The American Journal of Psychiatry

Integrating Multi-Omics and Clinical Data to Predict SSRI Therapeutic Response in Adults with Major Depressive Disorder: A Data-Driven Machine Learning Approach

Journal:	*The American Journal of Psychiatry*
Manuscript ID	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	ATHREYA, ARJUN; University of Illinois at Urbana-Champaign, Electrical and Computer Engineering Neavin, Drew; Mayo Clinic Minnesota Monrad, Ditlev; University of Illinois at Urbana-Champaign Frye, Mark; Mayo Clinic, Skime, Michelle; Mayo Clinic, Psychiatry & Psychology Rush, Augustus; Curbstone Consultant LLC, Matson, Wayne; Counterpoint Health Solutions Inc Biernacka, Joanna; Mayo Clinic, Department of Health Sciences Klein, Teri; Stanford University Wang, Liewei; Mayo Clinic Minnesota Kaddurah-Daouk, Rima; Duke University Iyer, Ravishankar ; University of Illinois at Urbana-Champaign, Electrical and Computer Engineering Weinshilboum, Richard; Mayo Clinic, Molecular Pharmacology Bobo, William; Mayo Clinic, Depatment of Psychiatry and Psychology
Keywords:	Major Depressive Disorder, Selective serotonin reuptake inhibitor (SSRI), citalopram, metabolomics, pharmacometabolomics

TITLE PAGE

Word Count:	**3,500**
Figures:	**4**
Tables:	**1**
Supplementary Figures:	**8**
Supplementary Tables:	**5**

TITLE: Integrating Multi-Omics and Clinical Data to Predict SSRI Therapeutic Response in Adults with Major Depressive Disorder: A Data-Driven Machine Learning Approach

BYLINE: Prediction of Clinical Outcomes in MDD using multi-omics
PREVIOUS PRESENTATION: None

AUTHORS: Arjun P. Athreya MS[1], Drew Neavin BS[2], Ditlev Monrad PhD[3], Mark A. Frye MD[4], Michelle Skime CCRP[4], A. John Rush MD[5], Wayne R. Matson[6], Joanna M. Biernacka PhD[4], Teri E. Klein PhD[7], Liewei Wang MD, PhD[2], Rima Kaddurah-Daouk PhD[5], Ravishankar K. Iyer PhD[1], Richard M. Weinshilboum MD[2] and William V. Bobo MDP, MPH[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.
[7] Department of Biomedical Data Science, Stanford University, Palo Alto, CA.

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, Selective serotonin reuptake inhibitor (SSRI), citalopram, metabolomics, pharmacometabolomics, data-driven analytics, machine learning.

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 U19 GM61388, R01 GM28157, R01 MH108348, RC2 GM092729, R24 GM078233 and RC2 GM092729; 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 or the NIH. We thank Dr. Krishna Rani Kalari at Mayo Clinic for her insightful comments, and 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 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 Akili Inc., Brain Resource Ltd, Curbstone Consultant

LLC., Emmes Corp., Holmusk, Liva-Nova, National Institute of Drug Abuse, Santium Inc., Sunovion, Taj Medical, Takeda USA; speaking fees from Liva-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.

ABSTRACT

OBJECTIVE: There is large variation in drug response in major depressive disorder (MDD). Our objective was to predict antidepressant remission/response by integrating associated multi-omics and clinical measures using machine learning.

METHOD: The PGRN-AMPS study treated 290 MDD patients with citalopram/escitalopram for 8 weeks. Blood samples were taken for genomics (baseline), plasma drug levels (4 and 8 weeks), and targeted electrochemistry-based metabolomics (31 metabolites at all time points). Unsupervised (GMM clustering) learning was used to identify patient subgroups that were validated using STAR*D and ISPC datasets. Supervised (support vector machine) learning was then used to predict response/remission by integrating baseline metabolomics data with baseline depression severity scores (QIDS-C and 17-item, HDRS) and previously reported SNPs (in *DEFB1, ERICH3, AHR* and *TSPAN5* genes) associated with baseline serotonin and kynurenine concentrations.

RESULTS: Metabolomic profiles differed significantly by sex at all time points. Three distinct clusters ($p<1.3E-09$) were identified in men and women separately at each time point which replicated in STAR*D and ISPC datasets. Clusters were associated with baseline plasma metabolite concentrations, but not demographic/clinical factors, or plasma drug levels. For an individual patient's baseline metabolite concentrations and genomics data, our supervised learning prediction's accuracies were statistically significant ($p<0.05$) for remission (men – 85%; women – 80%) and response (men – 84%; women – 89%), representing an additional 10% improvement over using baseline metabolomics data alone.

CONCLUSIONS: Sex-specific prediction of antidepressant response using machine learning significantly improved when genomics data was combined with baseline metabolomics. This work highlights the importance of integrating multiple biological measures to predict antidepressant response.

INTRODUCTION

The application of machine learning methods in medicine outside of psychiatry has significantly improved predictive capabilities, especially when a variety of data types can be analytically integrated to provide a biological context (1, 2). An important step toward improving the predictive accuracies of machine learning approaches in psychiatry involves assessing whether such predictions can be driven by biological measures related to the clinical response to antidepressants, both overall and within biologically distinct subgroups (3, 4).

Only about half of major depressive disorder (MDD) patients respond to an initial therapeutic trial of SSRIs—the most commonly prescribed first-line pharmacotherapy for MDD – after 8 weeks of therapy (5), and an even lower proportion of SSRI-treated patients (25%-40%) achieve remission (5). Individual demographic and selected clinical risk factors can only weakly predict SSRI treatment outcomes (6). Previous machine learning models that only considered clinical and demographic factors in aggregate yielded a predictive accuracy of 65% (7). Our prior machine learning work showed that integrating baseline metabolomics data with clinical/demographics data can improve predictive accuracy over using clinical/demographic data alone (8), thus raising the possibility of improved predictive accuracies by integrating additional biological measures, such as genomics. However, different types of biological or clinical measures present varying complexities in their specificity and relevance to clinical outcomes (9). We assert that addressing a phenotype as complex as antidepressant response will require an analyses workflow that breaks the overall problem down into smaller but less complex problems, each characterized by unique time-varying data types, and addressed by specific machine learning methods.

Therefore, we developed a machine learning workflow that first stratified patients by sex and depression severity, and then seamlessly integrated baseline metabolomics with genomics data from the Pharmacogenomics Research Network-Antidepressant Medical Pharmacogenomic Study (PGRN-AMPS) (10), to predict antidepressant remission/response at 8 weeks. Key result is, using baseline metabolomics data associated with cluster's depression severity significantly improved sex-specific prediction of antidepressant remission/response over using clinical/demographic data; this predictability further improved when SNPs associated with baseline metabolomic concentrations were also included as predictor variables. Our results raise the possibility that SNPs linked with concentrations of specific metabolites used in our analyses may have biological effects that influence MDD outcomes that extend beyond their known effects on these metabolites.

DATA AND METHODS

DATA SOURCES

PGRN-AMPS (NCT 00613470) was an 8-week, single-arm, open trial designed to assess clinical outcomes in adults (aged 18-84 years) with MDD in response to citalopram/escitalopram, and then to examine metabolomic and genomic factors associated with those outcomes (10). Subjects were recruited from primary and specialty care settings from March 2005 to May 2013. All psychiatric diagnoses were confirmed using modules A, B (screen only version), and D of the Structured Clinical Interview for DSM-IV (SCID) administered by trained clinical research staff (11). Clinical and demographic variables from the PGRN-AMPS dataset used in the analyses (Supplementary Table 1) were assessed at baseline using standardized questionnaires.

Data from the initial phase of Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (NCT 00021528) and International SSRI Pharmacogenomics Consortium (ISPC) were used to validate depressive symptom response patterns found in the PGRN-AMPS study (5). Briefly, the initial phase of STAR*D was a 12-week clinical 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. For the present analyses, we utilized data from 788 Caucasian STAR*D subjects who had complete Phase 1 clinical response data and had provided DNA samples. ISPC comprised 7 member sites that contributed data from 7 clinical trials of SSRIs for depression (578 subjects total) carried out in North America, Europe, and Asia, and was established to examine genetic factors driving variation in clinical response to SSRIs (12). Details of STAR*D study procedures and a description of each

contributing study in ISPC have been published previously (5, 13), and are described in Supplementary Sec. 1.

CLINICAL OUTCOMES

In all three trials included in our study, treatment outcomes were established using the 16-item, clinician-rated version of the Quick Inventory of Depressive Symptomatology (QIDS-C (14)) or the 17-item Hamilton Depression Rating Scale (HDRS (15)). Remission was defined as a QIDS-C score ≤ 5 (14) (HDRS score ≤ 7 (15)) at 4 or 8 weeks. Response was defined as a ≥ 50% reduction in QIDS-C or HDRS total scores from baseline to either 4 or 8 weeks. Across the three datasets, 60-66% of subjects were classified as responders, and 40-42% were classified as remitters, at 8 weeks.

PLASMA METABOLOMIC AND DRUG CONCENTRATION ASSAYS

Plasma metabolite concentrations were assayed using samples from 306 randomly selected PGRN-AMPS patients who had samples obtained at baseline and after 4 and 8 weeks of SSRI therapy. As described previously (16), 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 2 lists the 31 metabolites assayed and their associated pathways. The analyses included metabolites with more than one redox state, resulting in a total of 31 metabolites and 35 "metabolite features". This particular metabolomic platform was selected because it is quantitative and highly sensitive for the assay of metabolites of monoamine neurotransmitters known to play a role in the

pathophysiology of MDD. As a result, the pathways that include many of these metabolites have known relevance to the pathophysiology of depression and/or the therapeutic effects of SSRI antidepressants (17-24). Finally, plasma SSRI drug and drug metabolite concentrations were also assayed in all 4 and 8 week samples using an HPLC MS/MS platform.

ANALYSIS WORKFLOW

We developed a three-stage machine learning workflow illustrated in Fig. 1 to predict antidepressant remission/response.

STAGE-1 – Establish sex differences in metabolomics profiles: Anticipating sex differences in metabolomic profiles (25), 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), validate the cluster patterns with STAR*D and ISPC data (Stage 2B), and identify factors associated with clusters (Stage 2C): In **Stage 2A**, we used mixture model-based unsupervised learning (8) with Gaussian Mixture Models to algorithmically identify the optimum number of clusters (minimum Gaussians sufficient to estimate the actual distribution of depressive symptom severity – see Supplementary Sec. 2) of patients based on depression symptom severity, as measured by the QIDS-C and HDRS at baseline, and after 4 or 8 weeks of treatment in the PGRN-AMPS study. In **Stage 2B**, the clustering approach developed in **Stage 2A** was subjected to STAR*D and ISPC datasets to investigate whether the distributions of depression severity (using Kolmogorov-

Smirnov test) and proportion of patients that move between the clusters from baseline to 4 weeks to 8 weeks were the same in the three independent datasets. In **Stage 2C**, Kolmogorov-Smirnov (continuous data) and two-way Chi-square (categorical data) tests were used to identify metabolomic, clinical and demographic factors that were associated with the depression severity clusters. Using a metabolomics-informed-genomics approach, top SNPs were identified using genome-wide association studies (GWAS) for concentrations of plasma metabolites that showed associations with baseline symptom severity (16-24, 26, 27).

STAGE-3 – Predict antidepressant remission/response: We used support vector machines with radial basis function (SVM-RBF), a non-parametric supervised learning method for handling correlated predictor variables, as a binary classifier to predict remission/response, using baseline symptom severity, baseline metabolomics, and genomics data from PGRN-AMPS (avoiding the effects of genotypes obtained from different platforms and imputation methods in STAR*D/ISPC datasets, and also lacking metabolomics data). The SVM family of classifiers maximizes separation between classes of data and has shown greater generalizability in several clinical applications when compared with regression models (28) and random forests (29). Due to limited samples with biological measures, we investigated if machine learning methods could better predict antidepressant remission/response at 8 weeks using baseline biological measures, irrespective of baseline clustering behavior.

Chances of model overfit was minimized in two phases. For training the classifier, a training dataset (80% of data) comprising predictor variables (baseline metabolomics, symptom severity and genomic data) and associated training labels (remitters/non-remitters or responders/non-responders) were used as inputs. 10-fold cross-validation with 5 repeats was used

to study the bias-variance tradeoffs (minimizing the chances of overfitting) to choose the classifier's parameters that maximized the AUC. Further minimizing the effects of class imbalance (e.g., unequal number of responders vs non-responders), we used the SMOTE algorithm which simulated patient profiles of the under-sampled class and up-sampled the under-sampled class to ensure that both classes had equal sizes (30). In the testing phase, the outcome labels of the test dataset (comprising data not used for training) were predicted using the trained classifier. Statistical significance of the classifier's prediction accuracy (fraction of correctly predicted labels in the test data set) was demonstrated using AUC and the null information rate (NIR – in this case 0.5, due to equal number of positive and negative test samples), which served as a proxy for chance.

RESULTS

SEX DIFFERENCES IN METABOLOMIC PROFILES

Plasma concentrations of several metabolites differed significantly by sex at baseline, 4 weeks and 8 weeks, regardless of response/remission status or the depression rating scale (QIDS-C/HDRS) used to define these outcomes (Supplementary Table 3). The specific metabolites that were changed after initiating citalopram/escitalopram also differed by sex and by depression rating scale (Fig. 2). There were no significant sex-differences in clinical/demographic factors (Supplementary Fig.1). Based on these results, we proceeded through the remaining stages of the workflow using separate strata defined by sex and by depression rating scale.

DEPRESSIVE SYMPTOM 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 (Fig. 3, Supplementary Fig. 2). The 9 depressive symptom clusters in men and women were labeled using a convention wherein the numeral 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 were unable to achieve the same clustering pattern using individual scale item scores (see Fig. 4).

In both STAR*D (QIDS-C) and ISPC (HDRS) datasets, our unsupervised learning approach algorithmically also identified three clusters of men and women at all time-points, that 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) conformed to accepted clinical definitions of remission, response and non-response, respectively, on both depression rating scales, as observed in the PGRN-AMPS data. Furthermore, the proportion of patients who moved between these respective clusters from baseline to 4 weeks and from 4 weeks to 8 weeks were not significantly different between independent trials. As an additional validation step, we subjected entire ISPC dataset to the clustering approach, which included both Asians and Caucasians. In this analysis, clustering patterns for both Asians and Caucasians together, and Asians only, did not differ ($p > 0.5$) from the clustering patterns in Caucasians alone (Supplementary Figs. 3-4). Our externally validated clusters then allowed us to identify associations of their depression severity with clinical, demographic and biological factors.

ASSOCIATION OF CLINICAL, DEMOGRAPHIC AND BIOLOGICAL FACTORS WITH SEVERITY-BASED CLUSTERS

There were no significant differences in any of the clinical or demographic factors listed in Supplementary Table 1 or plasma drug levels (at 4 weeks and 8 weeks) between clusters ($p > 0.1$, Supplementary Figs. 5-7). Only metabolite concentrations in baseline clusters and baseline depression severity were significantly correlated ($p < 0.05$) with response and remission status at 8 weeks, including baseline concentrations of 5HT, KYN, 4HBAC, TRP, TYR and PARAXAN. The fact that these particular metabolites were correlated with measures of response was

reassuring since they have been identified in the course of previous studies of metabolomics and many of them are also related to monoamine neurotransmitter pathways that have been associated with MDD and its treatment response (16-24, 26, 27). Those 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 (16, 27). These SNPs, together with metabolomics data, were then included in our prediction model.

OUTCOME PREDICTION PERFORMANCE

The classifiers that used baseline symptom severity and metabolomic concentrations to predict remission (C1) and response (C2) had significantly greater predictive performance (Table 1 – A) than using only clinical/demographic factors (7). The key findings are as follows:

- Remission in men: 70-75%, $p < 0.04$, AUC 0.63 – 0.69;
- Remission in women: 70 – 72%, $p < 0.008$, AUC 0.78-0.9;
- Response in men: 72-75%, $p < 0.03$, AUC 0.78-0.91; and
- Response in women: 81-87.5%, $p < 8.2E-05$, AUC 0.82-0.88.

The use of baseline symptom severity with genomic (i.e., SNPs) data generated similar predictive performance as did baseline symptom severity and metabolomics (Supplementary Table 4). However, the prediction performance improved further when genomic data were used in addition to baseline depressive symptom severity and metabolomics data for remission and response (Table 1 – B). The key findings are as follows:

- Remission in men: 80 – 90%, $p < 0.005$, AUC 0.74 – 0.92;
- Remission in women: 71 – 90%, $p < 0.006$, AUC 0.9 – 0.92;

- Response in men: 84%, $p < 0.008$, AUC 0.83 – 0.9; and
- Response in women: 87 – 91%, $p < 8.2E-07$, AUC 0.87 – 0.92.

Similar prediction performance was seen when random forest was used (Supplementary Table 5). The balance between sensitivity/specificity and PPV/NPV (Table 1 and Supplementary Table 4) indicate that the cross-validation approach minimized the chances of model overfit.

TOP BIOLOGICAL PREDICTORS OF CLINICAL OUTCOMES

Top predictors were chosen on the basis of relative importance (>80%) in their respective prediction models. Of the individual metabolites, baseline 5HT levels were among those with the highest relative contribution to the accuracy of the predictive model for all treatment outcomes, except for response in women defined by HDRS scores. At least one metabolite related to the tryptophan pathway, including 5HT and its precursors 5HTP and TRP, was among the top predictive metabolites for all treatment outcomes. Metabolites in the purine pathway were also among those with the highest relative contribution to the accuracy of the prediction model for all clinical phenotypes except for response in men defined by HDRS scores. Tyrosine metabolites, including those related to catecholamine biosynthesis, were included among the top predictors of HDRS-defined response and remission in women, whereas tocopherol metabolites were among the top predictors of response and remission in men. The *DEFB1* SNP, rs2702877, was included among the top predictors of QIDS-C-defined response in both men and women, whereas the *TSPAN5* SNP, rs10516436, was among the top predictors of QIDS-C-defined response in men.

DISCUSSION

MULTI-OMICS MEASURES and BIOLOGICAL SIGNIFICANCE

Our work represents a significant advance toward the subclassification of MDD patients based on underlying biological mechanisms—a major goal of psychiatry—by integrating baseline depressive symptom severity with metabolomic and genomic data to predict sex-specific remission and response after 8 weeks of SSRI treatment. It is clear that sex represents an important risk factor for MDD, with virtually all studies reporting that two thirds of these patients are women. Although sex has been reported to influence response to antidepressants in some studies (25, 31-33), prior machine learning approaches using clinical measures alone did not identify sex as a robust predictor of remission (7). The sex-specific differences in metabolomic profiles and some top predictors of treatment outcomes in our study, as well as previously reported sex differences in depressive symptom profiles (34), suggest that sex-specific biological mechanisms may play an important role in antidepressant response.

We included baseline depressive symptom scores in our prediction model because they are often the most robust predictor of antidepressive outcomes (35, 36). An earlier machine learning approach also identified baseline depression severity as their best predictor of remission among clinical factors (7). These observations are in agreement with with the fact that social/demographic factors individually or in aggregate cannot accurately predict antidepressant treatment outcomes (6, 7, 37, 38), and were not associated with any of the depressive symptom severity clusters in this work.

Previous metabolomics studies provided evidence that metabolites involved in the biosynthesis of 5HT and catecholamines (i.e., tryptophan, tyrosine, and methoxyindole pathways) were related to SSRI response, MDD pathophysiology, or both (16, 18, 39). The fact that baseline 5HT figured prominently in the accuracy of our predictive model is consistent with our recent observation that, of all of the metabolites that we measured, 5HT was most highly associated with SSRI outcomes (16). We previously identified and functionally validated two novel genes associated with variation in plasma 5HT concentrations, *TSPAN5* and *ERICH3* (16). The metabolite that was most highly related to baseline symptom severity was KYN. A subsequent GWAS performed for plasma KYN identified top-hit SNPs that mapped to *DEFB1*. Our functional studies demonstrated that DEFB1 could influence KYN biosynthesis (27). The fact that prediction performance improved after adding these SNPs to baseline metabolomics data suggests that these SNPs may influence clinical outcomes via mechanisms that extend beyond 5HT and KYN—as suggested by previous experimental results (16).

CLINICAL IMPLICATIONS OF THE PATIENT CLUSTERING

An important contribution of our work is in identifying clusters using total depressive symptom severity that replicated across three independent datasets, and conformed to accepted clinical definitions of remission, response and non-response at 8 weeks. Our results in comparison with using multivariate clustering (Fig. 4) provide strong evidence that our clustering approach was computationally and ecologically valid.

Others have attempted to cluster individual depression scale items to identify symptoms with similar responses (40). 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. Therefore, our clustering approach was designed to model how patients move between symptom clusters and to identify how specific depression items change with time, within subgroups defined by baseline characteristics, a process which we have previously referred to as "symptom dynamics" (8).

DATA-DRIVEN PATIENT-SPECIFIC PREDICTIONS

Lacking in psychiatric medicine are predictive biomarkers with sufficient validity to select effective medications, similar to the case of breast cancer, in which the presence of estrogen and HER2 receptors in the tumor makes it possible to cluster patients and predict clinical drug response (41). The complex nature of the antidepressive response phenotype requires an integration of multiple, patient-specific, heterogeneous biological and non-biological factors with time-varying effects (42, 43). Rather than pursuing a "one-size-fits-all" machine learning method, the workflow presented here broke down the complex antidepressive response phenotype into a series of analyses, each addressed by a specific machine learning method. We began with the use of unsupervised learning to cluster MDD patients into subgroups defined by sex and baseline depressive symptom severity. This subsequently allowed us to better predict clinical outcomes to SSRI therapy using supervised learning methods that integrated the biological factors that differed by sex and were shown to be associated with baseline depressive symptom severity clusters. We assert that this type of data-driven analyses workflow, which combines multiple machine learning approaches, will be needed to achieve valid predictive models for treatment response in patients with MDD. Finally, since our workflow was developed to be agnostic to time-scale, our approaches can be applied to antidepressive

treatments that exert therapeutic effects over minutes to hours, e.g., intravenous/intranasal ketamine/esketamine, as opposed to weeks.

STUDY LIMITATIONS

Our sample was largely made up of Caucasian subjects, thus limiting generalizability of prediction model with respect to the biological predictors of outcomes. Additional studies in racially diverse samples are needed. However, as is the case with many genomic studies, restricting our analyses may have reduced confounding by race (44). We had no direct measures of socioeconomic status and comorbid anxiety, factors associated with lower likelihood of benefit from antidepressants (38, 45). With the use of complete cases, we cannot exclude the possibility of confounding by patients who dropped out later in the trial. We did not obtain fasting samples for the metabolomic profiling. Although fasting status may not significantly affect laboratory variability for most metabolites (46), we cannot exclude the possibility of bias introduced by systematic differences between groups in fasting time or in sampling time relative to meals. Finally, patients were not excluded on the basis of BMI or specific general medical conditions that could have confounded the relationship between metabolomic profile and SSRI treatment response. However, bearing this list of limitations in mind, the clear separation of men and women and the clear value of the addition of metabolomic and genomic information to the outcomes prediction model all represent significant progress toward the goal of Precision Medicine as applied to the treatment of MDD.

REFERENCES

1. Winslow RL, Trayanova N, Geman D, Miller MI. Computational medicine: translating models to clinical care. Sci Transl Med. 2012;4:158rv111.

2. Androulakis IP. Systems engineering meets quantitative systems pharmacology: from low-level targets to engaging the host defenses. Wiley Interdiscip Rev Syst Biol Med. 2015;7:101-112.

3. Kaddurah-Daouk R, Weinshilboum RM, Pharmacometabolomics Research N. Pharmacometabolomics: implications for clinical pharmacology and systems pharmacology. Clin Pharmacol Ther. 2014;95:154-167.

4. Kaddurah-Daouk R, Weinshilboum R, Pharmacometabolomics Research N. Metabolomic Signatures for Drug Response Phenotypes: Pharmacometabolomics Enables Precision Medicine. Clin Pharmacol Ther. 2015;98:71-75.

5. Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, Norquist G, Howland RH, Lebowitz B, McGrath PJ, Shores-Wilson K, Biggs MM, Balasubramani GK, Fava M, Team SDS. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry. 2006;163:28-40.

6. Bagby RM, Ryder AG, Cristi C. Psychosocial and clinical predictors of response to pharmacotherapy for depression. J Psychiatry Neurosci. 2002;27:250-257.

7. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, Cannon TD, Krystal JH, Corlett PR. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3:243-250.

8. Athreya AP, Banerjee SS, Neavin D, Daouk RK, Rush AJ, Frye MA, Wang L, Weinshilboum R, Bobo WV, Iyer RK: Data-Driven Longitudinal Modeling and Prediction of Symptom Dynamics in Major Depressive Disorder: Integrating Factor Graphs and Learning Methods. in IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology, IEEE Computational Intelligence Society; 2017.

9. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet. 2015;16:85-97.

10. Ji Y, Biernacka JM, Hebbring S, Chai Y, Jenkins GD, Batzler A, Snyder KA, Drews MS, Desta Z, Flockhart D, Mushiroda T, Kubo M, Nakamura Y, Kamatani N, Schaid D, Weinshilboum RM, Mrazek DA. Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: genome-wide associations and functional genomics. Pharmacogenomics J. 2013;13:456-463.

11. First MB, Spitzer, Robert L, Gibbon Miriam, and Williams, Janet B.W. Structured Clinical Interview for DSM-IV Axis I Disorders, Clinician Version (SCID-CV). American Psychiatric Press Inc. 1996.

12. Biernacka JM, Sangkuhl K, Jenkins G, Whaley RM, Barman P, Batzler A, Altman RB, Arolt V, Brockmoller J, Chen CH, Domschke K, Hall-Flavin DK, Hong CJ, Illi A, Ji Y, Kampman O, Kinoshita T, Leinonen E, Liou YJ, Mushiroda T, Nonen S, Skime MK, Wang L, Baune BT, Kato M, Liu YL, Praphanphoj V, Stingl JC, Tsai SJ, Kubo M, Klein TE, Weinshilboum R. The International SSRI Pharmacogenomics Consortium (ISPC): a genome-wide association study of antidepressant treatment response. Transl Psychiatry. 2015;5:e553.

13. Rush AJ, Fava M, Wisniewski SR, Lavori PW, Trivedi MH, Sackeim HA, Thase ME, Nierenberg AA, Quitkin FM, Kashner TM, Kupfer DJ, Rosenbaum JF, Alpert J, Stewart JW,

McGrath PJ, Biggs MM, Shores-Wilson K, Lebowitz BD, Ritz L, Niederehe G, Group SDI. Sequenced treatment alternatives to relieve depression (STAR*D): rationale and design. Control Clin Trials. 2004;25:119-142.

14. Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB. 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. 2003;54:573-583.

15. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23:56-62.

16. Gupta M, Neavin D, Liu D, Biernacka J, Hall-Flavin D, Bobo WV, Frye MA, Skime M, Jenkins GD, Batzler A, Kalari K, Matson W, Bhasin SS, Zhu H, Mushiroda T, Nakamura Y, Kubo M, Wang L, Kaddurah-Daouk R, Weinshilboum RM. TSPAN5, ERICH3 and selective serotonin reuptake inhibitors in major depressive disorder: pharmacometabolomics-informed pharmacogenomics. Mol Psychiatry. 2016;21:1717-1725.

17. Ali-Sisto T, Tolmunen T, Toffol E, Viinamaki H, Mantyselka P, Valkonen-Korhonen M, Honkalampi K, Ruusunen A, Velagapudi V, Lehto SM. Purine metabolism is dysregulated in patients with major depressive disorder. Psychoneuroendocrinology. 2016;70:25-32.

18. Kaddurah-Daouk R, Bogdanov MB, Wikoff WR, Zhu H, Boyle SH, Churchill E, Wang Z, Rush AJ, Krishnan RR, Pickering E, Delnomdedieu M, Fiehn O. Pharmacometabolomic mapping of early biochemical changes induced by sertraline and placebo. Transl Psychiatry. 2013;3:e223.

19. Berman RM, Narasimhan M, Miller HL, Anand A, Cappiello A, Oren DA, Heninger GR, Charney DS. Transient depressive relapse induced by catecholamine depletion: potential phenotypic vulnerability marker? Arch Gen Psychiatry. 1999;56:395-403.

20. Park DI, Dournes C, Sillaber I, Uhr M, Asara JM, Gassen NC, Rein T, Ising M, Webhofer C, Filiou MD, Muller MB, Turck CW. Purine and pyrimidine metabolism: Convergent evidence on chronic antidepressant treatment response in mice and humans. Sci Rep. 2016;6:35317.

21. Ressler KJ, Nemeroff CB. Role of serotonergic and noradrenergic systems in the pathophysiology of depression and anxiety disorders. Depress Anxiety. 2000;12 Suppl 1:2-19.

22. Uher R, Tansey KE, Dew T, Maier W, Mors O, Hauser J, Dernovsek MZ, Henigsberg N, Souery D, Farmer A, McGuffin P. An inflammatory biomarker as a differential predictor of outcome of depression treatment with escitalopram and nortriptyline. Am J Psychiatry. 2014;171:1278-1286.

23. Wichers MC, Koek GH, Robaeys G, Verkerk R, Scharpe S, Maes M. IDO and interferon-alpha-induced depressive symptoms: a shift in hypothesis from tryptophan depletion to neurotoxicity. Mol Psychiatry. 2005;10:538-544.

24. Young SN, Leyton M. The role of serotonin in human mood and social interaction. Insight from altered tryptophan levels. Pharmacol Biochem Behav. 2002;71:857-865.

25. Krumsiek J, Mittelstrass K, Do KT, Stuckler F, Ried J, Adamski J, Peters A, Illig T, Kronenberg F, Friedrich N, Nauck M, Pietzner M, Mook-Kanamori DO, Suhre K, Gieger C, Grallert H, Theis FJ, Kastenmuller G. Gender-specific pathway differences in the human serum metabolome. Metabolomics. 2015;11:1815-1833.

26. Kaddurah-Daouk R, Boyle SH, Matson W, Sharma S, Matson S, Zhu H, Bogdanov MB, Churchill E, Krishnan RR, Rush AJ, Pickering E, Delnomdedieu M. Pretreatment metabotype as

a predictor of response to sertraline or placebo in depressed outpatients: a proof of concept. Transl Psychiatry. 2011;1.

27. Liu D, Neavin D, Ray B, Athreya AP, Biernacka J, Bobo WV, Hall-Flavin D, Skime M, Zhu H, Jenkins G, Batzler A, Kalari K, Boakye-Agyeman F, Matson W, Bhasin SS, Mushiroda T, Nakamura Y, Kubo M, Iyer RK, Wang L, Frye MA, Daouk RK, Weinshilboum R: Beta-Defensin 1, an Epithelial Antimicrobial Peptide, and Plasma Kynurenine in Major Depressive Disorder: Metabolomics-informed Genomics. in 72nd Scientific Convention and Meeting of the Society of Biological Psychiatry2017.

28. Fernandez-Delgado M, Cernadas E, Barro S, Amorim D. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? Journal of Machine Learning Research. 2014;15:3133-3181.

29. Koutsouleris N, Kahn RS, Chekroud AM, Leucht S, Falkai P, Wobrock T, Derks EM, Fleischhacker WW, Hasan A. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach. Lancet Psychiatry. 2016;3:935-946.

30. Tang Y, Zhang YQ, Chawla NV, Krasser S. SVMs modeling for highly imbalanced classification. IEEE Transactions on Systems, Man, and Cybernetics

. 2009;39:281-288.

31. Kornstein SG, Schatzberg AF, Thase ME, Yonkers KA, McCullough JP, Keitner GI, Gelenberg AJ, Davis SM, Harrison WM, Keller MB. Gender differences in treatment response to sertraline versus imipramine in chronic depression. Am J Psychiatry. 2000;157:1445-1452.

32. Morishita S, Kinoshita T. Predictors of response to sertraline in patients with major depression. Hum Psychopharmacol. 2008;23:647-651.

33. Young EA, Kornstein SG, Marcus SM, Harvey AT, Warden D, Wisniewski SR, Balasubramani GK, Fava M, Trivedi MH, John Rush A. Sex differences in response to citalopram: a STAR*D report. J Psychiatr Res. 2009;43:503-511.

34. Martin LA, Neighbors HW, Griffith DM. The experience of symptoms of depression in men vs women: analysis of the National Comorbidity Survey Replication. JAMA Psychiatry. 2013;70:1100-1106.

35. Khan A, Brodhead AE, Kolts RL, Brown WA. Severity of depressive symptoms and response to antidepressants and placebo in antidepressant trials. J Psychiatr Res. 2005;39:145-150.

36. Friedman ES, Davis LL, Zisook S, Wisniewski SR, Trivedi MH, Fava M, Rush AJ, Team C-MS. Baseline depression severity as a predictor of single and combination antidepressant treatment outcome: results from the CO-MED trial. Eur Neuropsychopharmacol. 2012;22:183-199.

37. Mulder RT, Joyce PR, Frampton CM, Luty SE, Sullivan PF. Six months of treatment for depression: outcome and predictors of the course of illness. Am J Psychiatry. 2006;163:95-100.

38. Jain FA, Hunter AM, Brooks JO, 3rd, Leuchter AF. Predictive socioeconomic and clinical profiles of antidepressant response and remission. Depress Anxiety. 2013;30:624-630.

39. Zhu H, Bogdanov MB, Boyle SH, Matson W, Sharma S, Matson S, Churchill E, Fiehn O, Rush JA, Krishnan RR, Pickering E, Delnomdedieu M, Kaddurah-Daouk R, Pharmacometabolomics Research N. Pharmacometabolomics of response to sertraline and to placebo in major depressive disorder - possible role for methoxyindole pathway. PLoS One. 2013;8:e68283.

40. Chekroud AM, Gueorguieva R, Krumholz HM, Trivedi MH, Krystal JH, McCarthy G. Reevaluating the Efficacy and Predictability of Antidepressant Treatments: A Symptom Clustering Approach. JAMA Psychiatry. 2017;74:370-378.

41. Turner NC, Neven P, Loibl S, Andre F. Advances in the treatment of advanced oestrogen-receptor-positive breast cancer. Lancet. 2016.

42. Klengel T, Binder EB. Gene x environment interactions in the prediction of response to antidepressant treatment. Int J Neuropsychopharmacol. 2013;16:701-711.

43. Serretti A, Chiesa A, Calati R, Perna G, Bellodi L, De Ronchi D. Common genetic, clinical, demographic and psychosocial predictors of response to pharmacotherapy in mood and anxiety disorders. Int Clin Psychopharmacol. 2009;24:1-18.

44. Lawton KA, Berger A, Mitchell M, Milgram KE, Evans AM, Guo L, Hanson RW, Kalhan SC, Ryals JA, Milburn MV. Analysis of the adult human plasma metabolome. Pharmacogenomics. 2008;9:383-397.

45. Fava M, Rush AJ, Alpert JE, Balasubramani GK, Wisniewski SR, Carmin CN, Biggs MM, Zisook S, Leuchter A, Howland R, Warden D, Trivedi MH. Difference in treatment outcome in outpatients with anxious versus nonanxious depression: a STAR*D report. Am J Psychiatry. 2008;165:342-351.

46. Townsend MK, Clish CB, Kraft P, Wu C, Souza AL, Deik AA, Tworoger SS, Wolpin BM. Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. Clin Chem. 2013;59:1657-1667.

FIGURES

Figure 1: The three-stage analysis workflow. Our analysis workflow proceeded in three stages. In STAGE-1, we established sex differences in metabolomics profiles. In STAGE-2, we identified depressive symptom severity clusters in the PGRN-AMPS dataset using a data-driven approach (Stage 2A) separately for men and women, which we then validated using data from STAR*D and ISPC (Stage 2B). Factors that differentiated the validated depressive symptom clusters were idenbtified in Stage 2C using PGRN-AMPS data. The final predictive model was then developed and tested in STAGE-3.

Figure 2: Metabolites exhibiting significant concentration changes between baseline and 4 weeks, and baseline and 8 weeks, in depressed patients, stratified by clinical outcome. There were significant changes from baseline in the concentrations of 5-HT 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. Differences between men and women by outcome type were observed for the remaining metabolites (shown in black colored text).

Figure 3: Depressive symptom based clusters identified by data-driven unsupervised learning using Gaussian Mixture Model (GMM). The distribution of total depression severity

scores are represented by box plots, and width of the box is proportional to the number of patients comprising the cluster.

Figure 4: Comparison of depressive symptom clustering behavior using various approaches. Shown here are the probability density functions (PDFs) of our clustering approach using univariate total depression severity scores at 8 weeks, with a Gaussian Mixture Model (GMM) (first column), as compared with clustering patients using hierarchical clustering approaches with multivariate data (comprising individual item responses at 8 weeks with [middle column] and without [third column] total depression severity scores at 8 weeks). The probability density in each figure are represented on the y-axis, and depressive symptoms 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-colored vertical line in each plot. The ecological validity of the GMM clustering approach using univariate depression severity scores is represented by the 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.

TABLES

Table 1: Prediction performance SVM-RBF used to predict clinical outcomes at 8 weeks: Table A – using baseline depression symptom severity and baseline metabolomics data, and Table B – using baseline depression symptom severity, baseline metabolomics data and SNP genotype data.

TABLE 1-A

Prediction performance support vector machine with radial basis functions (SVM-RBF) used to predict clinical outcomes at 8 weeks using baseline depression symptom severity and baseline metabolomics data

	MEN				WOMEN			
	QIDS-C		HDRS		QIDS-C		HDRS	
	REMISSION	RESPONSE	REMISSION	RESPONSE	REMISSION	RESPONSE	REMISSION	RESPONSE
Accuracy	0.7	0.72	0.75	0.76	0.71	0.875	0.7	0.81
95% Confidence Interval	(0.45,0.88)	(0.5,0.87)	(0.50,0.91)	(0.54,0.90)	(0.54,0.84)	(0.74,0.95)	(0.53,0.83)	(0.67,0.91)
NIR	0.5	0.52	0.5	0.52	0.5	0.54	0.5	0.54
p-value	0.05	0.03	0.02	0.01	0.0069	8.80E-07	0.008	8.20E-05
Sensitivity	0.8	0.833	0.9	0.5	0.94	0.72	0.8	0.63
Specificity	0.6	0.61	0.6	1	0.47	1	0.6	0.96
PPV	0.6667	0.67	0.69	1	0.64	1	0.67	0.93
NPV	0.75	0.8	0.85	0.68	0.9	0.81	0.75	0.75
AUC	0.69	0.78	0.63	0.91	0.9	0.88	0.78	0.82
Top-Predictors[a,b]	5HTP, I3PA, GTOCO3, GTOCOAVERAGE, THEOPHYLINE	5HTP, XAN, 5HT, I3PA, HX	I3PA, XAN, THEOPHYLINE, 5HTP, 5HIAA	5HT, GTOCO3, 4HBAC, 5HTP, I3PA	PARAXAN, 5HT, GUANOSINE, 3OHKY, 4HBAC	5HT, PARAXAN, ATOCO, HGA	CYS, 5HT, 4HBAC, 3OHKY, PARAXAN	4HBAC, URIC, 3OHKY, PARAXAN, HX

[a] See Supplementary Table 1 for definitions of abbreviated names for each metabolite.
[b] SNP rs numbers: TSPAN5 – rs-10516436, AHR – rs- 17137566, DEFB1_1 – rs- 5743467, DEFB1_2 – rs-2741130, DEFB1_3 – rs-2702877, ERICH3 – rs-696692
Key: AUC = area under the curve; NIR = null information rate; NPV = negative predictive value; PPV = positive predictive value.

TABLE 1-B

Prediction performance SVM-RBF used to predict clinical outcomes at 8 weeks using baseline depression symptom severity, baseline metabolomics data and SNP genotype data.

	MEN				WOMEN			
	QIDS-C		HDRS		QIDS-C		HDRS	
	REMISSION	RESPONSE	REMISSION	RESPONSE	REMISSION	RESPONSE	REMISSION	RESPONSE
Accuracy	0.8	0.84	0.9	0.84	0.71	0.87	0.9	0.91
95% Confidence Interval	(0.56,0.94)	(0.63,0.95)	(0.75,0.97)	(0.63,0.95)	(0.54,0.84)	(0.74,0.95)	(0.76,0.97)	(0.8,0.97)
NIR	0.5	0.52	0.5	0.52	0.5	0.54	0.5	0.54
p-value	0.005	0.0008	2.00E-05	0.0008	0.006	8.80E-07	9.28E-08	1.80E-08
Sensitivity	1	0.67	0.9	0.67	1	0.72	0.8	0.81
Specificity	0.6	1	1	1	0.42	1	1	1
PPV	0.7	1	1	1	0.63	1	1	1
NPV	1	0.74	0.9	0.76	1	0.81	0.83	0.86
AUC	0.74	0.9	0.92	0.83	0.9	0.87	0.92	0.92
Top-Predictors[a,b]	I3PA, AMTRP, GUANINE, GTOCO3, 5HTP	5HTP, DEFB1_2, 5HT, TSPAN5, XAN	I3PA, XAN, AMTRP, THEOPHYLINE, GTOCO3	5HT, GTOCO3, 4HBAC, 30HKY_BAC KWAVE, 5HTP	PARAXAN, 30HKY, GUANOSINE, THEOPHYLINE, 4HBAC	5HT, PARAXAN, ATOCO, DEFB1_2, HGA	5HT, 4HBAC, CYS, VMA, PARAXAN	4HBAC, MHPG, PARAXAN, 4HPLA, URIC

[a] See Supplementary Table 1 for definitions of abbreviated names for each metabolite.
[b] SNP rs numbers: TSPAN5 – rs-10516436, AHR – rs- 17137566, DEFB1_1 – rs- 5743467, DEFB1_2 – rs-2741130, DEFB1_3 – rs-2702877, ERICH3 – rs-696692
Key: AUC = area under the curve; NIR = null information rate; NPV = negative predictive value; PPV = positive predictive value.

SUPPLEMENTARY MATERIALS

SUPPLEMENTARY FIGURES

Supplementary Figure 1: Principal component analyses stratified by sex illustrating no visible sex-differences in clinical and demographic data. The first two principal components describe most of the observed variance in the clinical and demographic data.

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). Probability densities are proportional to the fraction of patients with the associated symptom severity scores.

Supplementary Figure 3: Clustering of Caucasian men in AMPS (Fig. A), Caucasian and Asian men in ISPC (Fig. B), Asian men in ISPC (Fig. C), and Caucasian men in ISPC (Fig. D). The distribution of total depression severity scores are represented by box plots, and width of the box is proportional to the number of patients comprising the cluster.

Supplementary Figure 4: Clustering of Caucasian women in AMPS (Fig. A), Caucasian and Asian women in ISPC (Fig. B), Asian women in ISPC (Fig. C), and Caucasian women in ISPC (Fig. D). The distribution of total depression severity scores are represented by box plots, and width of the box is proportional to the number of patients comprising the cluster.

Supplementary Figure 5: Comparison of mean ages for men and women in clusters with comparable symptom severity at baseline, 4 weeks and 8 weeks.

Supplementary Figure 6: 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.

Supplementary Figure 7: 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).

Supplementary Figure 8: 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 TABLES

SUPPLEMENTARY TABLE 1

Clinical and Demographic factors from AMPS 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 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
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 3

Sex differences in metabolomic profiles irrespective of outcome status at 8 weeks

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.

Supplementary Table 4

Prediction performance of support vector machines using radial basis function (SVM-RBF) used to predict clinical outcomes at 8 weeks using baseline depression symptom severity and SNP genotype data.

	MEN				WOMEN			
	QIDS		HAMD		QIDS		HAMD	
	REMISSION	RESPONSE	REMISSION	RESPONSE	REMISSION	RESPONSE	REMISSION	RESPONSE
Accuracy	0.7	0.76	0.75	0.84	0.88	0.85	0.6	0.83
95% Confidence Interval	(0.45,0.88)	(0.54,0.9)	(0.5,0.91)	(0.63,0.95)	(0.75,0.96)	(0.77, 0.93)	(0.43,0.75)	(0.69,0.95)
NIR	0.5	0.52	0.5	0.52	0.5	0.54	0.5	0.54
p-value	0.05	0.01	0.02	0.0008	3.00E-07	4.60E-06	0.13	2.10E-05
Sensitivity	0.7	0.67	0.7	0.67	0.94	0.81	0.7	0.81
Specificity	0.7	0.84	0.8	1	0.84	0.88	0.5	0.84
PPV	0.7	0.8	0.77	1	0.85	0.85	0.58	0.81
NPV	0.7	0.7	0.72	0.76	0.94	0.85	0.63	0.84
AUC	0.8	0.69	0.81	0.75	0.94	0.91	0.65	0.81

Key: AUC = area under the curve; NIR = null information rate; NPV = negative predictive value; PPV = positive predictive value.

Supplementary Table 5

Prediction performance random forest used to predict clinical outcomes at 8 weeks using baseline depression severity, baseline metabolomics data and SNP genotype data.

	MEN				WOMEN			
	QIDS		HAMD		QIDS		HAMD	
	REMISSION	RESPONSE	REMISSION	RESPONSE	REMISSION	RESPONSE	REMISSION	RESPONSE
Accuracy	75	0.84	0.85	0.8	0.84	0.85	0.825	0.875
95% Confidence Interval	(0.5,0.91)	(0.63,0.95)	(0.62,0.96)	(0.59,0.93)	(0.68,0.93)	(0.72,0.93)	(0.62,0.92)	(0.74, 0.95)
NIR	0.5	0.52	0.5	0.52	0.5	0.54	0.5	0.54
p-value	0.02	0.00089	1.00E-03	0.003	1.00E-05	4.60E-06	2.11E-05	8.80E-07
Sensitivity	0.9	0.83	0.9	0.67	0.94	0.72	0.9	0.81
Specificity	0.6	0.84	0.8	0.92	0.73	0.96	0.75	0.92
PPV	0.69	0.83	0.81	0.88	0.78	0.94	0.78	0.9
NPV	0.85	0.84	0.88	0.75	0.93	0.8	0.88	0.85
AUC	0.95	0.93	0.9	0.84	0.94	0.87	0.94	0.9

Key: AUC = area under the curve; NIR = null information rate; NPV = negative predictive value; PPV = positive predictive value.

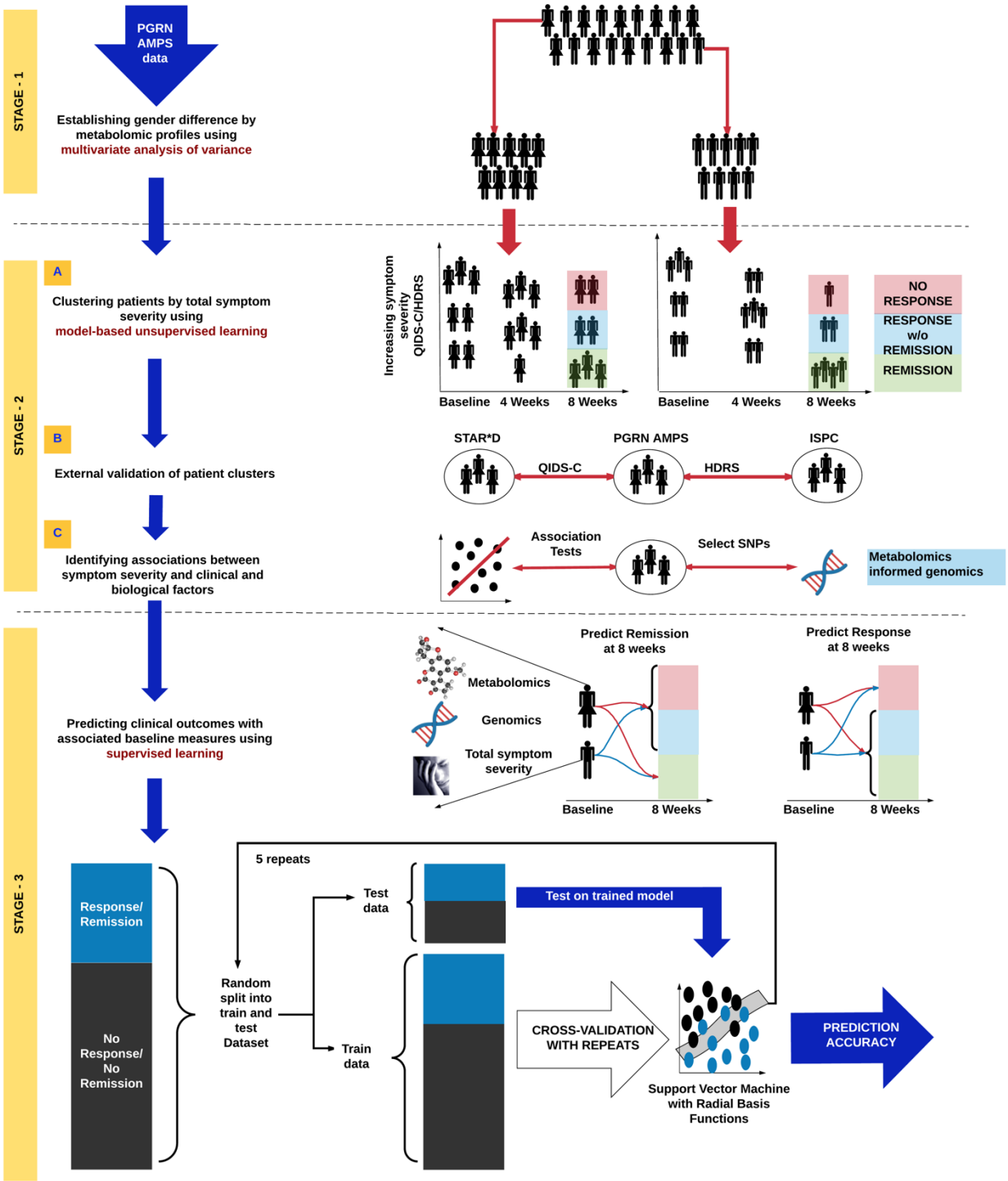

STAGE - 1
PGRN AMPS data
Establishing gender difference by metabolomic profiles using multivariate analysis of variance
STAGE - 2
A
Clustering patients by total symptom severity using model-based unsupervised learning
B
External validation of patient clusters
C
Identifying associations between symptom severity and clinical and biological factors
STAGE - 3
Predicting clinical outcomes with associated baseline measures using supervised learning
Increasing symptom severity QIDS-C/HDRS
Baseline 4 Weeks 8 Weeks
NO RESPONSE
RESPONSE w/o REMISSION
REMISSION
STAR*D
QIDS-C
PGRN AMPS
HDRS
ISPC
Association Tests
Select SNPs
Metabolomics informed genomics
Metabolomics
Genomics
Total symptom severity
Predict Remission at 8 weeks
Predict Response at 8 weeks
Baseline 8 Weeks
Response/ Remission
No Response/ No Remission
5 repeats
Random split into train and test Dataset
Test data
Train data
Test on trained model
CROSS-VALIDATION WITH REPEATS
Support Vector Machine with Radial Basis Functions
PREDICTION ACCURACY

Fig. 1

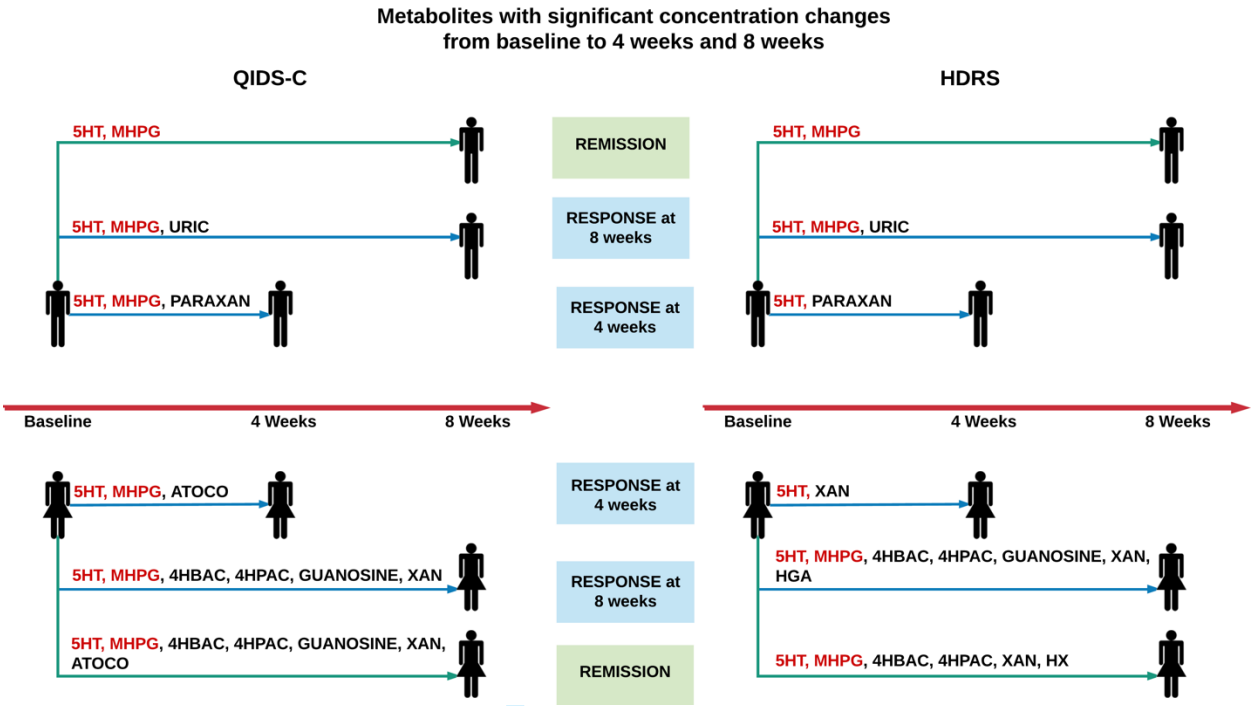


Fig. 2

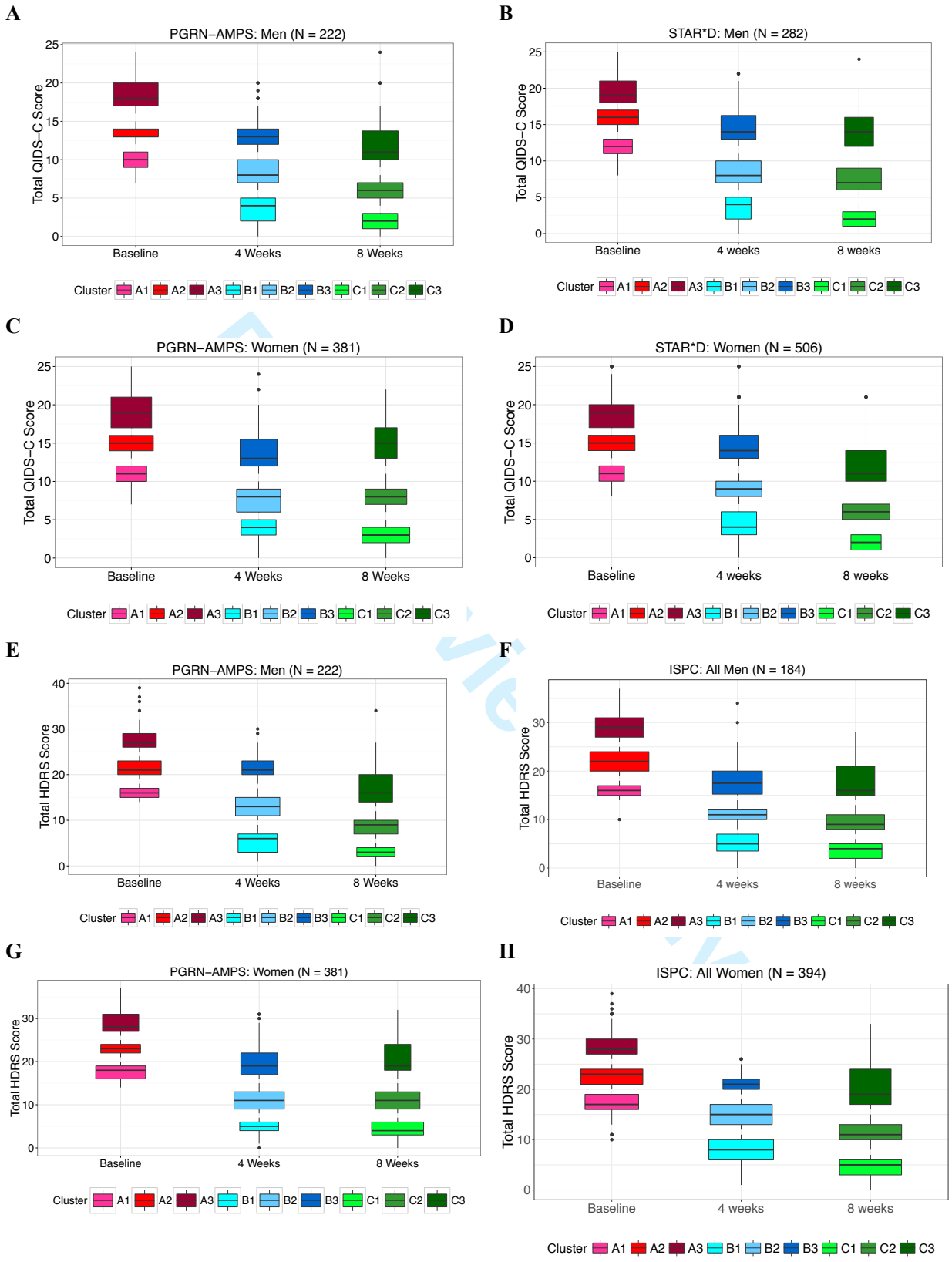

A
PGRN–AMPS: Men (N = 222)
Total QIDS–C Score
Baseline
4 Weeks
8 Weeks
B
STAR*D: Men (N = 282)
Total QIDS–C Score
Baseline
4 weeks
8 weeks
C
PGRN–AMPS: Women (N = 381)
Total QIDS–C Score
Baseline
4 Weeks
8 Weeks
D
STAR*D: Women (N = 506)
Total QIDS–C Score
Baseline
4 weeks
8 weeks
E
PGRN–AMPS: Men (N = 222)
Total HDRS Score
Baseline
4 Weeks
8 Weeks
F
ISPC: All Men (N = 184)
Total HDRS Score
Baseline
4 weeks
8 weeks
G
PGRN–AMPS: Women (N = 381)
Total HDRS Score
Baseline
4 Weeks
8 Weeks
H
ISPC: All Women (N = 394)
Total HDRS Score
Baseline
4 weeks
8 weeks
Cluster A1 A2 A3 B1 B2 B3 C1 C2 C3

Fig. 3

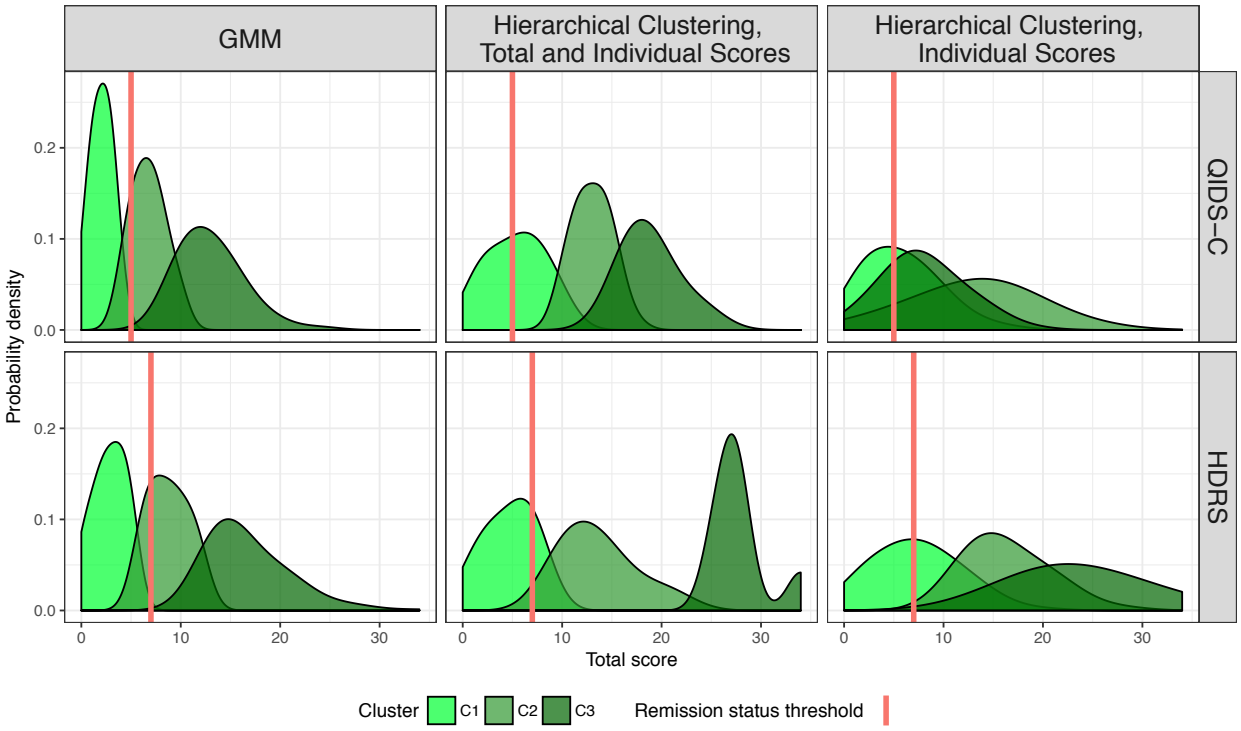
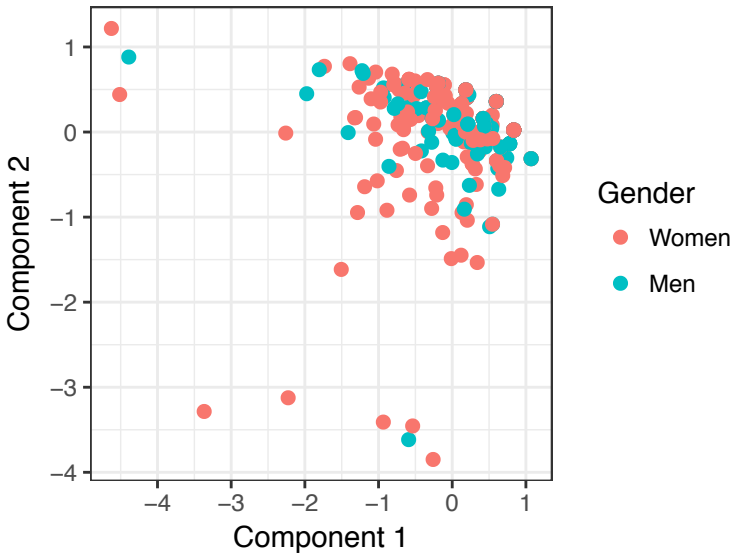
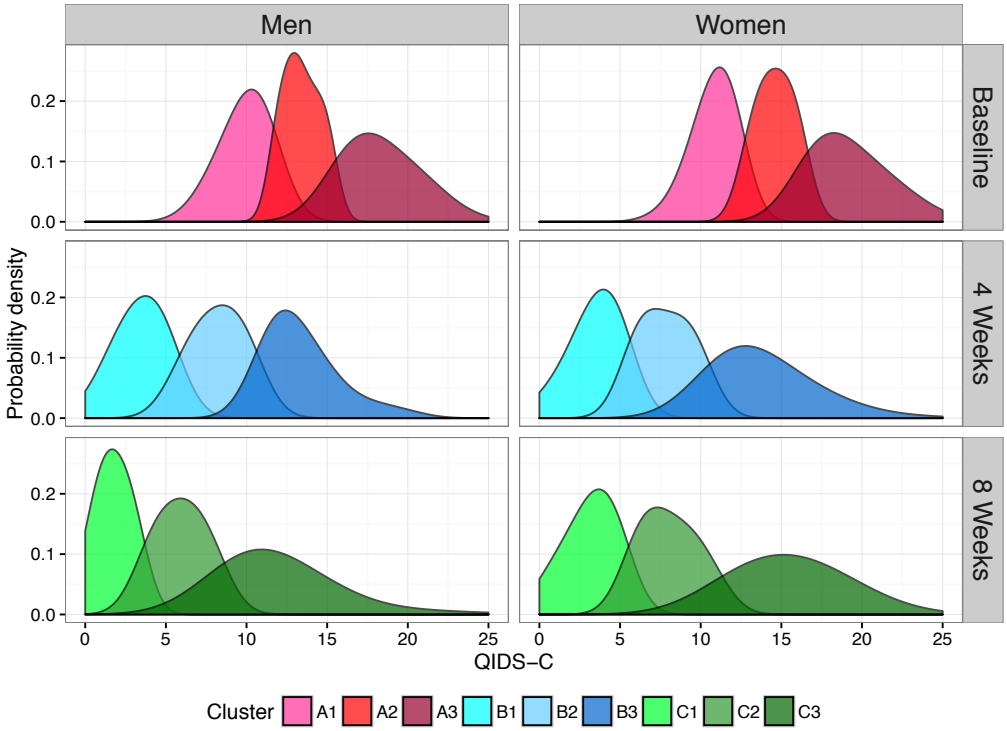


Fig. 4

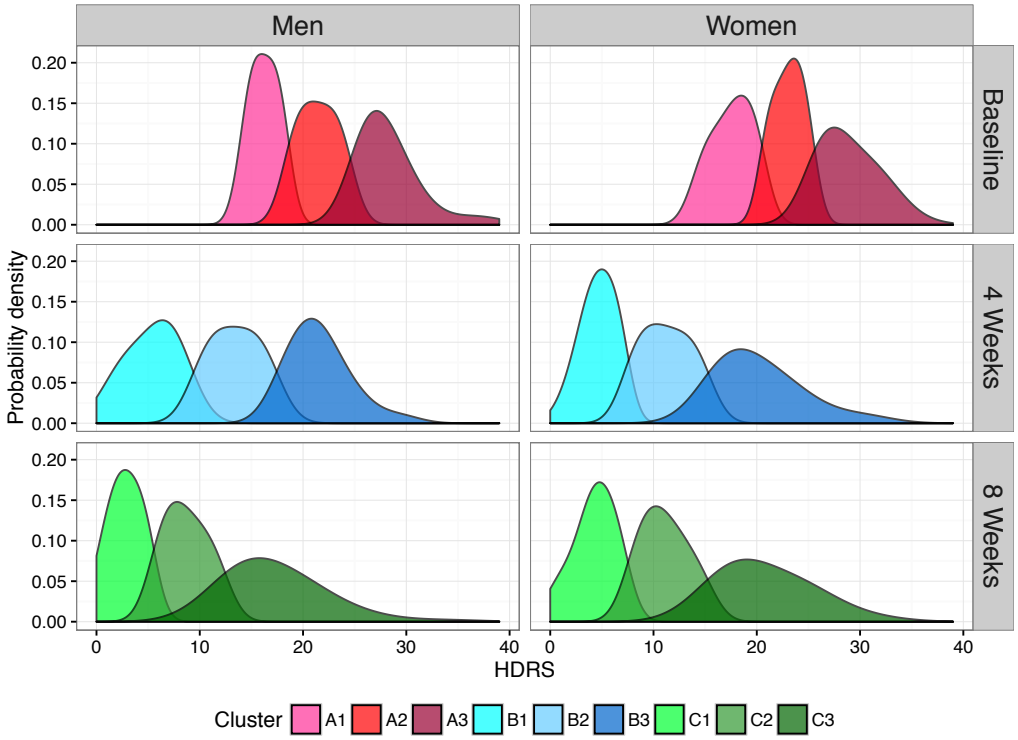


Supplementary Fig. 1

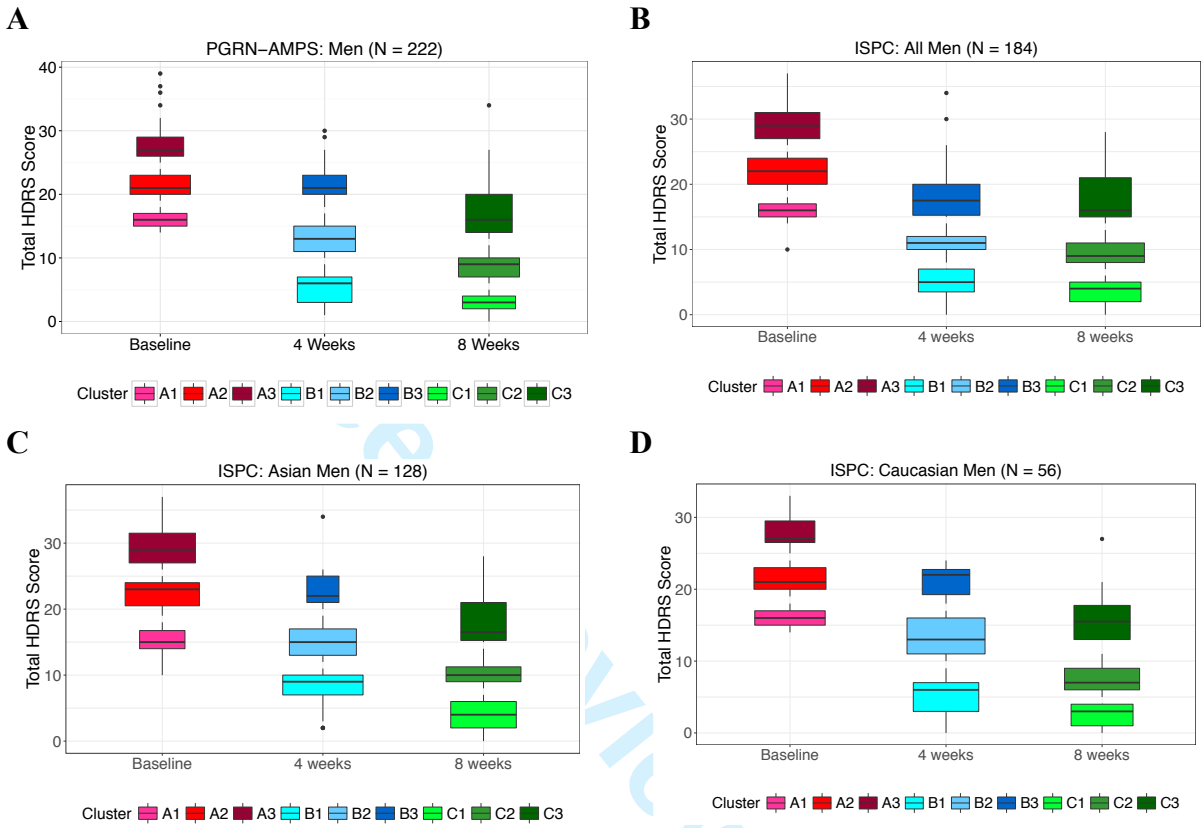
A



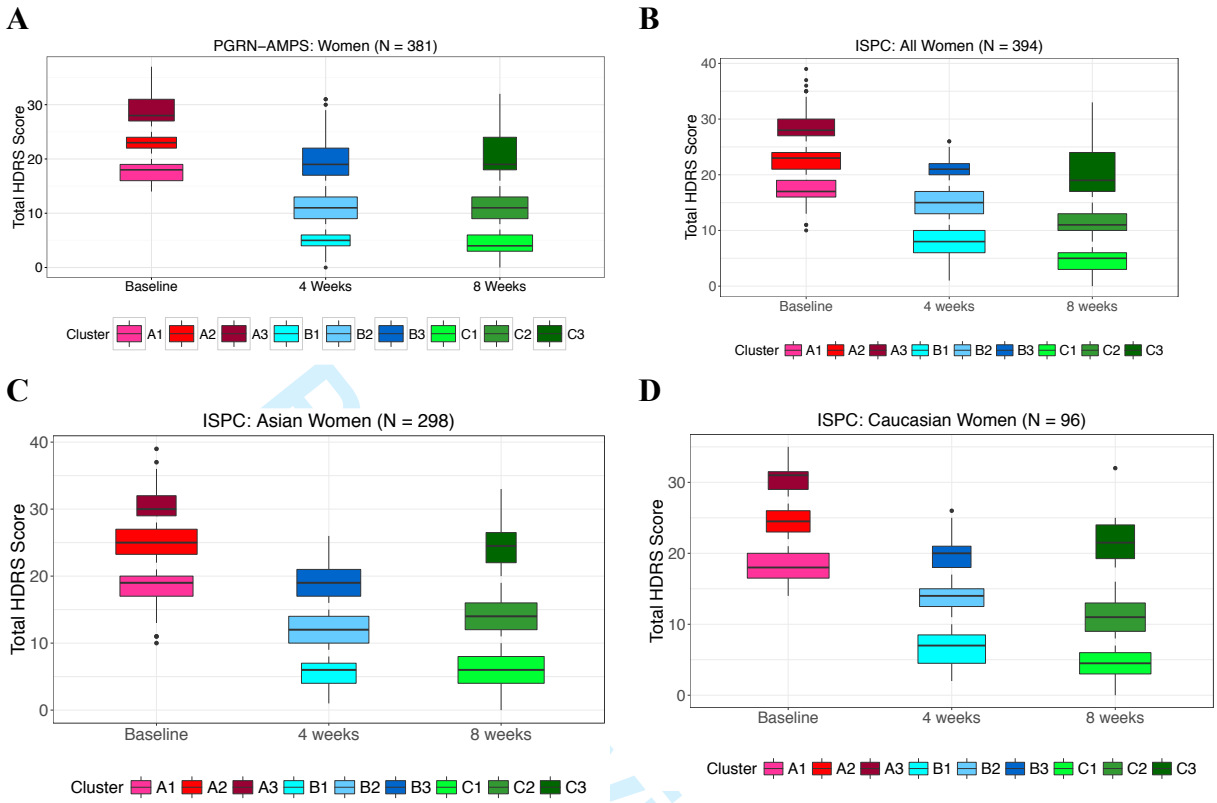
B



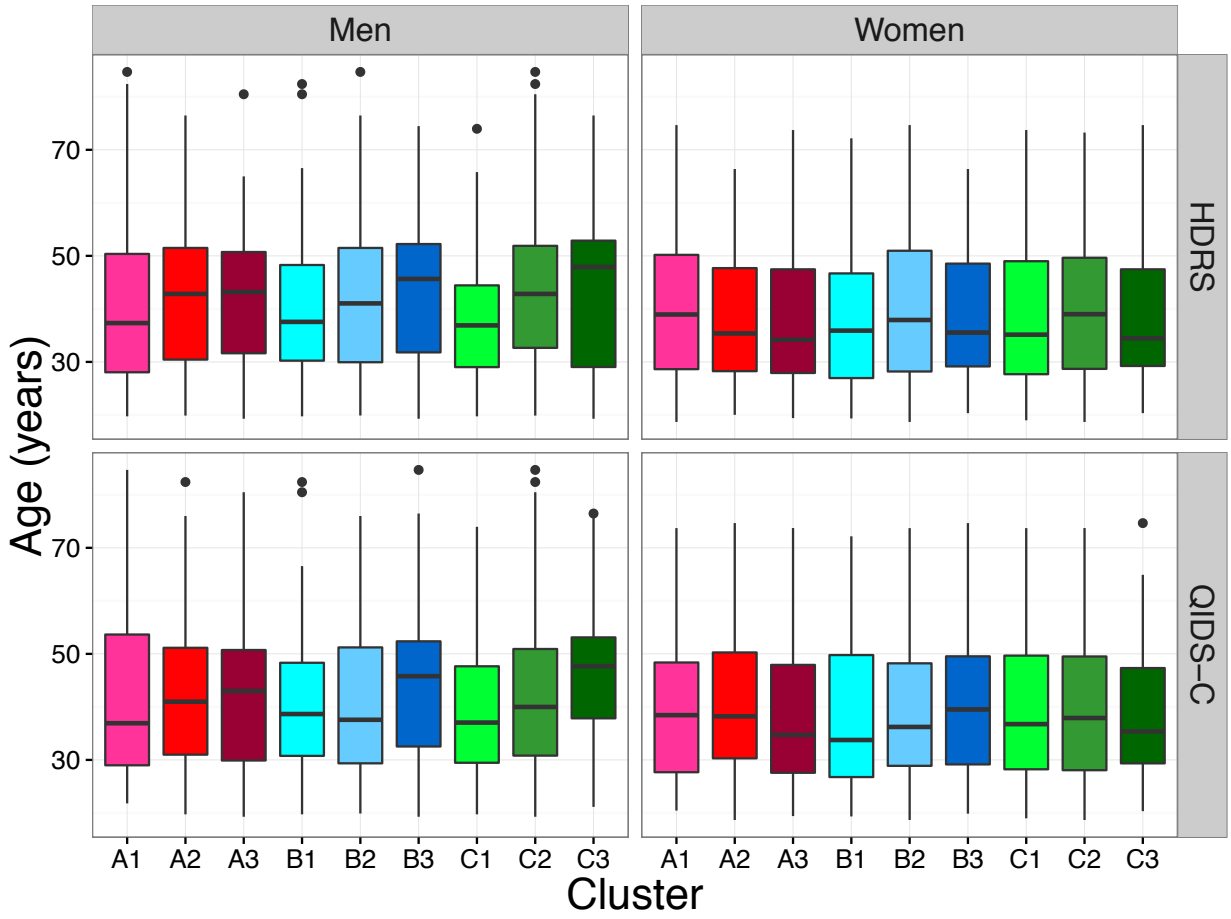
Supplementary Fig. 2



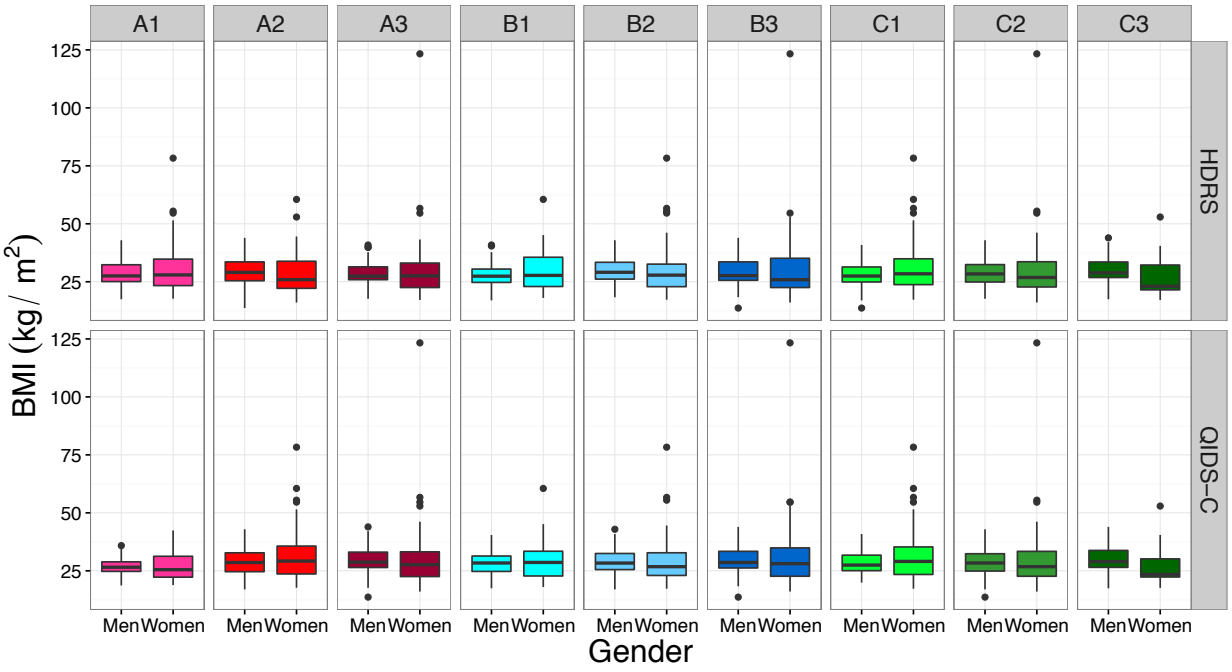
Supplementary Fig. 3



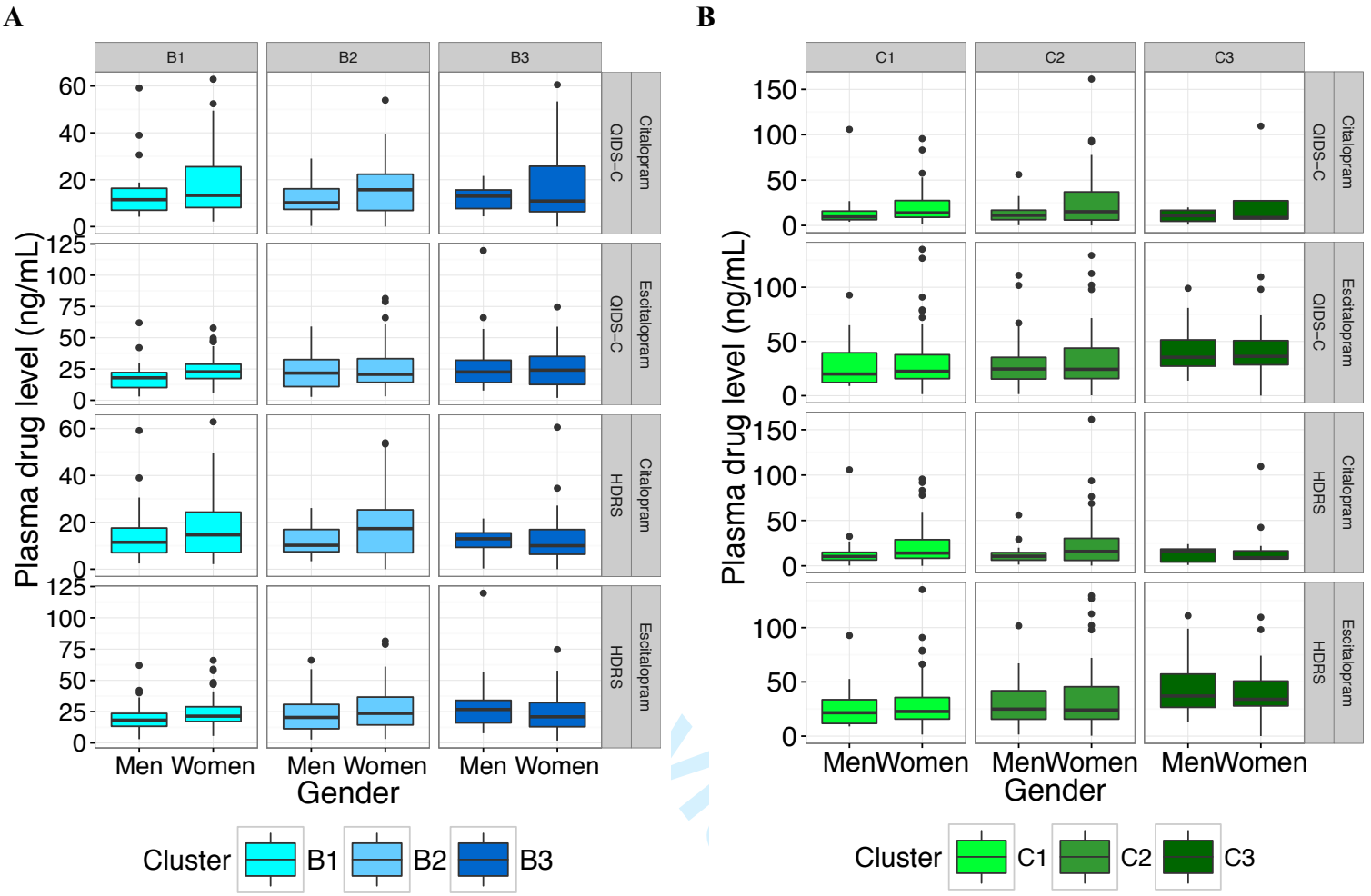
Supplementary Fig. 4



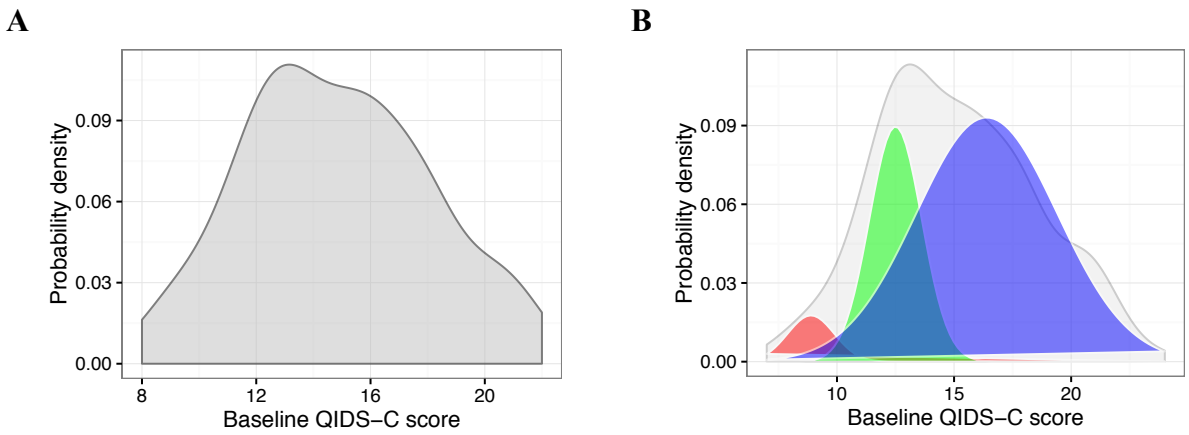
Supplementary Fig. 5



Supplementary Fig. 6



Supplementary Fig. 7



Supplementary Fig. 8