka, W. V. Bobo, D. K. Hall-Flavin, M. K. Skime, H. Zhu et al., "Beta-defensin 1, aryl hydrocarbon receptor and plasma kynurenine in major depressive disorder: Metabolomics-informed genomics," *Translational Psychiatry*, vol. 8, no. 1, p. 10, January 2018.

[33] M. Gupta, D. Neavin, D. Liu, J. Biernacka, D. Hall-Flavin, W. V. Bobo, M. A. Frye, M. Skime, G. D. Jenkins, A. Batzler et al., "TSPAN5, ERICH3 and selective serotonin reuptake inhibitors in major depressive disorder: Pharmacometabolomics-informed pharmacogenomics," *Molecular Psychiatry*, December 2016.

[34] Z. D. Cohen and R. J. DeRubeis, "Treatment selection in depression," *Annual Review of Clinical Psychology*, vol. 14, pp. 209–236, 2018.

[35] R. C. Kessler, H. S. Akiskal, M. Ames, H. Birnbaum, P. Greenberg, R. M. A. Hirschfeld, R. Jin, K. R. Merikangas, G. E. Simon, and P. S. Wang, "Prevalence and effects of mood disorders on work performance in a nationally representative sample of US workers," *American Journal of Psychiatry*, vol. 163, no. 9, pp. 1561–1568, September 2006.

[36] N. Olchanski, M. M. Myers, M. Halseth, P. L. Cyr, L. Bockstedt, T. F. Goss, and R. H. Howland, "The economic burden of treatment-resistant depression," *Clinical Therapeutics*, vol. 35, no. 4, pp. 512–522, April 2013.

[37] D. Chisholm, K. Sweeny, P. Sheehan, B. Rasmussen, F. Smit, P. Cuijpers, and S. Saxena, "Scaling-up treatment of depression and anxiety: A global return on investment analysis," *The Lancet Psychiatry*, vol. 3, no. 5, pp. 415–424, 2016.

[38] T. Kendrick, F. King, L. Albertella, and P. W. Smith, "GP treatment decisions for patients with depression: An observational study," *British Journal of General Practice*, vol. 55, no. 513, pp. 280–286, 2005.

[39] C. Hudon, M.-C. Chouinard, M.-F. Dubois, P. Roberge, C. Loignon, É. Tchouaket, M. Lambert, É. Hudon, F. Diadiou, and D. Bouliane, "Case management in primary care for frequent users of health care services: A mixed methods study," *The Annals of Family Medicine*, vol. 16, no. 3, pp. 232–239, 2018.

[40] E. Andreoulakis, T. Hyphantis, D. Kandylis, and A. Iacovides, "Depression in diabetes mellitus: a comprehensive review," *Hippokratia*, vol. 16, no. 3, p. 205, 2012.

[41] D. L. Musselman, D. L. Evans, and C. B. Nemeroff, "The relationship of depression to cardiovascular disease: Epidemiology, biology, and treatment," *Archives of General Psychiatry*, vol. 55, no. 7, pp. 580–592, 1998.

[42] Z. Bhagwagar, E. Rabiner, P. Sargent, P. Grasby, and P. Cowen, "Persistent reduction in brain serotonin 1a receptor binding in recovered depressed men measured by positron emission tomography with [11 C] WAY-100635," *Molecular Psychiatry*, vol. 9, no. 4, p. 386, 2004.

[43] P. J. Modrego and J. Ferrández, "Depression in patients with mild cognitive impairment increases the risk of developing dementia of alzheimer type: A prospective cohort study," *Archives of Neurology*, vol. 61, no. 8, pp. 1290–1293, 2004.

[44] A. Gotham, R. Brown, and C. Marsden, "Depression in Parkinson's disease: A quantitative and qualitative analysis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 49, no. 4, pp. 381–389, 1986.

[45] W. Bardwell, P. Nicassio, M. Weisman, R. Gevirtz, and D. Bazzo, "Rheumatoid arthritis severity scale: A brief, physician-completed scale not confounded by patient self-report of psychological functioning," *Rheumatology*, vol. 41, no. 1, pp. 38–45, January 2002.

[46] W. Kwong and D. Pathak, "Validation of the eleven-point pain scale in the measurement of migraine headache pain," *Cephalalgia*, vol. 27, no. 4, pp. 336–342, April 2007.

[47] M. Kosinski, M. Bayliss, J. Bjorner, J. Ware, W. Garber, A. Batenhorst, R. Cady, C. Dahlöf, A. Dowson, and S. Tepper, "A six-item short-form survey for measuring headache impact: The hit-6," *Quality of Life Research*, vol. 12, no. 8, pp. 963–974, December 2003.

[48] E. Shorter, "History of psychiatry," *Current Opinion in Psychiatry*, vol. 21, no. 6, p. 593, 2008.

[49] K. J. K. Harding, A. J. Rush, M. Arbuckle, M. H. Trivedi, and H. A. Pincus, "Measurement-based care in psychiatric practice: A policy framework for implementation." *The Journal of Clinical Psychiatry*, 2011.

[50] A. Frances, H. A. Pincus, M. First et al., *Diagnostic and Statistical Manual of Mental Disorders: DSM-IV*. American Psychiatric Association, Washington DC, 1994.

[51] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. American Psychiatric Association, Washington DC, 2013.

[52] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber et al., "The 16-item quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression," *Biological Psychiatry*, vol. 54, no. 5, pp. 573–583, September 2003.

[53] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery, and Psychiatry*, vol. 23, no. 1, p. 56, 1960.

[54] S. A. Montgomery and M. Åsberg, "A new depression scale designed to be sensitive to change," *The British Journal of Psychiatry*, vol. 134, no. 4, pp. 382–389, 1979.

[55] B. G. Katzung, S. B. Masters, and A. J. Trevor, *Basic and Clinical Pharmacology (LANGE Basic Science)*. McGraw-Hill Education, 2012.

[56] J. W. Smoller, "Psychiatric genetics and the future of personalized treatment," *Depression and Anxiety*, vol. 31, no. 11, pp. 893–898, 2014.

[57] P. Hamet and J. Tremblay, "Genetics and genomics of depression," *Metabolism*, vol. 54, no. 5, pp. 10–15, 2005.

[58] J.-P. Guilloux, S. Bassi, Y. Ding, C. Walsh, G. Turecki, G. Tseng, J. M. Cyranowski, and E. Sibille, "Testing the predictive value of peripheral gene expression for nonremission following citalopram treatment for major depression," *Neuropsychopharmacology*, vol. 40, no. 3, p. 701, 2015.

[59] A. M. Chekroud, R. Gueorguieva, H. M. Krumholz, M. H. Trivedi, J. H. Krystal, and G. McCarthy, "Reevaluating the efficacy and predictability of antidepressant treatments: A symptom clustering approach," *JAMA Psychiatry*, vol. 74, no. 4, pp. 370–378, 2017.

[60] D. Novick, J. Hong, W. Montgomery, H. Dueñas, M. Gado, and J. M. Haro, "Predictors of remission in the treatment of major depressive disorder: Real-world evidence from a 6-month prospective observational study," *Neuropsychiatric Disease and Treatment*, vol. 11, p. 197, 2015.

[61] R. Uher, R. Perlis, N. Henigsberg, A. Zobel, M. Rietschel, O. Mors, J. Hauser, M. Dernovsek, D. Souery, M. Bajs et al., "Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms," *Psychological Medicine*, vol. 42, no. 5, pp. 967–980, 2012.

[62] R. Uher, A. Farmer, W. Maier, M. Rietschel, J. Hauser, A. Marusic, O. Mors, A. Elkin, R. Williamson, C. Schmael et al., "Measuring depression: Comparison and integration of three scales in the GENDEP study," *Psychological Medicine*, vol. 38, no. 2, pp. 289–300, 2008.

[63] R. Uher, W. Maier, J. Hauser, A. Marušič, C. Schmael, O. Mors, N. Henigsberg, D. Souery, A. Placentino, M. Rietschel et al., "Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression," *The British Journal of Psychiatry*, vol. 194, no. 3, pp. 252–259, 2009.

[64] R. Iniesta, K. Hodgson, D. Stahl, K. Malki, W. Maier, M. Rietschel, O. Mors, J. Hauser, N. Henigsberg, M. Z. Dernovsek et al., "Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables," *Scientific Reports*, vol. 8, no. 1, p. 5530, 2018.

[65] E. Lin, P.-H. Kuo, Y.-L. Liu, Y. W. Yu, A. Yang, and S.-J. Tsai, "A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers," *Frontiers in Psychiatry*, vol. 9, p. 290, 2018.

[66] C. P. O'Connell, A. N. Goldstein-Piekarski, C. B. Nemeroff, A. F. Schatzberg, C. Debattista, T. Carrillo-Roa, E. B. Binder, B. W. Dunlop, W. E. Craighead, H. S. Mayberg et al., "Antidepressant outcomes predicted by genetic variation in corticotropin-releasing hormone binding protein," *American Journal of Psychiatry*, vol. 175, no. 3, pp. 251–261, 2017.

[67] G. I. Papakostas, R. H. Perlis, M. J. Scalia, T. J. Petersen, and M. Fava, "A meta-analysis of early sustained response rates between antidepressants and placebo for the treatment of major depressive disorder," *Journal of Clinical Psychopharmacology*, vol. 26, no. 1, pp. 56–60, 2006.

[68] R. S. McIntyre, P. Gorwood, M. E. Thase, C. Liss, D. Desai, J. Chen, and M. Bauer, "Early symptom improvement as a predictor of response to extended release quetiapine in major depressive disorder," *Journal of Clinical Psychopharmacology*, vol. 35, no. 6, p. 706, 2015.

[69] P. A. Kudlow, D. S. Cha, and R. S. McLntyre, "Predicting treatment response in major depressive disorder: The impact of early symptomatic improvement," *The Canadian Journal of Psychiatry*, vol. 57, no. 12, pp. 782–788, 2012.

[70] S. Wagner, A. Engel, J. Engelmann, D. Herzog, N. Dreimüller, M. B. Müller, A. Tadić, and K. Lieb, "Early improvement as a resilience signal predicting later remission to antidepressant treatment in patients with major depressive disorder: Systematic review and meta-analysis," *Journal of Psychiatric Research*, vol. 94, pp. 96–106, 2017.

[71] Y. A. de Vries, A. M. Roest, E. H. Bos, J. G. Burgerhof, H. M. van Loo, and P. de Jonge, "Predicting antidepressant response by monitoring early improvement of individual symptoms of depression: Individual patient data meta-analysis," *The British Journal of Psychiatry*, pp. 1–7, 2018.

[72] N. Ram and K. J. Grimm, "Methods and measures: Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups," *International Journal of Behavioral Development*, vol. 33, no. 6, pp. 565–576, 2009.

[73] T. Jung and K. Wickrama, "An introduction to latent class growth analysis and growth mixture modeling," *Social and Personality Psychology Compass*, vol. 2, no. 1, pp. 302–317, 2008.

[74] H. Tokuoka, H. Takahashi, A. Ozeki, A. Kuga, A. Yoshikawa, T. Tsuji, and M. M. Wohlreich, "Trajectories of depression symptom improvement and associated predictor analysis: An analysis of duloxetine in double-blind placebo-controlled trials," *Journal of Affective Disorders*, vol. 196, pp. 171–180, 2016.

[75] M. E. Kelley, B. W. Dunlop, C. B. Nemeroff, A. Lori, T. Carrillo-Roa, E. B. Binder, M. H. Kutner, V. A. Rivera, W. E. Craighead, and H. S. Mayberg, "Response rate profiles for major depressive disorder: Characterizing early response and longitudinal nonresponse," *Depression and Anxiety*, vol. 35, no. 10, pp. 992–1000, 2018.

[76] R. Gueorguieva, C. Mallinckrodt, and J. H. Krystal, "Trajectories of depression severity in clinical trials of duloxetine: Insights into antidepressant and placebo responses," *Archives of General Psychiatry*, vol. 68, no. 12, pp. 1227–1237, 2011.

[77] R. Gueorguieva, A. M. Chekroud, and J. H. Krystal, "Trajectories of relapse in randomised, placebo-controlled trials of treatment discontinuation in major depressive disorder: An individual patient-level data meta-analysis," *The Lancet Psychiatry*, vol. 4, no. 3, pp. 230–237, 2017.

[78] K. L. Nylund, T. Asparouhov, and B. O. Muthén, "Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study," *Structural Equation Modeling*, vol. 14, no. 4, pp. 535–569, 2007.

[79] D. J. Bauer and P. J. Curran, "Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes." *Psychological Methods*, vol. 8, no. 3, p. 338, 2003.

[80] P. Cuijpers, E. Weitz, J. Twisk, C. Kuehner, I. Cristea, D. David, R. J. DeRubeis, S. Dimidjian, B. W. Dunlop, M. Faramarzi et al., "Gender as predictor and moderator of outcome in cognitive behavior therapy and pharmacotherapy for adult depression: An 'individual patient data' meta-analysis," *Depression and Anxiety*, vol. 31, no. 11, pp. 941–951, 2014.

[81] R. Keers and K. J. Aitchison, "Gender differences in antidepressant drug response," *International Review of Psychiatry*, vol. 22, no. 5, pp. 485–500, 2010.

[82] C. Berlanga and M. Flores-Ramos, "Different gender response to serotonergic and noradrenergic antidepressants: A comparative study of the efficacy of citalopram and reboxetine," *Journal of Affective Disorders*, vol. 95, no. 1-3, pp. 119–123, 2006.

[83] J. D. Banfield and A. E. Raftery, "Model-based gaussian and non-gaussian clustering," *Biometrics*, pp. 803–821, September 1993.

[84] B. J. Frey, F. R. Kschischang, H.-A. Loeliger, and N. Wiberg, "Factor graphs and algorithms," in *Proceedings of The Annual Allerton Conference on Communication Control and Computing*, vol. 35. UNIVERSITY OF ILLINOIS, 1997, pp. 666–680.

[85] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Processing Magazine*, vol. 21, no. 1, pp. 28–41, 2004.

[86] S. Huang, K. Chaudhary, and L. X. Garmire, "More is better: Recent progress in multi-omics data integration methods," *Frontiers in Genetics*, vol. 8, p. 84, 2017.

[87] A. A. Onitilo, J. M. Engel, R. T. Greenlee, and B. N. Mukesh, "Breast cancer subtypes based on ER/PR and Her2 expression: Comparison of clinicopathologic features and survival," *Clinical Medicine & Research*, vol. 7, no. 1-2, pp. 4–13, June 2009.

[88] Y. Ji, J. M. Biernacka, S. Hebbring, Y. Chai, G. D. Jenkins, A. Batzler, K. A. Snyder, M. S. Drews, Z. Desta, D. Flockhart et al., "Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: Genome-wide associations and functional genomics," *The Pharmacogenomics Journal*, vol. 13, no. 5, p. 456, 2013.

[89] J. Krumsiek, K. Mittelstrass, K. T. Do, F. Stückler, J. Ried, J. Adamski, A. Peters, T. Illig, F. Kronenberg, N. Friedrich et al., "Gender-specific pathway differences in the human serum metabolome," *Metabolomics*, vol. 11, no. 6, pp. 1815–1833, 2015.

[90] C. Bouveyron and C. Brunet-Saumard, "Model-based clustering of high-dimensional data: A review," *Computational Statistics & Data Analysis*, vol. 71, pp. 52–78, 2014.

[91] C. Genolini and B. Falissard, "Kml: k-means for longitudinal data," *Computational Statistics*, vol. 25, no. 2, pp. 317–328, 2010.

[92] R. Uher, O. Mors, M. Rietschel, A. Rajewska-Rager, A. Petrovic, A. Zobel, N. Henigsberg, J. Mendlewicz, K. J. Aitchison, A. Farmer et al., "Early and delayed onset of response to antidepressants in individual trajectories of change during treatment of major depression: A secondary analysis of data from the genome-based therapeutic drugs for depression (GENDEP) study." *The Journal of Clinical Psychiatry*, 2011.

[93] R. Uher, B. Muthén, D. Souery, O. Mors, J. Jaracz, A. Placentino, A. Petrovic, A. Zobel, N. Henigsberg, M. Rietschel et al., "Trajectories of change in depression severity during treatment with antidepressants," *Psychological Medicine*, vol. 40, no. 8, pp. 1367–1377, 2010.

[94] M. Chang, G. Tybring, M.-L. Dahl, and J. D. Lindh, "Impact of cytochrome p450 2c19 polymorphisms on citalopram/escitalopram exposure: A systematic review and meta-analysis," *Clinical Pharmacokinetics*, vol. 53, no. 9, pp. 801–811, 2014.

[95] M. Nassan, W. T. Nicholson, M. A. Elliott, C. R. R. Vitek, J. L. Black, and M. A. Frye, "Pharmacokinetic pharmacogenetic prescribing guidelines for antidepressants: A template for psychiatric precision medicine," in *Mayo Clinic Proceedings*, vol. 91, no. 7. Elsevier, 2016, pp. 897–907.

[96] M.-H. Tsai, K.-M. Lin, M.-C. Hsiao, W. W. Shen, M.-L. Lu, H.-S. Tang, C.-K. Fang, C.-S. Wu, S.-C. Lu, S. C. Liu et al., "Genetic polymorphisms of cytochrome p450 enzymes influence metabolism of the antidepressant escitalopram and treatment response," *Pharmacogenomics*, vol. 11, no. 4, pp. 537–546, 2010.

[97] K. Hodgson, K. Tansey, M. Z. Dernovšek, J. Hauser, N. Henigsberg, W. Maier, O. Mors, A. Placentino, M. Rietschel, D. Souery et al., "Genetic differences in cytochrome P450 enzymes and antidepressant treatment response," *Journal of Psychopharmacology*, vol. 28, no. 2, pp. 133–141, 2014.

[98] E. J. Peters, S. L. Slager, J. B. Kraft, G. D. Jenkins, M. S. Reinalda, P. J. McGrath, and S. P. Hamilton, "Pharmacokinetic genes do not influence response or tolerance to citalopram in the STAR* D sample," *PloS One*, vol. 3, no. 4, p. e1872, 2008.

[99] V. Florio, S. Porcelli, A. Saria, A. Serretti, and A. Conca, "Escitalopram plasma levels and antidepressant response," *European Neuropsychopharmacology*, vol. 27, no. 9, pp. 940–944, 2017.

[100] M. Olfson and S. C. Marcus, "National patterns in antidepressant medication treatment," *Archives of General Psychiatry*, vol. 66, no. 8, pp. 848–856, 2009.

[101] O. Q. Yin, Y.-K. Wing, Y. Cheung, Z.-J. Wang, S.-L. Lam, H. F. Chiu, and M. S. Chow, "Phenotype-genotype relationship and clinical effects of citalopram in Chinese patients," *Journal of Clinical Psychopharmacology*, vol. 26, no. 4, pp. 367–372, 2006.

[102] R. M. Bagby, A. G. Ryder, and C. Cristi, "Psychosocial and clinical predictors of response to pharmacotherapy for depression," *Journal of Psychiatry and Neuroscience*, vol. 27, no. 4, p. 250, July 2002.

[103] F. A. Jain, A. M. Hunter, J. O. Brooks, and A. F. Leuchter, "Predictive socioeconomic and clinical profiles of antidepressant response and remission," *Depression and Anxiety*, vol. 30, no. 7, pp. 624–630, July 2013.

[104] R. T. Mulder, P. R. Joyce, C. M. Frampton, S. E. Luty, and P. F. Sullivan, "Six months of treatment for depression: Outcome and predictors of the course of illness," *American Journal of Psychiatry*, vol. 163, no. 1, pp. 95–100, 2006.

[105] H. Zhu, M. B. Bogdanov, S. H. Boyle, W. Matson, S. Sharma, S. Matson, E. Churchill, O. Fiehn, J. A. Rush, R. R. Krishnan et al., "Pharmacometabolomics of response to sertraline and to placebo in major depressive disorder–possible role for methoxyindole pathway," *PLoS One*, vol. 8, no. 7, p. e68283, 2013.

[106] R. Kaddurah-Daouk, M. Bogdanov, W. Wikoff, H. Zhu, S. Boyle, E. Churchill, Z. Wang, A. Rush, R. Krishnan, E. Pickering et al., "Pharmacometabolomic mapping of early biochemical changes induced by sertraline and placebo," *Translational Psychiatry*, vol. 3, no. 1, p. e223, 2013.

[107] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2001.

[108] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, p. 1, 2010.

[109] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, February 2002.

[110] J. Friedman, T. Hastie, and R. Tibshirani, *The Elements of Statistical Learning*. Springer series in statistics Springer, Berlin, 2009, vol. 2.

[111] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 281–288, February 2009.

[112] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002.

[113] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: An overview of methods, challenges, and prospects," *Proceedings of The IEEE*, vol. 103, no. 9, pp. 1449–1477, September 2015.

[114] B. Ray, M. Henaff, S. Ma, E. Efstathiadis, E. R. Peskin, M. Picone, T. Poli, C. F. Aliferis, and A. Statnikov, "Information content and analysis methods for multi-modal high-throughput biomedical data," *Scientific Reports*, vol. 4, p. 4411, March 2014.

[115] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "–omic and electronic health record big data analytics for precision medicine," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 263–273, February 2017.

[116] J. J. Schildkraut, "Neuropsychopharmacology and the affective disorders," *New England Journal of Medicine*, vol. 281, no. 6, pp. 302–308, August 1969.

[117] J. Axelrod and R. Weinshilboum, "Catecholamines," *New England Journal of Medicine*, vol. 287, no. 5, pp. 237–242, August 1972.

[118] M. D. Ritchie, E. R. Holzinger, R. Li, S. A. Pendergrass, and D. Kim, "Methods of integrating data to uncover genotype-phenotype interactions," *Nature Reviews Genetics*, vol. 16, no. 2, pp. 85–97, Feb 2015.

[119] M. Bersanelli, E. Mosca, D. Remondini, E. Giampieri, C. Sala, G. Castellani, and L. Milanesi, "Methods for the integration of multi-omics data: Mathematical aspects," *BMC Bioinformatics*, vol. 17, no. Suppl 2, p. 15, December 2016.

[120] M. H. Trivedi, M. Fava, S. R. Wisniewski, M. E. Thase, F. Quitkin, D. Warden, L. Ritz, A. A. Nierenberg, B. D. Lebowitz, M. M. Biggs et al., "Medication augmentation after the failure of SSRIs for depression," *New England Journal of Medicine*, vol. 354, no. 12, pp. 1243–1252, March 2006.

[121] R. Hirschfeld, J. M. Russell, P. L. Delgado, J. Fawcett, R. A. Friedman, W. M. Harrison, L. M. Koran, I. W. Miller, M. E. Thase, R. H. Howland et al., "Predictors of response to acute treatment of chronic and double depression with sertraline or imipramine," *The Journal of Clinical Psychiatry*, vol. 59, pp. 669–675, July 1998.

[122] A. C. Altamura, C. Montresor, D. Salvadori, and E. Mundo, "Does comorbid subthreshold anxiety affect clinical presentation and treatment response in depression? A preliminary 12-month naturalistic study," *International Journal of Neuropsychopharmacology*, vol. 7, no. 4, pp. 481–487, December 2004.

[123] K. Martinowich, D. Jimenez, C. Zarate, and H. Manji, "Rapid antidepressant effects: Moving right along," *Molecular Psychiatry*, vol. 18, no. 8, pp. 856–863, August 2013.

[124] A. Athreya, R. Iyer, D. Neavin, L. Wang, R. Weinshilboum, R. Kaddurah-Daouk, J. Rush, M. Frye, and W. Bobo, "Augmentation of physician assessments with multi-omics enhances predictability of drug response: A case study of major depressive disorder," *IEEE Computational Intelligence Magazine*, vol. 13, no. 3, pp. 20–31, 2018.

[125] C. A. Webb, M. H. Trivedi, Z. D. Cohen, D. G. Dillon, J. C. Fournier, F. Goer, M. Fava, P. J. McGrath, M. Weissman, R. Parsey et al., "Personalized prediction of antidepressant v. placebo response: Evidence from the EMBARC study," *Psychological Medicine*, pp. 1–10, 2018.

[126] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[127] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[128] T. M. Ball, M. B. Stein, H. J. Ramsawh, L. Campbell-Sills, and M. P. Paulus, "Single-subject anxiety treatment outcome prediction using functional neuroimaging," *Neuropsychopharmacology*, vol. 39, no. 5, p. 1254, 2014.

[129] S. Haase, A. Haghikia, N. Wilck, D. N. Müller, and R. A. Linker, "Impacts of microbiome metabolites on immune regulation and autoimmunity," *Immunology*, vol. 154, no. 2, pp. 230–238, 2018.

[130] K. V.-A. Johnson and K. R. Foster, "Why does the microbiome affect behaviour?" *Nature Reviews Microbiology*, p. 1, 2018.

[131] S. G. Kornstein, A. F. Schatzberg, M. E. Thase, K. A. Yonkers, J. P. McCullough, G. I. Keitner, A. J. Gelenberg, S. M. Davis, W. M. Harrison, and M. B. Keller, "Gender differences in treatment response to sertraline versus imipramine in chronic depression," *American Journal of Psychiatry*, vol. 157, no. 9, pp. 1445–1452, 2000.

[132] E. A. Young, S. G. Kornstein, S. M. Marcus, A. T. Harvey, D. Warden, S. R. Wisniewski, G. Balasubramani, M. Fava, M. H. Trivedi, and A. J. Rush, "Sex differences in response to citalopram: A STAR* D report," *Journal of Psychiatric Research*, vol. 43, no. 5, pp. 503–511, 2009.

[133] L. A. Martin, H. W. Neighbors, and D. M. Griffith, "The experience of symptoms of depression in men vs women: Analysis of the national comorbidity survey replication," *JAMA Psychiatry*, vol. 70, no. 10, pp. 1100–1106, 2013.

[134] F. A. Jain, A. M. Hunter, J. O. Brooks III, and A. F. Leuchter, "Predictive socioeconomic and clinical profiles of antidepressant response and remission," *Depression and Anxiety*, vol. 30, no. 7, pp. 624–630, 2013.

[135] M. Fava, A. J. Rush, J. E. Alpert, G. Balasubramani, S. R. Wisniewski, C. N. Carmin, M. M. Biggs, S. Zisook, A. Leuchter, R. Howland et al., "Difference in treatment outcome in outpatients with anxious versus nonanxious depression: A STAR* D report," *American Journal of Psychiatry*, vol. 165, no. 3, pp. 342–351, 2008.

[136] J. D. Clapp, A. L. Grubaugh, J. G. Allen, J. Mahoney, J. M. Oldham, J. C. Fowler, T. Ellis, J. D. Elhai, and B. C. Frueh, "Modeling trajectory of depressive symptoms among psychiatric inpatients: A latent growth curve approach," *The Journal of Clinical Psychiatry*, vol. 74, no. 5, p. 492, 2013.

[137] S. F. Smagula, M. A. Butters, S. J. Anderson, E. J. Lenze, M. A. Dew, B. H. Mulsant, F. E. Lotrich, H. Aizenstein, and C. F. Reynolds, "Antidepressant response trajectories and associated clinical prognostic factors among older adults," *JAMA Psychiatry*, vol. 72, no. 10, pp. 1021–1028, 2015.

[138] W. Maier and M. Philipp, "Improving the assessment of severity of depressive states: A reduction of the hamilton depression scale," *Pharmacopsychiatry*, vol. 18, no. 01, pp. 114–115, 1985.

[139] P. Bech, "Rating scales for affective disorders: Their validity and consistency." *Acta Psychiatrica Scandinavica*, 1981.

[140] R. McIntyre, S. Kennedy, R. M. Bagby, and D. Bakish, "Assessing full remission," *Journal of Psychiatry and Neuroscience*, vol. 27, no. 4, p. 235, 2002.

[141] N. De La Garza, A. John Rush, B. D. Grannemann, and M. H. Trivedi, "Toward a very brief self-report to assess the core symptoms of depression (vqids-sr5)," *Acta Psychiatrica Scandinavica*, vol. 135, no. 6, pp. 548–553, 2017.

[142] M. J. Taylor, N. Freemantle, J. R. Geddes, and Z. Bhagwagar, "Early onset of selective serotonin reuptake inhibitor antidepressant action: Systematic review and meta-analysis," *Archives of General Psychiatry*, vol. 63, no. 11, pp. 1217–1223, 2006.

[143] A. Szegedi, W. T. Jansen, A. P. van Willigenburg, E. van der Meulen, H. H. Stassen, and M. E. Thase, "Early improvement in the first 2 weeks as a predictor of treatment outcome in patients with major depressive disorder: A meta-analysis including 6562 patients." *The Journal of Clinical Psychiatry*, 2009.

[144] M. A. Posternak, L. Baer, A. A. Nierenberg, and M. Fava, "Response rates to fluoxetine in subjects who initially show no improvement." *The Journal of Clinical Psychiatry*, vol. 72, no. 7, pp. 949–954, 2011.

[145] N. I. for Health and C. Excellence, "Depression in adults: Recognition and management. nice guideline (cg90)," 2009.

[146] F. M. Quitkin, P. J. McGrath, J. W. Stewart, K. Ocepek-Welikson, B. P. Taylor, E. Nunes, D. Deliyannides, V. Agosti, S. J. Donovan, E. Petkova et al., "Chronological milestones to guide drug change: When should clinicians switch antidepressants?" *Archives of General Psychiatry*, vol. 53, no. 9, pp. 785–792, 1996.

[147] M. Trivedi, D. W. Morris, B. D. Grannemann, and S. Mahadi, "Symptom clusters as predictors of late response to antidepressant treatment." *The Journal of Clinical Psychiatry*, 2005.

[148] A. P. Association et al., "Practice guideline for the treatment of patients with major depressive disorder. 2010," *Washington, DC: American Psychiatric Association*, 2017.

[149] S. J. Fredman, M. Fava, A. S. Kienke, C. N. White, A. A. Nierenberg, and J. F. Rosenbaum, "Partial response, nonresponse, and relapse with selective serotonin reuptake inhibitors in major depression: a survey of current "next-step" practices," *Journal of Clinical Psychiatry*, vol. 61, no. 6, pp. 403–408, 2000.

[150] S. H. Kennedy, R. W. Lam, R. S. McIntyre, S. V. Tourjman, V. Bhat, P. Blier, M. Hasnain, F. Jollant, A. J. Levitt, G. M. MacQueen et al., "Canadian network for mood and anxiety treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: Section 3. Pharmacological treatments," *The Canadian Journal of Psychiatry*, vol. 61, no. 9, pp. 540–560, 2016.

[151] R. S. McIntyre, "When should you move beyond first-line therapy for depression?" *The Journal of Clinical Psychiatry*, vol. 71, pp. 16–20, 2010.

[152] D. Mischoulon, A. A. Nierenberg, L. Kizilbash, J. F. Rosenbaum, and M. Fava, "Strategies for managing depression refractory to selective serotonin reuptake inhibitor treatment: A survey of clinicians," *The Canadian Journal of Psychiatry*, vol. 45, no. 5, pp. 476–481, 2000.

[153] A. Cleare, C. M. Pariante, A. H. Young, I. M. Anderson, D. Christmas, P. J. Cowen, C. Dickens, I. Ferrier, J. Geddes, S. Gilbody et al., "Evidence-based guidelines for treating depressive disorders with antidepressants: A revision of the 2008 british association for psychopharmacology guidelines," *Journal of Psychopharmacology*, vol. 29, no. 5, pp. 459–525, 2015.

[154] P. A. Kudlow, R. S. McIntyre, and R. W. Lam, "Early switching strategies in antidepressant non-responders: Current evidence and future research directions," *CNS Drugs*, vol. 28, no. 7, pp. 601–609, 2014.

[155] E. I. Fried, "Are more responsive depression scales really superior depression scales?" *Journal of Clinical Epidemiology*, vol. 77, pp. 4–6, 2016.

[156] T. Ali-Sisto, T. Tolmunen, E. Toffol, H. Viinamäki, P. Mäntyselkä, M. Valkonen-Korhonen, K. Honkalampi, A. Ruusunen, V. Velagapudi, and S. M. Lehto, "Purine metabolism is dysregulated in patients with major depressive disorder," *Psychoneuroendocrinology*, vol. 70, pp. 25–32, 2016.

[157] J. T. Arnedt, L. M. Swanson, R. R. Dopp, H. S. Bertram, A. J. Mooney, E. D. Huntley, R. F. Hoffmann, and R. Armitage, "Effects of restricted time in bed on antidepressant treatment response: A randomized controlled trial," *The Journal of Clinical Psychiatry*, vol. 77, no. 10, p. e1218, 2016.

[158] M. A. Dew, C. F. Reynolds, P. R. Houck, M. Hall, D. J. Buysse, E. Frank, and D. J. Kupfer, "Temporal profiles of the course of depression during treatment: Predictors of pathways toward recovery in the elderly," *Archives of General Psychiatry*, vol. 54, no. 11, pp. 1016–1024, 1997.

[159] M. Gilthorpe, D. Dahly, Y.-K. Tu, L. Kubzansky, and E. Goodman, "Challenges in modelling the random structure correctly in growth mixture models and the impact this has on model mixtures," *Journal of Developmental Origins of Health and Disease*, vol. 5, no. 3, pp. 197–205, 2014.

[160] Y.-K. Tu, K. Tilling, J. A. Sterne, and M. S. Gilthorpe, "A critical evaluation of statistical approaches to examining the role of growth trajectories in the developmental origins of health and disease," *International Journal of Epidemiology*, vol. 42, no. 5, pp. 1327–1339, 2013.

[161] A. P. Athreya, S. S. Banerjee, D. Neavin, R. Kaddurah-Daouk, A. J. Rush, M. A. Frye, L. Wang, R. M. Weinshilboum, W. V. Bobo, and R. K. Iyer, "Data-driven longitudinal modeling and prediction of symptom dynamics in major depressive disorder: Integrating factor graphs and learning methods," in *Proceedings of The 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, August 2017, pp. 1–9.

[162] B. Bao, Z. Wang, S. Ali, A. Ahmad, A. S. Azmi, S. H. Sarkar, S. Banerjee, D. Kong, Y. Li, S. Thakur et al., "Metformin inhibits cell proliferation, migration and invasion by attenuating csc function mediated by deregulating mirnas in pancreatic cancer cells," *Cancer Prevention Research*, vol. 5, no. 3, pp. 355–364, 2012.

[163] H. A. Hirsch, D. Iliopoulos, and K. Struhl, "Metformin inhibits the inflammatory response associated with cellular transformation and cancer stem cell growth," *Proceedings of The National Academy of Sciences*, vol. 110, no. 3, pp. 972–977, 2013.

[164] C. A. Hudis and L. Gianni, "Triple-negative breast cancer: An unmet medical need," *The Oncologist*, vol. 16, pp. 1–11, 2011.

[165] B. C. Litzenburger and P. H. Brown, "Advances in preventive therapy for estrogen-receptor-negative breast cancer," *Current Breast Cancer Reports*, vol. 6, no. 2, pp. 96–109, 2014.

[166] T. Griss, E. E. Vincent, R. Egnatchik, J. Chen, E. H. Ma, B. Faubert, B. Viollet, R. J. DeBerardinis, and R. G. Jones, "Metformin antagonizes cancer cell proliferation by suppressing mitochondrial-dependent biosynthesis," *PLoS Biology*, vol. 13, no. 12, p. e1002309, 2015.

[167] M. Cazzaniga and B. Bonanni, "Breast cancer metabolism and mitochondrial activity: The possibility of chemoprevention with metformin," *BioMed Research International*, vol. 2015, 2015.

[168] S. E. Weinberg and N. S. Chandel, "Targeting mitochondria metabolism for cancer therapy," *Nature Chemical Biology*, vol. 11, no. 1, pp. 9–15, 2015.

[169] M. Van Gisbergen, A. Voets, M. Starmans, I. de Coo, R. Yadak, R. Hoffmann, P. Boutros, H. Smeets, L. Dubois, and P. Lambin, "How do changes in the mtdna and mitochondrial dysfunction influence cancer and cancer therapy? Challenges, opportunities and models," *Mutation Research/Reviews in Mutation Research*, vol. 764, pp. 16–30, 2015.

[170] J.-W. Min, W. J. Kim, J. A. Han, Y.-J. Jung, K.-T. Kim, W.-Y. Park, H.-O. Lee, and S. S. Choi, "Identification of distinct tumor subpopulations in lung adenocarcinoma via single-cell RNA-seq," *PloS One*, vol. 10, no. 8, p. e0135817, 2015.

[171] F. Buettner, K. N. Natarajan, F. P. Casale, V. Proserpio, A. Scialdone, F. J. Theis, S. A. Teichmann, J. C. Marioni, and O. Stegle, "Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells," *Nature Biotechnology*, vol. 33, no. 2, pp. 155–160, 2015.

[172] J.-W. Min, W. J. Kim, J. A. Han, Y.-J. Jung, K.-T. Kim, W.-Y. Park, H.-O. Lee, and S. S. Choi, "Identification of distinct tumor subpopulations in lung adenocarcinoma via single-cell RNA-seq," *PloS One*, vol. 10, no. 8, p. e0135817, 2015.

[173] L. Haghverdi, F. Buettner, and F. J. Theis, "Diffusion maps for high-dimensional single-cell analysis of differentiation data," *Bioinformatics*, vol. 31, no. 18, pp. 2989–2998, 2015.

[174] R. Suzuki and H. Shimodaira, "Pvclust: An R package for assessing the uncertainty in hierarchical clustering," *Bioinformatics*, vol. 22, no. 12, pp. 1540–1542, 2006.

[175] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[176] J. M. Irish, R. Hovland, P. O. Krutzik, O. D. Perez, Ø. Bruserud, B. T. Gjertsen, and G. P. Nolan, "Single cell profiling of potentiated phospho-protein networks in cancer cells," *Cell*, vol. 118, no. 2, pp. 217–228, 2004.

[177] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan et al., "Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells," *Nature Structural & Molecular Biology*, vol. 20, no. 9, pp. 1131–1139, 2013.

[178] D. Ramsköld, S. Luo, Y.-C. Wang, R. Li, Q. Deng, O. R. Faridani, G. A. Daniels, I. Khrebtukova, J. F. Loring, L. C. Laurent et al., "Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells," *Nature Biotechnology*, vol. 30, no. 8, pp. 777–782, 2012.

[179] K. R. Kalari, A. A. Nair, J. D. Bhavsar, D. R. OBrien, J. I. Davila, M. A. Bockol, J. Nie, X. Tang, S. Baheti, J. B. Doughty et al., "MAP-RSeq: Mayo analysis pipeline for RNA sequencing," *BMC Bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[180] S. Wold, K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, pp. 37–52, 1987.

[181] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, no. 2579-2605, p. 85, 2008.

[182] N. K. Wilson, D. G. Kent, F. Buettner, M. Shehata, I. C. Macaulay, F. J. Calero-Nieto, M. S. Castillo, C. A. Oedekoven, E. Diamanti, R. Schulte et al., "Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations," *Cell Stem Cell*, vol. 16, no. 6, pp. 712–724, 2015.

[183] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[184] C. Fraley and A. E. Raftery, "MCLUST: Software for model-based cluster analysis," *Journal of Classification*, vol. 16, no. 2, pp. 297–306, 1999.

[185] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[186] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised Learning," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[187] L.-D. Li, H.-F. Sun, X.-X. Liu, S.-P. Gao, H.-L. Jiang, X. Hu, and W. Jin, "Down-regulation of NDUFB9 promotes breast cancer cell proliferation, metastasis by mediating mitochondrial metabolism," *PloS One*, p. e0144441, 2015.

[188] S.-P. Gao, H.-F. Sun, H.-L. Jiang, L.-D. Li, X. Hu, X.-E. Xu, and W. Jin, "Loss of COX5B inhibits proliferation and promotes senescence via mitochondrial dysfunction in breast cancer," *Oncotarget*, vol. 6, no. 41, pp. 43 363–43 374, 2015.

[189] F. Sotgia, D. Whitaker-Menezes, U. E. Martinez-Outschoorn, A. F. Salem, A. Tsirigos, R. Lamb, S. Sneddon, J. Hulit, A. Howell, and M. P. Lisanti, "Mitochondria "fuel" breast cancer metabolism: Fifteen markers of mitochondrial biogenesis label epithelial cancer cells, but are excluded from adjacent stromal cells," *Cell Cycle*, vol. 11, no. 23, pp. 4390–4401, 2012.

[190] S. De Flora, R. Balansky, F. D'agostini, C. Cartiglia, M. Longobardi, V. E. Steele, and A. Izzotti, "Smoke-induced microRNA and related proteome alterations. modulation by chemopreventive agents," *International Journal of Cancer*, pp. 2763–2773, 2012.

[191] R. Weinberg, *The Biology of Cancer*. Garland Science, 2013.

[192] M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen et al., "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PloS One*, p. e1651, 2008.

[193] C. G. A. R. Network et al., "Comprehensive molecular profiling of lung adenocarcinoma," *Nature*, pp. 543–550, 2014.

[194] E. Y. Lee and W. J. Muller, "Oncogenes and tumor suppressor genes," *Cold Spring Harbor Perspectives in Biology*, p. a003236, 2010.

[195] N. Howlader, A.-M. Noone, M. Krapcho, J. Garshell, D. Miller, S. Altekruse, C. Kosary, M. Yu, J. Ruhl, Z. Tatalovich et al., "SEER Cancer Statistics Review, 1975-2013, National Cancer Institute, Bethesda, MD," 2016.

[196] J. M. Smith, *Evolution and the Theory of Games*. Cambridge university press, 1982.

[197] M. A. Nowak, *Evolutionary Dynamics*. Harvard University Press, 2006.

[198] F. Michor, Y. Iwasa, and M. A. Nowak, "Dynamics of cancer progression," *Nature Reviews Cancer*, vol. 4, no. 3, pp. 197–205, 2004.

[199] L. Yu, N. W. Todd, L. Xing, Y. Xie, H. Zhang, Z. Liu, H. Fang, J. Zhang, R. L. Katz, and F. Jiang, "Early detection of lung adenocarcinoma in sputum by a panel of microRNA markers," *International Journal of Cancer*, pp. 2870–2878, 2010.

[200] T. Basar and G. J. Olsder, *Dynamic noncooperative game theory*. SIAM, 1995, vol. 200.

[201] S. Devarakonda, D. Morgensztern, and R. Govindan, "Genomic alterations in lung adenocarcinoma," *The Lancet Oncology*, pp. e342–e351, 2015.

[202] A. Drilon, H. Sugita, C. S. Sima, M. Zauderer, C. M. Rudin, M. G. Kris, V. W. Rusch, and C. G. Azzoli, "A prospective study of tumor suppressor gene methylation as a prognostic biomarker in surgically resected stage I to IIIA non–small-cell lung cancers," *Journal of Thoracic Oncology*, pp. 1272–1277, 2014.

[203] J. F. Nash et al., "Equilibrium points in n-person games," *Proceedings of The National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[204] D. G. Beer, S. L. Kardia, C.-C. Huang, T. J. Giordano, A. M. Levin, D. E. Misek, L. Lin, G. Chen, T. G. Gharib, D. G. Thomas et al., "Gene-expression profiles predict survival of patients with lung adenocarcinoma," *Nature Medicine*, pp. 816–824, 2002.

[205] P. Cao, E. Badger, Z. Kalbarczyk, R. Iyer, and A. Slagell, "Preemptive intrusion detection: Theoretical framework and real-world measurements," in *Proceedings of The 2015 Symposium and Bootcamp on the Science of Security*. ACM, 2015, p. 5.

[206] "Centers for medicaid and services," https://www.cms.gov/Medicare/E-health/EHealthRecords/index.html, accessed: 2016-10-10.

[207] X. Wang, F. Wang, and J. Hu, "A multi-task learning framework for joint disease risk prediction and comorbidity discovery," in *Proceedings of The 22nd International Conference on Pattern Recognition (ICPR)*. IEEE, 2014, pp. 220–225.

[208] H. Wu, C. Cheng, X. Han, Y. Huo, W. Ding, and M. D. Wang, "Post-surgical complication prediction in the presence of low-rank missing data," in *Proceedings of The 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 6808–6811.

[209] Y. Wang, K. Ng, R. J. Byrd, J. Hu, S. Ebadollahi, Z. Daar, S. R. Steinhubl, W. F. Stewart et al., "Early detection of heart failure with varying prediction windows by structured and unstructured data in electronic health records," in *Proceedings of The 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2015, pp. 2530–2533.

[210] M. A. Vedomske, D. E. Brown, and J. H. Harrison, "Random forests on ubiquitous data for heart failure 30-day readmissions prediction," in *12th International Conference on Machine Learning and Applications (ICMLA)*, vol. 2. IEEE, 2013, pp. 415–421.

[211] S. Karnik, S. L. Tan, B. Berg, I. Glurich, J. Zhang, H. J. Vidaillet, C. D. Page, and R. Chowdhary, "Predicting atrial fibrillation and flutter using electronic health records," in *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2012, pp. 5562–5565.

[212] R. Moskovitch, H. Choi, G. Hripcsak, and N. Tatonetti, "Prognosis of clinical outcomes with temporal patterns and experiences with one class feature selection," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 14, no. 3, pp. 555–563, 2017.

[213] R. Harpaz, S. Vilar, W. DuMouchel, H. Salmasian, K. Haerian, N. H. Shah, H. S. Chase, and C. Friedman, "Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions," *Journal of the American Medical Informatics Association*, vol. 20, no. 3, pp. 413–419, 2013.

[214] B. A. O'Mara, T. Byers, and E. Schoenfeld, "Diabetes mellitus and cancer risk: A multisite case-control study," *Journal of Chronic Diseases*, vol. 38, no. 5, pp. 435–441, 1985.

[215] C. La Vecchia, E. Negri, S. Franceschi, B. D'avanzo, and P. Boyle, "A case-control study of diabetes mellitus and cancer risk." *British Journal of Cancer*, vol. 70, no. 5, p. 950, 1994.

[216] "Medicaid.gov," https://www.medicaid.gov/medicaid-chip-program-information/by-topics/data-and-systems/icd-coding/icd.html, accessed: 2016-10-10.