ARTICLE OPEN

Check for updates

Development of a differential treatment selection model for depression on consolidated and transformed clinical trial datasets

Kelly Perlman[1,2,3,7]✉, Joseph Mehltretter[3,7], David Benrimoh[1,2,3,7], Caitrin Armstrong[3], Robert Fratila[3], Christina Popescu[3,4], Jingla-Fri Tunteng[2,3], Jerome Williams[2,3], Colleen Rollins[2,5], Grace Golden[3,6] and Gustavo Turecki[1,2]

Major depressive disorder (MDD) is the leading cause of disability worldwide, yet treatment selection still proceeds via "trial and error". Given the varied presentation of MDD and heterogeneity of treatment response, the use of machine learning to understand complex, non-linear relationships in data may be key for treatment personalization. Well-organized, structured data from clinical trials with standardized outcome measures is useful for training machine learning models; however, combining data *across* trials poses numerous challenges. There is also persistent concern that machine learning models can propagate harmful biases. We have created a methodology for organizing and preprocessing depression clinical trial data such that transformed variables harmonized across disparate datasets can be used as input for feature selection. Using Bayesian optimization, we identified an optimal multi-layer dense neural network that used data from 21 clinical and sociodemographic features as input in order to perform differential treatment benefit prediction. With this combined dataset of 5032 individuals and 6 drugs, we created a differential treatment benefit prediction model. Our model generalized well to the held-out test set and produced similar accuracy metrics in the test and validation set with an AUC of 0.7 when predicting binary remission. To address the potential for bias propagation, we used a bias testing performance metric to evaluate the model for harmful biases related to ethnicity, age, or sex. We present a full pipeline from data preprocessing to model validation that was employed to create the first differential treatment benefit prediction model for MDD containing 6 treatment options.

Translational Psychiatry (2024)14:263 ; https://doi.org/10.1038/s41398-024-02970-4

INTRODUCTION

The World Health Organization estimates that major depressive disorder (MDD) impacts more than 300 million people worldwide [1]. Despite the availability of several effective depression treatments, many patients undergo the inefficient "trial and error" approach to treatment selection, which can result in lost time and worse disease outcomes [2, 3] Considering the heterogeneity of depression and treatment response, it would be of significant value to identify the optimal treatment specific to a particular patient's characteristics [4, 5].

Machine learning (ML) methods are well-suited for the challenge of developing personalized treatment approaches in psychiatry [4, 5]. Moreover, deep learning, an ML technique that uses artificial neural networks to learn high-level representations from raw data and model complex relationships between variables, is particularly suited for predicting depression treatment response and facilitating effective personalized treatment (see review of previous work in [6, 7]).

A recent review summarizing eight studies using deep learning methods to predict treatment response in depression found that models generally reached AUCs (area under the curve, a threshold-free measure of a model's ability to classify successfully) between 0.66 and 0.82 [7]. While some of these studies incorporated functional and structural magnetic resonance imaging, genetics, and epigenetics as model features, it is not currently feasible to collect biomarker data in routine practice which leaves symptom/clinical and sociodemographic features as the most practical potential predictors [8]. One key limitation of most of the studies reviewed is that they aimed to predict treatment response between two treatments or one treatment at a time, whereas the clinical reality that clinicians and patients face is selecting a treatment from many options: over 20 antidepressants, psychological therapies, and neuromodulation treatments [9].

Another limitation of machine learning models is the potential for propagation or amplification of harmful biases [5, 10]. For example, it is crucial to be conscious of, and address, the possibility that a model that learns to predict worse remission rates for patients from a particular background than what is actually observed for those patients in the data (assuming that the data is reasonably representative of the intended use population, which may not always be the case).

In previous work, we demonstrated the use of a neural network capable of differential treatment benefit prediction (that is, the generation of remission probabilities) for a number of treatments

[1]Douglas Mental Health University Institute, Montreal, QC, Canada. [2]McGill University, Montreal, QC, Canada. [3]Aifred Health Inc., Montreal, QC, Canada. [4]University of Alberta, Edmonton, AB, Canada. [5]University of Cambridge, Cambridge, UK. [6]University of Waterloo, Waterloo, ON, Canada. [7]These authors contributed equally: Kelly Perlman, Joseph Mehltretter, David Benrimoh. ✉email: kelly.perlman@mail.mcgill.ca

Received: 28 September 2023 Revised: 9 May 2024 Accepted: 29 May 2024
Published online: 21 June 2024

[5]. In this work, we expand on our previous work in two key ways. The first is by addressing the problem of dataset merging. Datasets must be merged for two reasons: to generate a sufficient sample size for model training and to provide the model with examples of patients on a number of different treatments in order for the model to be able to learn to differentiate between treatments. When working with data from a number of different sources, a significant challenge is the heterogeneity of study design and data collection: different studies used a variety of scales and questionnaires to assess depression symptoms and other psychopathology, physical health, and well-being outcomes (i.e., comorbid psychiatric symptoms and quality of life). Driven by shortcomings in the clinical utility of diagnostic categories and the phenomenological overlap of symptoms and traits across diagnoses, Waszczuk et al., produced a hierarchical system focused on the dimensionality of emotional disorders (e.g., somatoform, internalizing, detachment) and the various manifestations of these disorders that fall within each category [11]. For example, OCD is classified under the "fear" category, which itself is classified under the "internalizing" category [12]. While HiTOP shows potential advantages in both clinical utility and research, this taxonomy structure is limited to psychopathology dimensions; it does not represent other patient-level characteristics that are vital for predicting treatment response, such as demographic information (e.g., years of education, socioeconomic status), personal history (e.g., trauma), or physical health (e.g., body weight [13], comorbid health problems) [8]. It similarly does not capture the health outcomes necessary for understanding a patient's well-being and response/remission to treatment, such as daily functioning and quality of life. Here we present a detailed methodology for a novel taxonomy (inspired by the HiTOP method) and variable transformation procedure that was created in order to facilitate data collation and model training.

The second expansion of our previous work is to include a more robust assessment of learned model bias, in order to ensure that models put into clinical practice do not propagate harmful biases. Below, we present our differential treatment benefit prediction model results, subgroup analyses arising from our bias testing, and the patient features retained by the model.

METHODS

The ethics committee/IRB of the Douglas Mental Health University Institute gave ethical approval for this analysis. The end goal of our work was to produce a model capable of predicting remission and generating differential treatment benefit predictions for a number of different treatments. We specifically selected remission, based on a cutoff of 10 points on the MADRS and 7 points on the HAM-D, as the main outcome measure because it is both binary and the gold-standard objective in depression treatment [3].

Data

After signing a data sharing agreement, de-identified patient-level data from clinical trials of depression treatment were provided by GlaxoSmithKline and Eli Lilly, via the Clinical Study Data Request (CSDR) platform, along with the relevant study protocols. These datasets were chosen because of their accessibility in a digitized format and their heterogeneity compared to the data used in our previous work. Our inclusion criteria for data were simply that the primary indication was depression (comorbidities are allowed) and that outcomes were measured using standardized rating scales. Studies were excluded if they measured a pediatric population, if the patients had bipolar disorder, or if their depression was caused by substances or by another medical condition. After the study selection process illustrated in Fig. 1, we were left with 17 included studies. Supplementary Table 1 shows the breakdown of the studies included and the sample size (n) of the patients given each medication in each study. Our total sample size ($n = 5032$) was considered sufficient by comparison with previous studies conducted by our group and others in the field. Our current sample size exceeds that of our previous analyses, which had a total sample size of 3222 subjects [5].

Data preprocessing and taxonomization

The main steps taken to preprocess and prepare the data for merging are summarized in Fig. 2. We began by extracting and standardizing all the questionnaires from the datasets corresponding to these studies. This involved creating standard question texts and response values for each individual question of each questionnaire. This resulted in 57 standard questionnaires. According to the associated study protocols, as well as the raw data itself, we identified the version of the questionnaires that were used in the respective studies, in cases where several versions existed across studies (e.g., short form vs long form) or where differences in question phrasing/text were found [14].

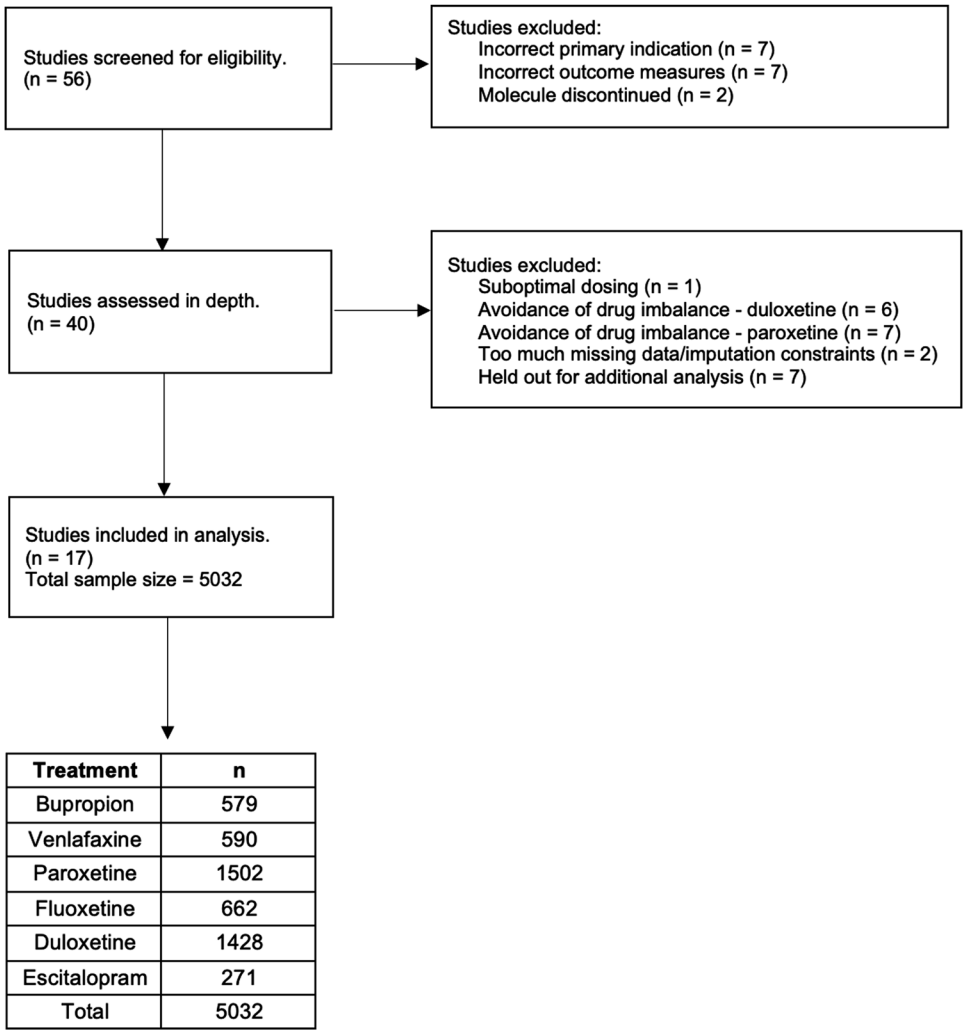
In parallel, we created a custom taxonomic system to categorize our data spanning across different clinical and demographic dimensions. The taxonomy was originally inspired by the work of Waszczuk et al. [11]. While including some of the emotional symptom-based categories defined by Waszczuk et al. [11], examples of additional higher-order clusters in our taxonomy were those defining sociodemographic, physiological, cognitive, and quality of life features. Overall, our taxonomy included 17 roots (i.e., higher-order categories), each containing a number of branches and leaves. As can be seen in Fig. 3, the quality of life root itself has further leaves and another sub-branch (relationships), which itself has leaves (family, social, romantic).

With this taxonomy system, we tagged each individual standardized question with a root category, and then used the branches to provide further categorical resolution. However, if the question was not able to be categorized using the lowest dimension it would be incorporated at the level at which it reflected the semantic meaning or clinical dimension of that feature, occasionally belonging to a higher-order category as necessary. For example, for a question about functional impairment in the context of personal relationships, it cannot be classified into any of the "family", "social", or "romantic" leaves, as the specific relationship is not specified; therefore, the most granular category it can accurately be attributed to is "relationships". We also allowed questions to be labeled with a secondary category in cases where this was deemed appropriate by 2 raters. Finally, we created flags that represent a certain characteristic of a question (e.g., whether it was specifically patient or clinician-rated, or referring to a past time point, etc…). See Supplementary Table 2 for a glossary of terms involved in the data preprocessing and taxonomy.

During this exercise, the categories assigned to each question were assessed by at least 2 raters. Disagreements were resolved by discussion and group consensus, which ensured there was reliability and consistency of categorization. With the questions sorted by category, we created "transformed questions" that would allow for semantically similar or identical standard questions to be combined into the same feature, which could then serve as input to the predictive model. We then matched the raw data to these standards and inserted it into the standardized database, generating a unique identifier for each standardized question.

This taxonomical organization helped to group all matched questions that belonged to the same high-level category. We could thus visualize all the questions within this high-level category and combine questions according to their lower-level category when appropriate. For example, if a question asked about overall functional impairment resulting in depression, it would be tagged in the quality of life → functional impairment category, but if a question asked specifically about functional impairment as it relates to the ability to care for oneself, it would be tagged in the quality of life → functional impairment → self-care category (Fig. 3).

While attempts were made to produce features at the lowest level of categorization (i.e., the most granular level), we did not combine questions where semantic differences existed and where the increased resolution would come at the expense of feature validity as assessed by raters and group discussion. Rather, we went back to the next highest level of the tree to encompass the questions at the more granular level, when this was appropriate (specifically, avoid over-grouping items in any way that comprises feature validity). One priority was to combine questions that had categorical or continuous, rather than binary, response values, in order to minimize the loss of resolution that would occur when a continuous-response question was binarized. Questions that were to be combined were first rescaled by equating the smaller scale variables to the largest scale variable. In doing so all variables that could be combined based on semantic similarity were all on the same scale. The specific equating method used was dependent on the value distribution of the variables to be combined. We iteratively compared each grouped variable with the max-scale variable and if both variables to be equated had a normal distribution then linear equating was the method used; otherwise, equipercentile equating was used [15]. Once questions within a category were all on a similar scale, we created our new transformed variable



Treatment	n
Bupropion	579
Venlafaxine	590
Paroxetine	1502
Fluoxetine	662
Duloxetine	1428
Escitalopram	271
Total	5032

Fig. 1 Flowchart summarizing study selection. Adapted PRISMA diagram indicating how individual studies were reviewed and selected for analysis, along with reasons for exclusion. "Avoidance of drug imbalance" refers to withholding studies to avoid having an overrepresentation of a given drug, reducing the risk of potential bias.

by taking the average value of all the variables for each patient. During scaling and transformation of categorical variables, as a validity check, we confirmed that the variance among the categorized values was not larger than 1; this value was chosen because, for the type of variables we were merging, a change in the value of 1 indicates a new level of severity (i.e., a change from 'often' to 'very often' on a question about a given symptom). Additionally, we confirmed using bar plots that the distribution of the transformed variable did not significantly vary from the variables that were averaged. By testing the variance of grouped variables after rescaling we could identify if any single variable within a group did not belong. Once equating was verified, we scaled all continuous variables using a standard scaler and added the minimum value plus a constant of 0.01 to variables where a value of less than or zero occurred. This additional constant was added due to how some neural network activation functions treat negative and zero values irregularly [16].

If all questions within a feature were binary, no transformation was necessary. In cases where a mixture of binary and categorical questions existed within a particular semantic category, and there was a preponderance of binary questions, the entire semantic category was "binarized", meaning that each response value was given a value of either 0 or 1, depending on the response value text and how it relates to the transformed question semantics. As a validity check, we verified that the binarized version of the categorical questions followed similar response distributions as the native binary questions in the same category. That way, we could be more confident that our binarization procedure did not introduce artifactual variability. Any conflict was settled by group consensus. Moreover, having a binary cutoff on a categorical question allowed for thresholding by symptom dimension intensity when necessary. The response value transformations for all of the included transformed questions can be found in Supplementary file 1 (spreadsheet). In summary, with all the questions in our dataset matched to standards, and these standards taxonomized, we could group semantically similar questions—those querying the same dimension—into a common "feature".

Feature selection

There exist many methods for feature selection, many of which rely on tree, linear, or logistic regression algorithms. However, in some situations, the optimal feature set selected is only optimal when used by the underlying algorithm. As we are using neural networks for our final classification model (based on the superior performance of these model types in previous work [5]) we decided to use neural networks for our feature selection task in the form of a new layer called CancelOut [17]. CancelOut is a fully connected layer that allows us to create a classification model with the same task of training with a target of remission. The CancelOut layer has a custom loss function that works as a scoring method so that by the end of training we can view and select features based on their score.

We first trained one neural network with the same specifications as the network we were testing with the additional CancelOut Layer as the first layer. This layer cannot be used in the final neural network model that will be used for testing and inference in the clinic, as this model must necessarily input only the chosen features to reduce the burden of data collection on patients and clinicians. The number of features retained was a feature fed into our Bayesian optimization framework (see below) that was then used for choosing the optimal set of hyperparameters.

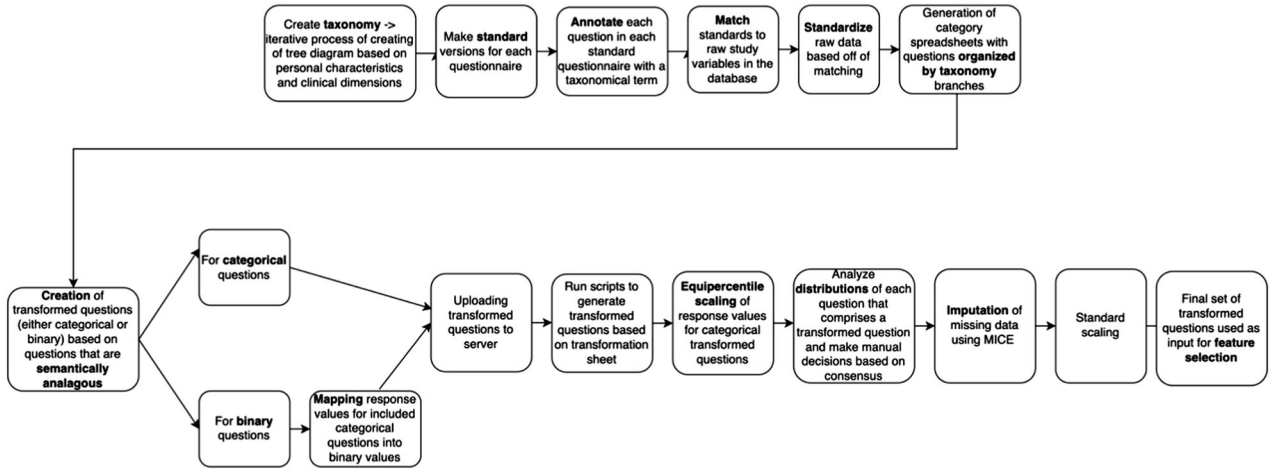


Fig. 2 Flowchart for custom data preprocessing and transformation. The process illustrated here covers the main steps from creating a standardized classification system for questions based on a taxonomy tree all the way to having a finalized set of transformed and scaled input features to use for feature selection.

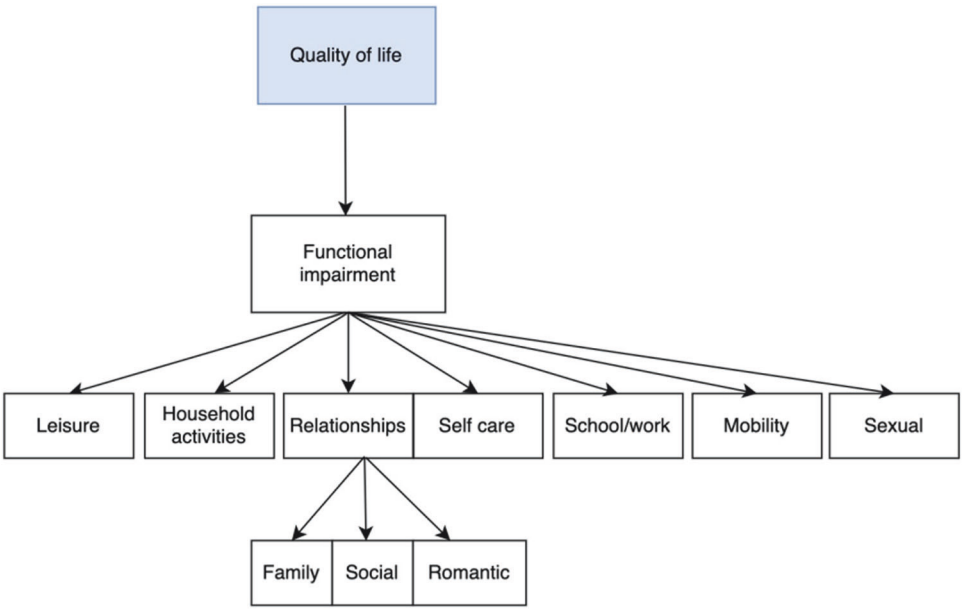


Fig. 3 Example of taxonomical category. The functional impairment category is a branch of the quality of life root that itself has further leaves and another sub-branch (relationships), which itself has leaves (family, social, romantic).

Performance metrics

Our model selection process using Bayesian optimization focused on optimizing the highest Area Under the Receiver Operating Curve (AUROC, often abbreviated as AUC), as it allowed us to understand if we were able to well separate between patients expected to remit or not remit. Since this metric is scale-invariant (i.e. more focused on the ranking than on the prediction of absolute values) and classification-threshold-invariant (focused on the effectiveness of the predictions irrespective of where the classification threshold is set), we can get a holistic and well-rounded view on the quality of the model and its performance [18]. Finally, positive predictive value (PPV), negative predictive value (NPV), sensitivity, and specificity are also used as these are commonly used in clinical studies are interpretable to both machine learning engineers and clinicians, and are clinically relevant. Multi-component metrics such as these (e.g., PPV + NPV and sensitivity + specificity) provide additional granularity on the model's ability to identify the positive and negative class independently. This granularity can aid in determining the clinical tolerance for false positive and false negative samples so that the model can be tailored correctly. All of these metrics leverage true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The PPV is calculated as TP/(TP + FP) and the PPV as TN/(TN + FN). The sensitivity is calculated as TP/(TP + FN) and the specificity as TN/(TN + FP) [19].

Sensitivity analyses were performed wherein we systematically varied the inputs of just one variable while holding all other variables constant [20]. Examining the resulting impact on model output predictions is useful for examining the directionality and impact of individual variables on predictions and ensuring that these cohere with established literature.

Bias testing

Finally, we consider bias testing as an important performance metric. We, therefore, compare the actual population remission rate to the average predicted remission rate for each of our accessible demographic factors (age, race, and sex) and confirm that there is not more than a 5% under-prediction of remission rates and produce subgroup analyses which examine the performance of the model on subgroups in different data splits.

Table 1. Sociodemographic data for all patients across datasets.

Total N	**5032**
Age	**mean ± sd**
Overall	43.6 ± 13.9
Sex	***n***
Male	1830 (36.34%)
Female	3202 (63.63%)
Race	***n***
White	3801 (75.53%)
Asian	758 (15.06%)
African Descent	246 (4.88%)
Hispanic	185 (3.67%)
Other	42 (0.83%)

RESULTS

Selected studies and final dataset

A total of 56 studies were screened for inclusion in our analysis (Fig. 1). Of these, 16 were excluded on the basis of having either an incorrect primary indication ($N = 7$) or outcome measure ($N = 7$) or having involved the study of a drug that has since been discontinued ($N = 2$). Of the remaining 40 studies, 23 were excluded on the basis of demonstrating suboptimal dosing regimens [3] ($N = 1$), and to avoid overrepresentation of certain medications for which large amounts of data were available (Duloxetine $N = 6$, Paroxetine, $N = 7$). A further 2 studies were removed as a result of too much missing data, and 7 were held-out for additional analysis. A total of 17 studies remained, containing data on 5032 patients and six treatments/medications, with absolute data point numbers ranging from 271 (escitalopram) to 1502 (paroxetine). Sociodemographic data (age, sex, race/ethnicity) for the population can be found in Table 1. These studies were then processed according to the procedures outlined in the methods section of this paper (Fig. 2). The dataset splits had the following sample sizes: train = 4099, validation = 422, test = 511. These studies were then processed according to the procedures outlined in the methods section of this paper (Fig. 2). The treatments covered three major drug classes: selective serotonin reuptake inhibitors (SSRI), serotonin and norepinephrine reuptake inhibitors (SNRI), and a norepinephrine and dopamine reuptake inhibitor (NDRI). Supplementary Table 1 shows the drug breakdown of the 17 studies included. The real population remission rate prior to modeling was 43.18%.

To identify the optimal structure for our dense neural network (DNN) we employed Bayesian optimization with our package, Vulcan. Bayesian optimization allowed us to test various network configurations and feature sets to test for optimality based on a set accuracy metric. For our testing purposes, we want to find a DNN structure and feature set that maximizes the AUC value. Our Bayesian optimization testing produced an optimal model that had two hidden layers both with 40 nodes that used exponential linear units (ELU) and a dropout value of 0.15 to assist with generalization [21]. The prediction layer determined our remission probabilities through the use of the softmax function. During training the network parameters were tuned using the Adam optimization algorithm with a learning rate of 0.001. This architecture was trained with early stopping to prevent the network from overfitting. Specifically, we had 300 epochs set as the maximum with an early stopping patience of 100, which resulted in the model using all 300 epochs to train. Table 3 demonstrates the basic statistical metrics used to analyze our data, stratified by the 3 respective test sets used.

Model performance

Table 2 lists the 26 features (both categorical and binary), including those developed with our custom taxonomy and transformation process, which were included in our final model. Please see Supplementary Fig. 1 for a visual representation of the missingness score per feature by study. We achieved accuracies of 65-66% on all data splits and AUCs of 0.65–0.7 (Table 3), results which are in line with previous work (see refs. 22, 23) despite the inclusion of a larger number of medications and the merging of several datasets. We maintained reasonable F1 scores, indicating that the model achieved a balance of precision and recall. For comparison, a logistic regression model was run using the same features and data splits; it achieved an AUC of 0.62 on the test set, underperforming the neural network model (see Supplementary results). It is important to note that these predictors for the logistic regression were selected by the deep learning model as it trained which is in line with common practices of using the same model for feature selection and model training.

Table 2. The feature table depicts the top 26 features included in the model.

Features	
Sociodemographic	Race/ethnicity
	Sex
	Age
Symptoms	Negative thoughts—binary
	Concentration difficulties—binary
	Suicidal ideation and planning—binary
	Excessive guilt—binary
	Early insomnia
	Hypochondriasis
	Psychomotor agitation
	Late insomnia
	Overall suicidal ideation
	Weight loss
	Genital symptoms (loss of libido, menstrual disturbances)
	Guilt
	Decreased appetite
	Middle insomnia
	Feelings of worthlessness
	Anxiety—somatic (gastrointestinal, indigestion, cardiovascular, palpitation, headaches, respiratory, genito-urinary, etc.)
	Insight
	Gastrointestinal disturbances
	Anxiety—psychic
	Hopeless outlook on the future
	Anxiety/tension/inability to relax—binary
	Anxious state
	Leaden paralysis

Examples of sensitivity analysis results are presented in Fig. 4. We found that these cohered with previous literature and generally with clinical experience [8]. For example, the proportion of patients remitting decreased as we increased the suicidal ideation score; this coheres with a systematic metareview finding that suicidality was a predictor of treatment response [8]. Similarly, as concentration difficulty and leaden paralysis scores increase (indicating worsening symptomatology) the proportion of patients remitting decreases. We note that cognitive impairment, related to poor concentration, has been found to be predictive of reduced response to antidepressants, and that psychomotor retardation has been linked with worse

Table 3. Demonstration of basic statistical metrics, stratified by the 3 sets of data used.

Data split	Accuracy	Sensitivity	Specificity	PPV	NPV	F1	AUC
Train	0.66	0.535	0.75	0.615	0.68	0.57	0.7
Validation	0.65	0.51	0.76	0.62	0.67	0.56	0.65
Test	0.66	0.52	0.76	0.63	0.67	0.57	0.7

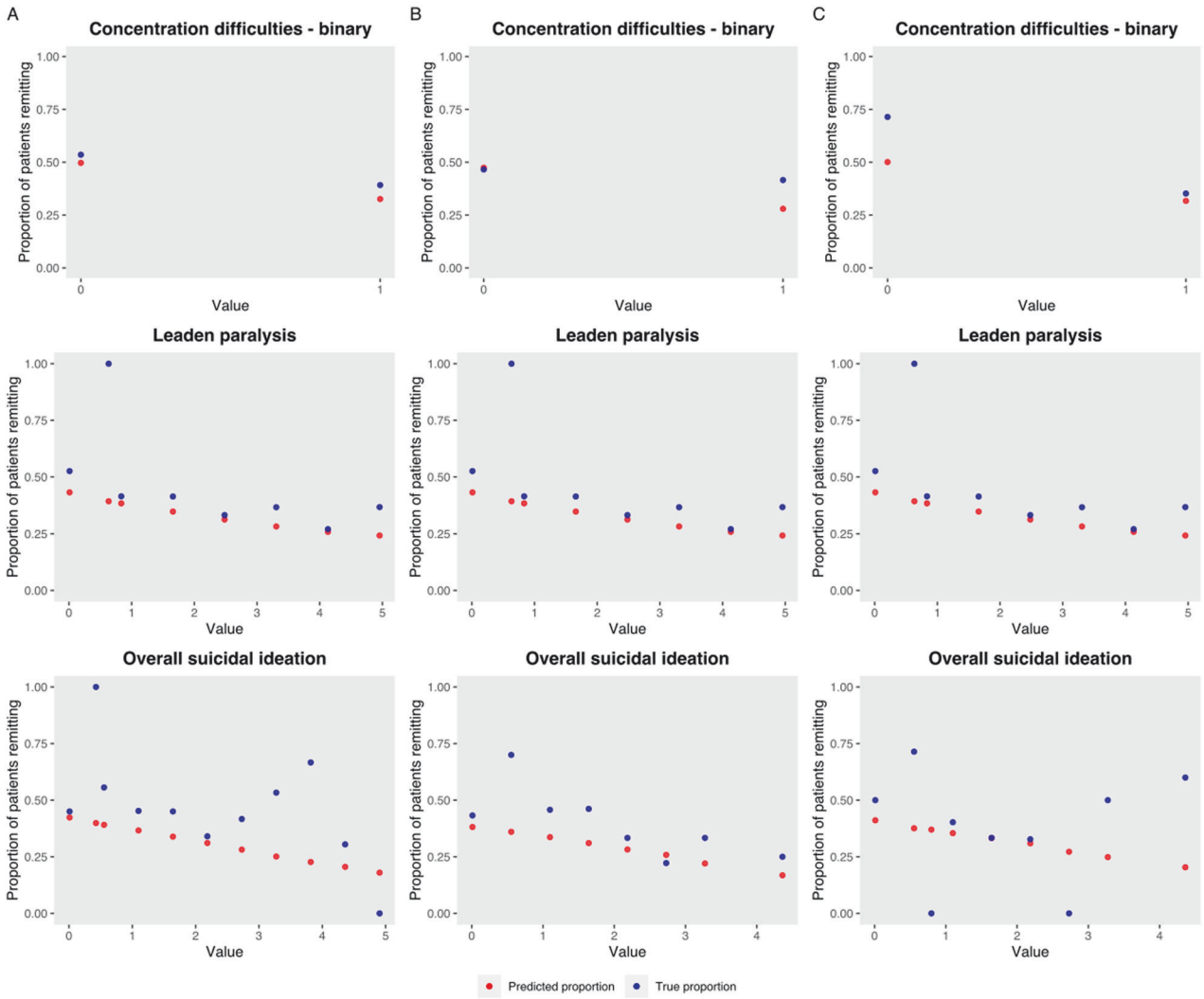


Fig. 4 Sensitivity test output examples. These figures show, that for each of our **A** train ($n = 4099$), **B** validation ($n = 422$), and **C** test ($n = 511$) sets, the proportion of patients remitting according to the target variable is varied. Of these 3 features, two (recent suicidal ideation, leaden paralysis) possess categorical values, while one (concentration) is an example of a binary feature. The number of data points reflects the possible values that were created during the transformation and standardization process—in some cases values were created between whole numbers in order to best represent partially overlapping scales. Blue dots represent the observed proportion of patients remitting who truly had those values; red dots represent the predicted probabilities when all patients in the dataset have the feature set to the given value. Note that in smaller datasets like in validation or training, where very few patients may naturally have a given value, a larger variance occurs in the observed values.

response to SSRIs [8]. Note that we see greater amounts of variance for the validation and test sets than the train set, as is expected given their smaller size of 422 and 511 subjects, respectively.

Bias testing

Figure 5 provides the results of our post hoc bias tests examining the difference in the predicted versus the observed (real) probability for each of the subgroups. No group had an under-prediction of remission rate, compared to the true rate, of more than 5%, indicating that the model did not learn to amplify biases for any one group. In Supplementary Tables 3–5, the basic statistical metrics used to analyze our subgroup data (race, sex, age) are listed for each of the respective sets (train, validation, test); reasonable metrics are preserved for each subgroup, except when the n for that subgroup is small.

DISCUSSION

In this study, we present a full pipeline from data preprocessing to model validation that was harnessed to create the first-ever

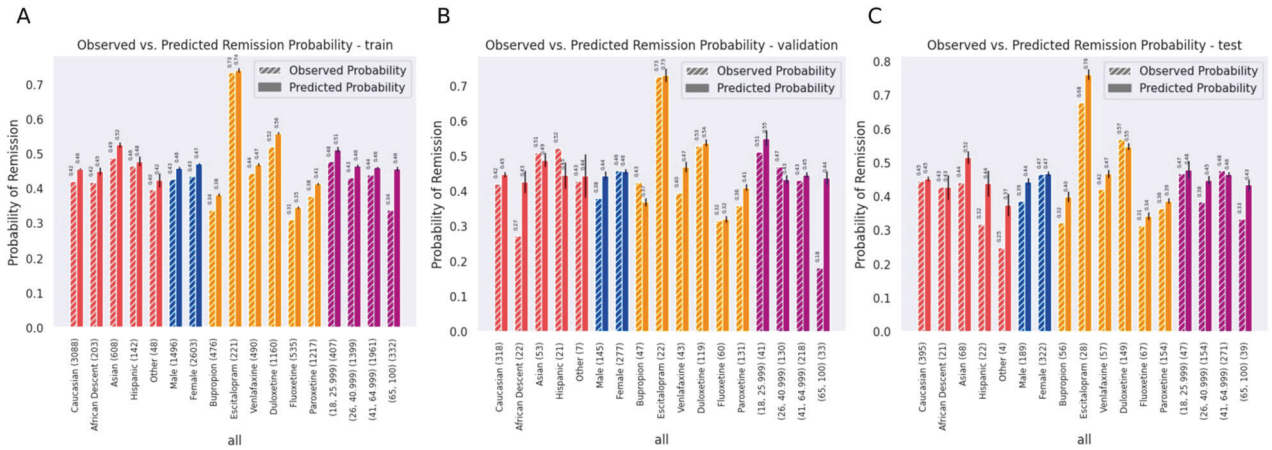


Fig. 5 Report on bias testing. This figure shows both the observed and mean predicted probability, including the standard error, of patients remitting based on their race (Caucasian, African Descent, Asian, Hispanic, Other), sex (Male, Female), antidepressant treatments (Bupropion, Escitalopram, Venlafaxine, Fluoxetine, Paroxetine), and age groups (18–25.9; 26–40.9; 41–64.9; 65–130.9) in **A** train, **B** validation, **C** test sets. The error bars represent the standard error for the predicted probability.

differential treatment selection for MDD containing 6 treatment options. We introduce our novel methodology for tackling the heterogeneity of available datasets—particularly the creation of a custom taxonomy and transformation process that expanded upon previous taxonomies which, alone, were unable to capture and integrate the breadth of variables required. We also discuss other challenges, such as the missingness of data and the need to transform categorical and binary variables differently. In addition, we demonstrate bias testing through the analysis of model metrics and the comparison of predicted and observed remission rates in key subgroups.

In the literature, there is a general lack of detailed mechanistic understanding regarding factors that impact treatment remission in depression and how they interact with one another [8]. Recent work has focused on anxious depression and its correlation with worse prognosis [8], and studies have found that certain variables such as psychic anxiety are more negatively predictive of treatment outcome [8]. Our model also found psychic anxiety to be an important predictive feature (Table 2). Iniesta et al. [24] found other predictive variables including questions related to apparent sadness, pessimism, and indecisiveness. Sadness, perhaps due to being a common symptom and therefore unlikely to be predictive of differential treatment response, was not identified as an important feature by our model; however, pessimism—captured as a "hopeless outlook of future"—did appear on our list of features, as did difficulties with concentration, which may be linked to indecisiveness [25, 26]. Sleep is another feature identified as predictive by our model, which has also been found to be predictive in previous work (see [8] for a review). More specifically, work has shown that prolonged sleep latency and insomnia, both alone and in combination, are predictive of nonremission [27]. Moreover, a sleep profile consisting of reduced REM latency, increased REM density, and poor sleep continuity described in older work, was associated with poorer outcomes following psychotherapy (interpersonal and cognitive behavioral therapy [28, 29]) and pharmacotherapy (fluoxetine or imipramine [29]).

Race was found to be an important predictive feature. Our interpretation of this finding stems from the notion that race is often confounded with a number of important sociodemographic features—not available in the dataset and that this may be driving the effect. Previous studies, such as the STAR*D study, have concluded that black individuals had lower remission rates than white individuals [30] and hispanic individuals were somewhere in between; crucially, this was before adjusting for social factors and the statistical significance was lost following the adjustments, although black individuals did still have a lower remission rate [31]. A follow-up study argued for the role of genetic ancestry—not race—as accounting for much of the residual disparity even after accounting for socioeconomic and baseline clinical factors [32]. As such, clinicians interpreting the results of models trained on datasets such as this one should consider not just race, but other social determinants of health often confounded with race when treating patients.

Sex was also found to be an important feature. This is despite the fact that there is no clear consensus regarding sex-related differences in remission with antidepressant treatment [33]. There does seem to be some evidence that serotonergic antidepressants yield better responses in females than males due to the modulating role of estrogen [34]. What is most likely, however, is that sex is interacting with other features while the prediction is made, as noted for sleep above. These complex, non-linear interactions can be a challenge to interpret, though some headway can be made using classical techniques, though this would be out of scope for this paper (see [5]).

The reader may note that, at the population level, clear trends emerge in our data with respect to the ranking of antidepressants in terms of their predicted effectiveness. These predictions reflect the underlying data (see Fig. 5) and are indeed reminiscent of the order of the ranking of treatment efficacy in two large meta-analyses [9, 35]. At the individual patient level, however, the ranking of all treatments changes, and, crucially, the remission rate predictions for each treatment vary. The fundamental clinical utility of the model lies in this variation. However, the good performance of some drugs (e.g., escitalopram) may positively skew the predicted benefit result at the population level (this was despite the fact that most trials included here, including escitalopram, had similar inclusion and exclusion criteria). As such, clinical trials are needed to assess the real-world impact of the model on treatment outcomes.

There has been an encouraging recent effort to standardize the assessment instruments used in both clinical practice and clinical studies—for example, the use of the PHQ-9, GAD-7, and WHODAS 2.0 [36]. While these questionnaires were not available in the majority of the datasets we had access to for this study, their inclusion in future research should facilitate the development of the next generation of predictive models and simplify pipelines required for their development. At the same time, efforts to standardize assessments may also be informed by work such as that which we present here. For example, the PHQ-9 has a single item covering both reduced and increased appetite, whereas our model identified reduced appetite as a predictive symptom; as

such, future work on harmonized instruments might benefit from including more precise items but only in cases where these have been shown to be predictive.

Compared to imaging or genetic-based predictive tools, questionnaire-based assessments may be more easily integrated into clinical practice and may provide results with greater speed as they can be generated as soon as the patient and clinician respond to any required items. However, there can still be significant barriers to implementing questionnaire-based assessments and predictive tools into clinical practice. These barriers can include assessments that are not designed to fit into the clinical workflow, which are too time-consuming for clinicians or patients to complete, or which are done in an unwieldy manner (e.g., on paper and then uploaded to a computer database, or through a poorly designed computer program). In previous work, we have described a participatory iterative design process in which clinicians and patients are engaged in the development and validation of a computer-based decision support system [37–39]. The platform was designed to be used rapidly by clinicians within a clinical appointment, and by patients at home via mobile application with reminders to complete assessments. With this computerized system, both patients and clinicians were shown to use the platform in a consistent manner over a 12-week follow-up period [38].

Once a model has been trained and tested in the manner we describe in this article, it is important to consider how it could then be implemented into clinical practice. The first step would be to define the process for collecting data from future patients. The feature selection, performed by the pipeline we describe, defines the features that must be collected from patients in order to generate predictions. Once the features are known, a representative question could then be chosen from among the existing validated items used to create the feature. This could be done based on selecting the most common question in the dataset from among those comprising the feature, or based on the question whose distribution best matches that of the final feature. The questions selected in this manner could then be administered to patients via a computer interface, and the responses to these questions would then be fed through the pipeline we describe in order to generate data consistent with the features in the model training data. This data could then be used by the model to generate predictions, which would take the format of remission probabilities for each treatment the model was trained on. These remission probabilities would then be presented to clinicians using a computerized decision support platform, such as the one described in [37, 38], as one more piece of clinical information that could be used to, in collaboration with the patient, make a treatment decision. Indeed, we have proceeded to do this in a clinical trial (NCT04655924) whose results will be reported separately.

Our study has a number of limitations. Firstly, several features had to be discarded because they did not meet the sample size limitations. It remains possible that these features may have predictive power; however, they could not be included in our model given that questions must have a representative population for each treatment. Furthermore, the results of individual studies are not representative of populations that are not included in the data. The use of clinical trial data in this case results in a dataset with a number of exclusion criteria, limiting generalizability. Common exclusion criteria for the included studies are psychiatric comorbidities, many of which, such as personality disorders for instance, are predictors of treatment response. Consequently, the results may not be generalizable to certain patient groups with more complex courses of illness or with treatment-resistant depression. There are also some concerns with the basic demographic information that was included. Notably, race and ethnicity categories are inconsistent across studies. For instance, some studies have a specific category for Native Americans whereas others may consider this population as part of the 'Other' category. Nevertheless, our data did have a wide spread of ages and considerable representation of different ethnicities, though future work needs to include more diverse populations. Moreover, some studies are missing important sociodemographic data such as level of education and socioeconomic status, which are both key variables in predicting remission [40]. Additionally, the model does not include all possible treatments due to the constraints of the data available; a notable example is mirtazapine. Finally, it is worth noting that our taxonomization process depended on subjective judgements of raters and the larger group; despite the fact that the merged data was validated in the method described above, this initial reliance on a qualitative process does introduce the possibility for biases or errors (see Supplementary Methods for details on how this process was handled, see Supplemental Material for a discussion on the bias).

In conclusion, we present a complete pipeline used to produce the first-ever differential treatment benefit prediction model for MDD containing 6 first-line medication options. Mental health data taken from different studies is remarkably heterogeneous, making attempts at merging datasets in order to facilitate the generation of a differential treatment benefit prediction model inherently difficult. To our knowledge, we are the first to create a detailed pipeline for the merging and transformation of heterogeneous clinical trial dataset variables in order to facilitate the generation of a treatment benefit prediction model. We hope that other researchers in the field of psychiatry and beyond can use this generalizable framework to harness the utility of these highly variable yet crucial datasets.

DATA AVAILABILITY

All GlaxoSmithKline and Eli Lilly data were obtained via the Clinical Study Data Request (CSDR) platform—we do not own any of the clinical trial data. Studies with the following IDs were utilized in this paper: gsk_29060_128, gsk_29060_115, US-HMCR, gsk_ak130939, gsk-WXL101497, gsk_29060_874, lilly_F1J-MC-HMCQ, lily_F1J-AA-HMCV, gsk_29060_810, gsk_AK1113351, lilly_FJ1-MC-HMAYa, lilly_FJ1-MC-HMAYb, lilly_FJ1-MC-HMATA, lilly_FJ1-MC-HMATb, lilly_FIJ-MC-HMAQb, lilly_F1J-MC-HMBV, lilly_FJ1-MC-HMAQa. All information about the model, including model weights and scaling factors can be found at the following github link: https://github.com/Aifred-Health/pharma_research_model.

CODE AVAILABILITY

Vulcan framework for rapid deep learning model prototyping and analysis is available for public use: https://github.com/Aifred-Health/VulcanAI.

REFERENCES

1. World Health Organization. Depression and other common mental disorders: global health estimates. World Health Organization; 2017. https://apps.who.int/iris/bitstream/handle/10665/254610/W?sequence=1.
2. Kraus C, Kadriu B, Lanzenberger R, Zarate CAJ, Kasper S. Prognosis and improved outcomes in major depression: a review. Transl Psychiatry. 2019;9:127.
3. Kennedy SH, Lam RW, McIntyre RS, Tourjman SV, Bhat V, Blier P, 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.
4. Benrimoh D, Fratila R, Israel S, Perlman K, Mirchi N, Desai S, et al. Aifred health, a deep learning powered clinical decision support system for mental health. In: The NIPS '17 Competition: Building Intelligent Systems. California USA: Spinger International Publishing; 2018. pp. 251–87.
5. Mehltretter J, Fratila R, Benrimoh DA, Kapelner A, Perlman K, Snook E, et al. Differential treatment benet prediction for treatment selection in depression: a deep learning analysis of STAR*D and CO-MED data. Computational Psychiatry. 2020;4:61.
6. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. Mol Psychiatry. 2019;24:1583–98.
7. Squarcina L, Villa FM, Nobile M, Grisan E, Brambilla P. Deep learning for the prediction of treatment response in depression. J Affect Disord. 2021;281:618–22.
8. Perlman K, Benrimoh D, Israel S, Rollins C, Brown E, Tunteng J-F, et al. A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. J Affect Disord. 2019;243:503–15.
9. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. Lancet. 2018;391:1357–66.

10. Tanguay-Sela M, Benrimoh D, Popescu C, Perez T, Rollins C, Snook E, et al. Evaluating the perceived utility of an artificial intelligence-powered clinical decision support system for depression treatment using a simulation center. Psychiatry Res. 2022;308:114336.
11. Waszczuk MA, Kotov R, Ruggero C, Gamez W, Watson D. Hierarchical structure of emotional disorders: from individual symptoms to the spectrum. J Abnorm Psychol. 2017;126:613–34.
12. Ruggero CJ, Kotov R, Hopwood CJ, First M, Clark LA, Skodol AE, et al. Integrating the hierarchical taxonomy of psychopathology (HiTOP) into clinical practice. J Consult Clin Psychol. 2019;87:1069–84.
13. Puzhko S, Aboushawareb SAE, Kudrina I, Schuster T, Barnett TA, Renoux C, et al. Excess body weight as a predictor of response to treatment with antidepressants in patients with depressive disorder. J Affect Disord. 2020;267:153–70.
14. Fenton C, McLoughlin DM. Usefulness of Hamilton rating scale for depression subset scales and full versions for electroconvulsive therapy. PLoS One. 2021;16:e0259861.
15. Kolen MJ, Brennan RL. Test equating, scaling, and linking. New York, NY: Springer New York; 2014.
16. Lederer J. Activation functions in artificial neural networks: a systematic overview. https://doi.org/10.48550/ARXIV.2101.09957. 2021.
17. Borisov V, Haug J, Kasneci G. CancelOut: a layer for feature selection in deep neural networks. In: Tetko IV, Kůrková V, Karpov P, Theis F, editors. Artificial neural networks and machine learning—ICANN 2019: deep learning. Cham: Springer International Publishing; 2019. pp. 72–83
18. Jin Huang, Ling CX. Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng. 2005;17:299–310.
19. Tharwat A. Classification assessment methods. Appl Comput Inf https://api.semanticscholar.org/CorpusID:59212480. 2020.
20. Engelbrecht AP, Cloete I, Zurada JM. Determining the significance of input parameters using sensitivity analysis. In: Mira J, Sandoval F, editors. From natural to artificial neural computation. Berlin, Heidelberg: Springer; 1995. pp. 382–8.
21. Clevert D-A, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (ELUs). https://doi.org/10.48550/ARXIV.1511.07289. 2015.
22. Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH, et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry. 2016;3:243–50.
23. Mehltretter J, Rollins C, Benrimoh D, Fratila R, Perlman K, Israel S, et al. Analysis of features selected by a deep learning model for differential treatment selection in depression. Front Artif Intell. 2019;2:31.
24. Iniesta R, Malki K, Maier W, Rietschel M, Mors O, Hauser J, et al. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res. 2016;78:94–102.
25. American Psychiatric Association. DSM-5 diagnostic classification. Diagnostic and statistical manual of mental disorders. 10th ed. Washington, D.C.: American Psychiatric Association; 2013.
26. Lauderdale SA, Oakes K. Factor structure of the revised indecisiveness scale and association with risks for and symptoms of anxiety, depression, and attentional control. J Rat-Emo Cognitive-Behav Ther. 2021;39:256–84.
27. Troxel WM, Kupfer DJ, Reynolds CF III, Frank E, Thase ME, Miewald JM, et al. Insomnia and objectively measured sleep disturbances predict treatment outcome in depressed patients treated with psychotherapy or psychotherapy-pharmacotherapy combinations. J Clin Psychiatry. 2012;73:478–85.
28. Thase ME. Abnormal electroencephalographic sleep profiles in major depression: association with response to cognitive behavior therapy. Arch Gen Psychiatry. 1996;53:99.
29. Thase ME. Which depressed patients will respond to interpersonal psychotherapy? The role of abnormal EEG sleep profiles. Am J Psychiatry. 1997;154:502–9.
30. Panaite V, Bowersox NW, Zivin K, Ganoczy D, Kim HM, Pfeiffer PN. Individual and neighborhood characteristics as predictors of depression symptom response. Health Serv Res. 2019;54:586–91.
31. Lesser IM, Castro DB, Gaynes BN, Gonzalez J, Rush AJ, Alpert JE, et al. Ethnicity/race and outcome in the treatment of depression: results from STAR*D. Medical Care. 2007;45:1043–51.
32. Murphy E, Hou L, Maher BS, Woldehawariat G, Kassem L, Akula N, et al. Race, genetic ancestry and response to antidepressant treatment for major depression. Neuropsychopharmacol. 2013;38:2598–606.
33. Sramek JJ, Murphy MF, Cutler NR. Sex differences in the psychopharmacological treatment of depression. Dialogues Clin Neurosci. 2016;18:447–57.
34. Berlanga C, Flores-Ramos M. Different gender response to serotonergic and noradrenergic antidepressants. A comparative study of the efficacy of citalopram and reboxetine. J Affect Disord. 2006;95:119–23.
35. Cipriani A, Furukawa TA, Salanti G, Geddes JR, Higgins JP, Churchill R, et al. Comparative efficacy and acceptability of 12 new-generation antidepressants: a multiple-treatments meta-analysis. Lancet. 2009;373:746–58.
36. Waheed A, Afridi AK, Rana M, Arif M, Barrera T, Patel F, et al. Knowledge and behavior of primary care physicians regarding utilization of standardized tools in screening and assessment of anxiety, depression, and mood disorders at a large integrated health system. J Prim Care Community Health. 2024;15:21501319231224710.
37. Benrimoh D, Tanguay-Sela M, Perlman K, Israel S, Mehltretter J, Armstrong C, et al. Using a simulation centre to evaluate preliminary acceptability and impact of an artificial intelligence-powered clinical decision support system for depression treatment on the physician–patient interaction. BJPsych Open. 2021;7:e22.
38. Popescu C, Golden G, Benrimoh D, Tanguay-Sela M, Slowey D, Lundrigan E, et al. Evaluating the clinical feasibility of an artificial intelligence–powered, web-based clinical decision support system for the treatment of depression in adults: longitudinal feasibility study. JMIR Form Res. 2021;5:e31862.
39. Golden G, Popescu C, Israel S, Perlman K, Armstrong C, Fratila R, et al. Applying artificial intelligence to clinical decision support in mental health: what have we learned? Health Policy Technol. 2024.
40. Carter GC, Cantrell RA, Victoria Zarotsky, Haynes VS, Phillips G, Alatorre CI, et al. Comprehensive review of factors implicated in the heterogeneity of response in depression: review: heterogeneity in depression. Depress Anxiety. 2012;29:340–54.

ACKNOWLEDGEMENTS

We would like to thank GlaxoSmithKline and Eli Lilly for providing the de-identified individual patient raw data and information for the clinical trials.

AUTHOR CONTRIBUTIONS

DB, JM, RF, CA, KP, and GT conceived of and designed the study. KP, CP, JFT, JW, CR, GG, DB, and JM performed the preprocessing and transformation JM, CA, RF, and DB performed the machine learning analyses and validation. GT supervised the study. All authors contributed to the writing and/or editing of the manuscript. KP, JM, and DB contributed equally to this study.

FUNDING

This work was supported by a grant from ERA-Permed Vision 2020 supporting IMADAPT.

COMPETING INTERESTS

KP, JM, DB, CA, RF, CP, JFT, JW, CR, and GG are current or former shareholders, option holders, employees, and/or officers of Aifred Health. GT has no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41398-024-02970-4.

Correspondence and requests for materials should be addressed to Kelly Perlman.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.