Elsevier Editorial System(tm) for Biological Psychiatry

Manuscript Draft

Manuscript Number:

Title: Prediction of Clinical Outcomes after SSRI Therapy for Major Depressive Disorder Using Clinical and Metabolomics Data: A Data-Driven Machine Learning Approach

Article Type: Archival Report

Keywords: Major depressive disorder (MDD); Selective serotonin reuptake inhibitor (SSRI); citalopram; metabolomics; pharmacometabolomics; clinical outcomes prediction; data-driven analytics; machine learning

Corresponding Author: Mr. ARJUN ATHREYA,

Corresponding Author's Institution: University of Illinois at Urbana-Champaign

First Author: ARJUN ATHREYA

Order of Authors: ARJUN ATHREYA; Drew Neavin; Ditlev Monrad; Mark A Frye; Michelle Skime; John Rush; Wayne Matson; Liewei Wang; Rima Kaddurah-Daouk; Ravishankar Iyer; Richard Weinshilboum; William V Bobo

Abstract: BACKGROUND
There is large variation in drug response in major depressive disorder (MDD). Machine learning methods allow us to study this variation and predict clinical outcomes such as remission/response, using depression severity measures with plasma metabolomic profiles.

METHODS
The Mayo PGRN-AMPS study treated 800 MDD patients with citalopram/escitalopram for 8 weeks, with genomic and blood drug level analyses at baseline, 4 and 8 weeks plus targeted electrochemistry-based metabolomics to measure 31 metabolites in tryptophan, tyrosine, purine and related pathways. Response was defined as 50% reduction in symptom severity and remission involved achieving a QIDS score 5 or a HDRS score 7. Unsupervised (k-means clustering) and supervised (support vector machine) learning methods were used to identify patient subgroups and predict clinical outcomes, respectively.

RESULTS
Metabolomic profiles differed significantly between men and women at all time points. Three distinct clusters were identified in men and women separately at each time point ($p<1.3E-09$), and were validated using STAR*D data ($p>0.8$). Clustering behavior was associated with plasma metabolite concentrations, but not demographic and clinical factors or plasma drug levels. Machine learning prediction accuracies that included plasma metabolite concentrations were statistically significant for response: men-70%, women-88%, $p<0.04$; and remission: men-83%, women-75%, $p< 0.004$. Top predictor metabolites included several known to be associated with antidepressive response, including serotonin and metabolites of kynurenine and catecholamines.

CONCLUSIONS
Metabolites associated with acute SSRI response/remission differed in men and women. Machine learning approaches that include metabolomic measures enable better prediction of outcomes in MDD patients treated with SSRIs.

Suggested Reviewers: John H Krystal
Yale University
john.krystal@yale.edu

Roy Perlis
Harvard University
rperlis@partners.org

Boadie W Dunlop
Emory University
bdunlop@emory.edu

Rachel Tyndale
Center for Addiction and Mental Health
r.tyndale@utoronto.ca

Glenda MacQueen
Univ. of Calgary
gmmacque@ucalgary.ca

Robert Post
George Washington University School of Medicine
robert.post@speakeasy.net

Sid Kennedy
Univ. of Toronto
sidney.kennedy@uhn.ca

Opposed Reviewers:

MAYO CLINIC

200 First Street SW
Rochester, Minnesota 55905
507-284-2511

William V. Bobo, MD, MPH
Professor of Psychiatry
Department of Psychiatry & Psychology
PH 507-284-3143 FX 507-284-4158

19 May 2017

John H. Krystal, MD
Editor-in-Chief, *Biological Psychiatry*

Subj: Submission of manuscript

Dear Dr. Krystal,

Please find attached a manuscript bearing the title, "Prediction of Clinical Outcomes after SSRI Therapy for Major Depressive Disorder Using Clinical and Metabolomics Data: A Data-Driven Machine Learning Approach." It is submitted for your consideration as an archival report in *Biological Psychiatry*.

We applied a data-driven machine learning workflow for predicting response and remission to 8 weeks of SSRI treatment in depressed adults using data from the PGRN-AMPS study. To our knowledge our workflow is the first to utilize metabolomics data in addition to clinical and demographic variables for this purpose. Treatment response trajectories in PGRN-AMPS were classified into discrete outcome clusters that closely approximated widely-accepted definitions of remission and response, and were validated using STAR*D trial data. The predictive accuracies of the classifier that included baseline metabolomic measures ranged from 75%-88%, which is substantially higher than workflows that have been previously developed for predicting antidepressant response, which did not include biological measures. Additionally, the top predictive metabolites included several known to be associated with antidepressive responses, including serotonin, and metabolites of kynurinene and catecholamines. The metabolites associated with acute SSRI response and remission differed in men and women. We believe that our machine learning approach and our findings are both innovative and novel. Based on this work, it would appear that future efforts at predictive modeling of SSRI treatment outcome in depressed persons must account for sex differences and other biological factors in order to optimize predictive capability.

Our submission is original and unpublished. It is not being considered for publication elsewhere. All authors fulfil the journal's criteria for authorship, and have read over and approved the submitted version of the manuscript. We have disclosed any potential sources of conflicted interest.

We thank you for the time taken to review our work. We look forward to hearing from you in the future.

Respectfully,

William V. Bobo, MD, MPH

TITLE: Prediction of Clinical Outcomes after SSRI Therapy for Major Depressive Disorder Using Clinical and Metabolomics Data: A Data-Driven Machine Learning Approach

AUTHORS: Arjun P. Athreya[1], Drew Neavin[2], Ditlev Monrad[3], Mark A. Frye[4], Michelle Skime[4], A. John Rush[5], Wayne R. Matson[6], Liewei Wang[2], Rima Kaddurah-Daouk[5], Ravishankar K. Iyer[1], Richard M. Weinshilboum[2] and William V. Bobo[4]

AFFILIATIONS:
[1] Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL;
[2] Department of Molecular Pharmacology and Experimental Therapeutics, Mayo Clinic College of Medicine and Science, Rochester, MN;
[3] Department of Statistics, University of Illinois at Urbana-Champaign, Urbana, IL;
[4] Department of Psychiatry & Psychology, Mayo Clinic College of Medicine and Science, Rochester, MN;
[5] Department of Psychiatry & Behavioral Sciences, Department of Medicine and Duke Institute of Brain Sciences, Duke University School of Medicine, Durham, NC; Duke-NUS, Singapore.
[6] Department of Systems Biochemistry, Counterpoint Health Solutions Inc., Bedford, MA.

CORRESPONDENCE TO:
William Bobo, M.D., MPH
Department of Psychiatry and Psychology
Mayo Clinic College of Medicine and Science
Mayo Clinic
200 First Street SW, Generose 2A,
Rochester, MN 55905
USA
Tel.: +1 507 255 9412
E-mail: bobo.william@mayo.edu

KEY WORDS: Major depressive disorder (MDD), Selective serotonin reuptake inhibitor (SSRI), citalopram, metabolomics, pharmacometabolomics, outcomes prediction, data-driven analytics, machine learning.

SHORT TITLE: Prediction of Clinical Outcomes in MDD using Metabolomics

CONTENT DETAILS

MAIN TEXT

Word count (without abstract, figures legends, tables, supplementary materials)

Introduction	: 814
Methods	: 924
Results	:1087
Discussion	:1101
Total	**:3926**

Number of Figures: 3

Number of Tables: 2

SUPPLEMENTARY MATERIALS

Number of Supplementary Figures: 6

Number of Supplementary Tables: 5

Number of Supplementary Sections: 2

ABSTRACT

BACKGROUND

There is large variation in drug response in major depressive disorder (MDD). Machine learning methods allow us to study this variation and predict clinical outcomes such as remission/response, using depression severity measures with plasma metabolomic profiles.

METHODS

The Mayo PGRN-AMPS study treated 800 MDD patients with citalopram/escitalopram for 8 weeks, with genomic and blood drug level analyses at baseline, 4 and 8 weeks plus targeted electrochemistry-based metabolomics to measure 31 metabolites in tryptophan, tyrosine, purine and related pathways. Response was defined as 50% reduction in symptom severity and remission involved achieving a QIDS score ≤5 or a HDRS score ≤7. Unsupervised (k-means clustering) and supervised (support vector machine) learning methods were used to identify patient subgroups and predict clinical outcomes, respectively.

RESULTS

Metabolomic profiles differed significantly between men and women at all time points. Three distinct clusters were identified in men and women separately at each time point ($p<1.3E-09$), and were validated using STAR*D data ($p>0.8$). Clustering behavior was associated with plasma metabolite concentrations, but not demographic and clinical factors or plasma drug levels. Machine learning prediction accuracies that included plasma metabolite concentrations were statistically significant for response: men–70%, women–88%, $p<0.04$; and remission: men–83%, women–75%, $p< 0.004$. Top predictor metabolites included several known to be associated with antidepressive response, including serotonin and metabolites of kynurenine and catecholamines.

CONCLUSIONS

Metabolites associated with acute SSRI response/remission differed in men and women. Machine learning approaches that include metabolomic measures enable better prediction of outcomes in MDD patients treated with SSRIs.

INTRODUCTION

Major depressive disorder (MDD) affects over 350 million people worldwide (1, 2) and is a leading cause of disease burden and disability worldwide (3). MDD symptoms can often be managed with appropriate pharmacotherapy and targeted psychotherapy (4-6). However, antidepressant medications such as selective serotonin reuptake inhibitors (SSRIs)—the most widely prescribed first-line pharmacotherapy for MDD in adults—often require several weeks to take effect, and treatment choices are often made on a "trial and error" basis (7). Unfortunately, only about half of MDD patients respond (achieve a 50% reduction in depressive symptoms) to an initial therapeutic trial of SSRIs after 8 weeks of therapy (8), and an even lower proportion of SSRI-treated patients (25%-40%) achieve remission (near-absence of depressive symptoms) (8). Further complicating the wide inter-individual variability in SSRI treatment response is the fact that there are currently no validated biomarkers or other indicators that can be used to predict antidepressant treatment outcomes. Previous research has identified a series of factors associated with worse SSRI treatment outcomes including low socioeconomic status (9); low educational attainment (10); earlier age of depressive illness onset (11); comorbid psychiatric, personality, substance use disorders (12-14); anxious symptoms (15); and longer duration of the current depressive episode (16, 17). However, these factors do not have sufficient predictive validity to guide clinical decision making. There is currently no evidence-based approach to select the antidepressant that is most likely to be of benefit for an individual patient, and underlying molecular mechanisms that drive inter-individual variation in clinical outcomes are poorly understood. A predictive algorithm for SSRI response that includes biological information might help to enhance our understanding of mechanisms responsible for variation in medication effect as well as the underlying pathophysiology of MDD—thus using SSRIs as "molecular probes."

For example, in previous studies, we have used “pharmacometabolomics” to guide and inform “pharmacogenomics” studies to make it possible to identify novel genes and SNPs associated with SSRI response (42). Machine learning methods offer promising, data-driven approaches that can be used in an attempt to identify clinically useful predictors of antidepressant response. Unlike traditional logistic regression modeling approaches that test the independent predictive effects of individual variables on a given outcome, computational learning methods optimize predictions by integrating all available information across a wide range of variables. This type of approach has already been applied to data sets from large randomized trials of MDD antidepressant therapy to help predict therapeutic outcomes (9, 18, 19). For example, Chekroud et.al. used a variety of demographic and clinical variables derived from 1,985 participants in the first phase of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (8) and 476 patients from the Combining Medications to Enhance Depression Outcomes (COMED) trial (20) to predict clinical remission after 12 weeks of citalopram treatment (21). The accuracy of their machine learning model for predicting remission was 65%. This result was statistically significant with an acknowledgement by authors recognizing the absence of biologically-based variables (22) from the predictive algorithms. As a result, in addition to the use of machine learning techniques, we included metabolomic data in the analyses that are described subsequently.

Advances in computational and analytical tool development have enabled the adoption of a systems approach to clinical pharmacology, in part, by enabling the integration of high-dimensional, patient-specific metabolic signature (or “metabotype”) data that may serve as biomarkers for eventual response to antidepressant therapy (23) with other important patient-

specific characteristics and clinical response data. MDD prevalence rates and treatment response patterns vary according to sex, and the metabolome clearly varies between men and women (24) in the general population. Furthermore, previous work has shown that plasma metabolites are also clearly altered in MDD patients as compared to healthy controls (25-32) and specific metabolites have been associated with MDD severity (33) and clinical outcomes after antidepressant treatment (34-37). These findings suggest that the integration of patient-specific metabotype data with other patient-specific clinical measures might enhance clinical outcomes prediction accuracies for machine learning models.

Therefore, we have developed and tested a data-driven predictive model for MDD response and remission after 8 weeks of treatment with citalopram or escitalopram in adults using data from the Pharmacogenomics Research Network-Antidepressant Medical Pharmacogenomic (PGRN-AMPS) study (38). The PGRN-AMPS trial involved both validated clinical assessments and blood sampling for genomic, metabolomic and plasma drug level analyses, which made it possible to construct predictive models that integrated demographic data, clinical response information from standardized questionnaires, biological measures such as metabolomics data for 31 plasma metabolites and plasma drug levels. In the subsequent analysis, we initially observed that the biological profiles of MDD patients differed significantly between men and women and we then found that the addition of biological measures, ie. metabolomics data, significantly enhanced model prediction accuracies. Included among those metabolites, not surprisingly, were plasma compounds like serotonin (5HT), kynurenine also a tryptophan metabolite and MHPG, a major metabolite of norepinephrine in the central nervous system.

METHODS

Data Sources

This study represents a secondary analysis of individual patient data from the PGRN-AMPS study (NCT 00613470), a study that has been described in detail previously (39). PGRN-AMPS was designed to assess the clinical outcomes in adults (aged 18-84 years) with non-psychotic MDD in response to 4 and 8 weeks of open-label treatment with citalopram or escitalopram and to examine metabolomic and genomic factors associated with those outcomes. 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 A, B, and D of the Structured Clinical Interview for DSM-IV (SCID) administered by trained clinical research staff (40). Baseline assessment included structured patient-rated and clinician-administered questionnaires to ascertain demographic variables (age at enrollment, sex, race) and clinical characteristics (body mass index [BMI, kg/m^2], age at onset of first lifetime depressive episode, duration of current depressive episode, recurrent vs. single-episode depressive history, family psychiatric and substance use history, premenstrual (women only) or seasonal depression pattern, current involvement in psychotherapy, recent history of blood transfusions, and lifetime history of bariatric surgery. Blood samples for DNA, metabolomic and plasma drug level assays were obtained at baseline and at weeks 4 and 8.

Data from the initial phase of the STAR*D trial (NCT 00021528) were used to validate depressive symptom response patterns found in the PGRN-AMPS study. Details regarding the inclusion and exclusion criteria for STAR*D have been published previously (8, 41). Briefly, the initial phase of STAR*D consisted of a large, 12-week randomized trial of citalopram for adults

(aged 18-75 years) with MDD conducted in the United States from June 2001 to April 2004. Subjects were recruited from primary and specialty care settings. Specifically, 788 subjects from STAR*D who had complete Phase 1 clinical response data and had provided DNA samples were utilized for the analyses described subsequently (9).

The PGRN-AMPS and STAR*D study protocols were approved by the Institutional Review Boards of the Mayo Clinic, Rochester, MN and STAR*D participating study sites, respectively. All participants in both trials provided written informed consent.

Plasma metabolomic and drug concentration assays

Plasma metabolite concentrations were assayed using samples from 306 randomly selected PGRN-AMPS MDD patients who had samples obtained at baseline and after 4 and 8 weeks of SSRI therapy. As described previously (42), non-Caucasian patients and 10 non-adherent patients (as determined by plasma drug level assays) were excluded—leaving a total of 290 patients. Samples were assayed using a high performance liquid chromatography (HPLC) electrochemical coulometric array (LCECA) platform. Supplementary **Table 1** lists the 31 metabolites assayed and their associated pathways. Plasma drug and drug metabolite concentrations were also assayed in all 4 and 8 week samples using an HPLC MS/MS platform (see **Supplementary material** for details).

Dataset description

We utilized a complete cases approach, wherein only patients who had completed all study visits and provided blood samples for metabolomics data were considered for inclusion in the analysis

(n = 429). Of these patients, 290 had metabolomics data measured at all time points (baseline, week 4, week 8) for metabolite and plasma drug levels (Supplementary Table 1). For the STAR*D dataset, we only considered clinical ratings obtained at baseline and at 4 and 8 weeks for patients who completed Phase 1 of the trial and had complete depressive symptom score data at all three time points.

Clinical outcomes

In both the PGRN-AMPS and STAR*D studies, treatment outcomes were established using the 16-item, clinician-rated version of the Quick Inventory of Depressive Symptomatology (QIDS-C (43)) and the 17-item Hamilton Depression Rating Scale (HDRS (44)). Remission was defined as a QIDS-C score ≤ 5 (43) (HDRS score ≤ 7 (45)) at 4 or 8 weeks. Response was defined as a ≥ 50% reduction of QIDS-C or HDRS scores from baseline to either 4 or 8 weeks.

Analysis Workflow

We used a three-stage workflow of analyses as illustrated schematically in Fig. 1 to establish the predictability of MDD therapeutic outcome.

STAGE-1 – Establish sex differences in metabolomics profiles: We used multivariate analysis of variance (MANOVA) to determine sex differences in metabolite concentrations at baseline, and after 4- and 8 weeks of treatment.

STAGE-2 – Identify depressive symptom severity clusters (Stage 2a), and demographic, clinical, and biological factors that differentiated those clusters (Stage 2b): In **Stage 2a**, we

used unsupervised machine learning (k-means clustering) to identify clusters of patients based only on depression symptom severity, as measured by the QIDS-C and HDRS, using an unbiased approach, at baseline and after 4 or 8 weeks of treatment. In **Stage 2b**, Kolmogorov-Smirnov and two-way Chi-square tests were used to identify associated metabolomic, clinical and demographic factors that differentiated the depression symptom clusters.

STAGE-3 – Develop and test the final predictive algorithm: We used supervised (support vector machine) learning methods as a binary classifier to predict remission and response, using both baseline symptom severity and metabolomics data. Training the classifier required a training dataset with predictor variable data (in this case, baseline metabolomics and symptom severity data) and associated training labels (responders/non-responders or remitters/non-remitters) as inputs. The prediction accuracy of the resulting predictive algorithms was computed by evaluating the fraction of correctly predicted labels in a test dataset. Statistical significance of the prediction performance of the algorithms was established using the null information rate (NIR), which served as a proxy for chance. It should be emphasized once again that this algorithm included both clinical and biological (metabolomics) data as inputs.

Details of the methods used in the workflow shown in Figure 1 are discussed in **Supplementary Sec. 1**.

RESULTS

STAGE-1: Sex Differences in Metabolomics Response

Since our major goal was to integrate biological variables (i.e., metabolomics) joined with clinical and demographic data to predict MDD treatment outcomes, our first step involved the determination of important factors involved in variation of the metabolomics data. We observed significant differences between men and women in plasma concentrations of several metabolites at baseline or at 4 or 8 weeks of SSRI therapy (Table 1), regardless of response/remission status or the scale (QIDS-C or HDRS) used to define these outcomes ($p < 0.001$ from MANOVA). Therefore, as we moved to STAGE-2 in the analysis, we separated the analyses for men and women.

STAGE-2a: Response Profiles Inferred from Unsupervised Learning

We next attempted to identify clusters of patients based only on their QIDS-C or HDRS total scores at baseline and after 4 and 8 weeks, using data from the 603 PGRN-AMPS patients with complete QIDS-C and HDRS data at all three time points. The probability density functions (PDF) of the QIDS-C and HDRS total scores were not normally distributed at any time point (Shapiro-Wilk test, $p < 0.001$), as shown for illustrative purposes in Fig. S1-A for QIDS-C scores for men at baseline. Therefore, we applied an Expectation Maximization (EM) algorithm that assumed only one component in the mixture (a single bell-shaped curve distribution) and gradually increased the number of components (distributions with multiple peaks) until an adequate fit of the data was achieved (Kolmogorov-Smirnov test p-value < 0.05). The best fit

was achieved using an associated PDF made up of 3 components at all 3 time points for men and for women, as illustrated in Fig. S1-B for baseline QIDS-C scores in men.

Using the QIDS-C and HDRS scores at baseline, week 4, and week 8, this approach identified three depressive symptom clusters for both men and women at each time point ($p < 1.3e-09$; Figs. 2A and 2C). For illustrative purposes, the associated density functions for each of the 18 clusters (9 for men and 9 for women) are shown in Fig. S2. The distinct depressive symptom clusters were labeled using the following convention: baseline (A1, A2, A3), 4 weeks (B1, B2, B3), and 8 weeks (C1, C2, C3), wherein the numeral 1 represented mild depressive symptoms, 2 represented moderate depressive symptoms, and 3 represented severe depressive symptoms based on relative differences in mean QIDS-C and HDRS total scores between each of the clusters. Despite our use of an unbiased clustering approach, cluster B1 (mild depressive symptoms at 4 weeks) and cluster C1 (mild depressive symptoms at 8 weeks) contained all MDD patients who had achieved clinical remission. Furthermore, the majority (69%) of patients who achieved "response" but not "remission" were contained within the C2 cluster (moderate symptoms at 8 weeks).

In an attempt to replicate our findings, we applied the same clustering algorithm to 788 STAR*D subjects. As shown in Figs. 2B and 2D, the mean QIDS-C and HDRS total scores within each of the 18 STAR*D clusters were comparable to the 18 clusters in the PGRN-AMPS data set ($p > 0.8$). There was also a very high level of consistency between the PGRN-AMPS and the STAR*D datasets in the proportions of patients who moved between specific clusters during

follow-up. This step provided external validation by using STAR*D data of the algorithm-derived depressive symptom cluster patterns found in the PGRN-AMPS study (46).

STAGE-2b: Correlation of Demographic, Clinical, and Biological Factors with Symptom Response Profiles

To identify factors that might be driving the clustering behavior, we conducted a series of analyses that considered a variety of clinical, demographic, pharmacokinetic, and metabolomic factors (Supplementary Table 2). There were no statistically significant differences ($p > 0.3$) in plasma drug levels (Figs. 3), or in any of the demographic or clinical variables (see, for example, Figs. S3 and S4 for age and BMI respectively), between symptom clusters in men, between symptom clusters in women, or between men and women within a given symptom cluster at any time point. These observations indicated that variation in pharmacokinetic factors, selected demographic and clinical factors, and plasma drug levels could not explain the inter-individual variation in clinical outcomes reflected by the clusters. Only metabolite concentrations in the baseline clusters were significantly correlated ($p < 0.05$) with QIDS-C and HDRS total scores in the C2 cluster, and with both response and remission status at week 8. This finding provided the impetus for testing the usefulness of including only baseline metabolomics data and depressive symptom scores to predict clinical outcomes at 8 weeks, without additional clinical and demographic variables.

STAGE-3: Development and Testing of the Final Predictive Algorithm

The final classifier algorithm was trained separately for men and women and for QIDS-C and HDRS scores. It incorporated the concentrations of all 31 metabolites (see Supplementary Table

1) with depressive symptom scores at baseline to predict response (i.e., cluster C1 or C2) and remission status (cluster C1 only) at 8 weeks. The classifier that used metabolomics and baseline depressive symptoms scores had significantly greater predictive accuracy than chance alone for response (68-72%, $p < 0.01$ vs. chance in men; 80.3-95.8%, p=1.4E-10 in women) and remission (80-87.5%, $p < 0.05$ vs. chance in men; 72.5-78%, p=7.8E-4 in women) at 8 weeks, as listed in Table 2. However, the classifier that incorporated demographic and clinical variables in addition to metabolomics and baseline depressive symptom scores resulted in lower predictive accuracies that were not significantly greater than chance for either response (48-52% in men, 56-57% in women) or remission (52-58.33% in men, 50-57.5% in women) at 8 weeks. Of the individual metabolites, baseline serotonin (5HT) levels were among those with the highest relative contribution to the accuracy of the predictive model for all clinical phenotypes, except for response in women defined by HDRS total scores at 8 weeks, an observation compatible with our previous report (42). Furthermore, at least one metabolite related to the tryptophan pathway—including 5HT—was among the top predictive metabolites for all clinical phenotypes (Table 2). As shown in Table 2, cysteine (CYS) and indole-3-propionic acid (I3PA) levels were also highly predictive of response or remission at 8 weeks in women. However, in addition to 5HT itself, it is of note that the 5HT precursors 5HTP and TRP were included among the "Top Metabolites" listed in Table 2 as well as the typrosine pathway metabolites TYR and MHPG—metabolites related to the biosynthesis of catecholamine. Relative importance of all of the metabolites in prediction using metabolomics and baseline symptom severity data are illustrated graphically in Figs. S5 and S6.

DISCUSSION

Predictive biomarkers that can be used to select effective medications have revolutionized many aspects of modern medicine (47). For example, the presence of specific receptors e.g., estrogen and HER2 receptors can inform therapeutic options for breast cancer patients (48). However, it has proven difficult to identify similar biomarkers for MDD. Therapeutic response to SSRIs, the most frequently prescribed class of antidepressants, is highly variable, and it can take months to assess the effectiveness of these agents. Clearly, the availability of biomarkers to predict antidepressant response effectiveness for a given patient would both enhance our ability to treat those patients and might also provide insight into MDD pathophysiology. Previous attempts to identify predictive clinical, demographic, or biological measures have not been highly successful, due, at least in part, to the clinical and molecular heterogeneity of MDD and antidepressant treatment response (49). Previous metabolomics studies that we performed have provided evidence that metabolites in the tryptophan, tyrosine, methoxyindole and purine pathways were related to either SSRI response, MDD pathophysiology or both (34, 37, 42).

As a result, there has been increasing interest in the use of data-driven, machine-learning approaches to examine the effects of a very large number of variables—alone and in combination—on therapeutic outcomes for antidepressant therapy. The availability of powerful statistical modeling methods has had an enormous impact in numerous biological disciplines (50), including more recent applications in clinical pharmacology (51). The latter has enabled the creation of two important new fields, Quantitative and Systems Pharmacology and pharmacometabolics, that seek to improve the understanding of drug effects across biological

systems and improve treatment outcomes through individualized prescription of therapeutic agents (35). For example, Chekroud et.al's work (21) represented an important step forward by showing that machine-learning techniques applied to data collected via questionnaires administered to participants in two randomized trials of antidepressants for MDD resulted in predictive accuracy significantly greater than that due to chance alone. However, although statistically significant, the predictive accuracy of their model was modest relative to chance. In that report, the authors acknowledged this limitation, and suggested that the extension of their approach by including biomarker data in predictive modeling algorithms might result in enhance predictive ability.

We have extended this earlier work by using a data-driven clustering method to derive a predictive algorithm for SSRI antidepressant response that included data for 31 metabolites, primarily from tryptophan, purine, phenylalanine/tyrosine and cysteine/methionine pathways, joined with demographic and clinical predictor variables. (52-54). Based on known sex differences in the metabolome in the general population (24, 55, 56) and our ultimate goal of integrating metabolomics variables with clinical and demographic data to predict MDD treatment outcomes, we began by first showing that metabolomic profiles of men and women in our study differed significantly. This result provided the rationale for developing response clustering and final response prediction algorithms separately for men and women.

We then employed a data-driven method to cluster patients according to depressive symptom severity at 4 and 8 weeks, as measured by the HDRS and QIDS-C. Even though we took an unbiased, data-driven approach to characterizing the treatment response clusters, the clusters that

corresponded with the greatest improvement in depressive symptoms at 4 and 8 weeks included all subjects who achieved symptomatic remission. In addition, nearly 70% of subjects who achieved response (but not remission) were contained in the clusters associated with moderate improvement in depressive symptoms. Importantly, our data-driven unsupervised learning grouped PGRN-AMPS study participants into clinically meaningful categories, based on changes in mean HDRS and QIDS-C scores, that corresponded with commonly-employed definitions of remission (HDRS$\leq$7, QIDS-C$\leq$5), response ($\geq$50% decrease from baseline) and lack of response. Additionally, the symptom response clusters identified by our approach using PGRN-AMPS data were replicated in an independent SSRI-treated MDD cohort that consisted of participants in the STAR*D study. Taken together, our results showed that our clustering algorithm accurately identified biologically valid clinical outcomes groups across two large and independent samples of SSRI-treated depressed patients, and thus provided a valid platform upon which to test the use of baseline demographic, clinical, and metabolomics measures for predicting clinical outcomes during citalopram or escitalopram treatment.

Accordingly, we used a combination of demographic, clinical, and metabolomic measures at baseline in a data-driven algorithm to predict clinical outcomes (response or remission) during citalopram or escitalopram treatment over 8 weeks. Our decision to utilize baseline metabolomics data was designed to address some of the limitations of MDD treatment predictive algorithms based on clinical and demographic factors alone and on our desire to develop a predictive algorithm (thus directed at whether baseline characteristics can predict 8 week treatment outcomes). In our study, only plasma metabolite concentrations in baseline clusters and not any demographic or clinical variables were significantly associated with response or

remission status at week 8. Blood drug concentrations measured during therapy were also not predictive. In separate predictive models for men and women, the use of plasma metabolite concentrations in addition to baseline depressive symptom severity scores resulted in strikingly high and statistically significant predictive accuracies relative to chance for both response and remission. Particularly reassuring, as noted earlier, was the fact that many of those metabolites mapped to the tryptophan and tyrosine pathways that include the biosynthesis of both 5HT and catecholamines. Adding demographic and clinical variables to these models resulted in much lower predictive accuracies that were not significantly greater than chance for either response or remission.

The fact that baseline serotonin figured prominently in the accuracy of our predictive model fits well with our recent observation that, of all of the metabolites that we measured, serotonin was most highly associated with SSRI outcomes (42). Furthermore, in that recent study, we identified and functionally validated two novel genes associated with plasma serotonin concentrations, *TSPAN5* and *ERICH3* (42). The addition of biological measures such as plasma metabolite concentrations to predictive algorithms for SSRI anti-depressive response represents a step toward increased understanding, not only of mechanisms responsible for variation in SSRI response, but also the underlying pathophysiology of MDD. Future studies of MDD antidepressant response should include additional biological measures together with the application of data-driven methods such as those used in this study to help predict and clarify underlying pathophysiological processes in MDD patients that result in individual differences in antidepressant response. In summary, the power of data-driven analyses enabled the sub-classification of patients into reproducible subgroups and to identify sex-related differences in metabolomic profiles. The addition of biological data, in the present case metabolomics data, to

clinical measures of MDD symptom severity made it possible to better define differences between women and men in SSRI response, and to better predict drug response.

ACKNOWLEDGEMENTS

This material 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; National Science Foundation (NSF) under grants CNS 13-37732, CNS 16-24790 and CNS 16-24615; National Institutes of Health (NIH) under grants U19GM61388, R01GM28157, R01MH108348, RC2GM092729, R24GM078233 and RC2GM092729; and The Mayo Clinic Center for Individualized Medicine. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF and NIH. We thank Jenny Applequist at the Univ. of Illinois at Urbana-Champaign for her help in preparing the manuscript.

FINANCIAL DISCLOSURES

Dr. M.A. Frye has grant support from AssureRx, Mayo Foundation, Myriad, National Institute of Alcohol Abuse and Alcoholism (NIAAA), National Institute of Mental Health (NIMH), Pfizer, consults for Janssen, Mitsubishi Tanabe Pharma Corporation, Myriad, Neuralstem Inc., Otsuka American Pharmaceutical and Sunovion, Teva Pharmaceuticals. Drs. L. Wang and R. M. Weinshilboum Dr. Weinshilboum are co-founders and stockholders in OneOme. Dr. W. V. Bobo's research has been supported by the National Institute of Mental Health, the Agency for Healthcare Quality and Research, and the Mayo Foundation for Medical Education and Research. He has contributed chapters to UpToDate concerning the use of antidepressants and atypical antipsychotic drugs for treating adults with bipolar major depression. In the last 3 years, Dr. Rush has received consulting fees from the American Psychiatric Association, Brain Resource Ltd, Curbstone Consultant LLC., Eli Lilly, Emmes Corp., Holmusk, Liva-Nova,

National Institute of Drug Abuse, Santium Inc., Sunovion, Taj Medical, Takeda USA; speaking fees from Live Nova; royalties from Guilford Publications and the University of Texas Southwestern Medical Center. Dr. Kaddurah-Douk is an inventor on key patents in the field of metabolomics including applications for the study of CNS diseases. The remaining authors have no disclosures to make.

FIGURES

Figure 1: The three-stage analysis workflow for predicting response/remission in patients with MDD treated with citalopram/escitalopram.

Figure 2: Clusters of men (Figs. A, B) and women (Figs. C, D) at baseline, 4 weeks and 8 weeks for the Mayo PGRN-AMPS (Figs. A, C) and STAR*D (Figs. B, D) trials using QIDS-C as a measure of MDD symptom severity. Clusters of men (Fig. E) and women (Fig. F) at the same time points for the Mayo PGRN-AMPS trial using HDRS as a measure of MDD symptom severity. The box plots for each cluster reflect the variability of the cluster's associated symptom severity scores.

Figure 3: Comparison of citalopram and escitalopram plasma drug concentrations between men and women in clusters with comparable symptom severity at 4 weeks (Fig. A) and 8 weeks (Fig. B).

TABLES

Table 1: Mean plasma concentrations of metabolites that were identified to be significantly different in men and women at baseline, week 4, and week 8 based on the STAGE-1 analysis

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 Supplementary Table 1 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.

Table 2: Prediction performance at 8 weeks. Performance metrics in bold text indicate prediction performance with baseline metabolomics and symptom severity data alone in comparison with prediction performance in parenthesis using clinical, demographics data with metabolomics data. Top metabolites[a] of relative importance from the prediction model using only baseline metabolomics and symptom severity data are listed.

Gender	Metric	QIDS-C		HDRS	
		Remission	Response	Remission	Response
Men	Accuracy (%)	**80** (52)	**72** (52)	**87.5** (58.33)	**68** (48)
	p-value (accuracy > NIR)	**0.003** (0.16)	**0.03** (0.72)	**0.001** (0.27)	**0.07** (0.72)
	Sensitivity (%)	**76.2** (76)	**83.33** (33.33)	**75** (50)	**33.33** (16.67)
	Specificity (%)	**83.33** (50)	**61.54** (61.54)	**100** (66.67)	**100** (76.92)
	Top metabolites	AMTRP, I3PA, GTOCO3, 5HT	5HTP, XAN, 5HT, I3PA, HX, CYS	I3PA, 5HT, XANTH, GUANOSINE, AMTRP	5HT, HX, I3PA, 5HTP, XANTH, GTOCO3, CYS
Women	Accuracy (%)	**78** (50)	**80.30** (56.52)	**72.5** (57.5)	**95.83** (56.25)
	p-value (accuracy > NIR)	**7.8E-04** (0.58)	**0.0002** (0.58)	**0.003** (0.21)	**1.4E-10** (0.44)
	Sensitivity (%)	**100** (45.45)	**54** (0)	**90** (35)	**90** (18.18)
	Specificity (%)	**56.52** (54.54)	**100** (100)	**55** (80)	**100** (88.46)
	Top metabolites	5HT, PARAXAN, HGA, VMA, 3OHKY, CYS, THEOPHYLINE, 4HBAC, GUANOSINE	HGA, TRP, MET, 5HT, ATOCO, TYR, MHPG, 4HPLA	4HBAC, CYS, 5HT, HGA, PARAXAN, 3OHKY, I3AA, MHPG, I3PA	4HBAC, MHPG, PARAXAN, 3OHKY, 4HPLA, URIC, HX, THEOPHYLINE, CYS, MET

[a] See Supplementary Table 1 for definitions of abbreviated names for each metabolite.

REFERENCES

1. Kessler RC, Berglund P, Demler O, Jin R, Koretz D, Merikangas KR, et al. (2003): The epidemiology of major depressive disorder: results from the National Comorbidity Survey Replication (NCS-R). *JAMA*. 289:3095-3105.
2. Organization WH (2016): Depression Fact Sheets.
3. Murray CJ, Lopez AD (2013): Measuring the global burden of disease. *N Engl J Med*. 369:448-457.
4. Borges S, Chen YF, Laughren TP, Temple R, Patel HD, David PA, et al. (2014): Review of maintenance trials for major depressive disorder: a 25-year perspective from the US Food and Drug Administration. *J Clin Psychiatry*. 75:205-214.
5. Hansen R, Gaynes B, Thieda P, Gartlehner G, Deveaugh-Geiss A, Krebs E, et al. (2008): Meta-analysis of major depressive disorder relapse and recurrence with second-generation antidepressants. *Psychiatr Serv*. 59:1121-1130.
6. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, et al. (2016): Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments. *Can J Psychiatry*. 61:540-560.
7. Trivedi MH, Fava M, Wisniewski SR, Thase ME, Quitkin F, Warden D, et al. (2006): Medication augmentation after the failure of SSRIs for depression. *N Engl J Med*. 354:1243-1252.
8. Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, et al. (2006): Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. *Am J Psychiatry*. 163:28-40.
9. Jain FA, Hunter AM, Brooks JO, 3rd, Leuchter AF (2013): Predictive socioeconomic and clinical profiles of antidepressant response and remission. *Depress Anxiety*. 30:624-630.
10. Hirschfeld RM, Russell JM, Delgado PL, Fawcett J, Friedman RA, Harrison WM, et al. (1998): Predictors of response to acute treatment of chronic and double depression with sertraline or imipramine. *J Clin Psychiatry*. 59:669-675.
11. Perlis RH, Alpert J, Nierenberg AA, Mischoulon D, Yeung A, Rosenbaum JF, et al. (2003): Clinical and sociodemographic predictors of response to augmentation, or dose increase among depressed outpatients resistant to fluoxetine 20 mg/day. *Acta Psychiatr Scand*. 108:432-438.
12. Bagby RM, Ryder AG, Cristi C (2002): Psychosocial and clinical predictors of response to pharmacotherapy for depression. *J Psychiatry Neurosci*. 27:250-257.
13. Howland RH, Rush AJ, Wisniewski SR, Trivedi MH, Warden D, Fava M, et al. (2009): Concurrent anxiety and substance use disorders among outpatients with major depression: clinical features and effect on treatment outcome. *Drug Alcohol Depend*. 99:248-260.
14. Reich J (2003): The effect of Axis II disorders on the outcome of treatment of anxiety and unipolar depressive disorders: a review. *J Pers Disord*. 17:387-405.
15. Altamura AC, Montresor C, Salvadori D, Mundo E (2004): Does comorbid subthreshold anxiety affect clinical presentation and treatment response in depression? A preliminary 12-month naturalistic study. *Int J Neuropsychopharmacol*. 7:481-487.

16. Bukh JD, Bock C, Vinberg M, Kessing LV (2013): The effect of prolonged duration of untreated depression on antidepressant treatment outcome. *J Affect Disord*. 145:42-48.
17. Ghio L, Gotelli S, Marcenaro M, Amore M, Natta W (2014): Duration of untreated illness and outcomes in unipolar depression: a systematic review and meta-analysis. *J Affect Disord*. 152-154:45-51.
18. Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. (2016): Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *J Psychiatr Res*. 78:94-102.
19. Jakubovski E, Bloch MH (2014): Prognostic subgroups for citalopram response in the STAR*D trial. *J Clin Psychiatry*. 75:738-747.
20. Rush AJ, Trivedi MH, Stewart JW, Nierenberg AA, Fava M, Kurian BT, et al. (2011): Combining medications to enhance depression outcomes (CO-MED): acute and long-term outcomes of a single-blind randomized study. *Am J Psychiatry*. 168:689-701.
21. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. (2016): Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry*. 3:243-250.
22. Schmidt HD, Shelton RC, Duman RS (2011): Functional biomarkers of depression: diagnosis, treatment, and pathophysiology. *Neuropsychopharmacology*. 36:2375-2394.
23. Kaddurah-Daouk R, Weinshilboum RM, Pharmacometabolomics Research N (2014): Pharmacometabolomics: implications for clinical pharmacology and systems pharmacology. *Clin Pharmacol Ther*. 95:154-167.
24. Krumsiek J, Mittelstrass K, Do KT, Stuckler F, Ried J, Adamski J, et al. (2015): Gender-specific pathway differences in the human serum metabolome. *Metabolomics*. 11:1815-1833.
25. Ali-Sisto T, Tolmunen T, Toffol E, Viinamaki H, Mantyselka P, Valkonen-Korhonen M, et al. (2016): Purine metabolism is dysregulated in patients with major depressive disorder. *Psychoneuroendocrinology*. 70:25-32.
26. Liu X, Zheng P, Zhao X, Zhang Y, Hu C, Li J, et al. (2015): Discovery and validation of plasma biomarkers for major depressive disorder classification based on liquid chromatography-mass spectrometry. *J Proteome Res*. 14:2322-2330.
27. Papakostas GI, Shelton RC, Kinrys G, Henry ME, Bakow BR, Lipkin SH, et al. (2013): Assessment of a multi-assay, serum-based biological diagnostic test for major depressive disorder: a pilot and replication study. *Mol Psychiatry*. 18:332-339.
28. Xu HB, Fang L, Hu ZC, Chen YC, Chen JJ, Li FF, et al. (2012): Potential clinical utility of plasma amino acid profiling in the detection of major depressive disorder. *Psychiatry Res*. 200:1054-1057.
29. Zheng H, Zheng P, Zhao L, Jia J, Tang S, Xu P, et al. (2017): Predictive diagnosis of major depression using NMR-based metabolomics and least-squares support vector machine. *Clin Chim Acta*. 464:223-227.
30. Zheng P, Fang Z, Xu XJ, Liu ML, Du X, Zhang X, et al. (2016): Metabolite signature for diagnosing major depressive disorder in peripheral blood mononuclear cells. *J Affect Disord*. 195:75-81.
31. Zheng P, Wang Y, Chen L, Yang D, Meng H, Zhou D, et al. (2013): Identification and validation of urinary metabolite biomarkers for major depressive disorder. *Mol Cell Proteomics*. 12:207-214.

32. Liu Y, Yieh L, Yang T, Drinkenburg W, Peeters P, Steckler T, et al. (2016): Metabolomic biosignature differentiates melancholic depressive patients from healthy controls. *BMC Genomics*. 17:669.
33. Setoyama D, Kato TA, Hashimoto R, Kunugi H, Hattori K, Hayakawa K, et al. (2016): Plasma Metabolites Predict Severity of Depression and Suicidal Ideation in Psychiatric Patients-A Multicenter Pilot Analysis. *PLoS One*. 11:e0165267.
34. Kaddurah-Daouk R, Bogdanov MB, Wikoff WR, Zhu H, Boyle SH, Churchill E, et al. (2013): Pharmacometabolomic mapping of early biochemical changes induced by sertraline and placebo. *Transl Psychiatry*. 3:e223.
35. Rotroff DM, Corum DG, Motsinger-Reif A, Fiehn O, Bottrel N, Drevets WC, et al. (2016): Metabolomic signatures of drug response phenotypes for ketamine and esketamine in subjects with refractory major depressive disorder: new mechanistic insights for rapid acting antidepressants. *Transl Psychiatry*. 6:e894.
36. Woo HI, Chun MR, Yang JS, Lim SW, Kim MJ, Kim SW, et al. (2015): Plasma amino acid profiling in major depressive disorder treated with selective serotonin reuptake inhibitors. *CNS Neurosci Ther*. 21:417-424.
37. Zhu H, Bogdanov MB, Boyle SH, Matson W, Sharma S, Matson S, et al. (2013): Pharmacometabolomics of response to sertraline and to placebo in major depressive disorder - possible role for methoxyindole pathway. *PLoS One*. 8:e68283.
38. Ji Y, Biernacka JM, Hebbring S, Chai Y, Jenkins GD, Batzler A, et al. (2013): Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: genome-wide associations and functional genomics. *Pharmacogenomics J*. 13:456-463.
39. Mrazek DA, Biernacka JM, O'Kane DJ, Black JL, Cunningham JM, Drews MS, et al. (2011): CYP2C19 variation and citalopram response. *Pharmacogenet Genomics*. 21:1-9.
40. First MB, Spitzer, Robert L, Gibbon Miriam, and Williams, Janet B.W (1996): Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV). *American Psychiatric Press Inc*.
41. Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, et al. (2004): Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. *Control Clin Trials*. 25:119-142.
42. Gupta M, Neavin D, Liu D, Biernacka J, Hall-Flavin D, Bobo WV, et al. (2016): TSPAN5, ERICH3 and selective serotonin reuptake inhibitors in major depressive disorder: pharmacometabolomics-informed pharmacogenomics. *Mol Psychiatry*. 21:1717-1725.
43. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, et al. (2003): 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. *Biol Psychiatry*. 54:573-583.
44. Hamilton M (1960): A rating scale for depression. *J Neurol Neurosurg Psychiatry*. 23:56-62.
45. Frank E, Prien RF, Jarrett RB, Keller MB, Kupfer DJ, Lavori PW, et al. (1991): Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. *Arch Gen Psychiatry*. 48:851-855.

46. Rush AJ, Kraemer HC, Sackeim HA, Fava M, Trivedi MH, Frank E, et al. (2006): Report by the ACNP Task Force on response and remission in major depressive disorder. *Neuropsychopharmacology*. 31:1841-1853.
47. Chen R, Snyder M (2012): Systems biology: personalized medicine for the future? *Curr Opin Pharmacol*. 12:623-628.
48. Turner NC, Neven P, Loibl S, Andre F (2016): Advances in the treatment of advanced oestrogen-receptor-positive breast cancer. *Lancet*.
49. Marquand AF, Wolfers T, Mennes M, Buitelaar J, Beckmann CF (2016): Beyond Lumping and Splitting: A Review of Computational Approaches for Stratifying Psychiatric Disorders. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 1:433-447.
50. Winslow RL, Trayanova N, Geman D, Miller MI (2012): Computational medicine: translating models to clinical care. *Sci Transl Med*. 4:158rv111.
51. Androulakis IP (2015): Systems engineering meets quantitative systems pharmacology: from low-level targets to engaging the host defenses. *Wiley Interdiscip Rev Syst Biol Med*. 7:101-112.
52. Park DI, Dournes C, Sillaber I, Uhr M, Asara JM, Gassen NC, et al. (2016): Purine and pyrimidine metabolism: Convergent evidence on chronic antidepressant treatment response in mice and humans. *Sci Rep*. 6:35317.
53. Kaddurah-Daouk R, Boyle SH, Matson W, Sharma S, Matson S, Zhu H, et al. (2011): Pretreatment metabotype as a predictor of response to sertraline or placebo in depressed outpatients: a proof of concept. *Transl Psychiatry*. 1.
54. Renshaw PF, Parow AM, Hirashima F, Ke Y, Moore CM, Frederick Bde B, et al. (2001): Multinuclear magnetic resonance spectroscopy studies of brain purines in major depression. *Am J Psychiatry*. 158:2048-2055.
55. Ishikawa M, Maekawa K, Saito K, Senoo Y, Urata M, Murayama M, et al. (2014): Plasma and serum lipidomics of healthy white adults shows characteristic profiles by subjects' gender and age. *PLoS One*. 9:e91806.
56. Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, Prehn C, et al. (2011): Discovery of sexual dimorphisms in metabolic and genetic biomarkers. *PLoS Genet*. 7:e1002215.

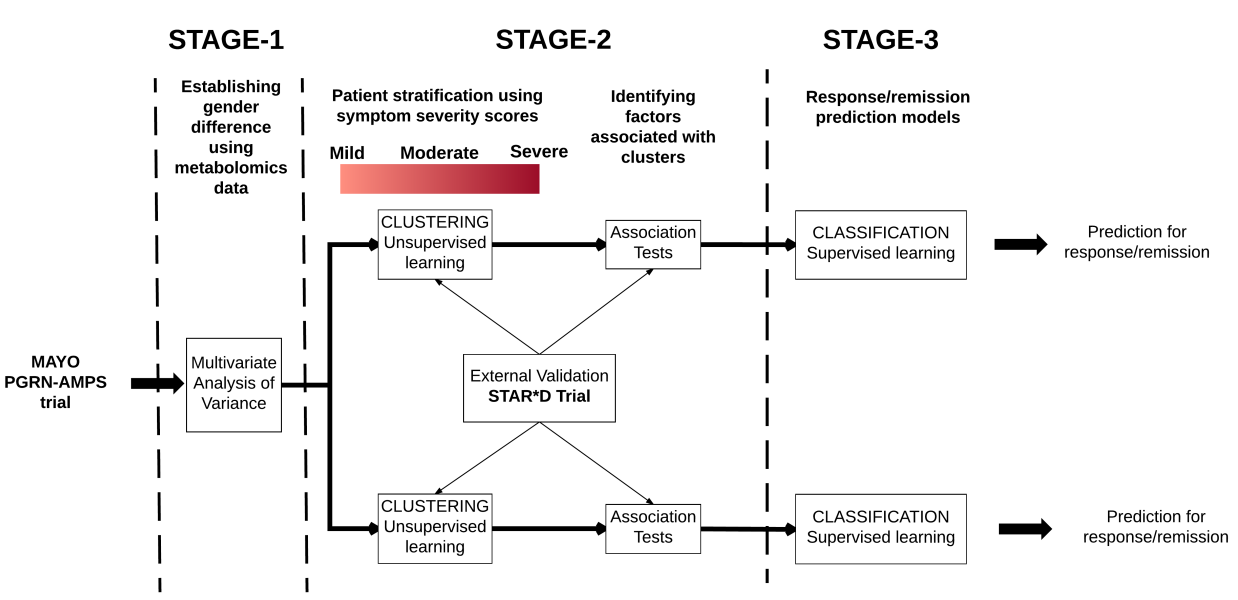

STAGE-1
STAGE-2
STAGE-3
Establishing gender difference using metabolomics data
Patient stratification using symptom severity scores
Mild
Moderate
Severe
Identifying factors associated with clusters
Response/remission prediction models
MAYO PGRN-AMPS trial
Multivariate Analysis of Variance
CLUSTERING Unsupervised learning
Association Tests
CLASSIFICATION Supervised learning
Prediction for response/remission
External Validation STAR*D Trial
CLUSTERING Unsupervised learning
Association Tests
CLASSIFICATION Supervised learning
Prediction for response/remission

Fig. 1

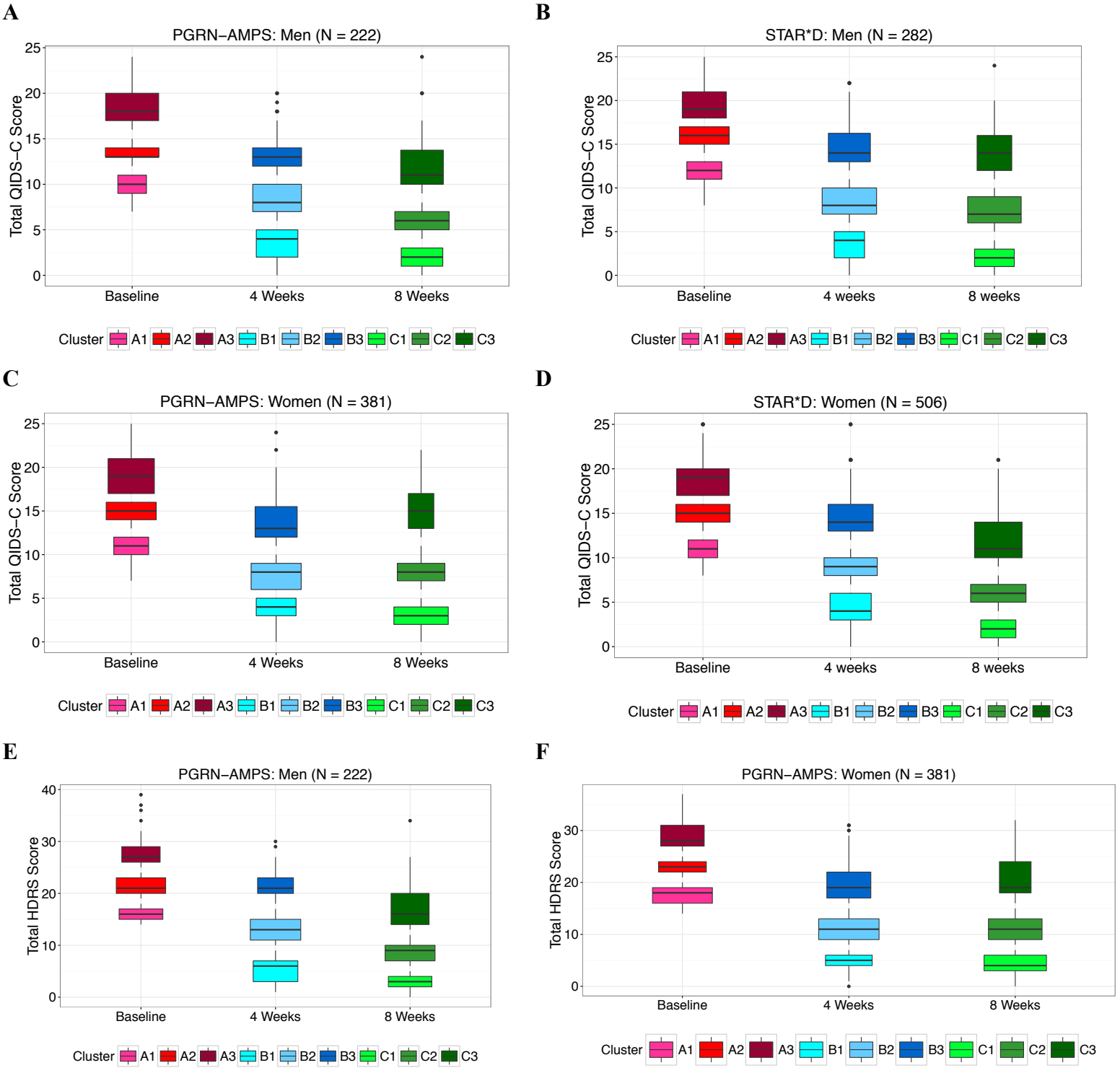

A
PGRN–AMPS: Men (N = 222)
Total QIDS–C Score
Baseline
4 Weeks
8 Weeks
B
STAR*D: Men (N = 282)
Total QIDS–C Score
Baseline
4 weeks
8 weeks
C
PGRN–AMPS: Women (N = 381)
Total QIDS–C Score
Baseline
4 Weeks
8 Weeks
D
STAR*D: Women (N = 506)
Total QIDS–C Score
Baseline
4 weeks
8 weeks
E
PGRN–AMPS: Men (N = 222)
Total HDRS Score
Baseline
4 Weeks
8 Weeks
F
PGRN–AMPS: Women (N = 381)
Total HDRS Score
Baseline
4 Weeks
8 Weeks
Cluster A1 A2 A3 B1 B2 B3 C1 C2 C3

Fig. 2

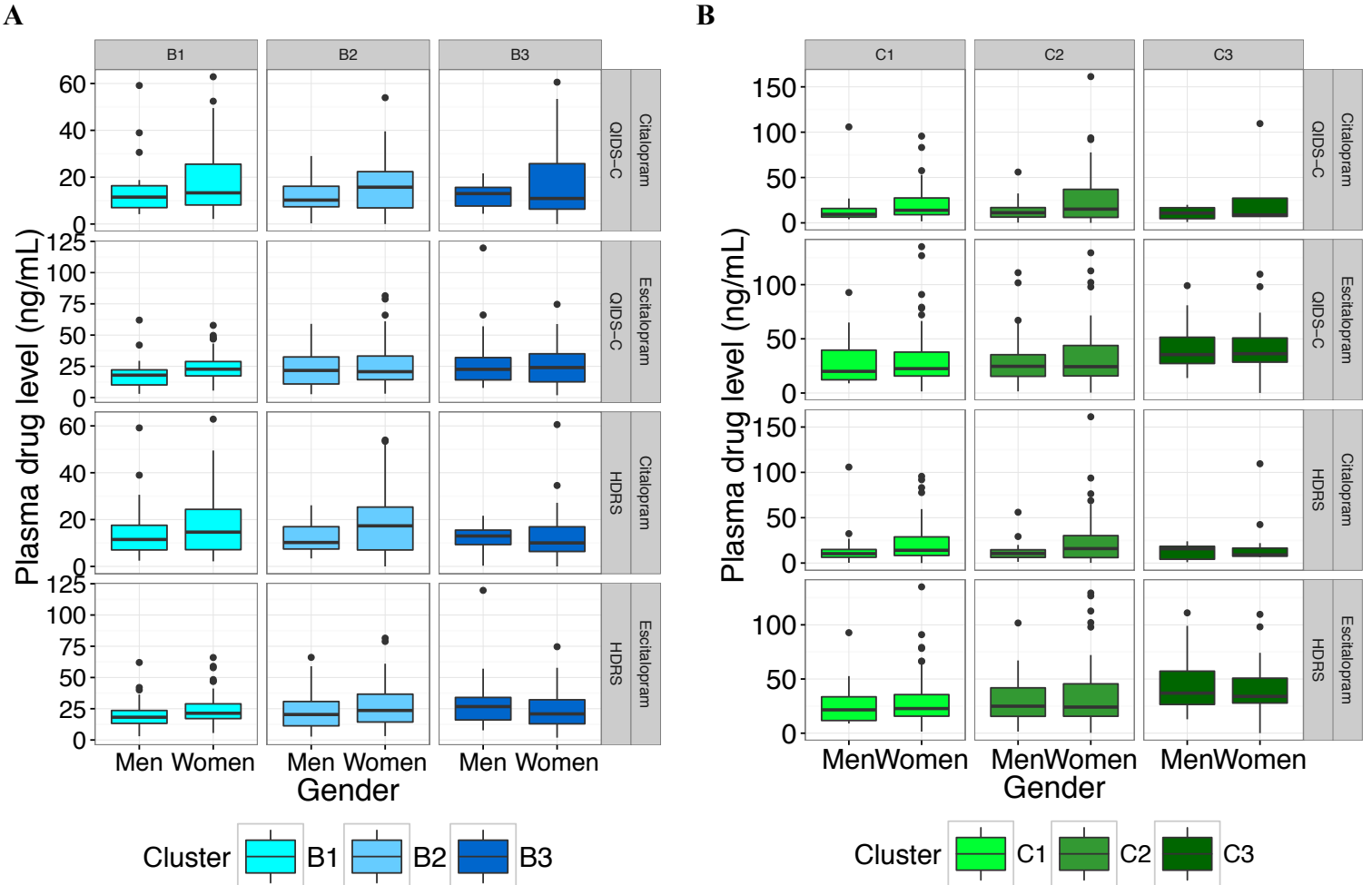
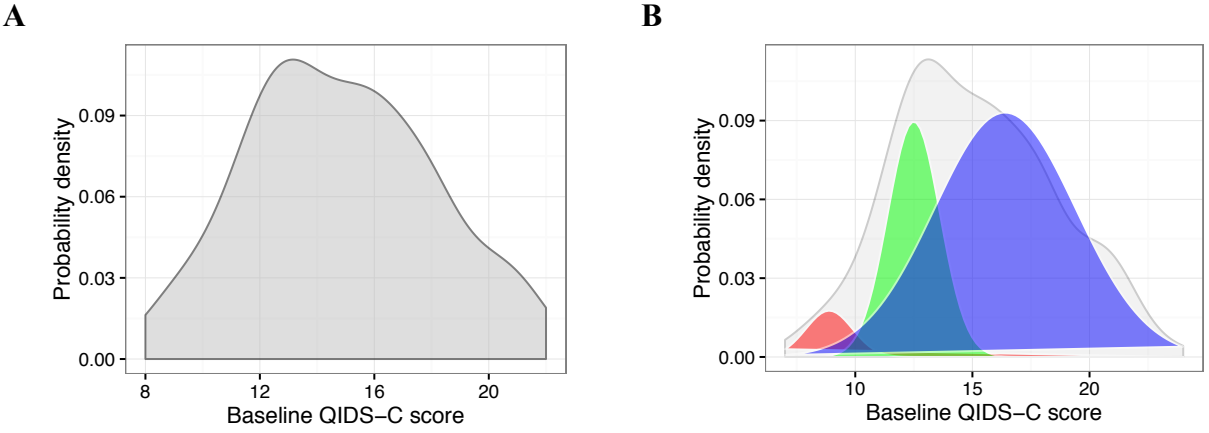
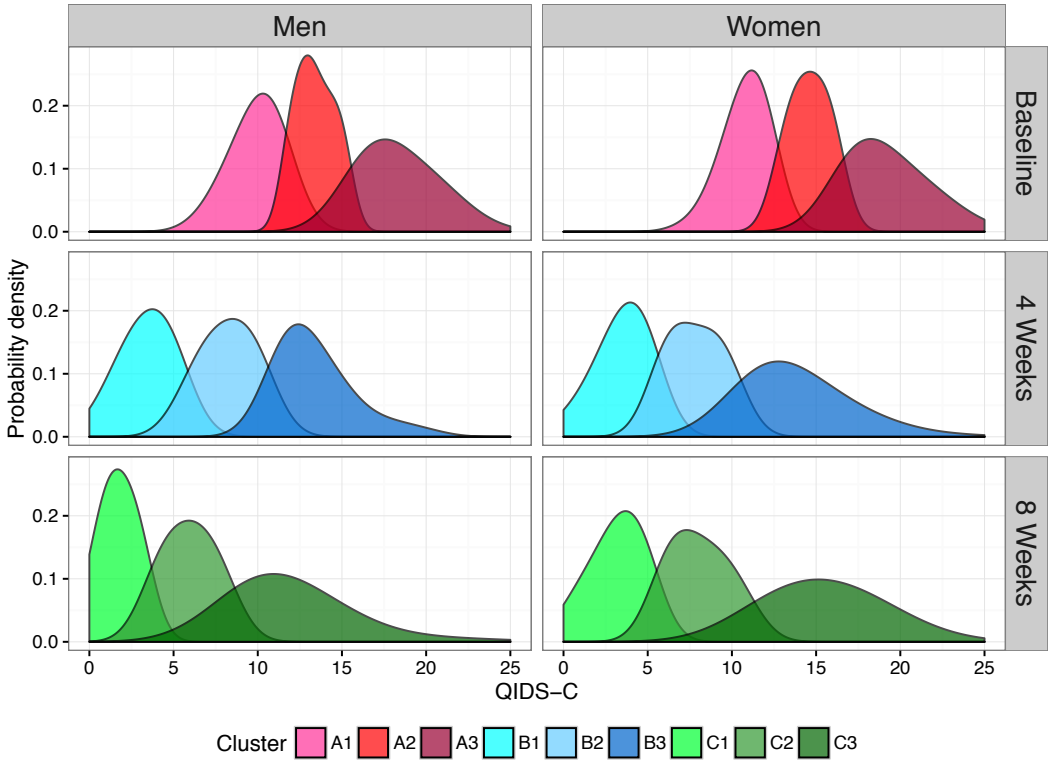

A
B1
B2
B3
Citalopram
Escitalopram
QIDS–C
HDRS
Plasma drug level (ng/mL)
Men Women
Gender
Cluster
B1
B2
B3
B
C1
C2
C3
Cluster
C1
C2
C3

Fig. 3

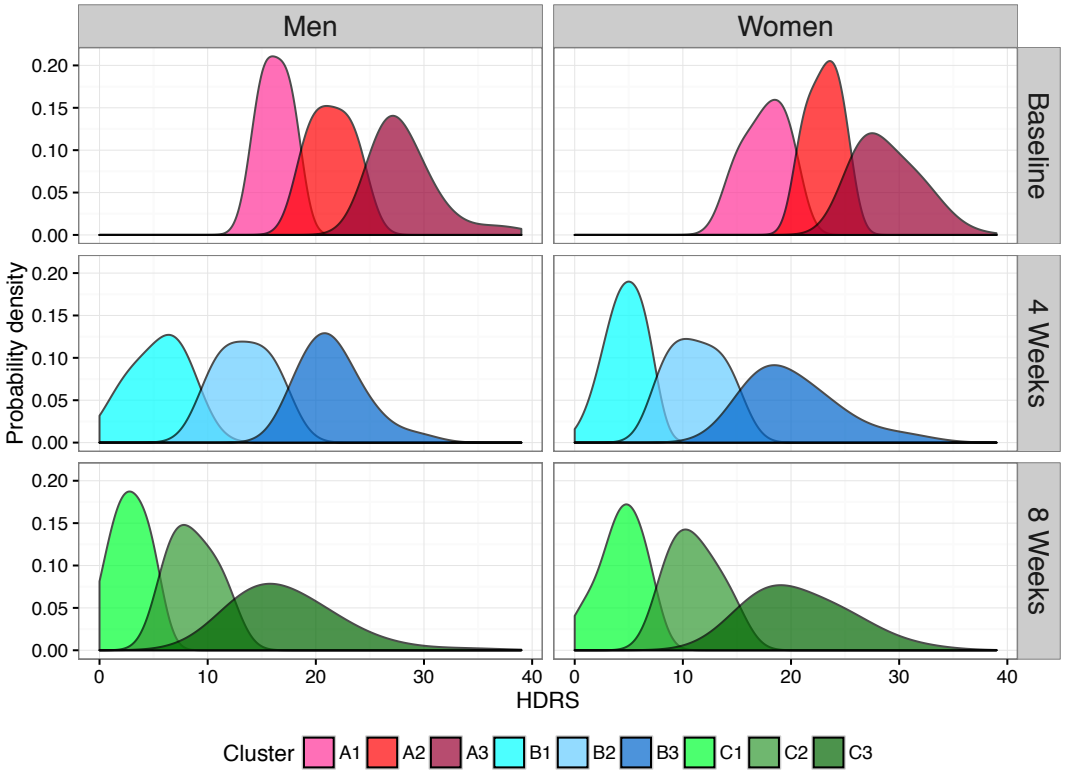


Supplementary Figure S1

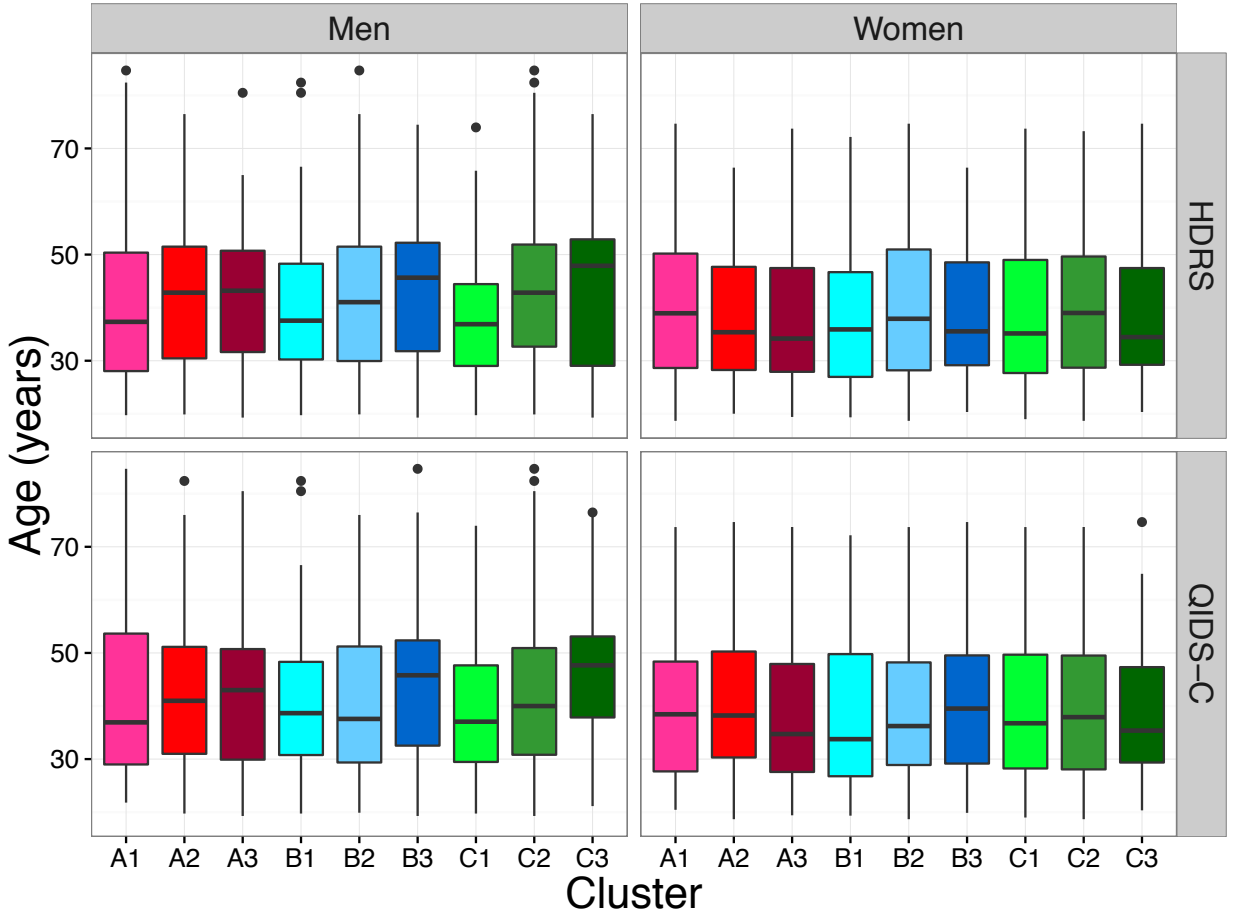
A



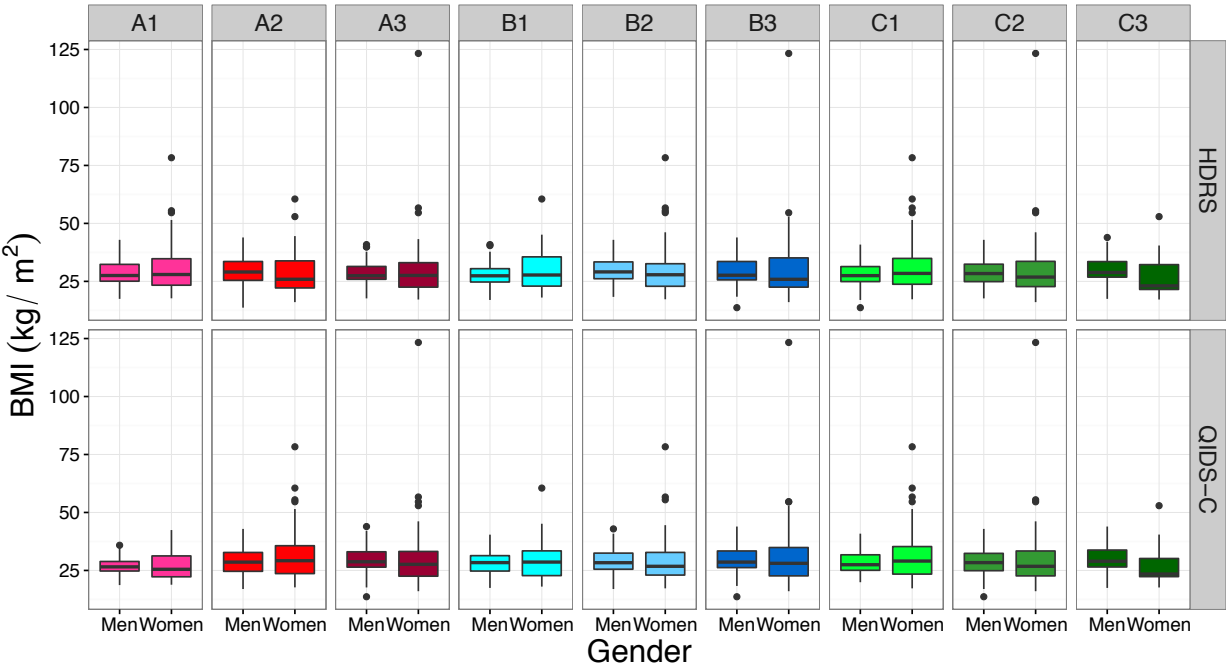
B



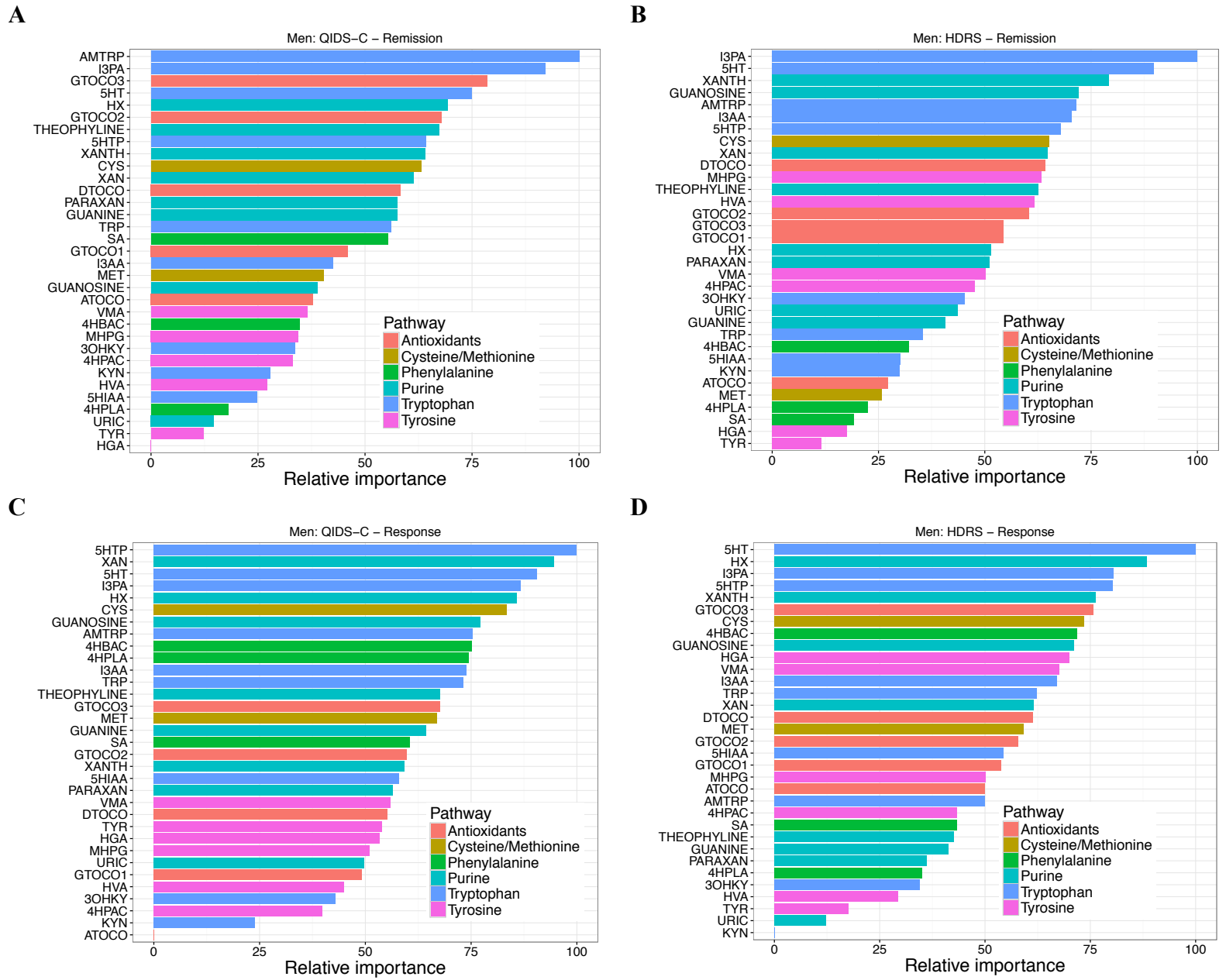
Supplementary Figure S2



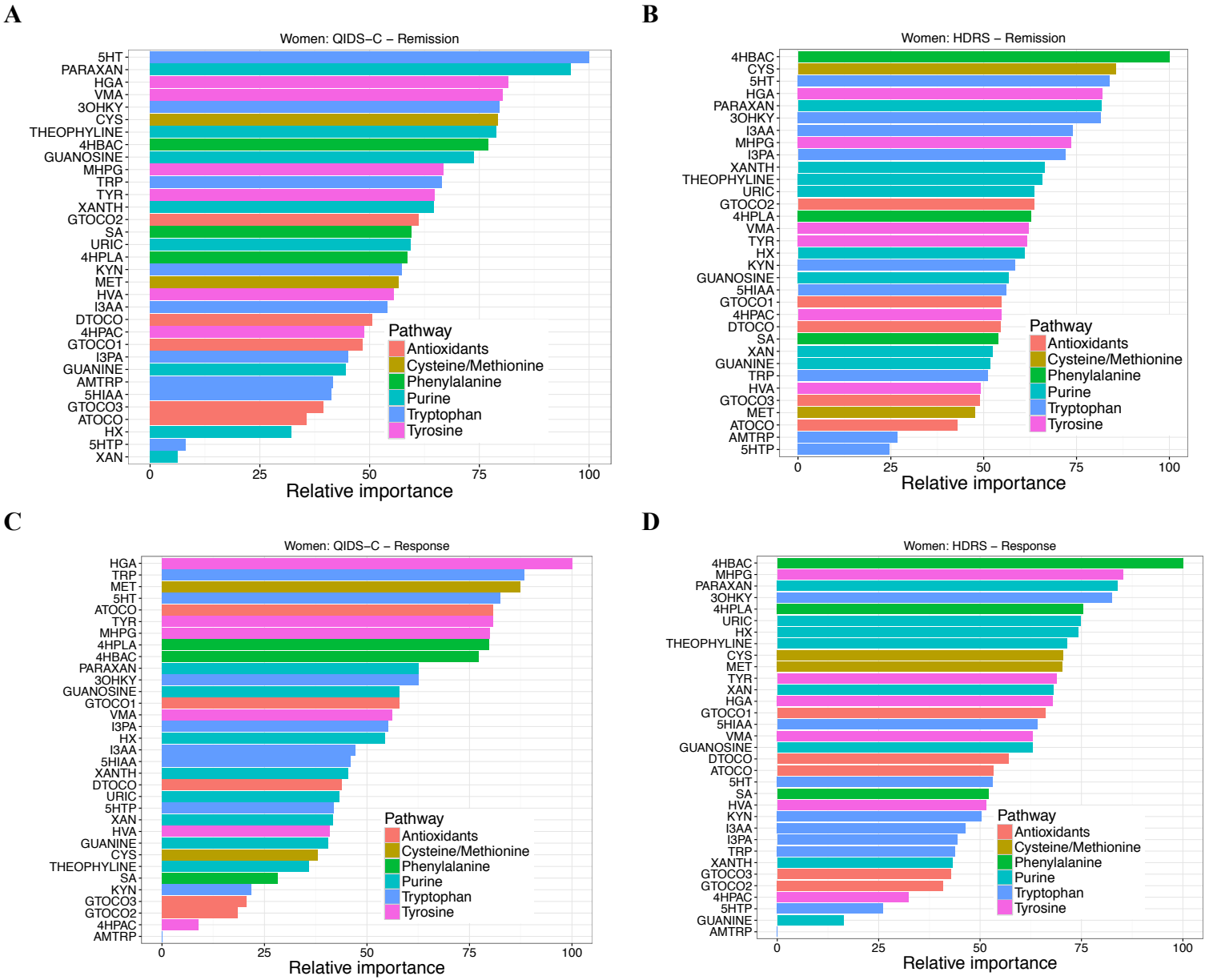
Supplementary Figure S3



Supplementary Figure S4



Supplementary Figure S5



Supplementary Figure S6

SUPPLEMENTARY MATERIALS

FIGURES

Supplementary Figure 1: Probability density function (PDF) of baseline QIDS-C symptom severity scores in men (Fig. A) and the estimated components of the PDF using an Expectation Maximization (EM) algorithm (Fig. B).

Supplementary Figure 2: Probability densities of symptom severity in clusters at baseline, 4 weeks and 8 weeks of the Mayo PGRN-AMPS trial for both QIDS-C (Fig. A) and HAM-D scales (Fig. B).

Supplementary Figure 3: Comparison of mean ages for men and women in clusters with comparable symptom severity at baseline, 4 weeks and 8 weeks.

Supplementary Figure 4: Comparison of mean body mass indes (BMI, kg/m^2) for men and women in clusters with comparable symptom severity at baseline, 4 weeks and 8 weeks.

Supplementary Figure 5: Relative importance of metabolites for the prediction of remission (QIDS-C – Fig A, HDRS – Fig B) and response response (QIDS-C – Fig C, HDRS – Fig D) at 8 weeks in men.

Supplementary Figure 6: Relative importance of metabolites for the prediction of remission (QIDS-C – Fig A, HDRS – Fig B) and response (QIDS-C – Fig C, HDRS – Fig D) at 8 weeks in women.

TABLES

SUPPLEMENTARY TABLE 1
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
Hypoxathine	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

SUPPLEMENTARY TABLE 2

Clinical and Demographic factors analyzed in this work

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
History of any other substance 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
Transplantation or transfusion	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 (GED) Some college Associate Degree/Technical Degree College Diploma Masters 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
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]

SUPPLEMENTARY TABLE 4

Metabolites that show significant differences between men and women with either remission/response status. e.g., performing MANOVA between metabolites men with remission at 8 weeks with metabolites women with remission at 8 weeks, both measured by QIDS-C.

Question Type Clinical outcome	Baseline	4 weeks	8 weeks
QIDS-C Response	4HPLA, GTOCO2, GUANOSINE, TRP, URIC	4HPLA, KYN, PARAXAN, TRP, URIC, XAN	4HPLA, 5HT,AMTRP,CYS,GUANOSINE, KYN, TRP,URIC,XAN
HDRS Response	4HPLA, DTOCO, GTOCO2, GUANOSINE, MET, TRP, URIC	3OHKY_BACKWAVE, 4HPLA, PARAXAN, TRP, RUIC, XAN	4HPAC, 4HPLA, 5HT, AMTRP, CYS, GUANOSINE, KYN, TRP, TYR, URIC,XAN
QIDS-C Remission	DTOCO, URIC	TRP, URIC	30HKY, 4HPLA, 5HT, AMTRP, ATOCO, TRP, URIC, XAN
HDRS Remission	DTOCO, URIC	GUANOSINE, TRP, URIC	4HPLA, 5HT, AMTRP, ATOCO, GTOCO3, GUANOSINE, KYN, TRP, URIC, XAN

SUPPLEMENTARY TABLE 5

Metabolites that showed significant differences between baseline and 8 weeks in men and women with response/remission status as measured by QIDS-C or HDRS. e.g., performing MANOVA between baseline metabolites of men defined remission at 8 weeks and their associated metabolites at 8 weeks, both measured by QIDS-C.

Question Type	**Outcome**	**MANOVA (baseline~8weeks)**	
		Men	**Women**
QIDS-C	**Response**	5HT, MHPG, URIC	4HBAC, 4HPAC, 5HT, GUANOSINE, MHPG, XAN
	Remission	5HT, MHPG	4HBAC, 4HPAC, 5HT, ATOCO, HGA, MHPG
HDRS	**Response**	5HT, MHPG, URIC	4HBAC, 4HPAC, 5HT, GUANOSINE, HGA, MHPG, XAN
	Remission	5HT, MHPG, VMA	4HPAC, 5HT, ATOCO, HX, MHPG, XAN

Supplementary Section 1: METHODS – Three-stage analyses workflow

We describe the methods used in the workflow used to perform these studies.

Multivariate Statistical Analysis: We used multivariate analysis of variance (MANOVA) to determine sex differences in metabolite concentrations at baseline, and after 4- and 8 weeks of citalopram treatment. When statistically significant differences were observed (at a threshold of $\alpha = 0.05$), analysis of variance (50) was used to identify the specific metabolites for which significant differences between the sexes in mean concentrations existed. Based on these results (discussed below), we analyzed the data separately for men and for women to avoid potential inaccuracies based on sex differences in metabolite concentrations during citalopram treatment in the prediction models.

Unsupervised Learning: We used unsupervised learning to identify clusters of patients (men and women separately) with similar symptom severity, as measured by the QIDS-C and HDRS, at baseline and after 4- and 8 weeks of treatment. The Shapiro-Wilk test was first used to test whether the distribution of symptom severity scores was normal (Gaussian). Because the data were not normally distributed (Fig. S1-A), we then assumed that the distribution of symptom severity scores was composed of a mixture of many Gaussians (components of the distribution), where each Gaussian distribution represented a cluster of patients with similar symptom severity. We applied an Expectation Maximization (EM) algorithm that assumed only one component in the mixture (a single bell-shaped curve distribution) and gradually increased the number of components (distributions with multiple peaks) until an adequate fit of the data was achieved. At

each step in this process, we generated 10,000 samples using the parameters (mean and variance of a Gaussian component) estimated by the EM algorithm and computed p-values using Kolmogorov-Smirnov and Wilcoxon-rank non-parametric tests. The process was stopped at $p > 0.05$ (failing to reject the null-hypothesis that estimated distribution and actual distributions are similar), and in this work, the process was stopped at three components (Fig. S1-B). Under the assumption of the distribution being a mixture of 3 components, patients were assigned a cluster based on which component their score belonged to, and clusters with the density functions from their associated scores are shown graphically in Fig. S2.

Using the chosen *k* and estimated mean of each component as seeds, *k*-means clustering was used to infer the clusters of patients with similar depressive symptom severity. Using the inferred clusters of patients, we correlated their symptom severities with clinical, demographic, and metabolomic data to determine the influence of these variables on clinical outcomes (movement of patients between depressive symptom severity clusters between baseline and follow-up time points) during citalopram treatment.

Supervised Learning: To predict clinical outcomes after SSRI treatment, we used supervised learning methods that required predictor variable data (e.g., metabolomics, clinical and demographics data) and training labels (responders/non-responders or remitters/non-remitters). Prediction models were trained separately for men and women, and also separately for response and remission as clinical outcomes. Since these clinical outcomes were binary, we used support vector machine (SVM) learning with radial kernels, gradient boosting machines, generalized linear models and random forests as classifiers. For the training and testing process described

next, we observed that SVM was the best classifier among others. To train the classifier, 10-fold cross-validation with 5-repeats was performed on a centered and scaled training data set which comprised a random 80% split of the patient cohort (men and women separately). Classifier performance was tested on the 20% of the patient cohort not used for training the classifier. We used area under curve (24), sensitivity, and specificity as metrics to evaluate the prediction accuracy. Statistical significance of the classifier's accuracy was computed using the null information rate (NIR), which is the fraction of labels in the test data. The p-value is computed using a one sided Chi-squared test for which the prediction accuracy is greater than the NIR.

Supplementary Section 2: Results

STAGE-1: Sex Differences in Metabolomics Response

There were no significant differences in metabolite concentrations between responders and non-responders (or remitters/non-remitters) at any time point, even when stratified by sex. However, there were significant changes in metabolite concentrations from baseline to 8 weeks in men and in women who were classified as responders ($p < 3.4E-06$ from MANOVA) as assessed by the QIDS-C or HDRS, and in men and women who were classified as remitters at 8 weeks as assessed by QIDS-C ($p<3.4E-06$), but not the HDRS ($p = 0.2$). The metabolites with significant changes between baseline and 8 weeks (irrespective of clinical outcomes) were different in men and women, except for 5HT and MHPG (Supplementary Table 5). These results indicated that men and women exhibited significant differences in plasma metabolite concentrations irrespective of clinical outcome, and they provided the quantitative rationale for building the predictive models separately for men and women.