IRB# 19-005341

TITLE

Clinical Pilot of Augmented Human Intelligence in Major Depressive Disorder
(AHI/Depression Pilot)

INVESTIGATORS/KEY STUDY PERSONNEL

William V. Bobo, M.D., M.P.H.,[1] Mohit Chauhan, M.D.,[1] William Leasure, M.D.,[2] John Presutti, M.D., Ph.D.,[1] Joshua Keith, M.D.,[1] Mark Williams, M.D.,[2] Arjun Athreya, Ph.D.,[2] Paul E. Croarkin, M.D.,[2] Jennifer Vande Voort, M.D.,[2] Balwinder Singh, M.D.,[2] Abd Moain Abu Dabrh, M.B., B. Ch., M.S.,[1] Michelle Skime,[2] Cynthia Stoppel,[2] Liewei Wang, M.D.,[2] Leila Jones, Ph.D.,[2] Ravishankar Iyer, Ph.D.,[3] and Richard Weinshilboum, M.D.,[2] Mark A. Frye, M.D.,[2] Katherine M. Moore, M.D.,[2] Hannah K. Betcher, M.D.[2]

[1] Mayo Clinic, Jacksonville, FL (USA)
[2] Mayo Clinic, Rochester, MN (USA)
[3] University of Illinois, Urbana-Champaign, IL (USA)

ABSTRACT

This study will pilot the clinical implementation of a validated Augmented Human Intelligence (AHI)-driven algorithm and clinical decision support tool to predict the 8-week response to selective serotonin reuptake inhibitor (SSRI) and serotonin-norepinephrine reuptake inhibitor (SNRI) treatment of adults with a clinical diagnosis of Major Depressive Disorder. This study will enroll 120 patients over two years.

Major depressive disorder (MDD) is a chronic and severe psychiatric illness that affects over 15 million people in the U.S. and is the leading cause of disability worldwide. Treatment choices are often made on a "trial and error" basis and an individual patient can receive several months of ineffective treatments before a clear "non-response" profile can be established. Investigators at Mayo Clinic Florida, Mayo Clinic Rochester, and the University of Illinois at Urbana-Champaign (UIUC) have developed an augmented human intelligence (AHI) algorithm for response prediction during short-term treatment with SSRIs and SNRIs, validated in three large, 8-week SSRI clinical trial datasets, and an additional pooled dataset of 4 SNRI clinical trials of 8-12 weeks duration. The algorithm takes as inputs clinical and demographic characteristics at baseline, depression severity at baseline and at interim treatment time points (including a smaller sub-set of highly-prognostic individual depression symptoms) measured using validated rating scales, and genomic information (when available). The output is a probability of eventual non-response and remission of depressive symptoms at an interim treatment time point when that information can be used to inform clinical judgment as to whether the current course of treatment should continue or be changed. A web-based tool was developed for clinical decision support in naturalistic treatment settings, based on the AHI algorithm. Mayo Clinic is now poised to implement and test the effectiveness of the clinical decision support tool with an 8-week prospective, naturalistic pilot in the clinical practice. The main hypothesis to be tested is that the eventual antidepressant treatment outcome (non-

response, remission) predicted by the decision support tool at 2- and 4 weeks will have high concordance with the observed outcome at 8 weeks.

INTRODUCTION

MDD is a chronic and severe psychiatric illness that affects over 15 million people in the U.S. (Kessler et al., 2005) and is the leading cause of disability worldwide (WHO 2017). MDD can generally be managed with antidepressants (Belmaker & Agam 2008). However, the most common antidepressants for treating MDD are ineffective for many patients (Frochtmann et al., 2005; Gartlehner et al., 2012; Rush et al., 2004, 2006), and multiple therapeutic trials of antidepressants (each lasting several weeks) are typically needed to achieve meaningful benefit (Trivedi et al., 2006). This means that an individual patient can receive several months of ineffective treatments before a clear "non-response" profile can be established (Rush et al., 2003). Hence, at an early point during treatment, there is a compelling clinical need in primary care and specialty care settings, where systematic recording of depressive symptoms may be accomplished through interactive devices (e.g., computers, tablets, smart-phones), to know, *"given a patient's depression severity change after X weeks, are there specific thresholds of change in Y depression symptoms, such that there is a Z% chance of patient not responding favorably to additional weeks of treatment with the current antidepressant medication."* The translational impact of generating quantitative prognoses that forecast poor response based on early response to antidepressant medication rests in the availability of objective measures to prompt an early change in treatment. Avoiding extended weeks of medication that is unlikely to benefit the patient could potentially reduce disease burden and suffering.

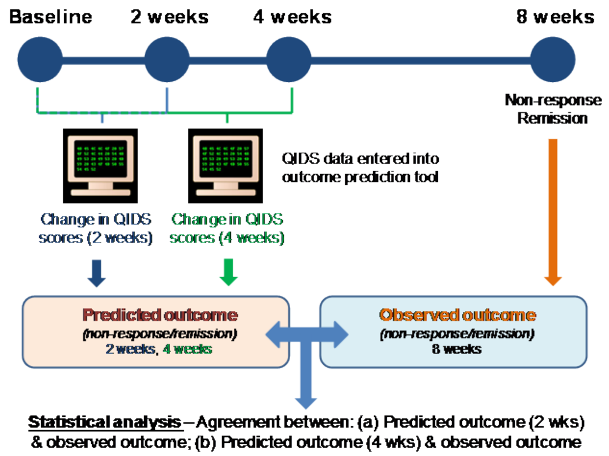
In a preliminary study, we developed a novel machine learning workflow that identified core symptoms from the full rating scales (QIDS-C and HDRS) using data from 3 large 8-week clinical trials of citalopram/escitalopram treatment in over 1,400 MDD patients. Supervised learning methods predicted sex-specific non-response and remission at 8 weeks with AUC 0.62–0.95 using the baseline severity of core symptoms and their associated changes at 4 weeks. The prognostic capabilities of the core symptoms identified using our approach replicated across all 3 clinical trials (manuscript in preparation). The algorithm was then applied to an additional pooled dataset of 4 randomized, placebo-controlled SNRI clinical trials, which resulted in a replication of core symptom identification and prognostic outcomes in the active treatment groups (manuscript in preparation). In both studies, specific thresholds of early improvement in core symptoms were identified that were highly-prognostic and predictive of non-response and remission after 8-12 weeks of antidepressant treatment.

Using the data from these two studies, we have since developed a web-based electronic decision support tool to derive predictions and prognostic evidence of likely non-response and remission based on early change in core depressive symptom severity. With the use of this AHI-based decision support tool, we are now poised to implement and test its effectiveness with an 8-week prospective, naturalistic pilot in the clinical practice.

STUDY OVERVIEW (Fig 1)

Adults (aged 18-64 years) who meet DSM-5 diagnostic criteria for non-psychotic unipolar major depressive disorder (MDD) and meet study eligibility criteria will receive 8 weeks of open-label treatment with an SSRI or SNRI antidepressant. Depressive symptoms will be assessed using the subject- and clinician-rated versions of the 16-items QIDS scale (QIDS-SR and QIDS-CR) and the 17-item HAMD at baseline, week 2 (via telephone), week 4, and week 8; with an additional phone contact at week 24. The QIDS scale scores at 2- and 4-weeks will be entered into the AHI-based clinical decision support tool, and the outcome predicted by the tool (e.g., the predicted eventual treatment outcome at 8 weeks) will be recorded. Clinicians and patients will be blinded to the clinical decision support tool prediction of outcome provided at the 2- and 4 weeks. The overarching hypothesis to be tested in this study is that the antidepressant treatment outcome (non-response, remission) predicted by the decision support tool at 2- and 4 weeks will have high concordance with the observed outcome at 8 weeks.

Figure 1. Study overview.



Note: As part of this protocol, blood will also be collected for genotyping at the baseline study visit. When available, a SNP panel will be performed to examine whether information on SNP profiles in four specific genes improve the accuracy of the predictions achieved using the AHI-based clinical decision support tool.

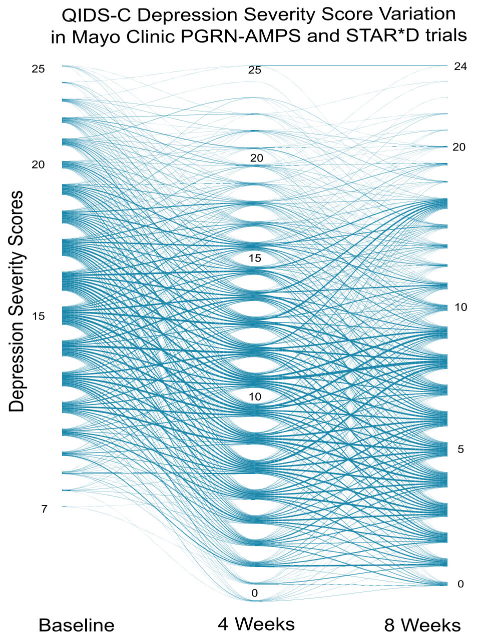
SPECIFIC AIMS

1. Evaluate the degree of statistical concordance between observed clinical outcomes (non-response/remission) after 8 weeks of treatment and the outcomes predicted by an AHI-based clinical decision support tool after 2 weeks of follow up (i.e., concordance between 2-week predicted outcome and 8-week observed outcome), as assessed using the QIDS-SR and QIDS-CR, in adults with DSM-5-defined MDD who receive prospective treatment with an SSRI or SNRI antidepressant.
2. Evaluate the degree of statistical concordance between observed clinical outcomes (non-response/remission) after 8 weeks of treatment and the outcomes predicted by an AHI-based clinical decision support tool after 4 weeks of follow up (i.e., concordance between 4-week predicted outcome and 8-week observed outcome), as assessed using the QIDS-SR and QIDS-CR, in adults with DSM-5-defined MDD who receive prospective treatment with an SSRI or SNRI antidepressant.
3.

BACKGROUND AND SIGNIFICANCE

For the reasons presented in the INTRODUCTION to this protocol, there is a crucial need for web-based tools that can be used to generate quantitative prognoses that can forecast a poor response to a given antidepressant medication in depressed patients, based on early responses to antidepressant medication. These objective measures can be used to augment clinician judgment in order to prompt an early change in treatment if the prognosis for non-response is high. Today, there is no evidence-based guidance on how much change in the overall depression score (or the scores for specific symptoms) is required to make those decisions.

Figure 2. Variation of depression severity during 8 weeks of SSRI treatment in 1,300 patients



The main challenge for developing clinically useful prognostic tools is the large heterogeneity (inter-individual variability) in response to antidepressant medication (Arnow et al., 2015; Fried 2017; Kessler et al., 2017; Musliner et al., 2016; Senn 2016). In measurement-based care of patients with MDD, standardized rating scales such as the Hamilton Depression Rating Scale (HDRS) (Hamilton 1960) and the Quick Inventory of Depressive Symptoms – clinician rated version (QIDS-CR) (Rush et al., 2003) are used to assess the total depression severity by adding up scores for scale items relating to sleep, appetite, mood, fatigability, etc. The ability to know whether an early change (e.g., at 2 or 4 weeks) in total depressive severity is prognostic of eventual outcomes (e.g., at 8 weeks) is challenged by the fact that patients with identical MDD severity respond differently to the same drug. In mathematical terms, if there are n levels of depression severity scores across k time-points of treatment, then there are n^k transitions – resulting in multiple hundreds of trajectories, as shown in Fig. 2. It is not tractable for even an experienced clinician to be able to memorize all possible trajectories of change in MDD severity to derive treatment prognoses at an individual patient level.

Others have attempted to utilize machine learning (ML) and artificial intelligence (AI) approaches to analyzing large datasets from clinical trials of antidepressants for depressed adults, with the goal of deriving more homogeneous patient subgroups or trajectories of clinical response (Chekroud et al., 2016, 2017; Iniesta et al., 2016). These approaches have relied on the use of baseline depressive symptoms and sociodemographic predictor variables. The clinical utility of these approaches is severely limited by the weak predictive effects of sociodemographic variables, either individually or in aggregate, for predicting outcomes of antidepressant treatment (Chekroud et al., 2016, 2017; Initesta et al., 2016). Furthermore, these ML/AI-based methods are unable to consider changes in depressive symptoms or other clinical assessments at intermediate time-points, before a therapeutic antidepressant trial is fully completed—a central concept in depression management (Crismon et al., 1999; Friedman et al., 2000; Kennedy et al., 2016).

To address the problem of antidepressant response heterogeneity while simultaneously considering changes in depressive symptoms at intermediate time points, we set out to derive more compact and data-driven representations of antidepressant response patterns in MDD patients and use this information to achieve interpretable (clinically useful) prognoses of antidepressant treatment outcome at an early stage of treatment. Briefly, we used probabilistic graphs to algorithmically explore the most-likely longitudinal variations (referred to as *symptom dynamic paths*) of total depression severity to achieve an eventual categorical treatment outcome using data from three large SSRI trials (n=1,846) and two rating scales—the HDRS and QIDS-CR (see PRELIMINARY RESULTS below). This approach provided the mathematical foundation needed to model conditional dependencies that follow a clinician's treatment logic, i.e., accounting for improvement in total depression severity, conditioned upon baseline depression severity and changes in depressive symptoms at intermediate time-points, in a purely data-driven manner, without a priori specification of trajectories (Athreya et al., 2017; Koller 2009). Second, using unsupervised machine learning, we identified a subset of individual depressive symptoms (referred to as *core depressive symptoms*) with homogeneous responses to symptom severity assessments, and found that their early (4-week) changes were prognostic and predictive of categorical treatment outcomes at 8 weeks (remission, response, non-response). Crucially, the symptom dynamic paths, core symptoms, and their predictive capabilities replicated across all three datasets and both rating scales. Specific thresholds of change in core symptoms were identified at 4 weeks that were highly prognostic of eventual outcome at 8 weeks. Subsequent application of this ML workflow to a large pooled dataset of duloxetine-treated adult patients with MDD (totaling 2,510 subjects) resulted in a replication of the same findings with SSRI antidepressants (manuscript in preparation). Additional methodological details of this preliminary work are provided below under the section entitled, PRELIMINARY STUDIES.

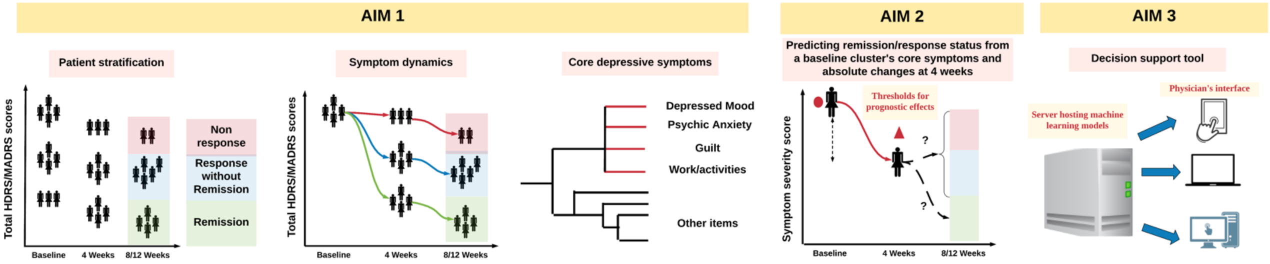
To increase the translational impact of our preliminary work, we built an interactive system that allows clinicians to input select demographic measures and depression severity scores, and obtain from the probabilistic graphical model output that conveys the predicted outcome and evidence supporting that prediction (see PRELIMINARY STUDIES below). While the methods developed in our preliminary work rely on retrospective analyses of clinical trial datasets, the aforementioned clinical impact is possible only when prognostic models reflect variability observed in naturalistic clinical settings and are shown to be valid prospectively in routine use settings. And this provides the motivation for the current study protocol.

PRELIMINARY STUDIES

Background

Preliminary studies have been conducted by Mayo Clinic and UIUC investigators in two phases. In the initial phase, we used data from 3 open trials of citalopram or escitalopram, two closely related SSRI antidepressants, to: (a) Cluster patients based on depressive symptoms severity, separately by sex, as a basis for modeling the change in depressive symptoms over time (symptom dynamic paths); (b) Identify a subset of core depressive symptoms with homogeneous longitudinal responses to study drugs; (c) Use core depressive symptom changes at 4 weeks to predict non-response, response (without remission), and remission at 8 weeks; and (d) Create a web-based decision support tool using the collected data. In the second phase, we replicated the findings from the first phase (with citalopram/escitalopram) using data from 9 randomized trials of duloxetine, and SNRI antidepressant, for treating adults with MDD, each with a follow-up duration of 8–12 weeks (NCT00406848, NCT00536471, NCT0073411, NCT00062673, NCT00036335, NCT02229825, NCT01000805, NCT010170329, and NCT02790970). Presented here are methods and results from the first phase (methods from the first phase were duplicated and the results were replicated in the second phase) and the third phase of our work. This workflow is summarized in Fig 3, where steps (a) and (b) are summarized under **Fig 3 Aim 1**, step (c) is summarized under **Fig 3 Aim 2**, and step (d) is summarized under **Fig 3 Aim 3**. We will conclude with a description of the successful integration of biological measures with our ML algorithm.

Figure 3. Overview of preliminary data workflow

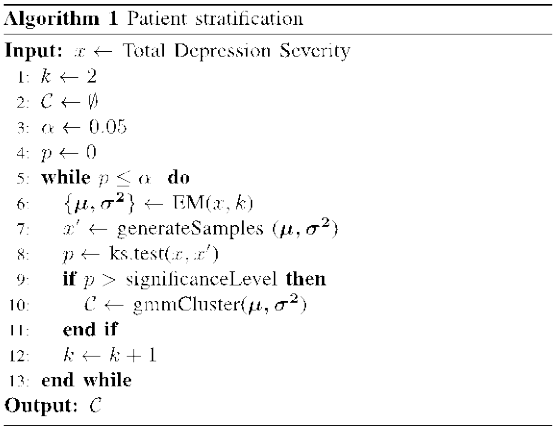


Data Sources

The primary source of data for our preliminary work in its first phase was the Pharmacogenomics Research Network Antidepressant Medical Pharmacogenomics Study (PGRN-AMPS, NCT 00613470). PGRN-AMPS was a single-arm, open trial designed to assess antidepressant effects of citalopram/escitalopram over 8 weeks in adults (aged 18−84 years) with MDD, and to examine metabolomic and genomic predictors of those outcome (Ji et al., 2013). Data from complete cases (baseline, 4-, and 8-week data) of step 1 of the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) trial (NCT 00021528) and International SSRI Pharmacogenomics Consortium (ISPC) were used to test the reproducibility of patterns of depressive symptom response inferred in the PGRN-AMPS study (Biernacka et al., 2015; Trivedi et al., 2006). The complete pooled dataset consisted of 1,846 adults, aged 18-75 years, with MDD.

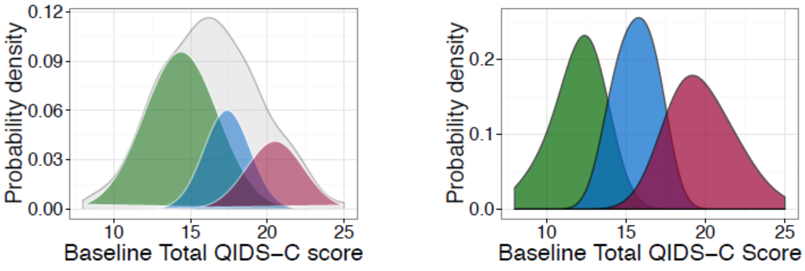
Fig 3 Aim 1 – Patient clustering and sex stratification

Unsupervised learning (Gaussian mixture models) were used to generate patient stratification (clusters) at baseline and at 4 and 8 weeks based on total QIDS-CR and HDRS17 scores (see Fig 4). The Shapiro-Wilk test was used to test whether the distribution of symptom severity scores is normal (Gaussian). Because total depression severity scores were not normally distributed (Fig. 4), we modeled the distribution of symptom severity scores as being composed of a mixture of Gaussians (k components of the distribution), where each Gaussian distribution represented a cluster of patients with similar symptom severity. We used an expectation-maximization (EM) algorithm that initially assumed only two (k=2, line 1 in Fig 4) components in the mixture (a single bell-shaped curve distribution) and gradually increased the number of components (multiple normal distributions) until an adequate fit of the data was achieved. For each value of k, we generated 10,000 samples (line 7 in Fig 4) using the parameters (mean and variance of a Gaussian component) estimated by the EM algorithm and compute p-values using Kolmogorov-Smirnov (line 8 in Fig 4) test. The process of increasing the components stopped at the first instance of p>0.05 (line 9 of Fig 4, i.e., test in line 8 fails to reject the null hypothesis that the estimated distribution and actual distribution of total depression severity scores are similar); otherwise, k was incremented by 1 (line 12 of Fig 4), and the process reverted back to line 5 of Fig 4. Under the assumption that the distribution was a mixture of k components (line 10 in Fig 4), patients were assigned to a cluster based on the components to which their scores belonged. Our solution yielded k=3 clusters, as shown in Fig 5, at baseline, 4 weeks, and 8 weeks. The 3 clusters of patients were labeled A1, A2, and A3; the clusters at 4 weeks were labeled B1, B2, and B3; and the clusters at 8 weeks were labeled C1, C2, and C3. At each time-point, the numeral 1 represented the mildest symptom cluster, 3 represented the most severe symptom cluster, and 2 represented an intermediate symptom cluster. Crucially, the clusters at 8 weeks (C1, C2, and C3) were shown to be ecologically (clinically) valid-- all patients in C1 achieved remission, all patients in C3 failed to achieve remission or response, and 87% of patients in C2 achieved response without remission.



```
Algorithm 1 Patient stratification
Input: $x \leftarrow$ Total Depression Severity
1: $k \leftarrow 2$
2: $C \leftarrow \emptyset$
3: $\alpha \leftarrow 0.05$
4: $p \leftarrow 0$
5: while $p \leq \alpha$ do
6:     $\{\mu, \sigma^2\} \leftarrow$ EM$(x, k)$
7:     $x' \leftarrow$ generateSamples $(\mu, \sigma^2)$
8:     $p \leftarrow$ ks.test$(x, x')$
9:     if $p >$ significanceLevel then
10:        $C \leftarrow$ gmmCluster$(\mu, \sigma^2)$
11:    end if
12:    $k \leftarrow k + 1$
13: end while
Output: $C$
```

Fig. 4: Patient stratification algorithm



(a) Estimating Distributions (b) Clusters from inferred GMM

Fig. 5: (a) shows the inference of mixtures comprising the distribution of symptom severity scores. (b) shows distributions of symptom severity within clusters inferred using the sufficient statistics of components inferred in (a).

Fig 3 Aim 1 – Modeling symptom dynamics pathways

Probabilistic Graphical Models (PGMs) with the forward algorithm were used to model how total QIDS-CR and HDRS17 scores changed during citalopram/escitalopram treatment by identifying the path each patient took starting from a given baseline cluster to a cluster at 4 weeks, and then from 4 weeks to 8 and 12 weeks. Using this approach, the nodes of the graph corresponded to symptom clusters at each time-point. We computed the likelihood of each path using the recursive forward algorithm by defining the graph as a hidden Markov model. The path likelihood of arriving at any other node was computed as a recursive sum of the product of the path likelihood of arriving at the precursor node, the transition probability from the precursor node to the current node (i.e., proportion of patients who moved between the two nodes), and the fraction of patients who achieved response status at the present node relative to their baseline severity. The paths for the pairs of baseline and 4- or 8-week clusters that had the highest likelihood scores were identified as the “most likely” paths, which we then referred to as symptom dynamics paths. The symptom dynamic paths (shown in Fig. 6 for women in PGRN-AMPS and ISPC) and distributions of total depression scores on the paths replicated in the PGRN-AMPS and STAR*D datasets (for QIDS-CR), and in the PGRN-AMPS and ISPC datasets (for HDRS17), including the most likely response status on these paths at 4 and 8 weeks. We thus demonstrated that a compact representation of antidepressant response was achieved by capturing a set of 9 significant trajectories of treatment response – as opposed to over 1,300 trajectories in Fig. 1.

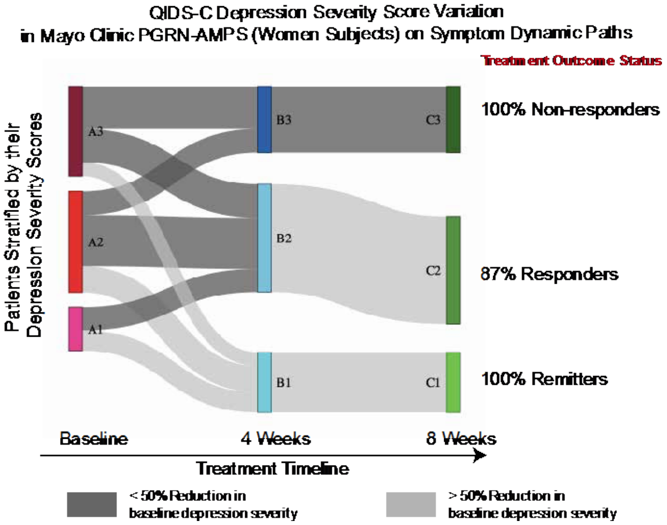


Fig. 6: Symptom dynamics, total depression severity.

Fig 3 Aim 1 – Extracting core depressive symptoms

To extract homogeneous patterns of citalopram/escitalopram response on the symptom dynamic paths, core depressive symptoms were defined using three criteria: (1) similar response patterns at all time-points, (2) low inter-individual variability, and (3) statistically distinct patterns of response at 4 weeks between symptom dynamic paths originating from a baseline cluster. Hierarchical clustering with complete linkage was used to identify individual depression scale items with similar rating patterns (clustered together with a common parent in the tree, other than the common parent of the entire hierarchy) within patient clusters at baseline, 4, and 8 weeks. Then, symptom clusters were identified wherein clinician ratings for each of the scale items at baseline had a nonzero median and low inter-individual variability (the chi-square test for the distribution of clinician ratings is significant after multiple comparisons, with the null hypothesis being that the distributions of ratings for that item are equal). Finally, the Kolmogorov-Smirnov test were used to determine if there were significant differences in the distributions of core symptom scores at 4 weeks between the symptom dynamic paths leading to non-response (<50% reduction in QIDS-CR or HDRS17 scores), response (>50% reduction in QIDS-CR or HDRS scores), and remission (QIDS-CR score ≤ 5, HDRS17 score ≤ 7) at 8 weeks, from a given baseline cluster. In the individual and combined datasets, 5 items in the QIDS-CR (sad mood, concentration/decision-making, self-outlook, involvement, and energy/fatigability) and 4 items in the HDRS17 (sad mood, psychic anxiety, guilt feelings/delusions, and work/activities) met the core depressive symptom criteria.

Fig 3 Aim 2 – Predicting non-response/remission after 8 weeks of SSRI treatment
We used random forests as a binary classifier to predict clinical outcomes at 8 weeks given a specific baseline cluster, using the baseline severity of core depressive symptoms and their changes at 4 weeks. AUC, PPV, NPV, sensitivity, and specificity were calculated as measures of algorithm performance, using the null information rate as a proxy for chance. To minimize the effects of overfit and information leak, nested cross-validation (nested-CV) with 5 repeats was used to train the classifiers and maximize the area under the curve. In our preliminary study of citalopram/escitalopram-treated adults, the predictive accuracies for non-response ranged from 71 to 84% (p<0.01, AUC 0.85–0.91) in women and from 89 to 92% (p<0.02, AUC 0.88−0.90) in men. For remission, predictive accuracies were 75–85% (p<0.04, AUC 0.84−0.89) in women and 63−95% (p<0.04, AUC 0.62–0.94) in men.

Fig 3 Aim 3 – Electronic decision support tool development
We have recently used data from our pilot work with SSRIs (and the SNRI, duloxetine) to build an interactive system (web-based decision support tool) that allows clinicians to input select demographic measures and depression severity scores, and obtain from the probabilistic graphical model output that conveys the predicted outcome and evidence supporting that prediction. The interactive system consists of a platform-independent front-end (i.e., it will work on desktops, laptops, smart phones and other hand-held devices) through which clinicians can input data and a back-end (e.g., a server hosting the ML framework and code) where results are computed. The front-end interacts with the trained prediction model in the server via a RESTful API that is integrated with iOS and Android mobile apps and HTTP-based websites, allowing clinicians to use a wide range of devices, including phones, tablets and workstations. The server is hosted at UIUC, which has vendor support from IBM Systems, Intel, and NVIDIA. Crucially, the analyses that provide the basis for the decision support tool's functionality are retrospective in nature; thus, prospective validation is needed—which serves as the motivation for the current study.

Additional Work – Integration of biological measures
Moreover, the ML framework described here allows for the integration of biological measures in future studies—which serves as the motivation for the blood collection as part of this protocol. We tested the capability of a basic version of our ML algorithm (using total depression scale scores) to include functionally validated pharmacogenomics

biomarkers joined with clinical measures to predict remission and response using data from 1,030 Caucasian MDD outpatients. The data source for this project, again, consisted of the PGRN-AMPS (n = 398), STAR*D (n = 467), and ISPC (n = 165) trials. GWAS for PGRN-AMPS plasma metabolites associated with SSRI response (serotonin) and baseline MDD severity (kynurenine) identified SNPs in DEFB1 (rs5743467, rs2741130, rs2702877), ERICH3 (rs696692), AHR (rs17137566), and TSPAN5 (rs10516436), which were tested as predictors. Supervised machine learning methods trained using SNPs and total baseline depression scores predicted remission and response at 8 weeks with AUC > 0.7 ($p<0.04$) in PGRN-AMPS patients, with comparable prediction accuracies >69% ($p<0.05$) in STAR*D and ISPC. None of the clinical or demographic predictor variables that were also tested in our ML algorithm were associated with depression severity at baseline or at 8 weeks. There was also a lack of a significant association between CYP2C19 metabolizer phenotypes with depression severity clusters at baseline and 8 weeks. These results demonstrate that machine learning can achieve accurate and, importantly, replicable prediction of SSRI therapy response using total baseline depression severity combined with pharmacogenomic biomarkers (Athreya et al., in press).

PARTICIPANTS

Entry Criteria

We will enroll 120 participants, both male and female. They will meet the following criteria:

(1) Outpatients with nonpsychotic MDD. Outpatient status assumes that the subject is not psychiatrically hospitalized or in an active suicidal crisis requiring hospitalization.
(2) A total score of ≥10 on the QIDS-CR and on the QIDS-SR (equivalent to a HAMD17 score of 13 [ids-qids.org, accessed on April 12, 2019]) given that when medication exceeds the effect of placebo in primary care, participants have a 17-item HAMD score >12. We added 2 HAMD points to take into account the possibility of measurement error. This is a very similar approach taken to the successful recruitment of subjects into the PGRN-AMPS trial.
(3) Antidepressant treatment is deemed appropriate by the study clinician.
(4) Adults who are between 18-64 years of age.

(5) Ability to provide informed consent
(6) Ability to understand English

Exclusion Criteria

The following will be exclusionary for participation in this study:
(1) A medical contraindication that precludes SSRI or SNRI treatment.
(2) Presence of a general medical condition that, in the opinion of this study clinician, is the cause of their depressive symptoms, will be exclusionary.
(3) People with treatment-resistant depression will be excluded from participating. For this study, treatment resistance will be defined as failure to respond to two or more adequate therapeutic trials of SSRIs **and** at least on SNRI therapeutic trial (sufficient

antidepressant dose, for 6 weeks or longer) during the current depressive episode. Failure to tolerate a therapeutic trial of a given medication (resulting in discontinuation due to adverse effects) will not be counted as exclusionary. Persons who have failed to respond to two or more adequately designed and executed therapeutic trials of SSRIs but have no history of least one failure to respond to SNRI treatment during the current depressive episode will be eligible to receive treatment with an SNRI in this study.

(4) Diagnosis of schizophrenia, schizoaffective disorder, bipolar I or II disorder, or bipolar disorder NOS (including other specified or other unspecified bipolar disorders) or a primary psychiatric condition that requires a different initial treatment than an antidepressant.

(5) Currently taking an antidepressant medication with subtherapeutic results in terms of antidepressive efficacy and unwilling to undergo a medication taper and discontinuation prior to initiation of a study drug from this protocol. The subject will be closely monitored by the study clinician during the medication taper and discontinuation phase. The design of the medication taper will be at the discretion of the study clinician. Subjects who cannot be safely tapered from their medication or who experience adverse effects during the taper that make further tapering infeasible will be excluded from the study.

(6) Use of antidepressant medication primarily for management of nicotine dependence, chronic pain, migraine prophylaxis or other diagnoses.

(7) Active substance use disorder. Persons in sustained full remission (> 12 months) and a negative urine drug of abuse screen at the screening visit will be considered eligible. Note: An additional urine drug screen will not be necessary for individuals with a negative urine drug screen documented in the medical record where the date of testing occurred within 12 weeks (84 days) of the screening/baseline study visit. However, study clinicians can still obtain a urine drug screen based on their clinical judgement even in participants with a negative drug screen within the 12 weeks preceding the screening/baseline study visit.

(8) Trazadone, melatonin, low-dose quetiapine (<100 mg QHS), z-drugs (zolpidem, zopiclone, eszopiclone, etc.), ramelteon, and diphenhydramine may be used as rescue medications for insomnia. Benzodiazepines may be used for treatment of anxiety, and atomoxetine may be used for the treatment of attention deficit disorder. Study subjects currently on antipsychotic medications (e.g., typical and atypical antipsychotic drugs, other than low-dose quetiapine for insomnia) and mood stabilizing agents (e.g., lithium, carbamazepine, valproate, lamotrigine) are not eligible for the study

(9) Pregnant subjects and those who are currently breastfeeding and who plan to continue breastfeeding will be excluded.

(10) Persons currently undergoing ECT, TMS or DBS as acute series or for maintenance.

(11) Patients currently psychiatrically hospitalized or in an active suicidal crisis requiring hospitalization in the opinion of the study clinician.

As an additional stipulation, individuals whose total QIDS-CR and total QIDS-SR scores are 10 or higher at the screening visit but decrease (improve) to total scores less than 10 on either the QIDS-CR or QIDS-SR at the baseline visit will be excluded.

METHODS

1. Patient Identification and Recruitment

Patient recruitment will be focused on the Mayo Clinic Rochester and Mayo Clinic Florida outpatient practices of the Department of Psychiatry and Psychology as well as Mayo Departments of Family Medicine, Preventive and Occupational Health and Community Internal Medicine and other medical providers also treat a large number of patients with MDD, so coordination will also occur with these departments. Study coordinators (SC) will coordinate recruitment efforts through daily contact with clinicians conducting new psychiatric evaluations.

2. Screening Visit

At the screening visit, potential study subjects will:

(1) Complete a brief form ascertaining basic demographic information and clinical information including prior course of the illness, prior suicide attempts, family history of MDD or bipolar disorder, current general medical illnesses, history of bone marrow or liver transplant or blood transfusion within the previous 6 weeks, and prior history of treatment during the current major depressive episode; and
(2) Complete a QIDS-SR.

At the screening visit, the study coordinator will:

(1) Obtain informed consent. Patients will have the opportunity to have the Mayo IRB patient advocate or a significant other present during the consent procedure if desired;
(2) Review the clinical and demographic information provided by the patient (see [1] above), including responses to items about prior course of the illness, prior suicide attempts, family history of MDD or bipolar disorder, current general medical illnesses, history of bone marrow or liver transplant or blood transfusion within the previous 6 weeks, and prior history of treatment during the current major depressive episode. The information regarding psychiatric and family history gathered via this process will be entered as a Research Note in the electronic health record to ensure that adequate information is available for follow-up after completion of the study;
(3) Measure the patient's height and weight;
(4) Ask participants for their permission to be contacted in the future to obtain additional information for the study;
(5) Complete the Structured Clinical Interview for DSM-IV (SCID) [First et al, 1995] to confirm a diagnosis of MDD;
(6) Administer the QIDS-C16; and
(7) Review the study inclusion/exclusion criteria.

Role of the study clinician: The study clinician will review the information provided by the patient and reviewed by the study coordinator and conduct a basic clinical interview to confirm eligibility for study participation. Safety issues, including suicidal ideation and risk for harming others, will be assessed.

When appropriate after a negative pregnancy test and/or urine drug of abuse screen, the study coordinator will invite subjects to participate in the study if entry criteria are met. If the subject agrees, he/she will receive treatment as described below. If the subject declines to participate, he/she will receive clinical care as usual from their primary physician or psychiatrist.

3. Research Evaluation Treatment Schedule

Participants selected for inclusion in the study will be seen at weeks 0 (baseline), 4, and 8, and telephone interviews will be scheduled at 2 and 24 weeks. For those subjects who are unable to attend the follow-up visits as scheduled, we will contact them by phone to complete the QIDS-CR and Frequency, Intensity, and Burden of Side Effects Ratings (FIBSER)/Antidepressant Side-Effect Checklist (ASEC) (Uher et al. 2009; Wisniewski et al., 2006). A summary of assessments and rating scales by follow-up time point is presented below in Table 1.

Table 1. Assessments and scales by follow-up time point.

Assessment	Baseline	Week 2 (telephone)	Week 4	Week 8	Week 24 (telephone)
QIDS-CR	•	•	•	•	•
QIDS-SR	•		•	•	
HAMD-17	•	•	•	•	•
CGI-S	•		•	•	
CGI-I			•	•	
C-SSRS*	•		•	•	
PETS			•	•	
FIBSER/ASEC		•	•	•	•
Side effects†		•	•	•	•
Vital signs	•		•	•	
ACE/MOSSS	•				
HCG/UDS‡	•				

* Refers to the C-SSRS screen version.

† Refers to spontaneous ascertainment of adverse effects by subject report.

‡ A urine pregnancy test (HCG) and/or urine drug screen will be ordered at baseline. The HCG will be ordered only for women under age 45 years.

Prior to Baseline Visit.

Some enrollees who are already taking antidepressants will undergo a medication taper and discontinuation prior to initiation of a study drug from this protocol. The rate of tapering will be determined by the study clinician, with the goal of taper completion within 2-8 weeks, depending on the current dose and antidepressant selected. Once the medication is fully tapered and discontinued, patients will be eligible for their baseline study visit, when study medication will be initiated. For patients taking fluoxetine prior to study entry, the medication will be able to be stopped without tapering in most cases.

Baseline Visit.
Note: If medication allows, the screening and baseline visits can occur on the same day. In other words, for patients will are identified by clinicians as needing to start antidepressants, this will be the first study visit for those participants.

(1) Clinical ratings: Before each visit with study clinicians, subjects will complete a QIDS-SR, ACE, and MOSSS; and the study coordinator will complete a QIDS-CR, 17-item Hamilton Depression Rating Scale (HAMD17) (Hamilton 1967), and Columbia-Suicide Severity Rating Scale (C-SSRS) screening version (Posner et al., 2011). QIDS-SR and –CR scale scores will be made available to study clinicians, as the latter will serve as a guide for clinical decisions about changing the dose of study drugs. At the end of the face to face visit, study clinicians will provide a CGI-S score (Guy 1976).
(2) Initial data entry into clinical decision support tool: Study clinicians will enter information about subject age, sex, QIDS-CR, and HAMD17 scores. This information will be used in conjunction with changes in depressive symptom scores at the week 4 face to face study visit to generate an output that specifies the prognoses of achieving eventual remission, response, and non-response at week 8.
(3) Sample collection: At the baseline visit, we will obtain blood for DNA using two EDTA tubes, DNA isolation-50ng/ml that will be sent to the BAP Lab for storage. All venipunctures will be performed using standard techniques.
(4) Vital signs: At the baseline visit, subjects will have their blood pressure, heart rate, height, and body weight measured.
(5) Study drug initiation: Study drugs will include all SSRI and SNRI antidepressants available for clinical use in the U.S. All study drugs will be initiated at standard doses, as shown in Table 2. For drugs with multiple indications, the doses chosen for this study are based on those recommended by the manufacturer.
(6) Scheduling next study visit(s): The study coordinator will schedule the week 2 telephone follow-up visit and the week 4 face to face study visit.

Table 2. Study drugs and standard starting doses and titration guidelines

Study drug name	**Starting dose (mg/day)**	**Titration guideline (optional)**
Citalopram (Celexa)	20	Increase to 40 mg if non-responder at wk 4
Escitalopram (Lexapro)	10	Increase to 20 mg if non-responder at wk 4
Fluoxetine	20	Increase to 40 mg if non-responder at wk 4
Fluvoxamine	50	Increase to 100mg if non-responder at wk 4
Paroxetine	20	Increase to 40 mg if non-responder at wk 4
Sertraline	50	Increase to 100 mg if non-responder at wk 4
Desvenlafaxine	50	Increase to 100 mg if non-responder at wk 4
Duloxetine	60*	Increase to 90 mg if non-responder at wk 4
Levomilnacipran	40†	Increase to 80 mg if non-responder at wk 4

Venlafaxine	75	Increase to 150 mg if non-responder at wk 4

* The total duloxetine dose can be given once daily, usually in the morning, or on a BID schedule (e.g., 30 mg BID). The study clinician may start the medication at a dose of 30 mg once daily and increase after 7 days as tolerated to an initial dose of 60 mg daily (or 30 mg BID).

† Levomilnacipran is initiated at a dose of 20 mg once daily for 2 days, then increased to 40 mg once daily.

Telephone Visit at Week 2.

At week 2, the study coordinator will conduct a telephone visit with each subject that will serve as an interim check regarding, mainly, tolerability of study drugs. Although they are regarded as self-report measures, the FIBSER and ASEC will be administered telephonically by the study coordinator, and spontaneous report of adverse effects will be elicited via open questioning. In addition, the study coordinator will administer the QIDS-CR and HAMD17, enter this information into the clinical decision support tool interface, record the support tool output (probability of a non-response/remission at 8 weeks) on a standard form, and confirm scheduling of the next study visit at week 4.

Although study clinicians will not be needed under routine circumstances for the telephone visit at week 2, they will be available for support when it is requested by study coordinators, or for special circumstances. Such circumstances include, but are not limited to, the following:

(1) The research coordinator will alert a study clinician if a patient expresses a wish to die (suicidal ideation), or if there are responses of 1 or higher on item 12 of the QIDS-CR or on item 3 of the HAMD17;

(2) The research coordinator will alert a study clinician if a patient complains of distressing medication side-effects, or if there is a FIBSER item 3 score of 4 (marked impairment) or higher.

During these circumstances, a study clinician will assess the clinical situation via the telephone with the patient and initiate appropriate next-step management. This may include, but not necessarily be limited to, stopping study medication, changing the dose of study medication, providing urgent or more frequent face to face clinical visits, or sending the patient to the emergency room/activating emergency response. Such decisions will be at the discretion of the study clinician. Additional details are provided below under the section entitled, SUICIDE RISK MANAGEMENT PLAN.

Face to Face Visit at Week 4.

(1) Clinical ratings: At the week 4 visit, subjects will complete a QIDS-SR, FIBSER, ASEC, and the Patient Experience with Treatment and Self-management scale (PETS); and the study coordinator will complete a QIDS-CR, HAMD17, and C-SSRS. All of these scale scores will be made available to study clinicians. At the end of the face to face visit, study clinicians will provide CGI-S and CGI-I scores.

(2) Determination of response status: For this study, a positive antidepressive response will be defined as a reduction (improvement) in QIDS-CR total score of 50% or more, relative to their baseline total QIDS-CR score. Subjects who achieve this threshold of improvement will be considered positive responders. Those who do not will be considered non-responders. This will serve as a clinical guide to dosing, as described in item (5) below. Some responders will be further classified as having

achieved symptomatic remission, which will be defined in this study as a QIDS-CR total score of 5 or lower (Rush et al., 2006).

(3) Quantification of the probability of non-response/remission at 8 weeks: In order to ensure the blinding of the clinical decision support tool prediction, the study coordinator will enter the week 4 QIDS-CR and HAMD17 core symptom scores into the clinical decision support tool. Using the information entered at baseline and at week 4, the tool will generate an output that specifies the prognoses of achieving eventual remission, response, and non-response at week 8. The study coordinator will record the output from the decision support tool on a standard form. Study clinicians will complete a short form in which they will select the outcome that they believe will be most likely at 8 weeks (non-response, response without remission, remission).

(4) Vital signs: At the week 4 visit, subjects will have their blood pressure, heart rate, and body weight measured.

(5) Dosing of study drugs: Study clinicians will be allowed to adjust the dose of study drug at the week 4 visit (but changing medications or adding adjunctive medication, will not be allowed). The decision to adjust the dose of study drug will be made at the discretion of the study clinician based on their judgment, in collaboration with the patient in a shared decision making framework, Table 2 provides guidance for the adjustment of the dose of study drugs).

(6) Scheduling next study visit: The research coordinator will schedule the week 8 face to face study visit, and will remind the subject that the next visit (week 8) will be the last face to face visit as part of this research.

Face to Face Visit at Week 8.

(1) Clinical ratings: At the week 8 visit, subjects will complete a QIDS-SR, FIBSER, ASEC, and PETS; and the study coordinator will complete a QIDS-CR, HAMD17, and C-SSRS. All of these scale scores will be made available to study clinicians. At the end of the face to face visit, study clinicians will provide CGI-S and CGI-I scores.

(2) Determination of response status: At week 8, subjects will be classified as having non-response (those who have not achieved a > 50% reduction from baseline in QIDS-CR total score), response without remission (> 50% reduction in QIDS-CR total score from baseline, but total score still adds up to > 5), or remission (QIDS-CR total score ≤ 5). This will again serve as a clinical guide to dosing, as described in item (4) below.

(3) Vital signs: At the week 8 visit, subjects will have their blood pressure, heart rate, and body weight measured.

(4) Dosing of study drugs: Subjects will be reminded that the week 8 visit will be the last face to face study visit. Further treatment will be provided by the subject's regular health care provider(s). For the participants who are classified as having symptomatic remission at week 8, no dose adjustment will typically be needed. For persons who have not achieved remission at week 8 (including non-responders) an increase in the dose of the antidepressant, augmentation/combination therapy, or therapeutic switch should be considered. Depending on the specific situation, study clinicians may elect to initiate this process themselves in close coordination with the subject's regular health care provider, with subsequent face to face follow up provided by the subject's

regular health care provider(s). Otherwise, study clinicians may provide specific next-step treatment recommendations to the subject so that they can discuss them with their regular health care provider.

(5) Subsequent follow-up: The research coordinator will schedule the week 24 telephone study visit.

Telephone Visit at Week 24.

At week 24, the study coordinator will conduct a telephone visit with each subject that will serve as final check regarding depressive symptom severity and subsequent treatment that they have received, using a standardized form. This telephone follow up visit will also provide an opportunity to answer any questions that subjects may have about their participation in the study. During this telephone follow-up visit, the study coordinator will administer the QIDS-CR, HAMD17, FIBSER, and ASEC, and will ascertain spontaneous reports of adverse effects of current treatments.

As was the case for the telephone visit at week 2, study clinicians will not be needed under routine circumstances; however, they will be available for support when it is requested by study coordinators, or for special circumstances. Again, such circumstances include, but are not limited to, the following:

(1) The research coordinator will alert a study clinician if a patient expresses a wish to die (suicidal ideation), or if there are responses of 1 or higher on item 12 of the QIDS-CR or on item 3 of the HAMD17;

(2) The research coordinator will alert a study clinician if a patient complains of distressing medication side-effects, or if there is a FIBSER item 3 score of 4 (marked impairment) or higher.

During these circumstances, a study clinician will assess the clinical situation via the telephone with the patient and initiate appropriate next-step management. This may include, but not necessarily be limited to, stopping study medication, changing the dose of study medication, providing urgent or more frequent face to face clinical visits, or sending the patient to the emergency room/activating emergency response. Such decisions will be at the discretion of the study clinician. Additional details are provided below under the section entitled, SUICIDE RISK MANAGEMENT PLAN.

CLINICAL DECISION SUPPORT TOOL

Overview.

We have built an interactive web-based system that allows clinicians to input select demographic measures and depression severity scores, and obtain from the probabilistic graphical model output (see PRELIMINARY STUDIES) that conveys the predicted outcome and evidence supporting that prediction. The interactive system will consist of a platform-independent front-end (i.e., it will work on desktops, laptops, smart phones and other hand-held devices) through which clinicians can input data and a back-end (e.g., a server hosting the artificial intelligence/machine learning framework and code—again, see PRELIMINARY STUDIES) where results are computed. The front-end will interact with the trained prediction model in the server via a RESTful API that is integrated with iOS and Android mobile apps and HTTP-based websites, allowing clinicians to use a wide range of devices, including phones, tablets and workstations. The server is hosted at UIUC, which has vendor support from IBM Systems, Intel, and NVIDIA.

Subject Confidentiality.

The tool will be accessed via password, and only study clinicians and research coordinators will be provided such access codes (passwords). The only inputs into the tool made by study clinicians or research coordinators will be unique study subject ID numbers (unique to each participant), subject sex, and scores for the QIDS-CR and HAMD-17. No protected health information will be entered into the tool, such as subject names, addresses, telephone numbers, social security numbers, Mayo Clinic patient numbers, or any other data that could serve as identifiers.

PREMATURE DISCONTINUATION FROM THE STUDY

Subjects who are prematurely discontinued from the study will then receive standard treatment from a primary physician or psychiatrist with different medications and/or therapeutic modalities of treatment (e.g. psychotherapy, etc.), if necessary. Patients do have the option to switch to another antidepressant at the end of the 8 week trial if they have financial or insurance concerns. And as specified earlier, all patients who complete the 8 weeks of the medication trial will be contacted by phone at week 24.

A subject will be withdrawn from the study for any of the following reasons:

- Loss to follow-up, defined as the failure to keep scheduled face to face or telephone study visits despite a minimum of 3 attempts (e.g., missed the week 4 study visit as originally scheduled but reachable by phone, then failed to show for a rescheduled visit, and then failed once more to show after rescheduling the visit a second time) or failure to contact a subject who misses a study visit despite a minimum of 3 attempts (e.g., failed to show for the week 4 study visit as originally scheduled, then 3 messages left to reschedule the visit with all calls unreturned).
- Withdrawal of consent
- Death
- Positive response between the screening visit and baseline visit without benefit of study drug (described earlier)
- Violation of protocol procedures, as per investigators' judgment
- Any rating of CGI-I or 5 (much worse) or 6 (very much worse)
- Emergence of worsening psychotic or perceptual changes or mania, as determined by the study clinician
- The investigator believes it is in the best interest of the subject to discontinue the study (e.g., for safety or tolerability reasons such as an adverse event)
- The subject becomes pregnant

If a subject withdraws from the study for reasons other than withdrawal of consent, an early termination visit will be conducted at the time of discontinuation. Reason(s) for withdrawal will be documented in the subjects' study documents as a brief clinical note.

Collected samples (blood draw obtained at baseline) will be retained and used in accordance with the subjects' original separate informed consent for research samples. The subject may withdraw consent for research samples, in which case the sample will be destroyed and no further testing will take place.

SUICIDE RISK MANAGEMENT PLAN

Suicide risk will be assessed in multiple domains both with brief surveys filled out by patients along with clinical interview and assessment according to the procedures noted above. Suicide specific items included in outcomes measures include suicide ratings in the HAMD17 and the QIDS. For more extensive assessment and analysis, the C-SSRS screening version will be used. Finally, patients will be asked in an open-ended manner if they have concerns related to their safety. There are other safety features that pertain to suicide risk in this protocol. Participants will be contacted by telephone at week two by the research coordinator to ensure medication compliance and patient safety. The research assistants will be trained to notify a study clinician at the time of the clinical visits or the phone interviews if the subject endorses any suicidal ideation or plans. After week 8 of the study, the subject will receive ongoing care from their regular health care provider(s) and will be off of the medication protocol, although we anticipate that the majority will continue to be treated with the study drugs.

If there are concerns raised that are related to suicide risk assessed by investigator/research assistants the patient will then be assessed by a study clinician for acute suicidal risk and suitability for continuation in the study. Specifically, this consultant will assess the patient and make determination of whether the patient needs more in-depth assessment, additional appointments to ensure or more closely monitor their safety, or emergent hospitalization. This will also apply if patients are having worsening of depressive symptoms without an increase in thoughts of suicide.

As an added precaution, subjects will be given listings of numbers to call if there are concerns in between assessments by research coordinators. These numbers will include study coordinators at each respective study site, Emergency Department Information/Triage Desk number (904-953-2000 in Jacksonville; 507-255-5385 in Rochester), and the National Suicide Prevention Lifeline at 1-800-273-TALK.

STATISTICAL PLAN

Sample Size Determination and Power Estimate.

We hypothesize the remission rate in the study cohort will be 45%, based on the remission rate (45.8%) reported for the Eligible for Analysis subset (n=463) in PGRN-AMPS trial (Mrazek et al., 2014). We anticipate the clinical decision support's prediction will have an overall accuracy of 80% and a sensitivity to predict remission of 95%. Under those assumptions, the study is investigating an off diagonal difference of 16% between the prediction of treatment response and the true (observed) remission rate at 8 weeks. The study has 97% power to detect that difference using a McNemar's test with 120 total subjects included in the study at the alpha=0.05 level of significance. Thus if the clinical decision support tool is under/over calling the 8 week response rate (i.e., generating discordant observations with the true clinical remission rate), there will be high power to detect it. We estimate that about 30 subjects will drop out of the study or be lost to follow-up after the baseline visit. The target sample size of 120 subjects takes into account the estimated dropout rate..

Descriptive Analyses.
Cohort clinical and demographic characteristics will be summarized using means and frequency (%), as required, for all consented participants and by analysis set (defined below). Data will also be visually inspected for measures of dispersion to assess potential outliers that warrant data clarification prior to the formal analyses.

Primary Outcome Analysis.
The main statistical objective is the quantification of agreement between the predicted outcome at 4 weeks using the clinical decision tool and the observed outcome at 8 weeks based on QIDS-CR total scores at 8 weeks. Remission is defined as an 8-week QIDS-CR total score ≤ 5. In addition to the remission endpoint, a secondary ordinal endpoint consisting of remission, response (a50% reduction from baseline in QIDS-CR total score at 8 weeks but not meeting the remission definition) vs. non-response will be established. The week 4 predictions of each of these clinical outcomes will be provided by the tool as the probability of the single "most likely" outcome. The observed outcomes at 8 weeks will also be unitary (remission OR response without remission OR non-response).

Agreement between the week 4 prediction of the outcome and the 8 week clinical outcome will be compared using McNemar's test (or Bowker's generalized test of symmetry for the ordinal secondary outcome). This will establish if there's a significant difference between the predicted outcome, and the actual observed outcome at 8 weeks. Kappa statistics will also be calculated to illustrate the agreement between the new approach and the 8 week clinical outcome. In addition the sensitivity for prediction of remission and specificity of the clinical decision support tool will be calculated to illustrate the ability of the new approach to estimate the 8 week clinical outcome as well as to provide clinical interpretable summaries of the results of the McNemar test (e.g., examination of false positive and false negatives).

We will conduct a descriptive analysis of the subjects in disagreement between predicted outcomes and actual outcomes. We will also assessing clinician prediction accuracy in this study, as well as the accuracy of the clinical decision support tool, although the objectives of this study and the reasons for developing the clinical decision support tool do not include comparing predictive accuracies of the decision support tool versus clinician assessment.

Analysis Sets
The primary analysis will be conducted under the intention to treat principle. Missing 8 week clinical outcomes will be imputed for the primary analysis using worst case imputation (assumed discordance of the algorithm with clinical data), chained equations and random forest imputation strategies if necessary. The results from these three methods will be compared qualitatively and reported separately in instances where the interpretation of the data varies between imputation strategies. In addition, a full analysis

set will also be constructed. This analysis set will include all randomized subjects that have the clinical decision support tool prediction available at 4 weeks (i.e., participated in the first half of the protocol sufficiently to have a predicted 8 week data available). This full analysis set may also require imputation of the 8 week outcome in the event of missing data. If needed, the same trio of imputation strategies will be applied to this analysis set. A completers analysis will also be conducted.

RATING SCALES AND INSTRUMENTS

(1) HDRS-17: A 17-item clinician rating of depressive symptoms, scored on a 4-point scale (0 to 4) (range 0–54). Anchors are provided for each of the numbered scale points. Higher scores represent higher levels of depression. Its psychometric properties have been studied extensively in adults (Cusin et al., 2010). The HDRS-17 in adults has high interrater reliability0.80-0.98) and high test-retest reliability (up to 0.81) (Williams 1998). Validity of the HDRS ranges from 0.65 to 0.90 with global depression measures and other well validated clinician-rated measures such as the IDS-C and MADRS (Hamilton 2000). Scores of 7 of less after treatment is a generally accepted definition of depressive symptom remission (Frank, et al., 1991). The following severity ranges for the HDRS-17 have been advocated: no depression, or remission (0–7); mild depression(8–16); moderate depression (17–23); and severedepression (≥24) (Zimmerman et al., 2013).

(2) QIDS-SR-16: The QIDS-SR - 16 has highly acceptable psychometric properties, which supports the usefulness of this brief rating of depressive symptom severity in both clinical and research settings (Rush et al., 2003). Assessed prior to infusion with MADRS and on post-infusion day assessment. This will be used for secondary outcome measure of self-report remission of symptoms.

(3) CGI: Overall clinical judgment symptom severity (I) and change for a preceding phase (II). Dropout criteria will be any 1 rating of CGI II – 5 (much worse) or 6 (very much worse) (Guy 1979).

(4) C-SSRS: The screening version of the full-scale C-SSRS is a shorter version of the overall scale designed to assess suicidal ideation and behavior over recent months with triage categories.

(5) FIBSER: The FIBSER is a standard self-reported side-effect measure that was designed to be easily adopted into clinical practice for patients receiving treatment for depression. Using data from STAR*D, the FIBSER as shown to be reliable, with high correlations between observations taken a short time apart, and correlations decreasing as time between observations increased (Wisniewski et al., 2006). There were also consistent relationships between items over time. The FIBSER has both face and construct validity.

(6) ASEC: The ASEC is a simple self-report measure that is used to describe adverse reactions to antidepressants. In a recent validation study, the ASEC and the psychiatrist-rated UKU Side Effect Rating scale were repeatedly administered to 811 depressed adults who received open-label treatment with escitalopram or nortriptyline (Uher et al., 2009). There was good agreement between self-report and psychiatrists' ratings.

(7) MOSSS: We will use the eight-item modified MOSSS, which assesses the quality of social support in medical patients and has been shown to good psychometric properties that are similar to the full 19-item version of the scale (Moser et al., 2012).

(8) ACE: The ACE questionnaire is a 10-item self-rated survey that assessed exposures to adverse childhood experiences that occurred prior to 18 years of age (Felitti et al., 1998). The questionnaire consists of 10 main-stem yes/no questions, some of which have additional questions.

(9) PETS: We will use the PETS scale as a patient-reported measure of treatment burden. The PETS consists of 78 items, divided into 15 domains that include learning about health conditions/care, medications, difficulty with medication taking, medical appointments, health monitoring, exercise/physical therapy, diet, use of medical equipment, interpersonal challenges, medical/healthcare expenses, confusion/concern about medical information, healthcare providers, difficulty with healthcare services, role functioning/social activity limitations, and physical/mental exhaustion. Individual scale items are rated on a 4- or 5-point Likert scale (e.g., very easy to very difficult, not at all to very much, strongly agree to strongly disagree, never to always). Recent validation work documented good internal consistency (alpha 0.79 – 0.95), with higher PETS scores associated with greater treatment burden, and significant correlation with more distress, less satisfaction with medications, lower self-efficacy, worse physical and mental health, and lower convenience of healthcare ($p<0.001$ for all correlations) (Eton et al., 2017).

HUMAN SUBJECTS

Common Adverse Effects.

The most commonly observed side effects of SSRIs and SNRIs are lightheadedness, fainting, dizziness, confusion, hallucinations, rhinitis, dry mouth, tremor, nausea, decreased libido, ejaculation disorder (primarily ejaculatory delay), erectile dysfunction, impotence, fatigue, nausea, diarrhea (or constipation), sleep disorders (somnolence), agitation, restlessness, anxiety, and sweating (Goldstein & Goodnick 1998). Some SNRIs, venalfaxine in particular, are associated with modest bust statistically significant increases in supine diastolic blood pressure, though typically at doses above 300 mg/day (Thase 1998). According to data provided by the manufacturer, 1.4% of patients treated with extended-release venlafaxine experienced a ≥15 mm Hg increase in supine diastolic blood pressure during short-term treamtent. In a meta-analysis of 3,744 patients with MDD who received venlafaxine for 6 weeks, active treamtent did not adversely affect the control of blood pressure for patients with pre-existing high blood pressure or elevated baseline values (Thase 1998).

Suicide Risk.

Although SSRIs have been associated with an idiosyncratic increase in suicidality in a FDA meta-analysis, the increase in risk is restricted to individuals under the age of 24 years--moreover, there is no significant increase in the risk of suicidal thinking in people aged 25 years and higher, and pharmacoepidemiological studies show a protective effect across the lifespan (Barbui et al., 2009; Brent 2016; Gibbons et al., 2012; Sharma et al., 2016; Stone et al., 2009). This protocol features extensive assessment of suicide risk, including telephone contact early in the course of follow-up after starting study drugs (week 2 telephone follow-up). These procedures, we believe, will mitigate the possible heightened risk of suicidality translating to suicide risk or behavior. Specific procedures for addressing suicide risk are described in detail above in the section entitled, SUICIDE RISK MANAGEMENT PLAN.

Other Risks.
Risks are also associated with a single venipuncture performed for DNA extraction at basleline, as well as the measurement of drug response by rating scales and questionnaires designed to assess depression severity and drug side effect frequency and intensity. While every effort will be made to schedule venipuncture at the times of clinically-indicated venipuncture, this cannot always be guaranteed.

Capacity for Consent.
Another common concern with psychiatric patients is the ability to provide informed consent. However, people with at least moderately severe recurrent major depression who are treated as outpatients show few impairments in their decision-making capacities related to research (Appelbaum et al., 1999), and depressive symptom severity and capacity do not appear to be significantly correlated in medical patients (Casarett et al., 2003). In our study, individuals with psychotic features and those whose depression is so severe that psychaitric hospitalization is indicated will be excluded. Individuals in acute suicidal crisis will also be excluded. All screening visits with study clinicians will include a judgment as the the capacity of subjects to provide valid consent. If there is any question about whether or not the patient can provide valid informed consent for this research, the patient will not be enrolled in this study.

Reproductive and Lactational Safety.
Regarding childbearing potential, the safety of SSRIs to the fetus has been extensively investigated and, with the possible exception of paroxetine and fluoxetine, SSRIs are not thought to be associated with clinically significant teratogenic potential (Myles et al., 2013; Wang et al, 2015). Far less is known about the reproductive safety of SNRIs (Richardson et al., 2019). In order to participate in this research, women of childbearing age must have a negative pregnancy test. Pregnant women will be excluded. Evidence and/or expert consensus suggests some potential for harmful infant effects of study drugs when used during breastfeeding, although most antidepressants are expected to produce low levels in breast milk with no clinical significance (Kronenfeld et al., 2017). Nevertheless, women who are currently breastfeeding and who plan to continue breastfeeding will be excluded.

BUDGET
Budget is attached.

CONSENT FORMS
Draft is attached.

REFERENCES

Altman DG, Bland JM. Measurement in Medicine: the Analysis of Method Comparison Studies. The Statistician. 1983;32:307–17.

Appelbaum PS, Grisso T, Frank E, et al. Competence of depressed patients to consent to research. Am J Psychiatry 1999;156:1380-4.

Arnow BA, Blasey C, Williams LM, et al. Depression Subtypes in Predicting Antidepressant Response: A Report From the iSPOT-D Trial. Am J Psychiatry. 2015;172:743-50.

Athreya AP, Banerjee SS, Neavin D, et al. Data-Driven Longitudinal Modeling and Prediction of Symptom Dynamics in Major Depressive Disorder: Integrating Factor Graphs and Learning Methods. Paper presented at: IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology 2017.

Athreya AP, Neavin D, Carrillo-Roa T, et al. Pharmacogenomics-Driven Prediction of Antidepressant Treatment Outcomes: A Machine Learning Approach with Multi-Trial Replication. Clin Pharmacol Ther. 2019; in press.

Barbui C, Esposito E, Cipriani A. Selective serotonin reuptake inhibitors and risk of suicide: a systematic review of observational studies. CMAJ 2009;180:291-7.

Belmaker RH, Agam G. Major depressive disorder. N Engl J Med. 2008;358:55-68.

Biernacka JM, Sangkuhl K, Jenkins G, et al. The International SSRI Pharmacogenomics Consortium (ISPC): a genome-wide association study of antidepressant treatment response. *Transl Psychiatry.* 2015;5:e553.

Brent DA. Antidepressants and suicidality. Psychiatr Clin North Am. 2016;39:503-12.

Casarett DJ, Karlawish JH, Hirschman KB. Identifying ambulatory cancer patients at risk of impaired capacity to consent to research. Journal of Pain and Symptom Management. 2003;26:615–624.

Chekroud AM, Gueorguieva R, Krumholz HM, et al. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry. 2017;74:370-8.

Chekroud AM, Zotti RJ, Shehzad Z, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3:243-50.

Crismon ML, Trivedi M, Pigott TA, et al. The Texas Medication Algorithm Project: report of the Texas Consensus Conference Panel on Medication Treatment of Major Depressive Disorder. J Clin Psychiatry. 1999;60:142-56.

Cusin C, Yang H, Young A, Fava M. Rating scales for depression. In: L Baer, MA Blais (eds.). Handbook of Clinical Rating Scales and Assessment in Psychiatry and Mental Health. Munich, Germany: Springer, 2010.

Eton DT, Yost KJ, Lai JS, et al. Development and validation of the Patient Experience with Treatment and Self-management (PETS): a patient-reported measure of treatment burden. Qual Life Res. 2017;26:489-503.
Felitti VJ, Anda RF, Nordenberg D, et al. Relationship of childhood abuse and household dysfunction to many of the leading causes of death in adults: the Adverse Childhood Experiences (ACE) Study. Am J Prev Med. 1998;14:245-258.

First Michael B., Williams Janet B.W., Spitzer Robert L., and Gibbon, Miriam: Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Clinical Trials Version (SCID-CT). New York: Biometrics Research, New York State Psychiatric Institute, 2007

Fochtmann LJ., Gelenberg AJ. 2nd ed. Washington, DC: American Psychiatric Association; Guideline Watch: Practice Guideline for the Treatment of Patients with Major Depressive Disorder. 2005

Frank E, Prien RF, Jarrett RB, et al. Conceptualization and rationale for consensus definitions of terms in major depressive disorder. Remission, recovery, relapse, and recurrence. Arch Gen Psychiatry 1991;48:851-5.

Freidman SJ, Fava M, Kienke AS, White CN, Nierenberg AA, Rosenbaum JF. Partial response, nonresponse, and relapse with selective serotonin reuptake inhibitors in major depression: a survey of current "next-step" practices. J Clin Psychiatry. 2000;61:403-8.

Fried EI. Moving forward: how depression heterogeneity hinders progress in treatment and research. Expert Rev Neurother. 2017;17:423-25.

Fruhling A, Lee S, 2005. Assessing the Reliability, Validity and Adaptability of PSSUQ. In Proceedings of the 9th Americas Conference on Information Systems, Omaha, Nebraska, USA, August 2005, http://aisel.aisnet.org/amcis2005/378.

Gartlehner G, Thaler K, Hill S, Hansen RA. How should primary care doctors select which antidepressants to administer? Curr Psychiatry Rep. 2012;14:360-9.

Gibbons RD, Brown CH, Hur K, et al. Suicidal thoughts and behavior with antidepressant treatment: reanalysis of the randomized placebo-controlled studies of fluoxetine and venlafaxine. Arch Gen Psychiatry 2012;69:580-7.

Goldstein BJ, Goodnick PJ. Selective serotonin reuptake inhibitors in the treatment of affective disorders—III. Tolerability, safety and pharmacoeconomics. J Psychopharmacol. 1998;12(3 suppl B):S55-87.

Grigoriadis S, VondePorten EH, Mamisashvili L, et al. Antidepressant exposure during pregnancy and congenital malformations: is there an association? A systematic review and meta-analysis of the best evidence. J Clin Psychiatry 2013;74:e293-308.

Guy W. ECDEU Assessment Manual for Psychopharmacology—Revised. Rockville, MD: U.S. Department of Health, Education, and Welfare; Public Health Service, Alcohol; Drug Abuse, and Mental Health Administration; National Institute of Mental Health; Psychopharmacology Research Branch; Division of Extramural Research Programs. pp. 218–222. OCLC 2344751. DHEW Publ No ADM 76–338.

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

Hamilton M. Development of a rating scale for primary depressive illness. Br J Soc Clin Psychol 1967; 6:278–96.

Hamilton M. Hamilton rating scale for Depression (Ham-D). In: Handbook of Psychiatric Measures. Washington, DC: American Psychiatric Association, 2000.

Iniesta R, Malki K, Maier W, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res. 2016;78:94-102.

Ji Y, Biernacka JM, Hebbring S, et al. Pharmacogenomics of selective serotonin reuptake inhibitor treatment for major depressive disorder: genome-wide associations and functional genomics. Pharmacogenomics J. 2013;13:456-463.

Kennedy SH, Lam RW, McIntyre RS, et al. Canadian Network for Mood and Anxiety Treatments (CANMAT) 2016 Clinical Guidelines for the Management of Adults with Major Depressive Disorder: Section 3. Pharmacological Treatments. Can J Psychiatry. 2016;61:540-60.

Kessler RC, Chiu WT, Demler O, et al. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005;62:617-27.

Kessler RC, van Loo HM, Wardenaar KJ, et al. Using patient self-reports to study heterogeneity of treatment effects in major depressive disorder. Epidemiol Psychiatr Sci. 2017;26:22-36.

Koller DF, N. Probabilistic Graphical Models: Principles and Techniques MIT Press; 2009.

Kronenfeld N, Berlin M, Shaniv D, Berkovitch M. Use of psychotropic medications in breastfeeding women. Birth Defects Res. 2017;109:957-97.

Lewis JR. Psychometric evaluation of the Post-Study System Usability Questionnaire: the PSSUQ. Proceedings of the Human Factors Society 36th Annual Meeting; Boca Raton, FL. 1992; pp. 1259-62.

McHugh ML. Interrater reliability: the kappa statistic. Biochem Med. 2012;22:276-82.

Moser A, Stuck AE, Silliman RA et al. The eight-item modified Medical Outcomes Study Social Support Survey: psychometric evaluation showed excellent performance. J Clin Epidemiol. 2012;65:1107-16.

Mrazek DA, Biernacka JM, McAlpine DE, et al. Treatment outcomes of depression: The pharmacogenomic research network antidepressant medication pharmacologenomic study. J Clin Psychopharmacol. 2014;43:313-7.

Musliner KL, Munk-Olsen T, Eaton WW, Zandi PP. Heterogeneity in long-term trajectories of depressive symptoms: Patterns, predictors and outcomes. J Affect Disord. 2016;192:199-211.

Myles N, Newall H, Ward H, Large M. Systematic meta-analysis of individual selective serotonin reuptake inhibitor medications and congenital malformations. Aust N Z J Psychiatry 2013;47:1002-12.

Posner K, Brown GK, Stanley B, et al. The Columbia–Suicide Severity Rating Scale: initial validity and internal consistency findings from three multisite studies with adolescents and adults. Am J Psychiatry. 2011;168:1266–77.

Richardson JL, Martin F, Dunstan H, et al. Pregnancy outcomes following maternal venlafaxine use: A prospective observational comparative cohort study. Reprod Toxicol. 2019;84:108-113.

Rush AJ, Bernstein IH, Trivedi, MH, et al. An Evaluation of the Quick Inventory of Depressive Symptomatology and the Hamilton Rating Scale for Depression: A Sequenced Treatment Alternatives to Relieve Depression Trial Report. Biol Psychiatry 2006;59:493-501.

Rush AJ, Thase ME, Dubé S. Research issues in the study of difficult-to-treat depression. Biol Psychiatry 2003;53:743-53.

Rush AJ, Trivedi MH, Ibrahim HM, et al. The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 2003;54:573-83.

Rush AJ, Trivedi M, Carmody TJ, Biggs MM, Shores-Wilson K, Ibrahim H, Crismon ML. One-year clinical outcomes of depressed public sector outpatients: a benchmark for subsequent studies. Biol Psychiatry. 2004;56:46-53.

Rush AJ, Trivedi MH, Wisniewski SR, et al. Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. Am J Psychiatry. 2006;163:1905-17.

Senn S. Mastering variation: variance components and personalised medicine. Stat Med. 2016;35:966-77.

Sharma T, Guski LS, Freund N, Gotzsche PC. Suicidality and aggression during antidepressant treatment: systematic review and meta-analyses based on clinical study reports. BMJ 2016;352:i65. Doi: 10.1136/bmj.i65.

Stone M, Laughren T, Jones ML, et al. Risk of suicidality in clinical trials of antidepressants in adults: analysis of proprietary data submitted to US Food and Drug Administration. BMJ 2009;339:b2880. Doi: 10.1136/bmj.b2880.

Thase ME. Effects of venlafaxine on blood pressure: a meta-analysis of original data from 3744 depressed patients. J Clin Psychiatry 1998;59:502-8.

Trivedi MH, Rush AJ, Wisniewski SR, et al. 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.

Uher R, Harmer A, Henigsberg N, et al. Adverse reactions to antidepressants. Br J Psychiatry 2009;195:202-10.

Wang S, Yang L, Wang L, et al. Selective Serotonin Reuptake Inhibitors (SSRIs) and the Risk of Congenital Heart Defects: A Meta-Analysis of Prospective Cohort Studies. J Am Heart Assoc. 2015;4:pii:e001681. Doi: 10.1161/JAHA.114.001681.

Williams JB. A structured interview guide for the Hamilton depression rating scale. Arch Gen Psychiatry 1988;45:742-7.

Wisniewski SR, Rush AJ, Balasubramani GK, et al. Self-rated global measure of the frequency, intensity, and burden of side effects. J Psychiatr Prac. 2006;12:71-9.

World Health Organization. Depression and Other Common Mental Disorders: Global Health Estimates. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO. Available online at: http://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf;jsessionid=5AA5A7106D619AF4891002ADAF41326E?sequence=1.

Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical Methods Used to Test for Agreement of Medical Instruments Measuring Continuous Variables in Method Comparison Studies: A Systematic Review. PLoS ONE 2012;v.7(5). PMC3360667.

Zimmerman M, Martinez JH, Young D, et al. Severity classificaiotn on the Hamilton depression rating scale. J Affect Disord 2013;150:384-8.