cpsy
an open access journal

Differential Treatment Benefit Prediction for Treatment Selection in Depression: A Deep Learning Analysis of STAR*D and CO-MED Data

Joseph Mehltretter[1,2], Robert Fratila[2], David Benrimoh[2,3,*], Adam Kapelner[4], Kelly Perlman[2,3], Emily Snook[2,3], Sonia Israel[2,3], Caitrin Armstrong[2], Marc Miresco[3], and Gustavo Turecki[3]

[1]Department of Computer Science, University of Southern California, Los Angeles, California, USA

[2]Aifred Health, Montreal, Quebec, Canada

[3]Department of Psychiatry, McGill University, Montreal, Quebec, Canada

[4]Department of Mathematics, Queens College, CUNY, Queens, NY, USA

Keywords: depression, treatment selection, artificial intelligence, machine learning, treatment prediction, mood disorders

ABSTRACT

Depression affects one in nine people, but treatment response rates remain low. There is significant potential in the use of computational modeling techniques to predict individual patient responses and thus provide more personalized treatment. Deep learning is a promising computational technique that can be used for differential treatment selection based on predicted remission probability. Using Sequenced Treatment Alternatives to Relieve Depression (STAR*D) and Combining Medications to Enhance Depression Outcomes (CO-MED) trial data, we employed deep neural networks to predict remission after feature selection. Treatments included were citalopram, escitalopram, bupropion SR plus escitalopram, and venlafaxine plus mirtazapine. Differential treatment benefit was estimated in terms of improvement of population remission rates after application of the model for treatment selection using two approaches: (1) using predictions generated directly from the model (the predicted improvement approach) and (2) using bootstrapping for sample generation and then estimating population remission rate for patients who actually received the drug predicted by the model compared to the general population (the actual improvement approach). Our deep learning model predicted remission in a pooled CO-MED/STAR*D dataset (including four treatments) with an area under the curve of 0.69 using 17 input features. Our actual improvement analysis showed a statistically significant 2.48% absolute improvement (corresponding to a 7.2% relative improvement) in population remission rate ($p = 0.01$, CI 2.48% ± 0.5%). Our model serves as proof-of-concept that deep learning approaches, with further refinement and work to address concerns about differences between studies when multiple datasets are used for training, may have utility in differential prediction of antidepressant response when selecting from a number of treatment options.

INTRODUCTION

Major depressive disorder (MDD) is the greatest cause of disability-adjusted life-years lost globally and affects more than 300 million people at any given time (World Health Organization, 2017). MDD strongly associates with suicide (Turecki & Brent, 2016) and early mortality

Citation: Mehltretter, J., Fratila, R., Benrimoh, D., Kapelner, A., Perlman, K., Snook, E., . . . Turecki, G. (2020). Differential treatment benefit prediction for treatment selection in depression: A deep learning analysis of STAR*D and CO-MED data. *Computational Psychiatry*, *4*, 61–75. https://doi.org/10.1162/cpsy_a_00029

DOI:
https://doi.org/10.1162/cpsy_a_00029

Supporting Information:
https://doi.org/10.1162/cpsy_a_00029

Received: 16 May 2019
Accepted: 19 October 2020

Competing Interests: The authors declare no conflict of interest.

Corresponding Author:
David A. Benrimoh
david.benrimoh@mail.mcgill.ca

The MIT Press

(Saint Onge, Krueger, & Rogers, 2014) and represents a significant cost to patients, families, health care systems, and the economy (Kessler, 2012; Stewart, Ricci, Chee, Hahn, & Morganstein, 2003). While clinical guidelines (Kennedy et al., 2016) define the optimal treatment outcome for depression to be full remission of symptoms, many patients will not reach remission after a first or second antidepressant treatment (Rush et al., 2006).

Professionals employ many different interventions to treat depression, but little is known about which patients will respond best to which treatments, and many possible patient characteristics are important in personalization (Perlman et al., 2019). The current gold standard of treatment is the application of treatment guidelines, such as the Canadian Network for Mood and Anxiety Treatments (CANMAT) (Kennedy et al., 2016). These guidelines help clinicians determine which types of treatment are evidence based and when they are to be used during the course of a patient's depression. They also help clinicians decide when to initiate pharmacotherapy, with the CANMAT guidelines recommending medication for patients with symptoms with moderate or greater severity. In this work, we explored the promise of using machine learning to develop personalized medicine approaches to MDD treatment to improve remission rates. We will elaborate on an approach that allows for not only prediction of remission but the selection of an optimal treatment from a set of options. We term this two-step process *differential prediction*.

Research in machine learning work has previously borne fruit in personalizing depression treatments (see Iniesta et al., 2016; Lee et al., 2019; Lin et al., 2018). Chekroud et al. (2016) used machine learning to predict remission for patients administered citalopram in the STAR*D study with roughly 65% accuracy using only clinical and demographic variables. Iniesta et al. (2018) used machine learning to predict remission in the Genome-based Therapeutic Drugs for Depression (GENDEP) study using clinical, demographic, and genetic information. In addition to predicting remission, there is great clinical utility in models that help assign patients to treatments in a way that improves their chances to remit. DeRubeis et al. (2014) produced the Personalized Advantage Index (PAI), which helps determine if patients are more likely to benefit from one treatment over another. In addition, specific alleles of genetic polymorphisms mediating processes such as stress response, immune regulation, and neurotransmission were relevant for predicting response to different antidepressants (Uher, 2011).

Most work to date has focused on predicting whether a patient would benefit more from one of two treatment options. However, personalized medicine models should be able to assess differential benefit between more than two, that is, *multiple* treatments. In addition, these models should be able to incorporate maximally accessible information, such as simple clinical and demographic information. In addition, modeling procedures should be flexible enough to accommodate additional multimodal data (such as information coming from genetics or neuroimaging) to take advantage of potential biomarkers. Finally, useful personalized medicine models must produce interpretable results—clinicians must be able to understand which specific patient characteristics drive remission.

Deep learning in a "deep neural network" is a collection of very simple mathematical operations called "artificial neurons," a classic model dating back to Rosenblatt's (1957) "perceptron," networked together and arrayed in layers. Modern computing allows for fitting these networks of simple artificial neurons to model very complicated relationships, making for powerful machine learning. Deep learning has recently gained popularity due to its superior predictive performance on various classification tasks, such as image recognition (Goodfellow, Bengio, & Courville, 2016). We set out to use deep learning to analyze data from two well-known datasets, Sequenced Treatment Alternatives to Relieve Depression (STAR*D) and Combining Medications to Enhance Depression Outcomes (CO-MED), to produce a differential treatment selection model that could help select between more than two treatments. Crucially, we used recently developed methodology to validate that model on data unseen by the algorithm. This provides a sense of what clinical benefit (in terms of population remission rates) could be expected when using the model to assign treatments to future real-world patients. We found that an algorithm that only required minimal clinical and demographic data could have a clinically significant impact on population remission rates. We also chose to create one model of differential prediction (instead of separate models for each drug). This was to ensure that we were truly capturing a differential prediction and not a prediction of general treatment response. We also chose not to create models specific to each drug as we wished to be able to capture differences between treatments directly, which allows us to use statistical techniques to estimate potential clinical utility of the model. To maximize accessibility and impact of our approach, we sought to create a model that did not rely on markers that are expensive to collect (such as genetics) but instead only incorporates clinical and demographic information. In this process, we strived to make our model interpretable at the individual patient level.

METHODS

Data

We analyzed patient-level data from two major trials: CO-MED (Rush et al., 2011) and the first level of STAR*D (Trivedi et al., 2006). CO-MED enrolled 665 outpatients with nonpsychotic depression who were randomized to three treatment arms: escitalopram and placebo, bupropion and escitalopram, or mirtazapine and venlafaxine. The purpose of this trial was to assess whether combination treatment was superior to monotherapy. The result was that similar remission rates were observed in each arm. STAR*D is the largest pragmatic trial of depression treatment to date. In the first of the four levels of the study, all patients were treated with citalopram, and the remission rate was 33% ($n = 2,757$; see case selection). Using these two studies was ideal for our analysis because they (a) included similar outcome measures, the Quick Inventory of Depressive Symptomatology Self-Report (QIDS-SR16); (b) recruited similar patients, those with at least moderately severe depression as determined by the Hamilton Depression Rating Scale (HDRS) and therefore for whom pharmacotherapy was appropriate; (c) included patients treated in both psychiatric and general practice settings; (d) collected similar clinical and demographic information; (e) treated patients for similar lengths of time, 12 weeks in the acute phase of CO-MED and 12–14 weeks in STAR*D Level 1; and (f) used measurement-based care protocols to adjust doses, that is, doses were adjusted based on patient symptom scores. Unless otherwise noted, we defined remission as being a score of 5 or less on the QIDS-SR16. Treatments included were citalopram (from STAR*D), escitalopram (plus placebo), bupropion SR plus escitalopram, and venlafaxine plus mirtazapine (all from CO-MED). Treatments included were citalopram (from STAR*D), escitalopram (plus placebo), bupropion SR plus escitalopram, and venlafaxine plus mirtazapine (all from CO-MED).

In STAR*D, our focus was to assess subjects at baseline and predict whether they went into remission after Level 1 treatment was administered (between Weeks 2 and 14). We removed 27 subjects who either did not have at least moderately severe depression (i.e., a score of at least 16 on the HDRS questionnaire), did not complete at least 1 week of treatment, did not return for a visit at Week 2, or did not return for a depression assessment in their final week.

We then merged the STAR*D dataset (2,757 subjects) with the CO-MED dataset (665 subjects), resulting in 3,222 patients and 213 numeric characteristics measured for each (all features that the two datasets did not have in common were dropped). The result of the merge can be thought of as an experiment with four different treatment groups (three from CO-MED and one from STAR*D). The variable that housed the patients' assignments to one of the four treatments we denote "drug assigned." This variable allows our algorithm to locate differential treatment benefit via modeling interactions with the 213 patient characteristics.

We chose to consider citalopram and escitalopram as separate treatments for two main reasons. First, previous evidence has shown that these treatments may have different levels of efficacy: the CANMAT 2016 guidelines for the treatment of depression review the literature and find some evidence for the superiority of escitalopram to citalopram (Kennedy et al., 2016). In addition, in CO-MED, escitalopram was provided alongside a placebo, which meant that the experience of treatment would likely have been different from patients receiving open-label citalopram monotherapy. Keeping these two treatments separate was therefore more prudent and further allowed us to evaluate differential benefit between a larger number of distinct treatments.

Data Processing and Feature Selection

From our combined dataset of 3,222 subjects, we first held out 200 subjects who were later used for the predicted improvement differential analysis, described later. We ensured that these were sampled to reflect the distribution found in the original datasets. This meant that 20% of the held-out 200 subjects were from the CO-MED study and 80% were from the STAR*D study. The remaining 3,022 subjects were then split into a training and a test set (80% training, 20% test). This 20% test set, as well as the initial 200 subjects extracted previously, were both held out of feature selection and final model training. Both our CO-MED and STAR*D models had imbalanced classes because the majority of subjects did not remit (only 34% of subjects in STAR*D and 36% of subjects in CO-MED remitted). Owing to this class imbalance, we used stratified sampling (Lang, Liberty, & Shmakov, 2016) when creating our training and test set for our assessment phase. This ensured that we were training and testing on mutually exclusive sets with similar endpoint distribution. This equalization prevents bias toward learning and predicting the majority class. We did not employ oversampling or undersampling, as this could carry risk of information loss or overrepresentation (He & Garcia, 2009). We emphasize that the 20% of the data on which the final model was tested, as well as the 200-patient holdout set used for the predicted improvement analysis, were true holdout sets; these data were not seen by any of our employed feature selection or model training algorithms prior to testing. In addition, these holdout sets were selected randomly and were not manually inspected or constructed.

After creating our training and test splits, we sent our training data through a *feature selection pipeline* to select a manageable subset of the original 213 patient characteristics. This pipeline began with expert inspection of the features' meta-information (not of the data values themselves) to remove duplicate or administrative variables of no clinical significance. We then employed a variety of methods to curate the remaining features to ensure the retention of those that are salient without sacrificing prediction performance. Our motivation for this curation is to both reduce computational complexity and to increase ultimate interpretability (Keogh & Mueen, 2017). Our variable curation procedure can be broken down into several steps, each progressively eliminating more variables: (a) variance thresholding (removing variables with less than a certain variance), (b) recursive feature elimination with cross-validation (RFECV), and (c) feature importance extraction. Each of these steps has tuning parameters. The parameter values were selected by using prediction metrics from the deep learning model. Optimistic bias is commonly an issue with thorough feature selection processes. To account for this, we extracted our test set of 200 subjects before performing these methods. These methods were implemented from the Python package Scikit-Learn (Pedregosa et al., 2011). Figure 1 details this process within the "Phase 1" subheader. Future analyses can use knockoff methods that control false discoveries (Candès, Fan, Janson, & Lv, 2018).

After expert inspection of variable meta-information, the next step was variance thresholding (removing variables with less than a certain variance). We ran various models under different variance thresholds and found that setting the threshold to 0.2 yielded the best performing model in terms of accuracy of remission prediction. This means that any feature with a variance less than 0.2 across all samples was removed. When testing various thresholds, we initially tested and analyzed CO-MED and STAR*D individually to ensure that known relevant features in the literature were not being removed. The 0.2 threshold found was chosen because it was optimal for CO-MED and STAR*D individually, as well as when the datasets were combined. We then performed RFECV using three folds with a random forest classifier. This method produced a subset of features considered the strongest with regard to predicting our target of remission. We utilized RFECV because there was a large number of features, and our objective was to reduce the number of those features based on a combined importance; that is, we wanted to identify a smaller group of features that could strongly predict remission (i.e., our objective was to capture the information-rich features). We opted for a random forest classifier owing to its robustness to hyperparameters (e.g., the number of classification trees and the number of variables to try for the split rules in the inner condition nodes). To further ensure the robustness of the important features selected, and reduce optimistic bias, we used threefold cross-validation. We then assessed the stability of the features selected by RFECV using randomized lasso. This methodology takes random subsets of subjects and a random subset of features and runs a feature selection algorithm on that subset to select the top features. It runs this process 200 times and, upon completion, calculates the percentage of the time a given feature was selected as a top feature. We then removed the features that did not exceed our "importance" threshold of 75%. This resulted in the final feature set described in Figure 2.

After the dataset was pared down to the features found in our feature selection pipeline, we performed the two main analyses: remission prediction and the differential treatment benefit analysis. The remission prediction accuracy analysis was itself divided into two independent steps: 10-fold cross-validation and, separately, training a model on the 80% training data and testing it on the 20% holdout data. For all model training, binary remission was the predicted target. The neural network architecture and configuration used for model training are described later.

We used 10-fold cross-validation to determine if the features selected were likely to create a useful model that could then be used in the differential analysis. Cross-validation was accomplished by combining the dataset through sequentially merging our training and validation sets and using that merged dataset to test our features and model configuration by training and validating our model in a 10-fold process. This analysis did not produce a single trained model but rather assessed how well our features and model configuration predicted remission. These metrics are shown in our results and are reported as macro metrics, which are calculated by calculating the metric for each class and then averaging these, taking into account performance on both classes. Once it was clear that the features selected would produce a useful model, we then performed our second analysis using the training and validation set we created before feature selection. It is important to note that these sets were not affected by the

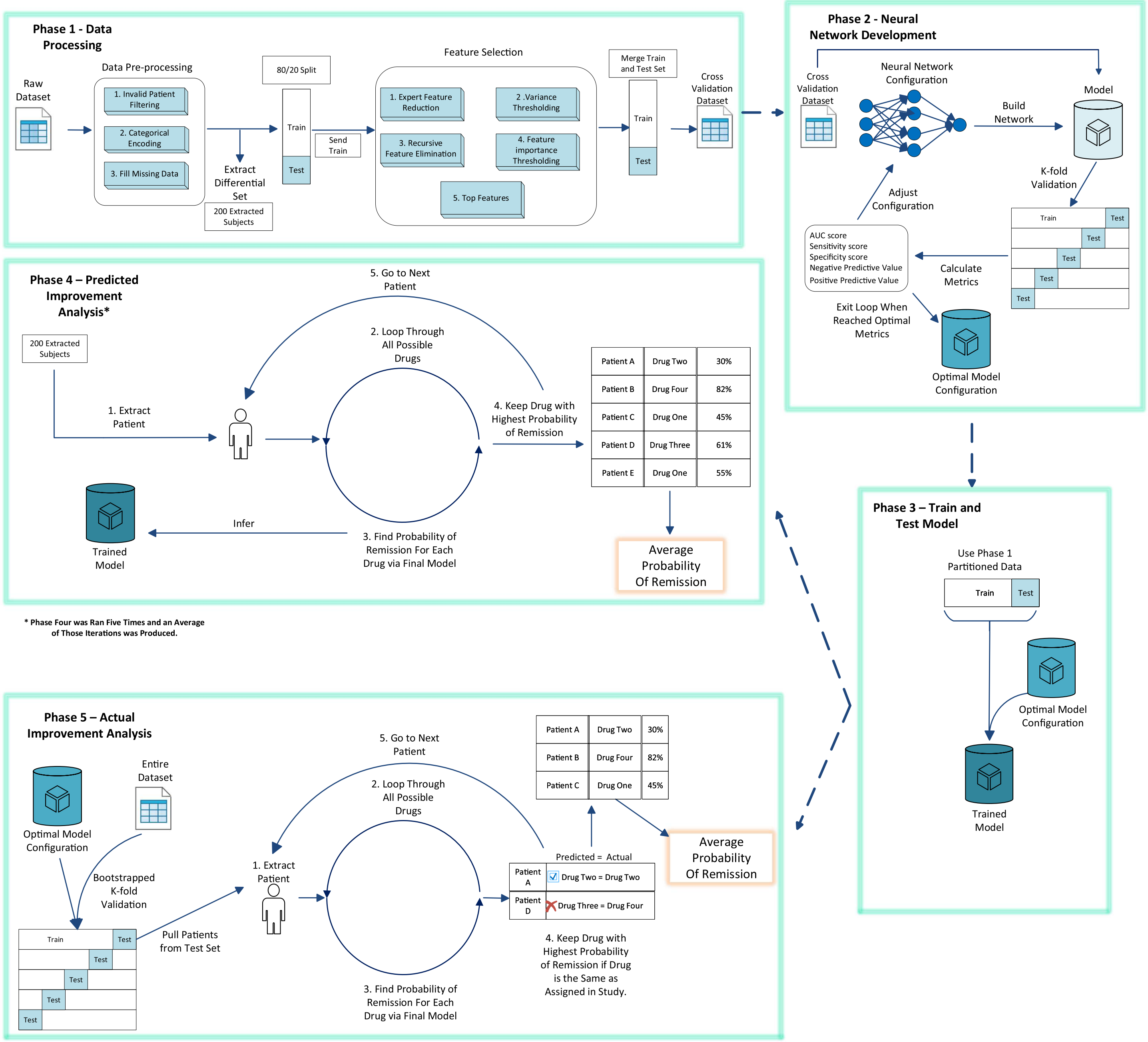
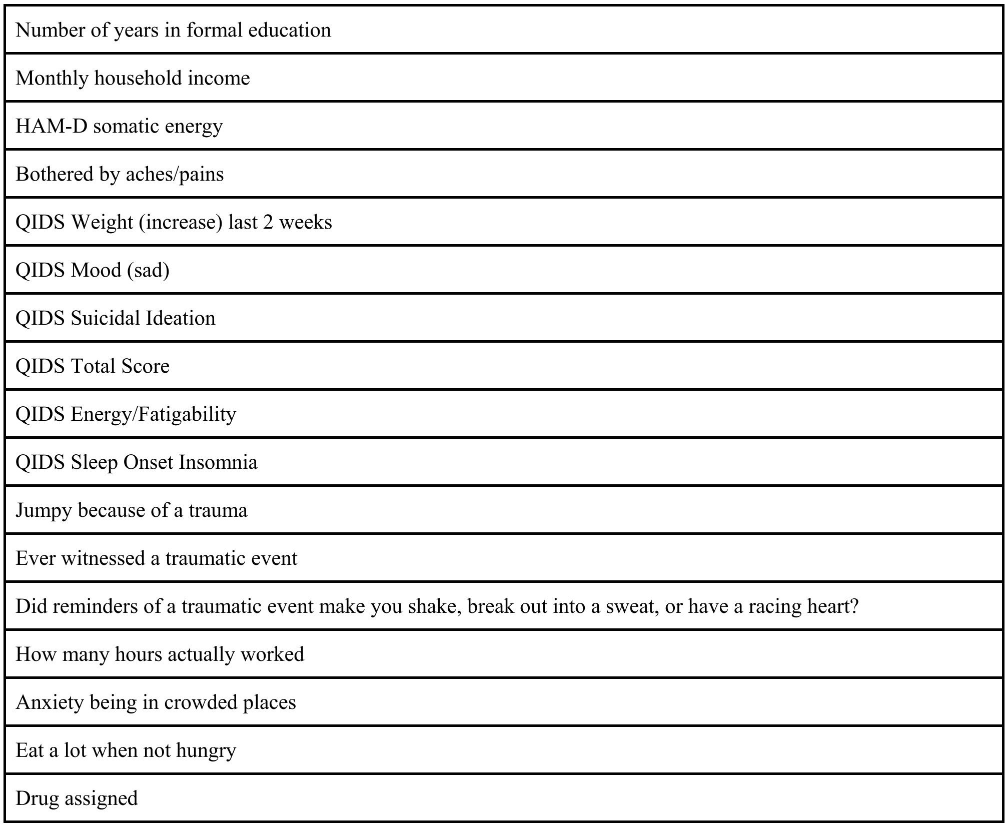

Phase 1 - Data Processing
Raw Dataset
Data Pre-processing
1. Invalid Patient Filtering
2. Categorical Encoding
3. Fill Missing Data
80/20 Split
Train
Test
Extract Differential Set
200 Extracted Subjects
Send Train
Feature Selection
1. Expert Feature Reduction
2 .Variance Thresholding
3. Recursive Feature Elimination
4. Feature importance Thresholding
5. Top Features
Merge Train and Test Set
Train
Test
Cross Validation Dataset
Phase 2 - Neural Network Development
Cross Validation Dataset
Neural Network Configuration
Model
Build Network
K-fold Validation
Adjust Configuration
AUC score
Sensitivity score
Specificity score
Negative Predictive Value
Positive Predictive Value
Calculate Metrics
Train
Test
Test
Test
Test
Test
Exit Loop When Reached Optimal Metrics
Optimal Model Configuration
Phase 4 – Predicted Improvement Analysis*
200 Extracted Subjects
1. Extract Patient
2. Loop Through All Possible Drugs
3. Find Probability of Remission For Each Drug via Final Model
4. Keep Drug with Highest Probability of Remission
5. Go to Next Patient
Patient A | Drug Two | 30%
Patient B | Drug Four | 82%
Patient C | Drug One | 45%
Patient D | Drug Three | 61%
Patient E | Drug One | 55%
Average Probability Of Remission
Infer
Trained Model
* Phase Four was Ran Five Times and an Average of Those Iterations was Produced.
Phase 3 – Train and Test Model
Use Phase 1 Partitioned Data
Train
Test
Optimal Model Configuration
Trained Model
Phase 5 – Actual Improvement Analysis
Optimal Model Configuration
Entire Dataset
Bootstrapped K-fold Validation
Train
Test
Test
Test
Test
Test
Pull Patients from Test Set
1. Extract Patient
2. Loop Through All Possible Drugs
3. Find Probability of Remission For Each Drug via Final Model
5. Go to Next Patient
Patient A | Drug Two | 30%
Patient B | Drug Four | 82%
Patient C | Drug One | 45%
Predicted = Actual
Patient A ☑ Drug Two = Drug Two
Patient D ✗ Drug Three = Drug Four
4. Keep Drug with Highest Probability of Remission if Drug is the Same as Assigned in Study.
Average Probability Of Remission

Figure 1. Drug differential analysis. Phase 1 (data processing): The raw dataset was preprocessed; split into a test, train, and validation set; and then fed through a feature selection procedure to produce a separate final dataset. This final dataset was our train and test set combined and only included our final features. Phase 2 (neural network development): We configured our neural network and used our final dataset to perform k-fold validation to produce metrics. During training, we iteratively optimized our neural network tuning parameters. Phase 3 (model training and testing): We then used our model configuration from Phase 2 and our train and test set from Phase 1 with our top features also from Phase 1 to train and validate a model. Phase 4 (predicted improvement analysis): We iterated through each subject of our differential analysis set. For each patient, we used our neural network with each possible drug to find the probability of remission with that drug. Once the patient had a probability of remission for each drug, the drug with the highest probability was retained. Once all 200 patients had a probability of remission, we took an average of those probabilities. This process was run five times, and the average was computed. Phase 5 (actual improvement analysis): We used k-fold validation with our entire dataset. This entire dataset is a combination of the training set and both holdout sets. The patients within the test set for each fold were used to perform the differential analysis. This analysis was conservative, as we only retained subjects if the drug for which our neural network produced the highest probability of remission was actually the drug they received in the study. We then took the average of all patients kept after all folds from the k-fold validation process.



Number of years in formal education
Monthly household income
HAM-D somatic energy
Bothered by aches/pains
QIDS Weight (increase) last 2 weeks
QIDS Mood (sad)
QIDS Suicidal Ideation
QIDS Total Score
QIDS Energy/Fatigability
QIDS Sleep Onset Insomnia
Jumpy because of a trauma
Ever witnessed a traumatic event
Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart?
How many hours actually worked
Anxiety being in crowded places
Eat a lot when not hungry
Drug assigned

Figure 2. The 17 most important features in the differential treatment prediction model.

cross-validation exercise. We trained a deep neural network model using the training data; that trained model was then validated with our 20% testing set. This trained model yielded similar results to our cross-validation model (see results) and was then used to perform our predicted improvement differential analysis on the 200 subject test set.

Differential Treatment Benefit Analysis

Our first “naive” experiment was aimed at estimating the performance of the model if it was applied blindly to a “new” clinical population (our holdout sample of 200 patients). Once we had completed this process once, we then repeated it five times (without repeating feature selection) to generate five different holdout sets of 200 patients. Our predicted improvement differential analysis results are reported as the average of these five repeats. To get a probability of remission, we passed each subject in the holdout sample four times through the final model, once for each possible drug in the dataset. The output of the forward pass was the probability of remission for that subject for that given drug. For each of the 200 subjects, we took the drug with the highest probability of remission and obtained the mean remission rate of the 200 subjects. Finally, we took the difference between the mean remission rate of the entire dataset and our mean remission rate for the 200 test subjects.

In the predicted improvement version of the analysis, we looked at hypothetical cases in which we did not necessarily know the outcome of patient–drug pairing. Our second “actual improvement ” analysis only considered the nonhypothetical cases in which we knew the real outcome of giving a specific treatment to a patient. This second analysis was inferential, and details of its methods are found in Kapelner et al. (2020). This analysis answered the question, Does our deep neural network personalization model outperform a null model—does it improve patient outcomes more than chance treatment allocation? For this null allocation, we simulated 1,000 bootstrap samples: we resampled from the original dataset with replacement, trained models through 10-fold cross-validation and computed an "improvement score" that compares the chance remission rate with the improved remission rate found using the personalized allocations from our deep neural network model. To exclude hypothetical cases, we compared improvement scores between patients who actually received, by chance, the optimal drug predicted by the model to the rest of the study population.

Neural Network Architecture and Configuration

Given the structured nature of the data collected in these studies, we opted for fully connected dense neural networks (DNN). To build, train, and evaluate all of our DNN configurations, we employed the open source package Vulcan.[1] DNNs allow us to capture complex, nonlinear relationships likely present in psychiatric data (e.g., mediation and moderation effects, which are unknown a priori). We limited our model's learning capacity (with the use of a shallow network) to explore more of the solution space before finding an optimal location. Our optimal network had a single hidden layer with 17 nodes using the scaled exponential linear unit activation function (Klambauer, Unterthiner, Mayr, & Hochreiter, 2017), while the final prediction layer used the softmax function to determine remission probabilities. This simple structure provides insurance against overfitting. We one-hot encoded our response variable to create a multiclass problem that uses categorical cross-entropy as the optimizer cost function (Goodfellow et al., 2016). We used the Adam optimizer for learning the network parameters (Kingma & Ba, 2014) with a learning rate of 0.0001. To further help with model generalization, we used a 50% dropout rate (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014). During the evaluation phase, we used 10-fold cross-validation. Each fold allowed the model to train for 200 epochs. For each model trained, remission was considered the target prediction outcome. Once again, the variable "drug assigned" was included as part of the feature set to ensure that the DNN has an opportunity to learn how the drug assigned interacts with the 17 patient-level characteristics. Drug-specific remission probabilities were obtained from our final model by varying the drug assigned variable, holding other patient features constant.

Note that we chose in this article to compare deep learning to the existing technique described by Chekroud et al. (2016) to determine if it was a viable technique on this—and potentially similar—datasets. As such, there was no search over algorithm types, as we already had a comparator to deep learning available.

RESULTS

In our differential treatment prediction analysis, we conducted two separate experiments to determine the relative increase in population remission rates when using the treatment recommended by our model compared to the treatment patients were assigned in the study. These experiments were done on the combined CO-MED and STAR*D datasets using the same

[1] The Vulcan platform used for this work is open source and can be found online at https://github.com/Aifred-Health/Vulcan. The data are available through the National Institute of Mental Health data request platform. Using the descriptions of the model available within this article, the publicly available data, and the open source Vulcan platform, investigators will be able to reproduce this model.

trained DNN model built after feature curation (section “Data Processing and Feature Selection”). After DNN training, the out-of-sample performance was 0.69 macro area under the curve (AUC), 0.70 macro positive predictive value (PPV), 0.70 macro negative predictive value (NPV), 0.72 macro sensitivity, 0.72 macro specificity on 10-fold cross-validation, and a 0.69 macro AUC, 0.60 macro sensitivity, 0.60 macro specificity, 0.63 macro NPV, and 0.63 macro PPV, indicating that the selected features could produce a useful model in a robust fashion.

We now move to the results of the predicted improvement analysis of differential treatment benefit. This analysis produced estimates of projected improvement in population remission rate, under the assumption that the model performs as expected in a new clinical population. It provides an “optimistic” view of the model’s potential benefit for improving remission rates, allowing for one bound on the range of effect sizes that might be expected in a clinical trial (the lower bound to this range is provided by the actual improvement analysis). Over five iterations (as described earlier), the predicted improvement analysis produced an $\sim$11% in the population remission rate, from 34.33% to 46.12%—a relative improvement of over 30%—when using our neural network to assign patients to drugs when compared with the baseline study drug assignation.

While the predicted improvement model produces one estimate of model differential treatment selection benefit that represents expected results if the model functions as expected, its estimate is hypothetical. In contrast, the actual improvement approach provides an estimate of differential benefit based on nonhypothetical cases. This inferential approach provides a convincing test of the value of a treatment selection model because it only considers the difference between patients who actually received the treatment selected by the model and those who were randomly allocated to treatment. We observed a significant improvement of 2.5%, $p = 0.01$ (CI 2.48% ± 0.5%) from 34.33% to 36.8% population remission rate, a relative improvement of 7.2%. The risk ratio for improvement was thus 1.07, and the odds ratio was 1.11, compared to baseline medication assignment.

The model target was remission prediction, but the purpose of the model is to be able to predict differences between treatments at the level of the individual. We tested its capacity to do the latter in two ways. First, we looked across all five test sets of 200 patients used for the predicted improvement analysis to see the average difference in remission probability between the highest and lowest probability treatments and found this to be 11% (±0.34%). Second, in Supplementary Table 6, a random sample of 10 patients from 1 of the 200 held-out sets is provided that demonstrates the varying remission probabilities assigned to each treatment, patient by patient.

DISCUSSION

We used deep learning to create a personalized medicine model that is useful as a proof-of-concept for differential treatment selection. We combined data from the CO-MED and STAR*D datasets to produce a pooled dataset with four different treatment types. Our main goal was to demonstrate that our model was capable of performing differential treatment selection and to estimate improvement in patient outcomes. We produced a well-validated model via cross-validation within the test set and validation on a held-out dataset, and only then did we apply the model to our holdout sample of 200 subsampled but mutually exclusive patients. By retaining the treatment selection (the drug class) as a feature in our model, we could then generate predictions for each drug class for all 200 held-out patients. With these predictions, we estimated the remission rate for these patients had they been assigned a drug based on our model, showing an 11% average increase in population remission rates. This potential remission increase of 11% is clinically significant, as it increases remission by one-third over measurement-based care alone in STAR*D and CO-MED. We then further validated our results by proving statistically that our personalized models beat chance remission via a conservative bootstrapping procedure, which demonstrated a significant increase in overall remission rate of 2.5%, $p = 0.01$. This actual improvement analysis's bootstrap estimate of out-of-sample improvement is only valid asymptotically and also has a large standard error. Thus the population remission improvement estimated using this analysis does not accurately depict the likely effect size of the differential prediction benefit; rather, it represents a "lower bound" of our approach's utility. We believe our true personalization benefit will likely lie between the numbers produced by the actual improvement and predicted improvement analyses.

A critical contribution of our work is extending these evaluation metrics to models that cover a number of treatment options. This is significant because most previous literature has focused on helping to select between two treatment options, when in clinical practice, clinicians are faced with more than a dozen treatment choices. Models that can help select between a number of different treatments may have significant clinical impact when properly validated and implemented.

As can be seen in the Supplementary Materials, a version of our model trained only on STAR*D generalizes to all three arms of CO-MED, in contrast to the model by Chekroud et al. (2016), which only generalized to two of the three arms. Importantly, there was significant overlap between the 25 features included in the Chekroud et al. model and the 14 included in our own (Supplementary Table 1). This initial analysis demonstrates that deep learning can provide improved results when compared to other machine learning techniques while potentially using fewer patient characteristics, allowing for easier clinical implementation. The deep learning advantage is small at present, but deep learning has often been found to significantly outperform other machine learning techniques as the size of the dataset increases (L'Heureux, Grolinger, & Capretz, 2017). Thus finding even a small advantage for deep learning in this fairly small (by deep learning standards) dataset leads us to speculate that deep learning will perform significantly better than other techniques when we have more subject data.

We examined the population remission rate, and as such it is difficult to determine the benefit for each individual patient. Future analyses, along the lines of the PAI (DeRubeis et al., 2014), may be able to help estimate the individual benefit with model-predicted treatment. However, what is intriguing is that despite equal effectiveness of these treatments at a population level, we are able to use individual patient differences in predicted remission generated by varying the assigned drug to improve the overall remission rate. If all treatments were truly equally effective for all individuals, we could have expected a model that approximated but did not improve upon the remission rate. The finding of a projected improvement supports a personalized medicine approach based on individualized prediction of response to treatment.

Deep learning has often been labeled as a "black box," meaning deep learning models can be a challenge to interpret (Samek, Wiegand, & Müller, 2017). Interestingly, when using only clinical and demographic measures, we find that deep learning systems provide a list of features that clinicians could interpret. As demonstrated in Box 1, the most important features used in the prediction for each individual patient can be recovered, providing insights personalized to that patient. This "personalized prediction report" demonstrates that deep learning–based tools may be able to provide information that is useful for understanding individual clinical cases.

Box 1. Examples of Differential Predictions for Two Subjects ("Personalized Prediction Reports")

Here we show "reports" for two patients showing the drug they received in the study and whether or not they went into remission; the drug the network predicted and the predicted remission rates for that drug and the drug the patient actually received; and the five most important features for that patient based on the neural network (all features were used for both patients, but the five listed were weighted more by the network for that patient).

Subject A was originally given escitalopram and went into remission. Our neural network found escitalopram to have an 87.57% chance of remission. The five most important features for this patient and the attendant responses were as follows:

- Total monthly income: $2,200
- Years of formal education: 12
- Experienced weight increase within the last 2 weeks? Has gained 5 pounds or more
- Have you been feeling down, blue, sad, or depressed? Feels sad less than half the time
- Do you eat a lot when not hungry? No

Subject B was originally given citalopram and did not go into remission. Our neural network found citalopram to have a 38.03% chance of remission, while bupropion SR & escitalopram had a 60.43% chance of remission, indicating that changing the medication may have been beneficial for this patient. The five most important features for this patient and attendant responses were as follows:

- Have you ever witnessed a traumatic event? No
- Did reminders of a traumatic event make you shake, break out into a sweat, or have a racing heart? No
- Have you had any trouble falling asleep when you go to bed? Takes at least 30 minutes to fall asleep, more than half the time
- Have you been feeling down, blue, sad or depressed? Feels sad less than half the time
- Do you eat a lot when not hungry? Yes

It is interesting to note that there are two categories of features in Figure 2: those likely to predict overall probability of remission and treatment-specific features. For example, level of education and income—which have both been found in other studies of remission prediction (Carter et al., 2012)—are both unlikely to be specifically related to any one drug's mechanism, as opposed to sleep pattern, which may be relevant to a particular drug's mechanism. It is important to review the features that have been identified in our model and compare them to models in previously published work, and a full discussion is included in the Supplementary Materials. Specific symptoms retained by our model may contribute to differential prediction. For example, in the Individualized Patient Reports (Box 1), Patient B was predicted to do better with a combination of bupropion and escitalopram, and one of the five most important features in that prediction was a tendency to eat a lot when not hungry, which is interesting given the fact that bupropion is often used clinically in cases of hyperphagia or when weight gain is to be avoided (Anderson et al., 2002; Patel et al., 2016).

Beyond the demonstration of a differential treatment prediction tool for more than two treatments, we also provide evidence, in accordance with recent papers (Lin et al., 2018), that deep learning can be readily applied to psychiatric datasets. This is an important development in the field of computational psychiatry, as deep learning is well suited to the analysis of multimodal data and may help bring together data from neuroimaging, genetics, or wearable technology with clinical and demographic measures.

It remains to be seen if the effect we estimate materializes in a clinical environment. That being said, our model's AUC of 0.69 is potentially clinically significant because of the low baseline response rate to antidepressants. Any tool that could help improve the accuracy with which clinicians can identify which patients are likely to benefit from which treatments may be welcomed by the clinical community. This is especially true given the low risk of choosing "wrong" (i.e., choosing an ineffective first-line antidepressant, which occurs commonly) and the high reward of choosing "right"—choosing the right medication for the individual patient and reducing time to remission. In this context, a model with a similar performance to our own might be clinically significant because it likely will outperform the status quo without significantly increasing the risk of adverse events. In addition, it is worth considering the model in terms of potential benefit to the population. Using our model's treatment recommendation, we estimate that between 80 and 354 more patients in the dataset would have reached remission after a single trial of pharmacotherapy (based on the actual improvement and predicted improvement analyses, respectively).

Even if AI-powered decision support tools continue to improve and demonstrate convincing accuracy and performance in clinical trials, great care will need to be taken to ensure that these models are implemented in a manner that is acceptable to patients, that is well integrated into the clinical workflow, and that does not infringe on—and ideally enriches—the integrity of the physician–patient relationship.

We note limitations to our study. In our pooled dataset, patients assigned citalopram vastly outnumbered those prescribed other treatments, limiting the extent to which we can be confident that we were predicting differences between four completely distinct treatment classes. Further studies analyzing datasets with smaller class imbalances are necessary. We recognize that while CO-MED was a randomized trial, all patients in STAR*D were initially assigned to citalopram. However, given the very similar patient populations and study protocols, we felt justified in combining the studies as if they were arms of the same study. One limitation is that we had far fewer patients taking drugs from the CO-MED study than those taking citalopram as part of STAR*D. This raises concerns about generalizability that could be addressed using other larger datasets that include these treatments, and this analysis is something we have planned for future work. However, the fact that our model performed well on a holdout set, and the fact that the model did not simply always predict that a patient should be prescribed citalopram (see Box 1), does raise hopes for the generalizability of this kind of model despite the class imbalance in terms of drug assigned. We do note that this particular model was intended as a proof-of-concept and would require further elaboration and validation using larger datasets prior to clinical implementation.

Another limitation of this work is the fact that treatments did not cross over between studies; that is, a patient could only receive escitalopram in CO-MED and citalopram in STAR*D, meaning that treatment may potentially be confounded with which study a patient was enrolled in. The choice of the CO-MED and STAR*D datasets was made with the intent to reduce this confound as much as possible, as CO-MED and STAR*D were very similar studies, with similar patient outcomes, allowing for the minimization of meaningful differences between studies that might affect patient-level response to treatment. The finding that differential prediction is possible using a set of reasonable predictive features, many of which have been previously found to predict antidepressant treatment response, does provide some evidence that treatment predictions were not simply due to confounding by study. However, future work, currently in progress, must work to address this problem of interstudy differences more directly, especially as the data used to train models in the future expand to include more and more dissimilar studies.

During our analyses, we identified unbalanced data in the treatment and remission variable (i.e., there was a preponderance of citalopram data with respect to treatment and of nonremission with respect to outcome). Unbalanced data can lead to models that do not generalize well on external data. Given this, as stated in the section "Data Preprocessing and Feature Selection," we extracted a 200-subject validation set that was representative of the data based on the treatment-assigned variable. For the remaining data, during model optimization, we stratified our data on the remission variable. Having an external validation set that was stratified by the treatment variable and a model that was trained and tested on data stratified by the remission variable increases confidence that the model will generalize. This was supported by the observation that in our tests on the holdout set, we saw an average 0.67 AUC (variance ± 0.03), similar to the k-fold results (0.69 AUC). Additionally, for the "actual improvement" analysis, our bootstrapping method used random sampling with replacement; therefore each sample was generated to be representative of the population. During the actual improvement analysis, the data were stratified based on the remission variable to ensure that for each sample, the model was not biased toward nonremission simply because this was the dominant class; this serves to increase our confidence that our model's predictions resulted from training on a dataset that matched the true study population and was not artificially enriched with either remitters or nonremitters.

We note that, at present, we demonstrate small advantages for deep learning compared to the comparator method, despite being a more complex algorithm. In future work on larger datasets, work will be done to determine if deep learning provides greater benefit than comparator methods, as has been noted in other areas of research (Goodfellow et al., 2016). There is some reason for optimism on this front, given prior evidence and the ability of neural networks to learn complex nonlinear interactions.

CONCLUSION

We have demonstrated proof-of-concept of using a deep learning model to predict remission and to guide differential treatment selection for more than two treatments. More data are required to validate these methods for more treatment types to create a clinically useful tool. Furthermore, clinical trials are needed to determine if our hypothetical treatment assignment is translatable to real patients. In addition, future work might examine the question of when beginning a pharmacological treatment is most beneficial, in order to augment clinical decision-making at this crucial juncture.

While we chose binary remission as our end point, as it is the gold standard recommended by treatment guidelines, future work will also focus on differential prediction of symptom reduction and the analysis of patient improvement, as well as quality of life and disability outcomes. Work currently in progress hopes to shed more light on why different patients respond differentially to treatments. This will be crucial both to increase clinician trust and to advance the study of depression pathophysiology and treatment. The flexibility and versatility of our model support the idea that deep learning will be a useful technique going forward in the field of personalized medicine.

AUTHOR CONTRIBUTIONS

Joseph Mehltretter: Conceptualization: Equal; Formal analysis: Equal; Methodology: Equal; Software: Equal; Writing – review & editing: Equal. Robert Fratila: Conceptualization: Equal; Formal analysis: Equal; Funding acquisition: Equal; Software: Lead; Supervision: Equal; Visualization: Equal; Writing – review & editing: Equal. David A Benrimoh: Conceptualization: Equal; Data curation: Equal; Funding acquisition: Equal; Project administration: Lead; Supervision: Equal; Validation: Equal; Writing – original draft: Lead; Writing – review & editing: Equal. Adam Kapelner: Methodology: Equal; Resources: Equal; Writing – review & editing: Supporting. Kelly Perlman: Investigation: Supporting; Project administration: Supporting; Writing – review & editing: Supporting. Emily Snook: Investigation: Supporting; Writing – review & editing:Supporting. Sonia Israel: Funding acquisition: Equal; Project administration: Supporting; Writing – review & editing: Supporting. Caitrin Armstrong: Writing – review & editing: Supporting. Marc Miresco: Conceptualization: Supporting; Writing – review & editing: Supporting. Gustavo Turecki: Data curation: Supporting; Resources: Equal; Supervision: Lead; Writing – review & editing: Equal.

ACKNOWLEDGMENTS

This work was approved by the Research Ethics Committee of the Douglas Mental Health University Institute.

FUNDING INFORMATION

Aifred Health.

REFERENCES

Anderson, J. W., Greenway, F. L., Fujioka, K., Gadde, K. M., McKenney, J., & O'Neil, P. M. (2002). Bupropion SR enhances weight loss: A 48-week double-blind, placebo-controlled trial. *Obesity Research*, *10*(7), 633–641. **DOI:** https://doi.org/10.1038/oby.2002.86, **PMID:** 12105285

Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: "Model-X" knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society, Series B*, *80*(3), 551–557. **DOI:** https://doi.org/10.1111/rssb.12265

Carter, G. C., Cantrell, R. A., Zarotsky, V., Haynes, V. S., Phillips, G., Alatorre, C. I., Goetz, I., Paczkowski, R., & Marangell, L. B. (2012). Comprehensive review of factors implicated in the heterogeneity of response in depression. *Depression and Anxiety*, *29*(4), 340–354. **DOI:** https://doi.org/10.1002/da.21918, **PMID:** 22511365

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *Lancet Psychiatry*, *3*(3), 243–250. **DOI:** https://doi.org/10.1016/S2215-0366(15)00471-X

DeRubeis, R. J., Cohen, Z. D., Forand, N. R., Fournier, J. C., Gelfand, L. A., & Lorenzo-Luaces, L. (2014). The Personalized Advantage Index: Translating research on prediction into individualized treatment recommendations—a demonstration. *PLoS ONE*, *9*, e83875. **DOI:** https://doi.org/10.1371/journal.pone.0083875, **PMID:** 24416178, **PMCID:** PMC3885521

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: MIT Press. https://www.deeplearningbook.org/

He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. **DOI:** https://doi.org/10.1109/TKDE.2008.239

Iniesta, R., Hodgson, K., Stahl, D., Malki, K., Maier, W., Rietschel, M., . . . Uher, R. (2018). Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Scientific Reports*, *8*, 5530. **DOI:** https://doi.org/10.1038/s41598-018-23584-z, **PMID:** 29615645, **PMCID:** PMC5882876

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., . . . Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research*, *78*, 94–102. **DOI:** https://doi.org/10.1016/j.jpsychires.2016.03.016, **PMID:** 27089522

Kapelner, A., Bleich, J., Levine, A., Cohen, Z. D., DeRubeis, R., & Berk, R. (2020). *Evaluating the effectiveness of personalized medicine with software*. arXiv:1404.7844v3.

Kennedy, S. H., Lam, R. W., McIntyre, R. S., Tourjman, S. V., Bhat, V., Blier, P., . . . CANMET Depression Work Group. (2016). Canadian

Network for Mood and Anxiety Treatments (CANMAT) 2016 clinical guidelines for the management of adults with major depressive disorder: Section 3. Pharmacological treatments. *Canadian Journal of Psychiatry*, *61*(9), 540–560. **DOI:** https://doi.org/10.1177/0706743716659061, **PMID:** 27486152, **PMCID:** PMC4994787

Keogh, E., & Mueen, A. (2017). Curse of dimensionality. In C. Sammut & G. I. Webb (Eds.), Boston, MA: Springer. **DOI:** https://doi.org/10.1007/978-1-4899-7687-1_192

Kessler, R. C. (2012). The costs of depression. *Psychiatric Clinics of North America*, *35*(1), 1–14. **DOI:** https://doi.org/10.1016/j.psc.2011.11.005, **PMID:** 22370487, **PMCID:** PMC3292769

Kingma, D., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv:1412.6980.

Klambauer, G., Unterthiner, T., Mayr, A., & Hochreiter, S. (2017). *Self-normalizing neural networks*. arXiv:1706.02515.

Lang, K., Liberty, E., & Shmakov, K. (2016). Stratified sampling meets machine learning. In *Proceedings of the 33rd International Conference on Machine Learning* (Vol. 48, pp. 2320–2329). New York, NY: JMLR.org.

Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutiliere, J. J., Rosenblat, J. D., Trevizol, A., . . . McIntyre, R. S. (2019). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. **DOI:** https://doi.org/10.1016/j.jad.2018.08.073, **PMID:** 30153635

L'Heureux, A., Grolinger, K., & Capretz, M. A. M. (2017). *Machine learning with big data: Challenges and approaches*. New York, NY: IEEE. **DOI:** https://doi.org/10.1109/ACCESS.2017.2696365

Lin, E., Kuo, P.-H., Liu, Y.-L., Yu, Y. W.-Y., Yang, A. C., & Tsai, S.-J. (2018). A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. *Frontiers in Psychiatry*, 9. **DOI:** https://doi.org/10.3389/fpsyt.2018.00290, **PMID:** 30034349, **PMCID:** PMC6043864

Patel, K., Allen, S., Haque, M. N., Angelescu, I., Baumeister, D., & Tracy, D. K. (2016). Bupropion: A systematic review and meta-analysis of effectiveness as an antidepressant. *Therapeutic Advances in Psychopharmacology*, *6*, 99–144. **DOI:** https://doi.org/10.1177/2045125316629071, **PMID:** 27141292, **PMCID:** PMC4837968

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perlman, K., Benrimoh, D., Israel, S., Rollins, C., Brown, E., Tunteng, J.-F., . . . Berlim, M. T. (2019). A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *Journal of Affective Disorders*, *243*, 503–515. **DOI:** https://doi.org/10.1016/j.jad.2018.09.067, **PMID:** 30286415

Rosenblatt, F. (1957). *The perceptron—a perceiving and recognizing automaton* (Report No. 85-460-1). Ithaca, NY: Cornell Aeronautical Laboratory.

Rush, A. J., Trivedi, M. H., Stewart, J. W., Nierenberg, A. A., Fava, M., Kurian, B. T., . . . Wisniewski, S. R. (2011). Combining medications to enhance depression outcomes (CO-MED): Acute and long-term outcomes of a single-blind randomized study. *American Journal of Psychiatry*, *168*(7), 689–701. **DOI:** https://doi.org/10.1176/appi.ajp.2011.10111645, **PMID:** 21536692

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., . . . Fava, M. (2006). Acute and longer-term outcomes in depressed outpatients requiring one or several treatment steps: A STAR*D report. *American Journal of Psychiatry*, *163*(11), 1905–1917. **DOI:** https://doi.org/10.1176/ajp.2006.163.11.1905, **PMID:** 17074942

Saint Onge, J. M., Krueger, P. M., & Rogers, R. G. (2014). The relationship between major depression and nonsuicide mortality for U.S. adults: The importance of health behaviors. *Journals of Gerontology, Series B*, *69*(4), 622–632. **DOI:** https://doi.org/10.1093/geronb/gbu009, **PMID:** 24569003, **PMCID:** PMC4049146

Samek, W., Wiegand, T., & Müller, K.-R. (2017). *Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models*. arXiv:1708.08296.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*, 1929–1958.

Stewart, W. F., Ricci, J. A., Chee, E., Hahn, S. R., & Morganstein, D. (2003). Cost of lost productive work time among US workers with depression. *JAMA*, *289*(23), 3135–3144. **DOI:** https://doi.org/10.1001/jama.289.23.3135, **PMID:** 12813119

Trivedi, M. H., Rush, A. J., Wisniewski, S. R., Nierenberg, A. A., Warden, D., Ritz, L., . . . Fava, M. (2006). Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: Implications for clinical practice. *American Journal of Psychiatry*, *163*(1), 28–40. **DOI:** https://doi.org/10.1176/appi.ajp.163.1.28, **PMID:** 16390886

Turecki, G., & Brent, D. A. (2016). Suicide and suicidal behaviour. *Lancet*, *387*(10024), 1227–1239. **DOI:** https://doi.org/10.1016/S0140-6736(15)00234-2

Uher, R. (2011). Genes, environment, and individual differences in responding to treatment for depression. *Harvard Review of Psychiatry*, *19*(3), 109–124. **DOI:** https://doi.org/10.3109/10673229.2011.586551, **PMID:** 21631158

World Health Organization. (2017). *Depression and other common mental disorders: Global health estimates*. Geneva, Switzerland: Author.