From the: Max Planck Institute of Psychiatry

Dissertation

zum Erwerb des Doctor of Philosophy (Ph.D.) an
der

Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

Using Machine Learning to Predict Treatment Outcome in Depression – Hype or Hope?

vorgelegt von:
Nicolas Rost

aus:

Bamberg, Germany

Jahr:

2023

Mit Genehmigung der Medizinischen Fakultät der

Ludwig-Maximilians-Universität zu München

First evaluator: *Prof. Dr. Dr. Elisabeth Binder*

Second evaluator: *Prof. Dr. Nikolaos Koutsouleris*

Third evaluator: *Priv. Doz. Dr. Daniela Eser-Valeri*

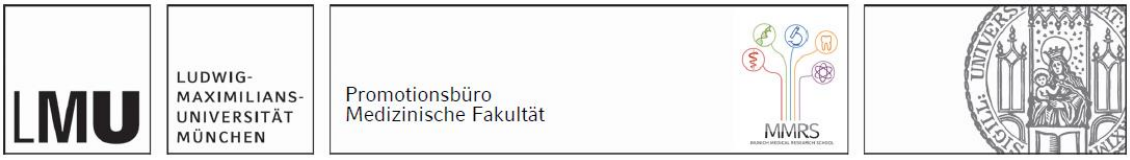
Fourth evaluator: *Priv. Doz. Dr. Florian Seemüller*

Dean: Prof. Dr. med. Thomas Gudermann

Datum der Verteidigung:

19.09.2023

Affidavit



Affidavit

Rost, Nicolas

Surname, first name

Kraepelinstr. 2-10

Street

80804, Munich, Germany

Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Using Machine Learning to Predict Treatment Outcome in Depression – Hype or Hope?

is my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

I further declare that the dissertation presented here has not been submitted in the same or similar form to any other institution for the purpose of obtaining an academic degree.

Munich, 1st March 2023
place, date

Nicolas Rost
Signature doctoral candidate

Confirmation of congruency

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Promotionsbüro Medizinische Fakultät

MMRS

Confirmation of congruency between printed and electronic version of the doctoral thesis

Rost, Nicolas

Surname, first name

Kraepelinstr. 2-10

Street

80804, Munich, Germany

Zip code, town, country

I hereby declare, that the submitted thesis entitled:

Using Machine Learning to Predict Treatment Outcome in Depression – Hype or Hope?

is congruent with the printed version both in content and format.

Munich, 1st March 2023
place, date

Nicolas Rost
Signature doctoral candidate

Table of content

List of abbreviations

CV cross-validation

HDRS Hamilton Rating Scale for Depression (17-item version)

ICD International Classification of Diseases

MARS Munich Antidepressant Response Signature

MDD major depressive disorder

ML machine learning

PRS polygenic risk scores

RFE recursive feature elimination

TAU treatment-as-usual

List of publications

Kaltwasser, L., **Rost, N.**, Ardizzi, M., Calbi, M., Settembrino, L., Fingerhut, J., Pauen, M., & Gallese, V. (2019). Sharing the filmic experience - The physiology of socio-emotional processes in the cinema. *PLoS ONE*, *14*(10), 1–19. https://doi.org/10.1371/journal.pone.0223259

Kappelmann, N., Arloth, J., Georgakis, M. K., Czamara, D., **Rost, N.**, Ligthart, S., Khandaker, G. M., & Binder, E. B. (2021). Dissecting the Association between Inflammation, Metabolic Dysregulation, and Specific Depressive Symptoms: A Genetic Correlation and 2-Sample Mendelian Randomization Study. *JAMA Psychiatry*, *78*(2), 161–170. https://doi.org/10.1001/jamapsychiatry.2020.3436

Kappelmann, N., Czamara, D., **Rost, N.**, Moser, S., Schmoll, V., Trastulla, L., Stochl, J., Lucae, S., Binder, E. B., Khandaker, G. M., & Arloth, J. (2021). Polygenic risk for immuno-metabolic markers and specific depressive symptoms: A multi-sample network analysis study. *Brain, Behavior, and Immunity*, *95*(March), 256–268. https://doi.org/10.1016/j.bbi.2021.03.024

Lopez, J. P., Brivio, E., Santambrogio, A., De Donno, C., Kos, A., Peters, M., **Rost, N.**, Czamara, D., Brückl, T. M., Roeh, S., Pöhlmann, M. L., Engelhardt, C., Ressle, A., Stoffel, R., Tontsch, A., Villamizar, J. M., Reincke, M., Riester, A., Sbiera, S., … Chen, A. (2021). Single-cell molecular profiling of all three components of the HPA axis reveals adrenal ABCB1 as a regulator of stress adaptation. *Science Advances*, *7*(5), 1–18. https://doi.org/10.1126/sciadv.abe4497

Rost, N., Binder, E. B., & Brückl, T. M. (2022). Predicting treatment outcome in depression: an introduction into current concepts and challenges. *European Archives of Psychiatry and Clinical Neuroscience*. https://doi.org/10.1007/s00406-022-01418-4

Rost, N., Brückl, T. M., Koutsouleris, N., Binder, E. B., & Müller-Myhsok, B. (2022). Creating sparser prediction models of treatment outcome in depression: a proof-of-concept study using simultaneous feature selection and hyperparameter tuning. *BMC Medical Informatics and Decision Making*, *22*(1), 181. https://doi.org/10.1186/s12911-022-01926-2

Rost, N., Dwyer, D. B., Gaffron, S., Rechberger, S., Maier, D., Binder, E. B., & Brückl, T. M. (2023). Multimodal predictions of treatment outcome in major depression: A comparison of data-driven predictors with importance ratings by clinicians. *Journal of Affective Disorders*, *327*, 330–339. https://doi.org/10.1016/j.jad.2023.02.007

Scherf-Clavel, M., Weber, H., Wurst, C., Stonawski, S., Hommers, L., Unterecker, S., Wolf, C., Domschke, K., **Rost, N.**, Brückl, T., Lucae, S., Uhr, M., Binder, E. B., Menke, A., & Deckert, J. (2022). Effects of Pharmacokinetic Gene Variation on Therapeutic Drug Levels and Antidepressant Treatment Response. *Pharmacopsychiatry*, *55*(5), 246–254. https://doi.org/10.1055/a-1872-0613

Schweizer, G., Furley, P., **Rost, N.**, & Barth, K. (2020). Reliable measurement in sport psychology: The case of performance outcome measures. *Psychology of Sport and Exercise, 48*. https://doi.org/10.1016/j.psychsport.2020.101663

Your contribution to the publications

Contribution to paper I

Paper I (Rost, Brückl, et al., 2022) was published in *BMC Medical Informatics and Decision Making* in 2022. NR developed the study design and methodology, conducted the statistical analyses, wrote the initial draft, and revised the manuscript. TMB and EBB curated and contributed data. NK and BMM contributed to the development of the study design and methodology. TMB, NK, EBB and BMM critically contributed to the writing of the manuscript and supervised the project.

Contribution to paper II

Paper II (Rost et al., 2023) was published in the *Journal of Affective Disorders* in 2023. NR developed the study design and methodology, conducted the statistical analyses, wrote the initial draft, and revised the manuscript. EBB and TMB curated and contributed data, contributed to the study design and the writing of the manuscript, and supervised the project. DM contributed to the study design and methodology. DBD, SG and SR contributed to the statistical analyses.

Contribution to paper III (Appendix A)

Paper III (Rost, Binder, et al., 2022) was published in *European Archives of Psychiatry and Clinical Neuroscience* in 2022. NR wrote the initial draft and revised the manuscript. EBB and TMB critically contributed to the writing of the manuscript and its revisions.

1. Introductory Summary

Major depressive disorder (MDD) is one of the most common and most severe illnesses worldwide by affecting over 300 million people (World Health Organization, 2017) and accounting for more than 14% of all years lived with disability (James et al., 2018). In addition to the high personal suffering of patients and the high mortality, indicated by nearly 800,000 suicide deaths per year (World Health Organization, 2017), MDD also results in a large economic burden (Greenberg et al., 2015; König et al., 2019). Given these circumstances, advancing antidepressant treatment and increasing its efficacy has been a major goal of psychiatric research ever since. However, response rates to antidepressant medication remain low. Depending on the respective trial, only about one third of patients achieve remission after the initial treatment (Rush et al., 2006), and about 50% do not show a sufficient response even after several treatment attempts (Souery et al., 2007; Thomas et al., 2013).

In general, patients with MDD should be treated according to evidence-based clinical practice guidelines (Hollon et al., 2014) or consensus papers that are available for different countries (for an overview, see Kraus et al., 2019). However, standardized guidelines are not always adhered to (Herzog et al., 2017) and antidepressant treatment is often administered in a trial-and-error fashion, influenced by clinician experiences and patient preferences (Cohen & DeRubeis, 2018; Maj et al., 2020). These current approaches to treatment selection may be one reason for the low response rates, among others. Moreover, they are not operating in the sense of the declared goal: a personalized psychiatry, also called “precision psychiatry”, i.e., the targeted and early matching of each patient to the best possible treatment based on individual patient characteristics (Cohen & DeRubeis, 2018). But is this goal even realistic, and if so, why has it not yet been achieved?

1.1 From associations to predictions of treatment outcome

Over many years, psychiatric research has examined and tested a wide variety of measures for associations with treatment outcome in order to find factors that were indicative of treatment success. Methodologically, this has mainly been done using hypothesis testing based on significance thresholds of linear relationships between response measures and the potential predictors (Rost, Binder, et al., 2022; see Appendix A). These approaches have led to the identification of numerous indicators from many different subfields and modalities, such as clinical characteristics, neuroimaging markers, blood parameters, or genetic information. Corresponding findings have been described and summarized in detail in several reviews (e.g., Bennabi et al., 2015; Kraus et al., 2019) and meta-reviews (Perlman et al., 2019).

However, the mere identification of a statistically significant effect of a given variable on treatment outcome does not mean that this variable can also prospectively and accurately *predict* treatment outcome. None of the factors identified so far has been able to provide sufficient predictive value to be relied upon for clinical prediction and treatment decisions (Chekroud et al., 2021). As a result, there is a growing view that prediction rather than association is necessary for advancing personalized psychiatry (Bzdok et al., 2021). With the increasing availability of large and high-dimensional patient datasets as well as advances in computational power, more and more studies have used predictive multivariable modeling by applying supervised machine learning (ML) techniques (Chekroud et al., 2021). Compared to previous studies, these approaches do not focus on detecting the effects of one or a few predictors but on combining the (often small) effects of many variables to maximize their predictive accuracy and the generalizability of the resulting predictions. In this context, generalizability is routinely assessed and ensured through the use of cross-validation (CV) during model training and subsequent external validation, i.e., testing the model predictions on a new dataset (Dwyer et al., 2018).

Although ML approaches represent a major methodological advancement and continue to grow in popularity, no model has yet been adopted in clinical practice for predicting treatment success in MDD, let alone providing personalized treatment recommendations (Chekroud et al., 2021; Kraus et al., 2019). The main reason seems to be the lack of highly predictive and robust (bio)markers. Apart from some sociodemographic and clinical features as well as some promising results from pharmacogenetic testing (Skryabin et al., 2022), there are no measures that have reliably and repeatedly demonstrated high prognostic value (Rost, Binder, et al., 2022; see Appendix A). Accordingly, the accuracies of well-designed and validated prognostic models average only approximately 63% in determining if a patient will benefit from antidepressant medication or not (Sajjadian et al., 2021). Whether such moderate prediction accuracies are clinically useful is debatable, especially because net benefit analyses are rare, as are prospective validation studies that include comparisons with treatment as usual (TAU) or clinician judgements, for instance. A recent clinical trial, however, could not show that algorithm-guided treatment resulted in greater symptom reduction than TAU (Browning et al., 2021). Such findings question the clinical benefits and highlight the shortcomings of current prediction models of MDD treatment outcome.

1.2 Aims of the thesis

The aim of this thesis was to identify and address several of the abovementioned shortcomings. A first publication (paper III, Appendix A) screened and summarized the current literature regarding predictions of MDD treatment outcome and translations into clinical practice in form of a narrative review (Rost, Binder, et al., 2022). It focused on different

operationalizations and measures of treatment outcome as well as on existing prediction models and issues with their implementation into clinical decision support systems. From this, we derived the main research questions that we addressed in the subsequent two studies. The first study (paper I) investigated ways to design predictive models of MDD treatment response as parsimoniously as possible without losing predictive power (Rost, Brückl, et al., 2022). The second study (paper II) built on this approach by comparing predictions of different definitions of treatment outcome and different sets, i.e., modalities, of predictor variables. In addition, the key data-derived predictors were compared with the results of an online survey among clinicians in which they reported and rated their most important indicators of antidepressant efficacy. The aims, procedures and results of paper I and paper II are described and discussed in more detail below. They are then followed by an overarching discussion, including limitations of the presented work as well as future directions and concluding remarks.

1.3 Optimizing prediction model sparsity

Prediction models of treatment outcome in MDD are commonly created on patient cohort data coming from longitudinal studies, such as randomized controlled trials or observational studies. These datasets are often quite rich in available baseline measurements, i.e., potentially predictive variables that can enter a model (also called "features"). With decreasing costs of complex biological characterizations, e.g., omics data, the number of features is becoming even larger. At the same time, sample sizes remain limited and do usually not exceed more than a couple of hundred patients (Sajjadian et al., 2021). Having substantially more features than samples can lead to models that are too well fitted to the data on which they were trained and therefore do not generalize well to new data, a phenomenon known as "overfitting" (Hastie et al., 2009). Further, models that require many measurements can lead to poor data quality. While only a single blood sample may be needed to generate omics data, for instance, a large proportion of clinical information are measured using psychometric assessments and questionnaires. These ratings and questions require time and cognitive resources which is why the quality of responses can suffer over time (Bowling et al., 2021; Rolstad et al., 2011). The inclusion of many features in a predictive model may also increase costs for the users, i.e., clinical institutions, especially if costly measures, such as neuroimaging, are included. These factors emphasize that clinical prediction models, including those predicting MDD treatment outcome, should be as sparse as possible (Sanchez-Pinto et al., 2018). In practice, however, not even half of the existing studies included in a recent meta-analysis used any kind of feature selection method (Sajjadian et al., 2021).

To tackle this issue, in the first paper of this thesis, we created a novel supervised ML pipeline with a feature selection method nested inside a CV framework. We tested this

pipeline on longitudinal MDD patient data as well as on a simulated dataset. As the real-world clinical dataset, we used the Munich Antidepressant Response Signature (MARS) project, a multicenter naturalistic observational study of MDD inpatients treated with antidepressant medication (Hennings et al., 2009). We included patients with a diagnosis of a depressive episode or a recurrent depressive disorder, i.e., an F32 or F33 diagnosis according to the International Classification of Diseases (ICD-10; World Health Organization, 1992), resulting in a total sample of 1,022 patients. Clinical characteristics and sociodemographic information measured at baseline, i.e., within one week after study inclusion and admission to the hospital, served as features for the predictive modeling (113 in total). Treatment response after six weeks was taken as the outcome to predict, defined by a symptom reduction of at least 50% on the total score of the 17-item version of the Hamilton Rating Scale for Depression (HDRS; Hamilton, 1960). The second, simulated dataset was created with similar dimensions and contained 1,000 samples and 125 features.

For feature selection, we used cross-validated recursive feature elimination (RFE), a wrapper method that selects the best-performing subset of features by iteratively removing the least important one for the prediction. RFE represents an extensive search over the entire feature space that additionally considers correlations between the features (Kubat, 2017). With an outer second CV used for hyperparameter tuning, this ML pipeline detects the best combination of hyperparameters and feature set in a completely data-driven manner. We tested our pipeline on both datasets using three different classification algorithms (elastic net regularized logistic regression, random forests, and support vector classifiers). In all six cases, it resulted in sparser prediction models, i.e., models that required fewer features than a reference pipeline without RFE where feature selection was only achieved by intrinsic feature selection of the classifiers. At the same time, in five of the six cases, the sparser models led to equally or more accurate predictions on the test sample. In general, accuracies for the MARS dataset ranged from 61% to 71%, which was in the range of previous studies (Lee et al., 2018; Sajjadian et al., 2021). Clinical information, such as prior hospitalizations, duration of the depressive episode, family history, and symptom severity were found to be the most robust and informative predictors of treatment response.

The results of this study carry several implications for research and clinical practice. First, by allowing a purely data-driven optimization of both input feature space and hyperparameters in one nested CV framework, the introduced ML pipeline does not require any a priori feature selection or threshold determination on the researcher's part. Previous studies involving feature selection have pre-selected potential predictors either based on the existing literature (e.g., Iniesta et al., 2016) or on intrinsic or filter methods that require the specification of a cutoff value, e.g., keeping the 25 most predictive features for the final model (Chekroud et al., 2016). While these approaches are still valid alternatives,

the method we introduced circumvents their additional researcher degrees of freedom. Second, as mentioned above, sparser prediction models may facilitate implementation into clinical practice by saving time and measurement costs. And third, fewer measurements reduce stress and fatigue in patients and may consequently lead to better data quality (Bowling et al., 2021; Rolstad et al., 2011).

Nevertheless, the lack of translation of prediction models into established clinical tools is mainly caused by the absence of strong and robust (bio)markers of antidepressant outcome. This issue also persisted in our first study, in which predictive accuracies remained limited regardless of the number of features required by the model. Hence, in the second paper of this thesis, we aimed at increasing model performance by applying multimodal modeling and testing alternative target definitions of treatment outcome. In addition, we conducted an online survey in which clinicians reported the indicators they use to assess treatment success in patients early on. We investigated whether their responses differed from the data-derived predictors and whether they might be informative for future modeling approaches.

1.4 Refining treatment outcome and multimodal modeling

Prognostic models in MDD research typically aim to predict one of two common measures of treatment success (Lee et al., 2018; Sajjadian et al., 2021): treatment response, usually defined by at least 50% symptom reduction on a depressive symptom scale compared to baseline (as in our first study; Figure 1A), or remission, defined by a specific cut-off score on a depressive symptom scale after a certain duration of treatment (Figure 1B). These two consensus definitions have become widely accepted (Rost, Binder, et al., 2022; see Appendix A), and corresponding binary classification models (response vs. non-response and remission vs. no remission, respectively) are easy to evaluate and interpret. However, they both rely on arbitrary threshold values, and binarizing semi-quantitative measures, such as symptom scale sum scores, always leads to a loss of information (Altman & Royston, 2006; Dawson & Weiss, 2012). Therefore, several studies have tried to avoid these definitions and have instead focused on identifying data-driven subgroups of patients with similar response patterns (e.g., Hartmann et al., 2018; Kelley et al., 2018; Paul et al., 2019; Uher et al., 2011). These outcome classes have mainly been generated using unsupervised ML, i.e., clustering methods. They are empirically supported and less governed by conventions than response and remission (Figure 1C). On the other hand, the number and patterns of the resulting classes strongly depend on the input data and the chosen clustering technique, which raises concerns about their generalizability. Furthermore, regarding clinical applicability, the question remains whether data-driven outcome definitions prove beneficial, i.e., whether they are easier to predict than response and remission.

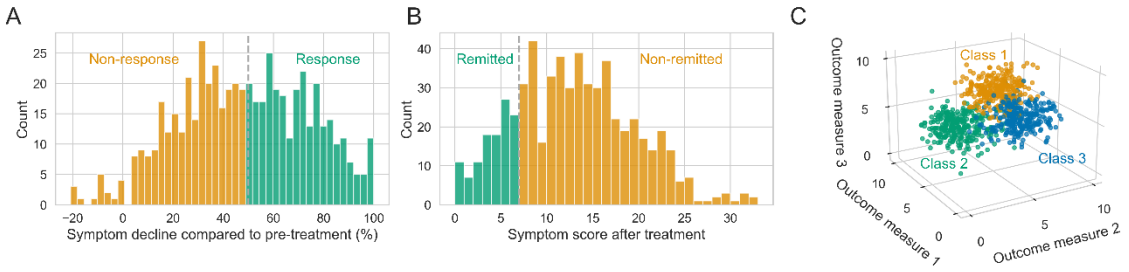


Figure 1. Different definitions of treatment outcome using simulated example data. The two histograms of (A) percentage change and (B) total score on a given symptom scale illustrate how cut-offs are used for the two common definitions of treatment response and remission, respectively. In contrast, (C) data-driven approaches aim to identify homogenous patient subgroups, i.e., outcome classes, based on selected outcome measures.

A second way to potentially improve predictive performance is through multimodal modeling. Instead of primarily using features from self-reports and clinician ratings, adding more objective and biological measures may be helpful (Chekroud et al., 2021). Multimodal modeling has already produced encouraging results, for instance in psychosis research (Koutsouleris et al., 2021), but has not yet been applied very often in MDD outcome prognosis. Recent studies could show that adding biological markers on top of clinical ratings and questionnaire data led to slight improvements in model accuracy (Dinga et al., 2018; Iniesta et al., 2018; Sajjadian et al., 2022). Given these findings, we investigated in the second study whether data-driven treatment outcome classes and features from further data modalities, such as biological information, could improve our predictions.

For this purpose, we used data from the MARS study again. Additionally, as an independent validation sample, data from a venlafaxine XR augmentation trial were taken. We used four different outcome variables, calculated from HDRS ratings at baseline and after 6 weeks of treatment, to generate treatment outcome classes. These variables were selected to include information on both relative changes in symptom severity and absolute symptom severity at week 6. Moreover, they did not only include HDRS sum scores but also a subscore of MDD core symptoms to identify potential deviations in their respective trajectories. To ensure that the resulting classes were not overly influenced by the chosen clustering technique, we applied two different algorithms. First, we used the Viscovery® SOMine® 7.2 software (Viscovery Software GmbH, 2021), which generated a two-dimensional representation of the input data based on self-organizing maps and subsequently applied Ward's hierarchical clustering to create clusters from these maps (Ward Jr., 1963). As a second method, we chose a consensus clustering approach on k-medoids partitions of the Gower distance matrix (Gower, 1971; Monti et al., 2003). Both methods aimed at forming patient subgroups that were as stable as possible. As a third data-driven outcome definition, we used the response classes by Paul et al. (2019), which were also

created on the MARS data, but via longitudinal clustering. Finally, response and remission, as defined above, were included as prediction targets.

For the multimodal modeling, we included all available baseline features from various data modalities with less than 50% missing values and divided them into biological and non-biological information. The non-biological feature set consisted of 136 features. The 65 biological features consisted of polygenic risk scores (PRS), physical parameters, and blood levels. These two feature sets as well as the complete set of all 201 features entered the prediction pipeline presented in the first paper to predict all of the aforementioned outcome definitions.

Our analyses led to two main results. First, with respect to the target outcomes, our clustering solutions did not lead to improved predictions, i.e., none of the treatment outcome classes were predicted more accurately than response or remission. In fact, the best-performing models across all three feature sets were predicting one of the latter two outcomes (response for non-biological and biological features, remission for the complete feature set). Second, similar to prior findings showing little incremental predictive value of biological information over clinical predictors, our prediction results revealed a comparable pattern. Models trained on the biological features did not perform consistently better than chance, with a few exceptions. Further, combining non-biological and biological features did not lead to better overall predictions than using non-biological features alone. Notably, model performances were generally moderate at best. The top-performing model achieved balanced accuracy scores of 63.9% in the MARS test sample, 59.5% in the MARS validation sample, and 56.9% in the external validation sample (venlafaxine XR trial). Thus, in our case, neither the incorporation of different data types nor variation in the targeted outcome measures could alleviate the challenges in predicting treatment outcome in MDD.

1.5 Comparing key predictors from the data to clinical expertise

As a next step, we wanted to compare the most predictive variables from the modeling results with indicators actually used by clinicians in their daily work. To date, only few studies have considered physicians' judgments of whether or not a patient will respond to treatment to benchmark the performance of a prediction algorithm against them (e.g., Koutsouleris et al., 2021). Even less frequently were clinicians additionally asked what experiences or patient characteristics they based their judgments on. This information, however, could be relevant for examining how evidenced-based clinician judgements are, i.e., how congruent they are with data-derived predictors. At the same time, it could also be interesting to know whether clinicians rely on information that is not typically assessed in clinical studies but might be useful for future predictive models.

To answer these questions, we performed an online survey among physicians and psychologists in five German psychiatric and psychosomatic clinics. Participants were first asked to indicate their professional experience. Afterwards, they reported in free text which indicators they use to estimate early on whether or not a patient will respond to antidepressant medication. Afterwards, they rated several selected patient characteristics that were available in the MARS data and included in the predictive modeling for their relevance, i.e., their predictive value for antidepressant treatment outcome.

The reports from 53 participants showed several congruencies with the results from the data. Treatment history, e.g., antidepressant non-response in previous episodes, and course of the disorder, e.g., quantity of previous episodes or duration of the disorder, were two of the most frequently mentioned and top-rated indicators, for instance. These factors were also highly influential in the predictive modeling, indicated by their high permutation importance for the best-performing model (predicting treatment response using only non-biological features). Furthermore, personality traits and environmental factors, such as childhood trauma and stressful life events, received high ratings by clinicians and had also been identified as predictive in prior studies (Nanni et al., 2012; Nelson et al., 2017; Takahashi et al., 2013; Williams et al., 2016). Unfortunately, in the MARS study, this information was not collected at baseline but at discharge and was therefore not included in our predictive analyses. Nonetheless, univariate analyses between outcome groups yielded several significant differences on personality traits and childhood trauma questionnaires, suggesting that they might indeed be able to contribute predictive value and should be included in future prediction models.

On the other hand, we found discrepancies between data-based predictors and clinician reports as well. In the free text answers, many participants mentioned patient attitude as an important indicator of treatment (non-)response. Concretely, they referred to a lack of trust in pharmacotherapy, a lack of openness to different forms of therapy, and a fixation on reporting side effects from the patient side. To our knowledge, these factors have not yet been measured or considered in any study of MDD outcome prediction, so it might be useful to examine their relevance in future investigations. Conversely, some of the most important predictors from the MARS data were underestimated by clinicians. These involved clinical variables in particular, such as symptom severity ratings, as well as sociodemographic information, such as marital or educational status. Since these characteristics have repeatedly been shown to be among the most reliable predictors of MDD treatment outcome (Chekroud et al., 2021; Rost, Binder, et al., 2022), clinicians should consider them when assessing a patient's likelihood of response and choosing the appropriate treatment.

1.6 Implications, limitations, and future directions

Overall, the results from our two studies confirmed that predicting treatment outcome in MDD patients remains an unresolved issue. Introducing data-derived outcome classes from symptom scales instead of relying on cut-off definitions, such as response and remission, does not seem to make this task any easier, neither do several multimodal modeling approaches. In addition, we contributed to the present state of knowledge by introducing a novel prediction pipeline that simultaneously optimizes accuracy and sparsity and by highlighting congruencies and divergences between model-based and clinician-rated predictors.

1.6.1 Measures of treatment outcome

With respect to treatment outcome definitions, the use of clustering algorithms and corresponding patient subgroups did not lead to better predictions than response and remission. These results challenge the usefulness of subgrouping in a prognostic context, especially since the concepts of response and remission are already well-known to clinicians and may be more intuitive to them. Decision support systems that predict response vs. non-response or remission vs. no remission may therefore require less training and less familiarization than systems that output a certain subgroup prediction (e.g., class X vs. all other classes). Particularly in cases where a clustering algorithm proposes many outcome classes, the respective symptom trajectories may become very similar and the clinical utility of distinguishing between them may decrease. For instance, if two classes showed similar trajectories of symptom severity over time, it is unlikely that physicians would treat patients from these groups differently. However, our investigation of treatment outcome classes was limited by the availability of longitudinal measures in MARS since the only weekly assessment was the HDRS. Therefore, we could only use this single symptom scale to inform the clustering. The reference outcome definitions, response and remission, were also derived from the HDRS which may have led to an assimilation of predictive accuracies. Even though the generated outcome classes were not based only on the HDRS total score, unlike response and remission, the common underlying measurement may have made strong differences in the prediction results unlikely. For future studies, the formation of data-driven outcome definitions may be of greater benefit if more than one outcome measure is available. When combining multiple measures, possibly even beyond MDD symptom ratings scales, clustering approaches may help to identify critical pathways of the disorder at several levels. This suggestion is consistent with a body of research proposing to broaden the definition of treatment outcome and to move away from mere symptom scales and corresponding total scores. Apart from valid concerns about their reliability (Bagby et al., 2004; Trajković et al., 2011) and comparability (Fried et al., 2022; Uher et al., 2008, 2012), a recent survey of patients, caregivers, and

clinicians on important aspects of treatment outcome revealed that domains of daily functioning matter as well (Chevance et al., 2020). Facets of elementary, social, and professional functioning are rarely measured in clinical trials but should be considered in future studies. Current concepts and challenges of MDD treatment outcome have been summarized and further discussed in our review paper (Rost, Binder, et al., 2022; see Appendix A).

1.6.2 Multimodal modeling

Regarding multimodal modeling, our investigations have led to similar results as previous studies showing limited additional predictive value of potential biomarkers (Dinga et al., 2018; Iniesta et al., 2018; Sajjadian et al., 2022). Adding PRS, physical parameters, and blood levels to basic clinical and sociodemographic features did not improve model performance overall. However, the predictive power of these measures was most likely limited by varying amounts of missing data. While PRS were available for almost all patients, other measures, such as waist circumference or cortisol levels, for instance, had to be imputed for many patients. Data from further modalities, e.g., DNA methylation, structural magnetic resonance imaging, or neuropsychological testing, were only available for smaller subsets of patients and were therefore not included in the predictive analyses at all. This limiting factor restricts the conclusiveness of our results regarding the lack of biological indicators. Still, even though the genetic data were almost complete, we could not confirm prior results showing that PRS can be successfully used to distinguish between different courses of MDD (Schultebraucks et al., 2021). If some PRS had carried predictive information in our data, models based on the biological features would have performed better than observed. Our results therefore suggest that the added value of multimodal modeling needs to be further investigated in future studies. Ideally, datasets should be large enough to handle high dimensionality and complete enough to include as many modalities as possible without critical amounts of missing values (Chekroud et al., 2021). Exploring novel types of data could additionally provide new insights. There may be some potential in the use of "digital phenotyping", i.e., data collected from sensors and digital devices, such as smartphones or smartwatches, for instance (Torous et al., 2016). A major advantage of these technologies is that they can collect large amounts of data at high resolution without the need of active patient engagement (Durstewitz et al., 2019). Promising initial results support the consideration of such data types in future research (Jacobson et al., 2019; Zarate et al., 2022). Advanced predictive methods, such as model stacking (Wolpert, 1992), could then be applied to additionally boost predictive performance of multimodal data (Dwyer et al., 2018).

1.6.3 Prediction model sparsity

With increasing availability of different types of data and thus increasing dimensionality, feature selection becomes more and more important in predictive modeling to avoid overfitting (Hastie et al., 2009). The results from our first paper highlight that feature selection methods should indeed be used during development of prediction models of MDD treatment outcome. Each feature should only be included as a predictor if its additional predictive value justifies the additional costs. This trade-off may thus vary, e.g., between an additional questionnaire that takes patients only a few minutes to complete and a biological measure that requires trained staff, laboratory equipment, and/or computational capacities. Even though the pipeline presented in our first study is not able to consider the costs of individual features, it can help to automatically select the best-performing feature set for the chosen classifier. By implementing it into future healthcare prediction models, it might thus reduce both stress for patients and costs for the user. Moreover, its utility is not restricted to datasets from MDD patients specifically, but can be transferred to other classification problems based on datasets with similar dimensions.

1.6.4 Clinical expertise

Comparisons between modeling results and reports from our online survey revealed both congruence and divergence on key predictors of treatment outcome. On the one hand, this strengthens the importance of some characteristics, such as course of the disorder and success (or failure) of prior treatment trials. On the other hand, our findings also provide new insights into which other factors clinicians consult for their predictions and how data acquisition and clinical expertise can inform each other in the future. Consequently, the top indicators from our clinical survey should be considered and evaluated in future studies to investigate their empirical validity and quantify their predictive value.

Clinician judgements may also be relevant for another reason. From a translational perspective, clinical prediction tools may be particularly useful when they outperform physicians in their daily tasks and lead to better outcomes than physician-guided treatment or TAU (Cohen & DeRubeis, 2018). Current research, however, suggests that combining clinical expertise and ML models may be the best way forward for future medicine (Gennatas et al., 2020; Topol, 2019). This has not only already led to improvements in psychosis prediction (Koutsouleris et al., 2021), but may additionally facilitate implementation and acceptance in clinical practice (Kilsdonk et al., 2017). Corresponding studies on MDD outcome prognosis are still to come but may provide further evidence.

1.7 Conclusion and outlook

In this thesis, we were able to contribute to ongoing attempts of predicting antidepressant treatment outcome with ML methods. We first introduced a classification pipeline that

focuses on model sparsity to build predictive models that are cost-efficient without losing performance. We were also able to identify factors that clinicians use to predict treatment outcome and illustrated how these overlapped with and differed from data-driven predictors. However, similar to most previous studies, our prediction results were largely within the range of moderate accuracies (Sajjadian et al., 2021), and the question whether a given patient with MDD will profit from antidepressant medication or not remains difficult to answer. Examining different outcome measures and multimodal modeling have not yet been able to overcome this challenge, including our studies. Are ML approaches to predicting treatment outcome in MDD thus more hype than hope?

ML has undoubtedly been in vogue for several years now and has contributed substantially to the identification and validation of various predictors. However, predictive performance is currently stuck somewhere in mediocrity and a major breakthrough for clinical application is still awaited. With a disorder as heterogeneous as MDD (Fried, 2017; Fried & Nesse, 2015), it seems increasingly unlikely to find a universal predictive model that works adequately well for all patients. Additional hope may therefore come from biological subtyping. Since the diagnosis of MDD itself may in fact be an agglomeration of many different pathophysiologies (Olbert et al., 2014), identifying homogenous subgroups of patients based on these pathomechanisms could lead to more specific, targeted and personalized treatment (Brückl et al., 2020). Subsequently, stratified prognostic models could recommend specific drug options tailored to a patient's individual biotype. This may not only apply to MDD but to other psychiatric disorders as well, paving the way for precision psychiatry.

2. Paper I

Rost, N., Brückl, T. M., Koutsouleris, N., Binder, E. B., & Müller-Myhsok, B. (2022). Creating sparser prediction models of treatment outcome in depression: a proof-of-concept study using simultaneous feature selection and hyperparameter tuning. *BMC Medical Informatics and Decision Making*, *22*(1), 181. https://doi.org/10.1186/s12911-022-01926-2

BMC Medical Informatics and Decision Making

RESEARCH

Open Access

Creating sparser prediction models of treatment outcome in depression: a proof-of-concept study using simultaneous feature selection and hyperparameter tuning

Nicolas Rost[1,2*], Tanja M. Brückl[1], Nikolaos Koutsouleris[3,4,5], Elisabeth B. Binder[1] and Bertram Müller-Myhsok[1,6]

Abstract

Background: Predicting treatment outcome in major depressive disorder (MDD) remains an essential challenge for precision psychiatry. Clinical prediction models (CPMs) based on supervised machine learning have been a promising approach for this endeavor. However, only few CPMs have focused on model sparsity even though sparser models might facilitate the translation into clinical practice and lower the expenses of their application.

Methods: In this study, we developed a predictive modeling pipeline that combines hyperparameter tuning and recursive feature elimination in a nested cross-validation framework. We applied this pipeline to a real-world clinical data set on MDD treatment response and to a second simulated data set using three different classification algorithms. Performance was evaluated by permutation testing and comparison to a reference pipeline without nested feature selection.

Results: Across all models, the proposed pipeline led to sparser CPMs compared to the reference pipeline. Except for one comparison, the proposed pipeline resulted in equally or more accurate predictions. For MDD treatment response, balanced accuracy scores ranged between 61 and 71% when models were applied to hold-out validation data.

Conclusions: The resulting models might be particularly interesting for clinical applications as they could reduce expenses for clinical institutions and stress for patients.

Keywords: Major depressive disorder, Treatment outcome, Predictive modeling, Feature selection, Precision psychiatry, Supervised learning

Background

Despite many efforts in psychiatric research, the question of which patient will respond to which treatment is still unanswered. Specifically for very heterogenous disorders, such as major depressive disorder (MDD), no reliable (bio-)markers have been uncovered yet and no validated tests are available that could match a patient to the treatment they would benefit from the most [1, 2]. Predicting how well patients will respond to medication in general would be an important improvement for psychiatric health care and a further step towards precision medicine in psychiatry. Given the complex pathogenesis of psychiatric disorders, including MDD, it is unlikely that a few single indicators will be sufficient to forecast a patient's response to pharmacotherapy. Rather, it will

*Correspondence: nicolas_rost@psych.mpg.de

[1] Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Kraepelinstraße 2-10, 80804 Munich, Germany
Full list of author information is available at the end of the article

be important to collect a variety of measurements and gather information from many potentially informative data modalities [2].

The need to combine information from many different sources is why prognostic multivariate clinical prediction models (CPMs) might be particularly important in psychiatry. CPMs, and precision psychiatry in general, are fueled by data: the more features (in terms of measured patient characteristics) are available, the higher the chances of finding predictive variables. And the more samples are available, the higher the chances to obtain robust and generalizable models. Most prediction models, including those targeting treatment outcome in MDD, use supervised machine leaning techniques in order to maximize predictive power and generalizability at the same time [3]. However, when there are more features than samples in the data, the risk of overfitting the model increases and its generalizability decreases. This is often the case for data sets from patient cohorts, especially when high-dimensional biological data, such as (epi-)genetics and brain imaging, are included [4].

With the increasing availability of large data sets and simultaneous advances in bioinformatics and computational power, several multivariate prognostic models for predicting treatment outcome have been developed. We will use research on MDD and treatment with antidepressant medication as an example here. In general, however, CPMs are relevant for any condition in which there is a need to combine a multitude of predictors because no sufficiently predictive single factors have been identified so far [5].

Chekroud et al. [6] used data from the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study [7] in order to train a supervised machine learning model that was able to predict patients' responses to the selective serotonin reuptake inhibitor escitalopram across different clinical trials with accuracies of 60–65%. Before training the model, they reduced the set of predictors by applying an elastic net regularized logistic regression [8] and kept the 25 most predictive variables (out of 164 initial variables). Dinga et al. [9] created a CPM of MDD long-term outcome based on observational data from the Netherlands Study of Depression and Anxiety [10]. The model was trained on different data modalities and included feature selection via elastic net regularization as well. It was able to differentiate between 3 patient groups (remission, improving, and chronic) with balanced accuracies of 60–66%. While these studies identified the most predictive variables using an entirely data-driven approach, i.e. via regularization techniques, other studies selected their variables a priori based on findings from previous research. Iniesta et al. [11], for instance, entered into their predictive models only demographic and clinical information that had been associated with treatment outcome in prior studies. They tested four different combinations of predictors, from a comparably sparse set of 60 variables up to 125, in order to evaluate the additional value of certain subgroups of variables. The best performing model predicted response to escitalopram with an area under the receiver operating characteristics curve of 0.75. Similarly, Athreya et al. [12] focused on previously identified factors in form of pharmacogenetic markers from genome-wide association studies. In combination with depression symptom scores, these markers predicted treatment response with accuracies between 71% and 86%. When applied to validation data sets, however, the model performances decreased below statistical significance. Further prediction models of MDD treatment outcome have been summarized in systematic reviews and meta-analyses [13, 14].

In general, CPMs are aimed at being translated and applied in clinical settings. They should be based on patient data that physicians can easily assess during their daily routine and should not require a lot of additional time and costs [15]. Consequently, the input data the model needs to make a prediction should be as sparse and cost-effective as possible [16]. If two models perform equally well, the simpler model should be preferred and will also be more likely to succeed as a clinical application, especially when the more complex model requires expensive additional measures. However, the majority of CPMs have either been constructed on a fixed, a priori selected feature set [6, 11, 12, 17], or included feature selection only in form of intrinsic regularization techniques [9]. None of the applied methods have used any further feature selection technique incorporated into the training process in order to develop sparser models. While regularization can effectively remove uninformative features from the final model, it cannot guarantee that an alternative model built on even less features would not perform equally well or even better when applied to new data. Hence, it might be beneficial to include an additional data-driven feature selection into the optimization framework in order to not just tune the model's hyperparameters but also the required input feature set.

Different feature selection methods exist that can be implemented into a predictive modeling pipeline. In general, apart from the abovementioned intrinsic feature selection, e.g., by adding regularization terms to a regression model, the two main selection methods are filters and wrappers [18]. Filter approaches use the relationship between features and target for selection by ranking features according to the strength of their association with the target variable. The top N features, where N is usually defined by a certain cut-off, are then retained for the

predictive modeling while the remaining features are discarded. A disadvantage of this technique is that relations between the features are not considered. Wrapper approaches, on the other hand, use searching techniques to find the most informative set of features. They create many different subsets of the input features and then select on the best performing subset according to a performance metric. These approaches can be more comprehensive, but also more computationally expensive [18]. Apart from feature selection methods, other techniques for dimensionality reduction exist, often including feature transformation, such as principal component analysis or multidimensional scaling. An overview over feature reduction methods for supervised learning problems is presented in Table 1.

In this study, we compared a standard predictive modeling pipeline, that is, a repeated cross validation (CV) framework, to the same pipeline with an additional wrapper method for feature selection, i.e., recursive feature elimination (RFE) nested within the CV. We investigated three commonly used classifiers applied to two different data sets: one real-world data set from an observational inpatient study on patients with MDD as well as one simulated data set with similar dimensions. Our research questions were threefold: First, does the combined hyperparameter tuning and feature selection approach lead to models with sparser feature sets than intrinsic feature selection alone? Second, are classification accuracies between the two pipelines comparable or does the additional feature selection lead to changes in model performance? Third, does permutation testing lead to accuracies around chance level and can thus confirm that there is no information leakage biasing the results?

Material & methods

Data sets

Two different data sets were included in our analyses. First, as a real-world clinical data set, we used data from the Munich Antidepressant Response Signature (MARS) project [19], a multicenter naturalistic inpatient study, in which patients diagnosed with a single depressive episode, recurrent depressive disorder, or bipolar disorder were observed during their hospitalization. Further information on the study protocol and exclusion criteria have been published elsewhere [19]. The MARS study was approved by the ethics committee of the Ludwig Maximilian University in Munich, Germany, and conducted according to the Declaration of Helsinki. For our analyses, clinical response after 6 weeks of treatment, defined by at least 50% symptom reduction on the 17-item Hamilton Rating Scale for Depression (HDRS-17) [20], was used as a binary target variable for the CPMs. Patient characteristics measured at baseline, i.e., within the first week after study inclusion, were eligible as features for the predictions. We limited the analysis to unipolar depression and excluded patients diagnosed with bipolar disorder as well as patients without HDRS-17 scores at week 6 and patients with at least 75% missing values across all baseline features. Data from the resulting 1022 patients were then randomly split into a training (80%, 817 patients) and validation set (20%, 205 patients). From initially 548 baseline features, we removed those with at

Table 1 Common feature reduction approaches for supervised machine learning

Method	Description	Examples	Evaluation
Feature selection			
Intrinsic/embedded methods	Feature selection is implemented into the learning algorithm and performed during training	Regularized regression models Decision trees	Computationally efficient Interconnected with learning algorithm No guarantee of optimal sparsity
Filter methods	Feature selection based on associations with target variable	Associations are calculated using, e.g., correlations or ANOVA; top N features (or N%) are retained for training	Computationally efficient Relations between features ignored Independent of learning algorithm
Wrapper methods	Selection of best performing subset of features	Recursive feature elimination Sequential forward selection	Extensive search over input feature space Interconnected with learning algorithm Consider relations between features Computationally expensive
Feature transformation			
Projection into lower-dimensional feature space	Data are transformed and new features are created	Principal component analysis Multidimensional scaling Matrix factorization	Further methods of dimensionality reduction Alternative approaches to feature selection

ANOVA, analysis of variance

least 30% missing values as well as strongly imbalanced binary variables (ratio of 95:5% or more extreme), resulting in a final number of 113 features. The final feature set included sociodemographic data as well as information on psychiatric symptom profiles, symptom severity, family history, history of MDD, and medication. An overview over all included clinical features is presented in Additional file 1: Table S1. A flow diagram of all preprocessing steps that led to the final sample and feature selection is depicted in Additional file 1: Fig. S1.

The second data set consisted of simulated data with similar characteristics. Using Python's *scikit-learn* package, we generated 1000 samples with 2 target classes and 125 features, consisting of 25 informative, 50 redundant, and 50 uninformative variables. Similar to the clinical data, the samples were randomly split into 800 training and 200 validation samples.

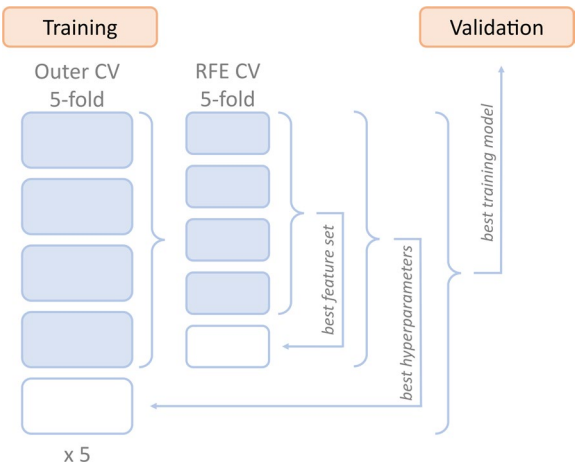


Fig. 1 Supervised machine learning pipeline based on repeated nested cross-validation combining hyperparameter tuning and feature elimination. RFE, recursive feature elimination

Predictive modeling pipelines

All analyses were performed in Python (version 3.8.5) using the *scikit-learn* package (version 0.23.1) [21] and additional custom functions. The predictive modeling consisted of three different methods: (1) the proposed repeated nested CV with a simultaneous optimization of hyperparameters and best performing feature set; (2) a reference pipeline without the nested feature selection method; (3) 100 runs of the complete proposed pipeline from method (1) but with randomly permuted target variables. The proposed nested CV pipeline is additionally illustrated in Fig. 1. It entails a repeated (5 times) nested 5-by-5-fold CV, where the outer CV is used for hyperparameter tuning and the inner CV is used for RFE, implemented with *scikit-learn*'s *RFECV()* function. The goal of RFE is to select features by iteratively testing smaller feature sets. Initially, the model is trained on the entire feature set and the importance of each feature is extracted. Then, in a stepwise process, the feature with the lowest predictive power is gradually removed from the feature set until the best performing set of features is found. In our approach, the performance of the model is evaluated on a test set using CV. Therefore, in this framework, feature selection could happen both intrinsically, e.g., by the tuning of regularizing hyperparameters, and by the RFE. The final model was then defined by the on average best performing combination of hyperparameters and feature sets across all test folds. The second method was included as a reference to represent a common supervised machine learning pipeline. It consisted of a repeated (5 times) 5-fold CV used for hyperparameter tuning. Hence, it was identical to the proposed pipeline except for the nested RFE, and feature selection was only possible through intrinsic selection. The final model was defined by the on average best performing combination of hyperparameters across all folds. The third method was included as a permutation test for the proposed first pipeline in order to rule out the possibility of information leakage. It consisted of 100 runs of the complete nested CV pipeline but with randomly permuted target variables.

All three methods were applied to the two data sets using three different types of classifiers: an elastic-net regularized logistic regression (LR), a random forest classifier (RF), and a linear support vector classifier (SVC). Elastic-net regularized LR combines two different kinds of penalties (L1 or Lasso and L2 or Ridge) on the model which are commonly used to reduce complexity when the number of features is large [22]. This way, the risk of overfitting can be reduced by shrinking the feature coefficients and reducing multicollinearity. The ratio between the two penalties is usually tuned as a hyperparameter. RF is an ensemble learner that uses the results of a large number of decision trees to make the best possible classification. Single decision trees are uncorrelated and make individual decisions on its own. From the set of individual decisions, the RF provides a final decision [23]. Linear SVCs try to find optimal separation lines between the samples of different classes that can then be used to assign new samples to the correct class. These decision boundaries are chosen to maximize the distance between the data points of the classes so that future data points can be classified with the greatest possible confidence [24]. The three classifiers were selected in order to cover linear (all three classifiers) and non-linear (RF) associations of the features with the target variable and because they provide measures of importance (coefficients/weights) for each feature. Furthermore, they have

frequently been used for various CPMs in psychiatry [6, 11, 25, 26]. Additional data preprocessing included k-nearest neighbors imputation of missing values [27] for all three classifiers and feature standardization for LR and SVC classification. Both steps were embedded into the (nested) CV, i.e., were created on the training folds and applied to the corresponding test fold of the CV loop. Hyperparameter tuning during model fitting was performed using Bayesian optimization [28]. After training, the resulting models were applied to the validation data set in order to get a final performance estimate. Crucially, the validation data set was completely left out of the training process and its CV loops. Such external validation on a hold-out data set is necessary to assess model performance independently of the training data on 'new' and 'unseen' data. Performance was primarily measured by Matthews correlation coefficient (MCC) [29] and the balanced accuracy score (BAC) [30]. Additionally, we extracted receiver operating characteristic curves and confusion matrices of all non-permuted classifiers. Since the MCC is a special form of the Pearson correlation coefficient, a value of 0 corresponds to chance level. For BAC scores, the chance level of a binary classifier is 0.5. MCC values from the permuted models across both data sets and all three classifiers were tested against their theoretical null distribution, that is, a t-distribution with n-2 degrees of freedom [31], using Kolmogorov–Smirnov tests. Statistical significance of the non-permuted models was tested using *p*-values derived from the same distribution. To compare the models with RFE to the models without RFE, we performed pairwise tests on the respective MCC values [32]. Further, for the non-permuted models, the importance of each feature was calculated by its permutation importance on the validation data, that is, by the average decrease in model performance when the feature was randomly permuted. The number of permutations for this procedure was set to 25.

Availability of data and materials

Data from the MARS study as well as the corresponding preprocessed data set that was used for the analyses can be requested by contacting Dr. Tanja Brückl (brueckl@psych.mpg.de). The TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) [33] checklist for the present study is presented in Additional file 1: Table S2. Analysis scripts are available at https://doi.org/10.5281/zenodo.6759730.

Results

In the clinical data set, 564 out of 1022 patients (55.19%) showed a clinical response, defined by at least 50% symptom reduction measured with the HRDS-17 sum score after 6 weeks of antidepressant treatment, whereas 458 patients (44.81%) did not respond. Hence, the outcome groups were slightly unequally large which is why the classifiers' class weights were balanced. Demographic data and basic clinical information for training and validation set are presented in Table 2. In the simulated data set, the outcome groups were created to be balanced with 500 samples in group 1 and 500 samples in group 2.

Table 2 Basic patient characteristics of the clinical data set (MARS study)

	Training data (N = 817)	Validation data (N = 205)	Overall (N = 1,022)	p
Gender				
Female	431 (52.8%)	105 (51.2%)	536 (52.4%)	0.753
Male	386 (47.2%)	100 (48.8%)	486 (47.6%)	
Age				
Mean (SD)	47.4 (14.0)	47.1 (14.4)	47.3 (14.1)	0.790
[Min, Max]	[18.0, 85.0]	[18.0, 87.0]	[18.0, 87.0]	
Diagnosis (ICD-10)				
F32	289 (35.4%)	61 (29.8%)	350 (34.2%)	0.152
F33	528 (64.6%)	144 (70.2%)	672 (65.8%)	
HDRS-17 baseline sum score				
Mean (SD)	24.0 (5.6)	23.4 (5.5)	23.8 (5.6)	0.185
[Min, Max]	[12.0, 40.0]	[10.0, 39.0]	[10.0, 40.0]	
Missing	11 (1.3%)	4 (2.0%)	15 (1.5%)	
HDRS-17 response				
Yes	454 (55.6%)	110 (53.7%)	564 (55.2%)	0.679
No	363 (44.4%)	95 (46.3%)	458 (44.8%)	

Two sample t-tests were computed for continuous variables, Chi-squared tests were used for categorical variables to compare training and test data set

HDRS-17, 17-item version of the hamilton rating scale for depression; *ICD-10*, international classification of diseases [34]

Model performances

Classification performances of the non-permuted models (with and without RFE) for the clinical data ranged from MCC values of 0.22 up to 0.43 (BAC scores: 0.61–0.71). For the simulated data, MCCs between 0.69 and 0.72 were observed (BAC scores: 0.84–0.86). Figure 2 shows the MCCs of the validation data for all computed models (for corresponding BAC scores, see Additional file 1: Fig. S2). Model performances of the non-permuted models are represented by vertical bars. Results from the 100 permutations are indicated by histograms, superimposed density curves and the respective average performance. Across all six comparisons, performances of the modeling pipeline with RFE and the pipeline without RFE were relatively similar. No significant differences were observed between the two pipelines (see Table 3). Interestingly, in four of the six cases, the models with RFE loop resulted in better predictions on the hold-out validation set than the models without RFE (all three classifiers on clinical data and SVC on simulated data). In one of the cases (LR on simulated data), MCCs and BAC scores were equal up to the second decimal place, and in one case (RF on simulated data), the model without RFE was superior. All non-permuted models both with and without RFE performed significantly better than chance, indicated by the p-values of the MCCs (all $p < 0.01$, see Additional file 1: Table S3). To further characterize the modeling results, we included the receiver operating characteristic curves and the corresponding areas under the curves in Additional file 1: Fig. S3. Confusion matrices and additional performance metrics, such as sensitivity and specificity

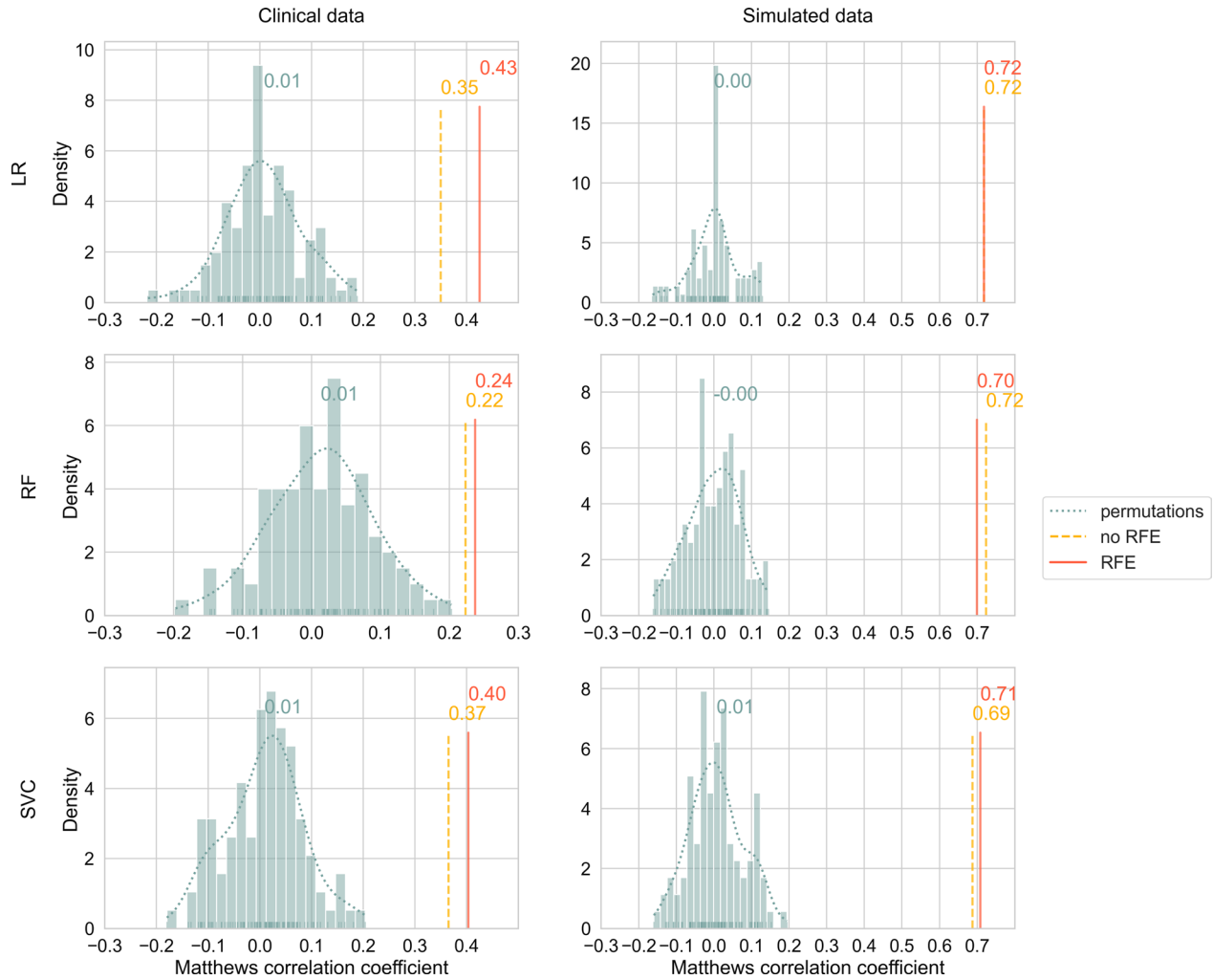


Fig. 2 Model performances for the three classifiers and the two data sets on the validation data. Matthews correlation coefficients are shown for the 100 permutations (annotations correspond to the respective means) as well as for the models with and without RFE. LR, logistic regression; RF, random forest classifier; RFE, recursive feature elimination; SVC, support vector classifier

Table 3 Pairwise statistical significance tests between model performances (MCC values) of the models with and without RFE on the validation data

	MCC		z	*p*
	RFE	No RFE		
Clinical data (N = 205)				
LR	0.425	0.350	0.888	0.375
RF	0.237	0.224	0.138	0.890
SVC	0.403	0.365	0.448	0.654
Simulated data (N = 200)				
LR	0.718	0.719	− 0.021	0.984
RF	0.700	0.724	− 0.483	0.629
SVC	0.709	0.688	0.407	0.684

LR, logistic regression; *MCC*, Matthews correlation coefficient; *RF*, random forest classifier; *RFE*, recursive feature elimination; *SVC*, support vector classifier

of the classifiers, are represented in Additional file 1: Table S4.

When the target class labels in the RFE pipeline were randomly permuted 100 times, the resulting performance metrics became distributed around their chance levels as expected (0 for MCC and 0.5 for BAC, respectively). For MCC values, Kolmogorov–Smirnov tests showed no significant deviations from the theoretical null distribution (all $p > 0.05$, see Additional file 1: Table S5). These results suggested no unintended information leakage from training to validation data. Quantile–quantile plots of empirical and theoretical MCC distributions are presented in Additional file 1: Fig. S4. None of the permutation runs led to better model performances than the corresponding non-permuted models (see Fig. 2).

Number of selected features

Overall, the RFE models resulted in sparser features sets than the models without RFE. Figure 3 shows the final numbers of features required by the models after intrinsic feature selection and selection via RFE. Across all six comparisons, the final models from the nested CV pipeline with RFE required less features than the equivalent models from the single CV pipeline without RFE. While for RFs, the RFE pipeline resulted in models requiring 76 and 96 features for the clinical and the simulated data, respectively, the models without RFE yielded 97 and 108 features with non-zero coefficients. Even stronger differences were obtained from the LR classifiers with differences of 50 features (clinical data) and 33 features (simulated data), and from the SVC models with differences of 31 features (clinical data) and 112 features (simulated data). Note that the pipeline without RFE could still lead to non-zero feature coefficients via intrinsic feature selection.

Figure 4 provides a combined overview over the main results by simultaneously depicting model performances (indicated by MCC on the y-axis) and numbers

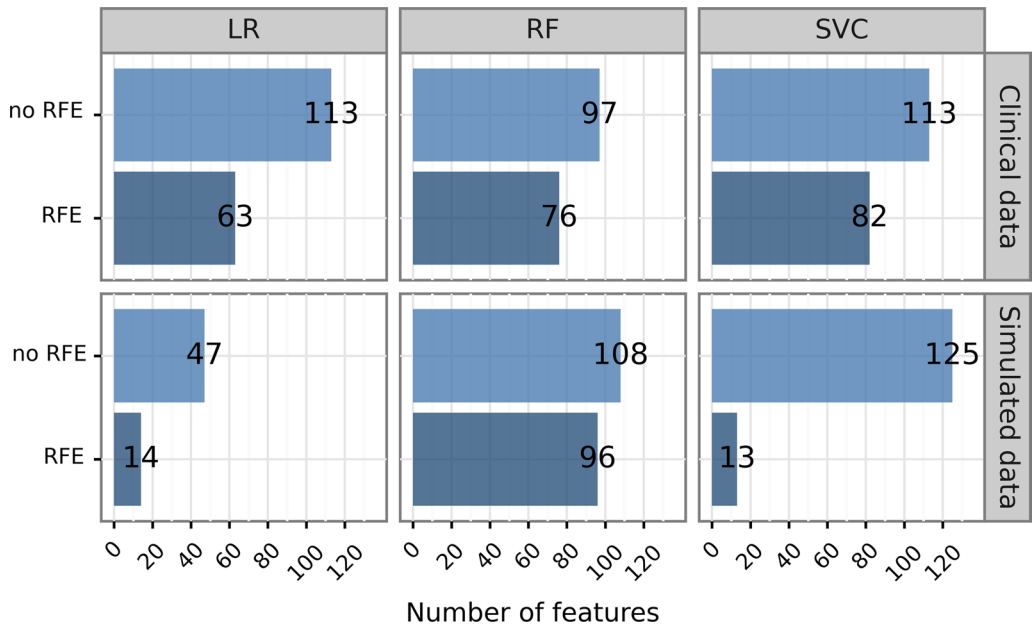


Fig. 3 Number of selected features for the non-permuted models. Across both data sets and all three classifiers, the nested cross-validation pipeline with RFE (lower rows) resulted in sparser models requiring less features than the reference method without RFE (upper rows). LR, logistic regression; RF, random forest classifier; RFE, recursive feature elimination; SVC, support vector classifier

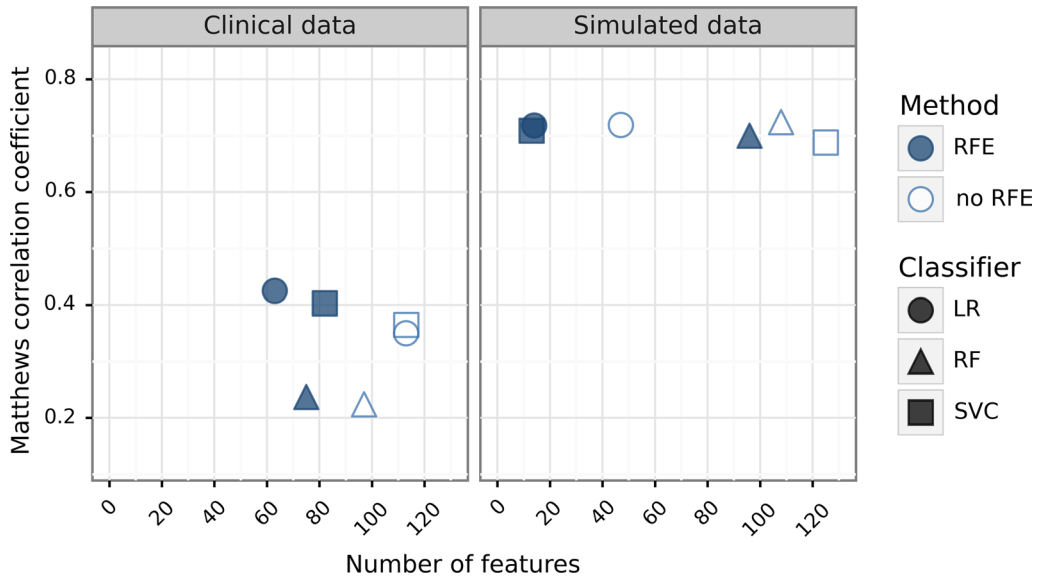


Fig. 4 Number of selected features for the non-permuted models plotted against the corresponding model performances. The left plot represents the values for the clinical data set, the right plot for the simulated data set. The different shapes indicate the three classifiers. LR, logistic regression; RF, random forest classifier; RFE, recursive feature elimination; SVC, support vector classifier

of selected features (on the x-axis) of all non-permuted models. Overall, the nested CV pipeline with RFE seemed to outperform the reference pipeline without RFE as it resulted on average in better performing models while also requiring less input features.

Clinical predictors of MDD treatment response

For the clinical data set, we were additionally interested in the most important predictors of MDD treatment response. Therefore, the permutation importance for each feature in each model was calculated using 25 permutations applied to the validation data set. The most informative features and their corresponding importance values, sorted by their importance (averaged over the three classifiers and the two pipelines), are illustrated in Fig. 5. The most informative features included information on the course of the disorder (e.g., number of prior hospitalizations, time since last hospitalization, duration of current episode), family history (of psychiatric disorders and MDD specifically) as well as symptom profiles and severity (e.g., various item scores from the HDRS and the Symptom Check-List-90-R [SCL90-R]) [35]. While several features showed rather consistent importance values (e.g., number of prior hospitalizations, nonviolent suicide attempts in medical history, psychiatric family history), regardless of which classifier or which pipeline was applied, other features varied in their permutation importance depending on the model that was used (e.g., preexisting dysthymia, SCL-90-R phobic anxiety, HDRS-17: total score). Note that negative importance values indicate that a feature was non-informative for a model but shuffling this feature led to a better model performance by chance.

A complete overview over the importance values of all features in alphabetical order is included in Additional file 1: Fig. S5. A more detailed description of the complete clinical feature set in is given in Additional file 1: Table S1. Corresponding feature importance values for the simulated data set are presented in Additional file 1: Fig. S6 (top predictors sorted by importance) and Additional file 1: Fig. S7 (complete feature set).

Discussion

In the present study, we tested whether a supervised machine learning pipeline that combined hyperparameter tuning and RFE in a repeated nested CV setup can lead to sparser but similarly accurate binary classification models than a default pipeline with only one CV loop for hyperparameter tuning. For this investigation, we used three different kinds of classification algorithms applied to two different data sets, one real-world data set on MDD treatment outcome and one simulated data set with similar dimensions. Our results showed that (1) the additional RFE loop led to sparser models that required less features for the classification; (2) although not statistically significant, the pipeline with RFE yielded equally well or better performing models on the validation data set in five of six cases; and (3) permutation tests suggested no unintended information leakage in the pipeline with RFE. Furthermore, all non-permuted models

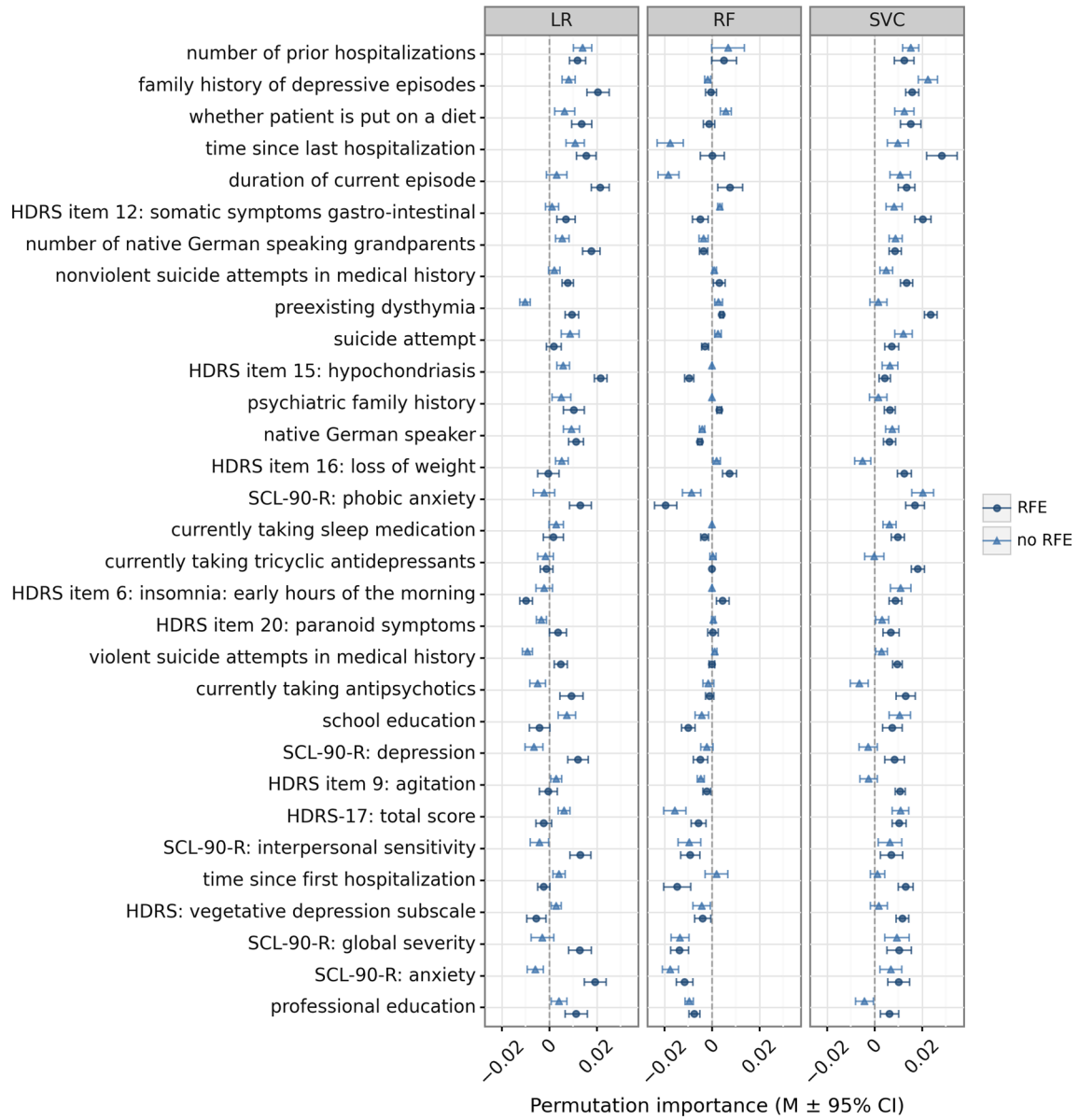


Fig. 5 Permutation importance from 25 permutations for the most informative clinical features, grouped by classifier and models with and without RFE. Only features that were selected by all 6 clinical models and showed a positive mean importance score (averaged over all 6 models) are presented. The scores show the average decrease in model performance on the test data when a feature was randomly permuted. Error bars represent 95% confidence intervals. LR, logistic regression; RF, random forest classifier; RFE, recursive feature elimination; SVC, support vector classifier

performed significantly better than chance, indicated by p-values < 0.01.

The results from the present study might be particularly relevant for classification tasks in clinical research. Clinical patient data sets are often based on comparably expensive measurements and sparser models requiring less features might not only decrease costs for clinical institutions but also stress for patients. Especially when expensive biological measures (e.g., brain imaging, -omics data) that need a lot of laboratory or computational capacities are included in data sets, it might be important to be rather strict on the inclusion of features into a predictive model. Measures that are not contributing strongly to the prediction should be omitted when there is a sparser model performing equally well or even better [16]. By using the pipeline proposed here, feature

selection, hyperparameter tuning and model fitting can be performed in one nested data-driven optimization process. Hence, this approach does not require any prior theory-driven feature selection but automatically selects the best performing feature set for each of the tested combinations of hyperparameters. Measurement time and costs can be reduced when applying such a reduced, sparser model in clinical practice. Sparser models also help to increase data quality because patients have to fill in less questionnaires which reduces respondent fatigue. In our analyses, the additional RFE loop reduced the number of features required by the final model by 12 features in the least extreme case (RF on simulated data) and 112 features in the most extreme case (SVC on simulated data). With respect to the MDD data set, features containing information on the patient's marital status, their gender, the origin of their grandparents and specific medication, for instance, were removed by the RFE across all three classifiers but had mostly non-zero feature coefficients in models created by the pipeline without RFE. By omitting these features, future applications of the model would require less information from the patients and could thus save time and efforts. While we have focused on RFE as a feature selection technique here, other filter or wrapper approaches might be similarly appropriate in general. In previous studies, different filter techniques have been successfully used for spam detection [36, 37], for instance, but have also been applied to biological human data [38, 39].

With respect to absolute performance of the predictive models, the observed performance values for the clinical data were within the expected range. The obtained MCCs of 0.22–0.43 and BAC scores of 0.61–0.71 were comparable to results from similar prior studies [13, 14]. Such classification accuracies of approximately 60–70% are far from ideal but might still be clinically relevant [40] and could provide support for clinicians in their treatment decisions. Our results underline that predicting antidepressant treatment outcome is a difficult and still unsolved endeavor, especially when the data set is as heterogenous as in our case. Since the MARS project was designed to be a naturalistic observational inpatient study, it included patients from various age groups with diverse symptom profiles and medical histories as well as different pharmacological treatments. On the other hand, it represents quite a realistic picture of the broad clinical spectrum of MDD. Regarding the simulated data (MCC: 0.69–0.72; BAC: 0.84–0.86), better performances compared to the clinical data were expected because 25 features were explicitly created to be informative for the target variable. The congruency of the main results across the two data sets highlights that the differences between the two pipelines do not depend on the overall informativeness of the features and might generalize to other data sets as well.

In addition to 'traditional' supervised machine learning algorithms, such as the classifiers applied in this study, deep learning in the sense of deep neural networks is becoming increasingly common in psychiatric research. So far, however, applications have rather focused on diagnosis than on prognosis or personalization of treatment [41]. A reason might be that deep learning usually requires large sample sizes and has an increased risk of overfitting due to the number of parameters fitted, especially in relatively small sample sizes that are common in psychiatric clinical trials [3, 42]. In addition, deep neural networks were shown to be not generally superior to other classifiers on many classification tasks [43–47], but come with comparatively high computational costs. However, for more complex features, such as brain imaging, time-series, or sensor-based data, prognostic research in psychiatry might benefit from deep learning [41, 42, 48]. There is also growing evidence that deep neural networks might be particularly useful for integration of multimodal data, e.g., from studies on stress detection [49] and diagnosing MDD [50] and Alzheimer's disease [51–53].

With respect to treatment outcome, we selected a reduction of $\geq 50\%$ on a symptom scale sum score after 6 weeks of treatment as the target variable for the clinical data set because it represents one of the most widely used definitions of treatment outcome in MDD research. Recently, more and more critique has come up on MDD measurement in general [54] and on symptom scale sum score-based outcome definitions in particular (for a review, see [55], for instance). The definition of response used here represents an artificial dichotomization of an ordinal scale and is therefore associated with loss of information [56]. While most MDD outcome classification models have aimed at such binary outcome definitions based on cut-off values [13, 14], others have used unsupervised learning to generate data-driven outcome classes beforehand [25]. So far, however, there is no evidence for the superiority of one outcome definition over another in terms of predictability.

Our study shows some limitations. First, our pipeline can only be applied to classification algorithms which provide some kind of feature coefficients, at least in the version of *scikit-learn* (0.23.1) that was used in the present study. SVCs with non-linear kernels, for instance, were not included in our analyses as they do not return feature coefficients required by the RFE. However, the applied classifiers represent a selection of commonly used classifiers for CPMs of MDD treatment outcome [14]. Second, it remains unclear how well our results generalize to data sets with very different dimensions, i.e., different sample-to-feature ratios. It is possible that

data sets with significantly more or less features compared to the number of samples might profit less from the nested pipeline with RFE. Still, we tested our pipeline both on real and simulated data with dimensions that are representative of many psychiatric patient cohorts and corresponding CPM studies (e.g., [6, 9, 25]). Third, the proposed pipeline with nested RFE is computationally expensive compared to a single CV pipeline or a nested CV without RFE. Hence, we restricted our analyses to 100 permutation runs even though a larger number of permutations might have resulted in a more precise empirical null distribution. In future applications, it might be worth to evaluate first if the benefits of a sparser CPM would outweigh the additional computational expenses needed during model development.

Conclusions

In conclusion, our nested supervised machine learning pipeline with simultaneous hyperparameter tuning and feature selection could lead to sparser CPMs without losses in accuracy. This approach might be particularly beneficial in scenarios in which a literature-based a priori feature selection is not possible, e.g., due to lack of evidence or, in contrast, due to a large number of potentially useful predictors, as observed in MDD, for instance [57]. If measurements that come with certain expenses are involved, sparser models could reduce both costs for users (e.g., clinical institutions) and stress for patients resulting in better data quality.

Abbreviations

BAC: Balanced accuracy; CPM: Clinical prediction model; CV: Cross-validation; LR: Logistic regression; MARS: Munich antidepressant response signature; MCC: Matthews correlation coefficient; MDD: Major depressive disorder; RF: Random forest classifier; RFE: Recursive feature elimination; STAR*D: Sequenced treatment alternatives to relieve depression; SVC: Support vector classifier.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-022-01926-2.

Additional file 1: **Table S1**. Baseline features used for predictive modeling in the clinical data set in alphabetical order. **Fig. S1**. Preprocessing workflow of samples and features from the clinical data set. **Table S2**. TRIPOD Checklist for Prediction Model Development and Validation. **Fig. S2**. Balanced accuracy scores for the three classifiers and the two data sets on the validation data. **Table S3**. Matthews correlation coefficients and corresponding *p*-values for non-permuted models. **Fig. S3**. Receiver operating characteristic curves and corresponding AUC values for all non-permuted models (with and without RFE) across the three classifiers and the two data sets on the validation data. **Table S4**. Confusion matrices and derived performance metrics including 95% confidence intervals for all non-permuted models on the validation data. **Table S5**. Results from Kolmogorov-Smirnov tests comparing the empirical MCC distributions of the permutation runs to the theoretical null distribution. **Fig. S4**. Quantile-quantile plots for the 100 permutation runs of each classifier and data set. **Fig S5**. Permutation importance from 25 permutations for all 113 clinical features, ordered alphabetically and grouped by classifier and model. **Fig. S6**. Permutation importance from 25 permutations for the most informative features from the simulated data set, grouped by classifier and models with and without RFE.. **Fig. S7**. Permutation importance from 25 permutations for all 125 features from the simulated data set, ordered by number and grouped by classifier and model (with and without RFE).

Acknowledgements

None.

Author contributions

NR conducted the statistical analyses and wrote the initial draft of the manuscript. TMB and EBB curated and contributed data. NR, NK and BMM developed the study design and methodology. TMB, NK, EBB and BMM critically contributed to the writing of the manuscript and supervised the project. All author read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. N.K. is supported by EU-FP7 project PRONIA ('Personalised Prognostic Tools for Early Psychosis Management') under the Grant Agreement No. 602152. N.R. received funding from the Bavarian Ministry of Economic Affairs, Regional Development and Energy (BayMED, PBN_MED-1711-0003). None of the funding programmes were directly used to fund writing of this manuscript.

Availability of data and materials

Data from the MARS study as well as the corresponding preprocessed data set that was used for the analyses can be requested by contacting Dr. Tanja Brückl (brueckl@psych.mpg.de).. Analysis scripts are available at https://doi.org/10.5281/zenodo.6759730.

Declarations

Ethics approval and consent to participate

The MARS study was approved by the ethics committee of the Ludwig Maximilian University in Munich, Germany. Written informed consent was obtained from all subjects.

Consent for publication

Not applicable.

Competing interests

The Authors declare that there is no conflict of interest.

Author details

[1]Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Kraepelinstraße 2-10, 80804 Munich, Germany. [2]International Max Planck Research School for Translational Psychiatry, Munich, Germany. [3]Department of Psychiatry and Psychotherapy, Ludwig Maximilian University, Munich, Germany. [4]Max Planck Institute of Psychiatry, Munich, Germany. [5]Institute of Psychiatry, Psychology and Neuroscience, King's College, London, UK. [6]Department of Health Data Science, University of Liverpool, Liverpool, UK.

Received: 29 April 2022 Accepted: 7 July 2022
Published online: 14 July 2022

References

1. Kapur S, Phillips AG, Insel TR. Why has it taken so long for biological psychiatry to develop clinical tests and what to do about it? Mol Psychiatry. 2012;17(12):1174–9. https://doi.org/10.1038/mp.2012.105.
2. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry. 2021;20(2):154–70.
3. Dwyer DB, Falkai P, Koutsouleris N. Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol. 2018;14(1):91–118.

4. Rutledge RB, Chekroud AM, Huys QJ. Machine learning and big data in psychiatry: toward clinical applications. Curr Opin Neurobiol. 2019;55:152–9. https://doi.org/10.1016/j.conb.2019.02.006.
5. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol. 2021;132:142–5. https://doi.org/10.1016/j.jclinepi.2021.01.009.
6. 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(3):243–50. https://doi.org/10.1016/S2215-0366(15)00471-X.
7. Trivedi MH, Rush AJ, Wisniewski SR, Nierenberg AA, Warden D, Ritz L, et al. Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry. 2006;163:28–40.
8. Zou H, Hastie T. Regaularization and variable selection via the elastic net. J R Stat Soc Ser B Statistical Methodol. 2005;67(2):301–20.
9. Dinga R, Marquand AF, Veltman DJ, Beekman ATF, Schoevers RA, van Hemert AM, et al. Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. Transl Psychiatry. 2018;8(1):241. https://doi.org/10.1038/s41398-018-0289-1.
10. Penninx BWJH, Nolen WA, Lamers F, Zitman FG, Smit JH, Spinhoven P, et al. Two-year course of depressive and anxiety disorders: results from the Netherlands study of depression and anxiety (NESDA). J Affect Disord. 2011;133(1–2):76–85. https://doi.org/10.1016/j.jad.2011.03.027.
11. 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(6):94–102.
12. Athreya AP, Neavin D, Carrillo-Roa T, Skime M, Biernacka J, Frye MA, et al. Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine-learning approach with multi-trial replication. Clin Pharmacol Ther. 2019;106(4):855–65.
13. Lee Y, Ragguett RM, Mansur RB, Boutilier JJ, Rosenblat JD, Trevizol A, et al. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord. 2018;241:519–32. https://doi.org/10.1016/j.jad.2018.08.073.
14. Sajjadian M, Lam RW, Milev R, Rotzinger S, Frey BN, Soares CN, et al. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. Psychol Med. 2021;51(16):2742–51.
15. Kilsdonk E, Peute LW, Jaspers MWM. Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis. Int J Med Inform. 2017;98:56–64. https://doi.org/10.1016/j.ijmedinf.2016.12.001.
16. Sanchez-Pinto LN, Venable LR, Fahrenbach J, Churpek MM. Comparison of variable selection methods for clinical predictive modeling. Int J Med Inform. 2018;116:10–7. https://doi.org/10.1016/j.ijmedinf.2018.05.006.
17. Maslej MM, Furukawa TA, Cipriani A, Andrews PW, Mulsant BH. Individual differences in response to antidepressants: a meta-analysis of placebo-controlled randomized clinical trials. JAMA Psychiat. 2020;77(6):607–17.
18. Kubat M. An introduction to machine learning. 2017. pp. 1–348.
19. Hennings JM, Owashi T, Binder EB, Horstmann S, Menke A, Kloiber S, et al. Clinical characteristics and treatment outcome in a representative sample of depressed inpatients - findings from the munich antidepressant response signature (MARS) project. J Psychiatr Res. 2009;43(3):215–29.
20. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry. 1960;23(1):56–62.
21. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in {P}ython. J Mach Learn Res. 2011;12:2825–30.
22. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction second edition. New York: Springer; 2009. p. 485–585.
23. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
24. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
25. Paul R, Andlauer TFM, Czamara D, Hoehn D, Lucae S, Pütz B, et al. Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Transl Psychiatry. 2019. https://doi.org/10.1038/s41398-019-0524-4.
26. Koutsouleris N, Dwyer DB, Degenhardt F, Maj C, Urquijo-Castro MF, Sanfelici R, et al. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. JAMA Psychiat. 2021;78(2):195–209.
27. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5.
28. Dewancker I, McCourt M, Clark S. Bayesian optimization primer. URL https//app sigopt com/static/pdf/SigOpt_ Bayesian_Optimization_ Primer pdf. 2015;
29. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975;405(2):442–51.
30. Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The balanced accuracy and its posterior distribution. Proc - Int Conf Pattern Recognit. 2010;3121:4.
31. Student. Probable error of a correlation coefficient. Biometrika. 1908;6(2–3):302–10.
32. Diedenhofen B, Musch J. Cocor: a comprehensive solution for the statistical comparison of correlations. PLoS ONE. 2015;10(4):1–12.
33. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. Br J Surg. 2015;102(3):148–58.
34. World Health Organization. The ICD-10 classification of mental and behavioural disorders: clinical descriptions and diagnostic guidelines. Geneva: World Health Organisation; 1992.
35. Derogatis LR, Savitz KL. The SCL-90-R, Brief Symptom Inventory, and Matching Clinical Rating Scales. In: The use of psychological testing for treatment planning and outcomes assessment, 2nd edn. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers; 1999. p. 679–724.
36. Sanghani G, Kotecha K. Incremental personalized E-mail spam filter using novel TFDCR feature selection with dynamic feature update. Expert Syst Appl. 2019;115:287–99. https://doi.org/10.1016/j.eswa.2018.07.049.
37. Zhang Y, Wang S, Phillips P, Ji G. Binary PSO with mutation operator for feature selection using decision tree applied to spam detection. Knowl Based Syst. 2014;64:22–31.
38. Zhang Y, Dong Z, Phillips P, Wang S, Ji G, Yang J, et al. Detection of subjects and brain regions related to Alzheimer's disease using 3D MRI scans based on eigenbrain and machine learning. Front Comput Neurosci. 2015;9:66.
39. Xuan P, Guo MZ, Wang J, Wang CY, Liu XY, Liu Y. Genetic algorithm-based efficient feature selection for classification of pre-miRNAs. Genet Mol Res. 2011;10(2):588–603.
40. Iniesta R, Hodgson K, Stahl D, Malki K, Maier W, Rietschel M, et al. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci Rep. 2018;8(1):1–9. https://doi.org/10.1038/s41598-018-23584-z.
41. Durstewitz D, Koppe G, Meyer-Lindenberg A. Deep neural networks in psychiatry. Mol Psychiatry. 2019;24(11):1583–98. https://doi.org/10.1038/s41380-019-0365-9.
42. Koppe G, Meyer-lindenberg A, Durstewitz D. Deep learning for small and big data in psychiatry. Neuropsychopharmacology. 2021. https://doi.org/10.1038/s41386-020-0767-z.
43. Wainer J. Comparison of 14 different families of classification algorithms on 115 binary datasets. arXiv Prepr arXiv160600930. 2016;
44. Zhang C, Liu C, Zhang X, Almpanidis G. An up-to-date comparison of state-of-the-art classification algorithms. Expert Syst Appl. 2017;82:128–50.
45. Makridakis S, Spiliotis E, Assimakopoulos V. Statistical and machine learning forecasting methods: concerns and ways forward. PLoS ONE. 2018;13(3): e0194889.
46. Gacto MJ, Soto-Hidalgo JM, Alcalá-Fdez J, Alcalá R. Experimental study on 164 algorithms available in software tools for solving standard non-linear regression problems. IEEE Access. 2019;7:108916–39.
47. Fernández-Delgado M, Cernadas E, Barro S, Amorim D. Do we need hundreds of classifiers to solve real world classification problems? J Mach Learn Res. 2014;15(1):3133–81.
48. Calhoun VD, Sui J. Multimodal fusion of brain imaging data: a key to finding the missing link (s) in complex mental illness. Biol psychiatry Cogn Neurosci Neuroimag. 2016;1(3):230–44.
49. Walambe R, Nayak P, Bhardwaj A, Kotecha K. Employing multimodal machine learning for stress detection. J Healthc Eng. 2021;2021:1–12.

50. Yang J, Yin Y, Zhang Z, Long J, Dong J, Zhang Y, et al. Predictive brain networks for major depression in a semi-multimodal fusion hierarchical feature reduction framework. Neurosci Lett. 2018;665:163–9.
51. Lu D, Popuri K, Ding GW, Balachandar R, Beg MF. Multimodal and multiscale deep neural networks for the early diagnosis of Alzheimer's disease using structural MR and FDG-PET images. Sci Rep. 2018;8(1):1–13.
52. Shi J, Zheng X, Li Y, Zhang Q, Ying S. Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. IEEE J Biomed Heal Inf. 2017;22(1):173–83.
53. Liu S, Liu S, Cai W, Che H, Pujol S, Kikinis R, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of Alzheimer's disease. IEEE Trans Biomed Eng. 2014;62(4):1132–40.
54. Fried EI, Flake JK, Robinaugh DJ. Revisiting the theoretical and methodological foundations of depression measurement. Nat Rev Psychol. 2022;1:358–68.
55. Rost N, Binder EB, Brückl TM. Predicting treatment outcome in depression: an introduction into current concepts and challenges. Eur Arch Psychiatry Clin Neurosci. 2022. https://doi.org/10.1007/s00406-022-01418-4.
56. Altman DG, Royston P. The cost of dichotomising continuous variables. Br Med J. 2006;332(7549):1080.
57. Perlman K, Benrimoh D, Israel S, Rollins C, Brown E, Tunteng JF, et al. A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. J Affect Disord. 2019;243:503–15. https://doi.org/10.1016/j.jad.2018.09.067.

Publisher's Note

3. Paper II

Rost, N., Dwyer, D. B., Gaffron, S., Rechberger, S., Maier, D., Binder, E. B., & Brückl, T. M. (2023). Multimodal predictions of treatment outcome in major depression: A comparison of data-driven predictors with importance ratings by clinicians. *Journal of Affective Disorders*, *327*, 330–339. https://doi.org/10.1016/j.jad.2023.02.007

Journal of Affective Disorders 327 (2023) 330–339

Contents lists available at ScienceDirect

Journal of Affective Disorders

journal homepage: www.elsevier.com/locate/jad

Multimodal predictions of treatment outcome in major depression: A comparison of data-driven predictors with importance ratings by clinicians

Nicolas Rost [a,b,*], Dominic B. Dwyer [c,d], Swetlana Gaffron [e], Simon Rechberger [e], Dieter Maier [f], Elisabeth B. Binder [a], Tanja M. Brückl [a]

[a] Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Munich, Germany
[b] International Max Planck Research School for Translational Psychiatry, Munich, Germany
[c] Department of Psychiatry and Psychotherapy, Ludwig Maximilian University, Munich, Germany
[d] Centre for Youth Mental Health, University of Melbourne, Melbourne, Australia
[e] Viscovery Software GmbH, Vienna, Austria
[f] Biomax Informatics AG, Planegg, Germany

ARTICLE INFO

Keywords:
Major depressive disorder
Treatment outcome
Clustering
Predictive modeling
Clinical expertise
Personalized psychiatry

ABSTRACT

Background: Reliable prediction models of treatment outcome in Major Depressive Disorder (MDD) are currently lacking in clinical practice. Data-driven outcome definitions, combining data from multiple modalities and incorporating clinician expertise might improve predictions.
Methods: We used unsupervised machine learning to identify treatment outcome classes in 1060 MDD inpatients. Subsequently, classification models were created on clinical and biological baseline information to predict treatment outcome classes and compared to the performance of two widely used classical outcome definitions. We also related the findings to results from an online survey that assessed which information clinicians use for outcome prognosis.
Results: Three and four outcome classes were identified by unsupervised learning. However, data-driven outcome classes did not result in more accurate prediction models. The best prediction model was targeting treatment response in its standard definition and reached accuracies of 63.9 % in the test sample, and 59.5 % and 56.9 % in the validation samples. Top predictors included sociodemographic and clinical characteristics, while biological parameters did not improve prediction accuracies. Treatment history, personality factors, prior course of the disorder, and patient attitude towards treatment were ranked as most important indicators by clinicians.
Limitations: Missing data limited the power to identify biological predictors of treatment outcome from certain modalities.
Conclusions: So far, the inclusion of available biological measures in addition to psychometric and clinical information did not improve predictive value of the models, which was overall low. Optimized biomarkers, stratified predictions and the inclusion of clinical expertise may improve future prediction models.

1. Introduction

Despite long term efforts in psychiatric research, the treatment of Major Depressive Disorder (MDD) remains a challenge. Success rates of antidepressant medication are still insufficient (Khan et al., 2017), and approximately 55 % of patients develop treatment resistance (Thomas et al., 2013). Given the high morbidity and mortality associated with MDD (Lépine and Briley, 2011) as well as its high economic burden (Greenberg et al., 2015), identifying non-responders at an *early stage* of treatment would be a major benefit for clinical decision-making and treatment success (Oluboka et al., 2018).

Many studies have identified factors associated with antidepressant treatment outcome (for a meta-review, see Perlman et al., 2019). However, single predictors from association studies often have low effect sizes limiting their translational value for clinical application. Therefore, the focus has shifted from hypothesis-driven experiments, searching for significant effects of specific indicators, towards data-driven multivariate approaches (Chekroud et al., 2021). Machine learning (ML) methods, which are capable of detecting complex patterns in large data sets, are promising approaches when the goal is to

* Corresponding author at: Max-Planck-Institute of Psychiatry, Kraepelinstraße 2-10, 80804 Munich, Germany.
E-mail address: nicolas_rost@psych.mpg.de (N. Rost).

https://doi.org/10.1016/j.jad.2023.02.007
Received 13 September 2022; Received in revised form 23 January 2023; Accepted 1 February 2023
Available online 6 February 2023

maximize predictive performance rather than discovering single associations (Dwyer et al., 2018).

With respect to forecasting treatment outcome in MDD, most ML approaches have focused on predicting standard dichotomized outcome definitions, predominantly response and remission, derived from symptom rating scales (e.g., Chekroud et al., 2016; Iniesta et al., 2018; Nie et al., 2018). According to a recent meta-analysis, the mean balanced accuracy of such prediction models was 63 % (Sajjadian et al., 2021). However, this approach is limited by the choice of arbitrary cut-off scores and ignores the diversity of clinically relevant courses that patients can experience (Rost et al., 2022a). As such, ML has also been used to first identify subgroups of patients with shared, clinically relevant symptom trajectories before the use of predictive algorithms. Paul et al. (2019), for example, used data from the Munich Antidepressant Response Signature (MARS) cohort (Hennings et al., 2009) and applied mixture modeling on weekly and biweekly symptom severity assessments in order to derive seven treatment outcome classes, ranging from a group showing rapid symptom improvement to a group with steadily high symptom scores. Based on 72 clinical baseline features and an early indicator of response after 2 weeks of treatment, a random forest algorithm classified patients with accuracies of 75–95 %, depending on the respective subgroup. Athreya et al. (2021) performed model-based clustering and probabilistic graphical modeling on single items from three Hamilton Rating Scale for Depression (HDRS; Hamilton, 1960) assessments (at baseline, week 4, and week 8), suggesting a cluster solution with 3 subgroups. Subsequently, treatment outcome after 8 weeks was predictable with an accuracy of 77 % using 4 HDRS items.

While the shift from *association* to *prediction* is crucial for advancing individualized treatments in psychiatry (Bzdok et al., 2021) and has led to many adequately performing prognostic models (for reviews, see Lee et al., 2018; Sajjadian et al., 2021), there is still a lack of (bio)markers or sets of markers that are predictive and robust enough for guiding treatment decisions in MDD. Hence, it has been suggested that future prediction models should focus on including multiple data modalities in order to combine multiple smaller effects of factors coming from different psychiatric subfields (Chekroud et al., 2021). So far, however, most studies have focused on feature sets coming from single data types (e.g., psychometric/clinical data (Chekroud et al., 2016; De Carlo et al., 2016; Paul et al., 2019), brain imaging (Frässle et al., 2020; Kang and Cho, 2020; Sämann et al., 2013), or genetic data (García-González et al., 2017; GENDEP Investigators et al., 2013)), even though multimodal modeling has led to promising results (Athreya et al., 2019; Iniesta et al., 2018; Koutsouleris et al., 2018; Lee et al., 2018; Sajjadian et al., 2022).

Furthermore, data-driven prediction models do not include clinician expertise. Greater involvement of clinician expertise in the variable selection process and model development may have a beneficial effect on resulting performance and facilitate acceptance of subsequent decision support systems (Jacobs et al., 2021; Kilsdonk et al., 2017). Therefore, it may be of additional value for future studies and related data collection to evaluate which indicators clinicians use to predict treatment response and to decide between different treatment options.

The aim of the present study was to identify early predictors of treatment outcome in patients with MDD that could guide medical treatment selection at a very early stage. We hereby focused on the prediction of treatment outcome to any kind of antidepressant drug in general rather than response to a specific antidepressant or predicting the most effective drug for an individual patient. Similar to Paul et al. (2019), we used data from the MARS study but included additional data modalities and a larger feature set. We attempted to expand previous research as follows: 1) We created data-driven treatment outcome classes using unsupervised learning and investigated whether this resulted in more accurate prediction models than the two standard outcome definitions of response and remission. 2) We used a strictly prospective approach to build the prediction models by including only variables that were assessed at baseline. 3) We combined data from various modalities which allowed us to evaluate the prognostic value of biological information beyond non-biological data. 4) We conducted an online survey among physicians and psychologists to investigate which indicators they use to assess a patient's likelihood of treatment success, and compared experts' responses with the results from the prediction models.

2. Materials and methods

2.1. Main sample

Data on MDD patients was obtained from the MARS project (Hennings et al., 2009), a multicenter observational inpatient study designed to create a broad characterization of patients admitted to the hospital with an acute depressive episode (according to the International Classification of Diseases (ICD-10; World Health Organization, 1992) criteria). Study inclusion happened within the first days after admission to the hospital and was accompanied by multiple psychometric and biological measurements. Patients received psychopharmacological medication at the discretion of the treating physician, resulting in the administration of various classes and combinations of antidepressants. The most frequently administered antidepressants during the first 6 weeks of treatment were serotonin-norepinephrine reuptake inhibitors (39.8 % of patients), selective serotonin reuptake inhibitors (38.2 % of patients), noradrenergic and specific serotonergic antidepressants (30.1 % of patients), and tricyclic antidepressants (28.3 % of patients). In the first 6 weeks of the study, weekly ratings of depressive symptoms were performed using the HDRS, followed by biweekly ratings afterwards. Missing HDRS values were estimated via linear interpolation from the two adjacent weeks. We selected patients with a main diagnosis of unipolar MDD (ICD-10 codes: F32 or F33) who had a baseline HDRS sum score ≥ 14 and HDRS data at treatment week 6, resulting in a sample of 1060 patients. To avoid overfitting in the ML analyses, the sample was initially randomly split into a training (80 %, $N = 848$) and a validation sample (20 %, $N = 212$). Table 1 shows a more detailed description of the two samples. The MARS project was approved by the ethics committee of the Ludwig Maximilian University, Munich, and informed consent forms were signed by all patients before participation. More details of the study are accessible elsewhere (Hennings et al., 2009).

2.2. External validation sample

To validate findings in a different sample, we used data from a double-blind randomized clinical trial that examined the augmentation of venlafaxine XR with the atypical neuroleptic quetiapine in treatment-resistant inpatients diagnosed with unipolar depression. Patients received 4 weeks of venlafaxine XR monotherapy, which was augmented starting at week 5 with either the atypical antipsychotic quetiapine or placebo in the event of non-response. This trial was selected because patients completed questionnaires that largely overlapped with measures used in the MARS cohort, making it possible to apply the predictive models to it. We used the same selection criteria as for the MARS sample (see above), resulting in 84 patients. The trial (registered in the European Clinical Trials database, EudraCT number: 2005-001217-17, and on ClinicalTrials.gov, NCT00253266) was funded by the German Federal Ministry of Education and Research (BMBF support code: 01KG0709) and approved by the ethics committee of the Bavarian State Medical Association (ethic number: 05059). Signed informed consent was obtained from each participant before entering the study. Further information is available in the Supplementary Methods.

2.3. Treatment outcome definitions

Main outcome definitions were based on ratings from the HDRS 17-item version (HDRS-17) from week 0 (baseline) and week 6. Treatment outcome classes were created using 4 outcome measures that entered 2

Table 1
Main sample characteristics from the Munich Antidepressant Response Signature project, split into training and validation set.

	Training sample (N = 848)	Validation sample (N = 212)	Overall (N = 1060)	Chi^2/t	p
Gender					
Female	443 (52.2 %)	112 (52.8 %)	555 (52.4 %)	0.006	0.939
Male	405 (47.8 %)	100 (47.2 %)	505 (47.6 %)		
Age					
Mean (SD)	47.7 (14.0)	46.4 (14.4)	47.5 (14.1)	1.213	0.226
Median [Min, Max]	48.0 [18.0, 87.0]	46.5 [19.0, 80.0]	48.0 [18.0, 87.0]		
Diagnosis (ICD-10)					
F32 (depressive episode)	288 (34.0 %)	77 (36.3 %)	365 (34.4 %)	0.320	0.572
F33 (recurrent depressive disorder)	560 (66.0 %)	135 (63.7 %)	695 (65.6 %)		
HDRS-17 baseline sum score					
Mean (SD)	23.8 (5.63)	24.0 (5.36)	23.8 (5.57)	−0.606	0.545
Median [Min, Max]	24.0 [10.0, 40.0]	23.5 [13.0, 35.0]	24.0 [10.0, 40.0]		
Missing	32 (3.8 %)	6 (2.8 %)	38 (3.6 %)		
Illness duration (years)					
Mean (SD)	11.1 (12.0)	10.1 (10.6)	10.9 (11.7)	1.154	0.249
Median [Min, Max]	6.98 [0, 66.9]	6.75 [0, 46.5]	6.89 [0, 66.9]		
Missing	60 (7.1 %)	11 (5.2 %)	38 (6.7 %)		

Note: For categorical variables, chi-squared tests were used, and for continuous variables, *t*-tests were used to test for differences between training and validation sample. HDRS-17, Hamilton Rating Scale for Depression (17-item version); ICD-10, International Classification of Diseases.

different clustering algorithms. The four outcome variables were specifically selected to 1) include both changes in symptoms severity relative to baseline and absolute symptom severity after 6 weeks of treatment; 2) focus not only on depressive symptom scale sum scores but also on depressive core symptoms as defined by the ICD-10 and the Diagnostic and Statistical Manual of Mental Disorders-IV (American Psychiatric Association, 1994) since a reduction of the sum scores is not necessarily accompanied by an improvement in core symptoms; and 3) to exclude variables incorporating arbitrary cut-off values such as response and remission. The selected variables that went into the clustering pipelines are presented in Table 2. A more detailed description and explanation of our outcome variable selection is available in the Supplementary Methods.

We decided to include two different clustering methods in order to compare resulting cluster solutions with respect to their reproducibility and predictability. With this, we wanted to ensure that predictability of the resulting outcome classes was not strongly depending on the selected clustering method. Both clustering procedures were chosen based on their ability to handle non-continuous data and to result in maximally robust cluster solutions. The first clustering was performed using Viscovery® SOMine® (Viscovery Software GmbH, 2021), a statistical learning software which uses self-organizing maps (SOMs; Kohonen, 1982) in combination with Ward's hierarchical clustering method

Table 2
Variables used for treatment outcome clustering and reference outcome definitions.

Variable description	Outcome type	Scale	Range (training sample)
Outcome variable set for clustering			
HDRS-17 percentage change in sum score from baseline to week 6	Change score	Interval	−100–85
HDRS-17 sum score at week 6	Absolute score	Ordinal	0–33
HDRS-17 percentage change in core symptom score (sum of items 1, 7, and 13) from baseline to week 6	Change score	Interval	−100–80
HDRS-17 core symptom score (sum of items 1, 7, and 13) at week 6	Absolute score	Ordinal	0–10
Binary reference outcome definitions			
Response (≥ 50 % reduction of HDRS-17 sum score from baseline to week 6)	Change score	Binary	0–1
Remission (HDRS-17 sum score < 8 at week 6)	Absolute score	Binary	0–1

Note: Outcome types indicated if the variable represents and absolute value at week 6 or a change score in relation to the respective baseline value. HDRS-17, Hamilton Rating Scale for Depression (17-item version).

(Ward, 1963). The best number of clusters is chosen based on a quality measure for each cluster count. The final cluster solution can be applied to a validation sample in order to receive the respective cluster allocations. As a second clustering method, a consensus clustering approach was implemented using Python version 3.8.5. A k-medoids algorithm applied to the Gower distance matrix of the respective training sample was chosen. Within the consensus framework, k = 2 to k = 9 clusters were fit 1000 times, each time on two thirds (N = 565) randomly subsampled from all patients in the training sample. The best number of clusters was selected based on the cophenetic correlation coefficient and the proportion of ambiguous clustering of the resulting consensus matrices. Hierarchical clustering was applied on the consensus matrix of the selected k in order to obtain cluster labels for all patients in the training sample. A random forest classifier trained on these patients was then used to assign cluster labels to the validation sample. More detailed descriptions of both clustering procedures are available in the Supplementary Methods.

Additionally, we included the outcome classes from Paul et al. (2019) as prediction targets which were created based not only on the information from weeks 0 and 6, but also from all other weekly and biweekly HDRS scores. This allowed us to evaluate whether predicting outcome classes derived from a different set of outcome variables would change predictive performance. We further included treatment response, defined as a minimum reduction of 50 % on the HDRS-17 sum score at week 6 compared to baseline, and remission, defined as an HDRS-17 sum score < 8 at week 6, as additional outcome measures to assess whether clustering solutions could outperform standard outcome measures in terms of predictability.

2.4. Available baseline measurements

Within the MARS study, several data modalities were measured at baseline, including social demographics, clinical data (e.g., comorbidities, information on prior treatments and medical history), and genotypes. Polygenic risk scores (PRSs) were calculated from genotype data in order to reduce dimensionality. For different subsets of the MARS patients, additional parameters from neuropsychological testing, physical characteristics and laboratory parameters, structural magnetic resonance imaging (sMRI), DNA methylation and gene expression were available. An overview over all included baseline variables (n = 401), divided into non-biological and biological information, is available in

Supplementary Table 1. Univariate analyses across all outcome definitions (outcome classes, response, and remission) and all baseline variables were performed using analysis of variance (ANOVA) for continuous and approximately normally distributed variables, Kruskal-Wallis tests for ordinal or skewed continuous variables, and Fisher's exact tests for categorical variables.

Apart from baseline measurements, several clinical patient characteristics were assessed at week 6 and at the end of study participation, respectively. These measures included personality questionnaires, information on stressful life events (e.g., childhood trauma), stress coping, somatic complaints, and metabolic abnormalities (see Supplementary Table 2 for a detailed list). Due to our strictly prospective approach for the predictive analyses and in contrast to Paul et al. (2019), we did not include these measures in the predictive modeling part but only in the univariate analyses.

2.5. Predictive modeling

Supervised machine learning in terms of classification was used to predict treatment outcome classes as well as treatment response and remission. Out of all 1060 MARS patients, we excluded those with > 60 % missing values across all baseline features, leaving us with 1018 patients (811 in the training, 207 in the validation sample). The training sample was additionally randomly split into a training and a test (or 'development') set, resulting in final sample sizes of $N = 604$ for the training data and $N = 207$ for the test and validation data each. Out of all 401 features, we excluded those with ≥ 50 % missing values ($n = 231$) and those which were binomially distributed and strongly skewed so that one category was highly underrepresented (≤ 5 %, $n = 30$). Since we were particularly interested in the predictive value of biological patient characteristics, the remaining 201 features were split into a biological ($n = 65$) and non-biological ($n = 136$) predictor set (see Supplementary Table 1).

Predictive modeling was performed in Python using the *scikit-learn* library (Pedregosa et al., 2011). As classification algorithms, we used elastic net regularized logistic regression (ENLR) models and random forest classifiers. Both classification methods have already been successfully applied in prior studies with similar goals (e.g., Athreya et al., 2019; Iniesta et al., 2018; Nie et al., 2018; Perlis, 2013). The classifiers were trained on the 604 training samples using a previously validated repeated nested cross-validation framework (Rost et al., 2022b), which included a recursive feature elimination (RFE) to generate prediction models that were as sparse as possible. All trained models were then applied to the test sample and to both the internal (MARS) and external (venlafaxine augmentation study) validation samples. Balanced accuracy (BAC) values (Brodersen et al., 2010) were used for model performance evaluation. To ensure comparability between different numbers of target classes, we scaled the BAC values by their respective chance levels (base rate). All classifications were run using three different feature sets: non-biological features only ($n = 136$), biological features only ($n = 65$) and the complete feature set ($n = 201$). The best-performing classifiers from each of the three feature sets were also applied to the internal (MARS) and external (venlafaxine augmentation trial) validation sample and their most important features, indicated by permutation importance, were extracted. A more detailed description of the predictive modeling is available in the Supplementary Methods.

2.6. Clinical online survey

To assess which patient characteristics physicians and psychologists use as early prognostic factors and to compare these results to the top predictors from the predictive modeling, we conducted an anonymous online survey in five psychiatric and psychosomatic clinics in Germany. The survey included a few questions on profession, experience, and clinical setting as well as two main parts containing 1) free text answers on early indicators of treatment outcome, 2) Likert scale ratings from 0 (no predictive value) to 4 (very high predictive value) on pre-selected patient features coming from 11 different categories (social demographics, vital parameters, clinical characteristics, comorbidities, medical history, treatment history, personality traits, substance use, stress coping, trauma and life events, genetics; for details, see Supplementary Methods). Free text answers were independently grouped into categories of indicators by two different authors (N.R., T.M.B.) and subsequently combined into a final categorization. The survey was approved by the ethics committee of the Ludwig Maximilian University, Munich (project number: 21–0175 KB).

3. Results

3.1. Treatment outcome classes

The combined SOM-Ward clustering in Viscovery SOMine suggested a solution with 3 clusters (Supplementary Fig. 1 A): one cluster ($N = 297$, 35.0 %) with comparably high symptoms scores and only minor improvements over the 6 weeks, a smaller cluster ($N = 236$, 27.8 %) showing moderate treatment response, and a third cluster ($N = 315$, 37.1 %) with moderate to strong symptom reduction and a high proportion of remission. Results from the consensus clustering proposed a 4-cluster solution (see Supplementary Fig. 1 B,C). Clusters were also characterized by different amounts of symptom reduction, from almost none ($N = 181$, 21.3 %), moderate ($N = 378$, 44.6 %) and strong ($N = 177$, 20.9 %) to complete symptom reduction ($N = 112$, 13.2 %). As shown in Fig. 1, the trajectories of HDRS-17 sum scores over the 6 treatment weeks (Fig. 1 A,B) resembled the trajectory of core symptoms (Fig. 1 B,D) for both clustering solutions, suggesting that overall improvement in HDRS symptoms paralleled improvement in depressive core symptoms. Contingency matrices between the SOM-Ward clusters and the consensus clusters are shown in Table 3. Furthermore, both cluster solutions showed significant associations with treatment response and remission (all Fisher's exact tests $p < .001$, see Table 4).

Univariate comparisons across all baseline variables between outcome classes as well as responders vs. non-responders and remitters vs. non-remitters showed that outcome groups mainly differed in clinical characteristics (e.g., initial symptom severity), medical history, and prior treatments (see Supplementary Tables 3–6). With respect to non-baseline measures, several significant differences in personality traits, stress coping, and childhood trauma were observed (see Supplementary Tables 7–10).

3.2. Predictions of treatment outcome

The combination of RFE and intrinsic feature selection of the algorithms led to considerable feature reduction. Supplementary Table 11 shows the reduction in features across all models. Scaled BAC scores of the prediction models for the MARS test sample are presented in Fig. 2. The best-performing models for all three feature sets were ENLR classifiers of treatment response and remission. The non-biological feature set targeting response had the highest overall BAC at 63.9 % (scaled BAC: 0.28). Using only biological features, the best model also targeted response with a BAC of 59.6 % (scaled BAC: 0.19). The best model from the complete feature set targeted remission and achieved a BAC of 62.5 % (scaled BAC: 0.25). When applied to the MARS validation sample, the corresponding BAC values reached 59.5 %, 49.8 %, and 70.4 %, respectively. On the external validation sample, values of 56.9 %, 40.6 %, and 54.3 % were observed. More detailed evaluations of these models are presented in Supplementary Table 12 and Supplementary Fig. 4. For completeness, an overview of the BAC scores of all computed models for both validation data samples is shown in Supplementary Fig. 5. Overall, data-driven outcome classes did not result in better predictions compared to standard definitions of treatment response and remission. Furthermore, predictions created from the biological feature set generally showed worse performance (scaled BAC values: −0.06–0.19) than

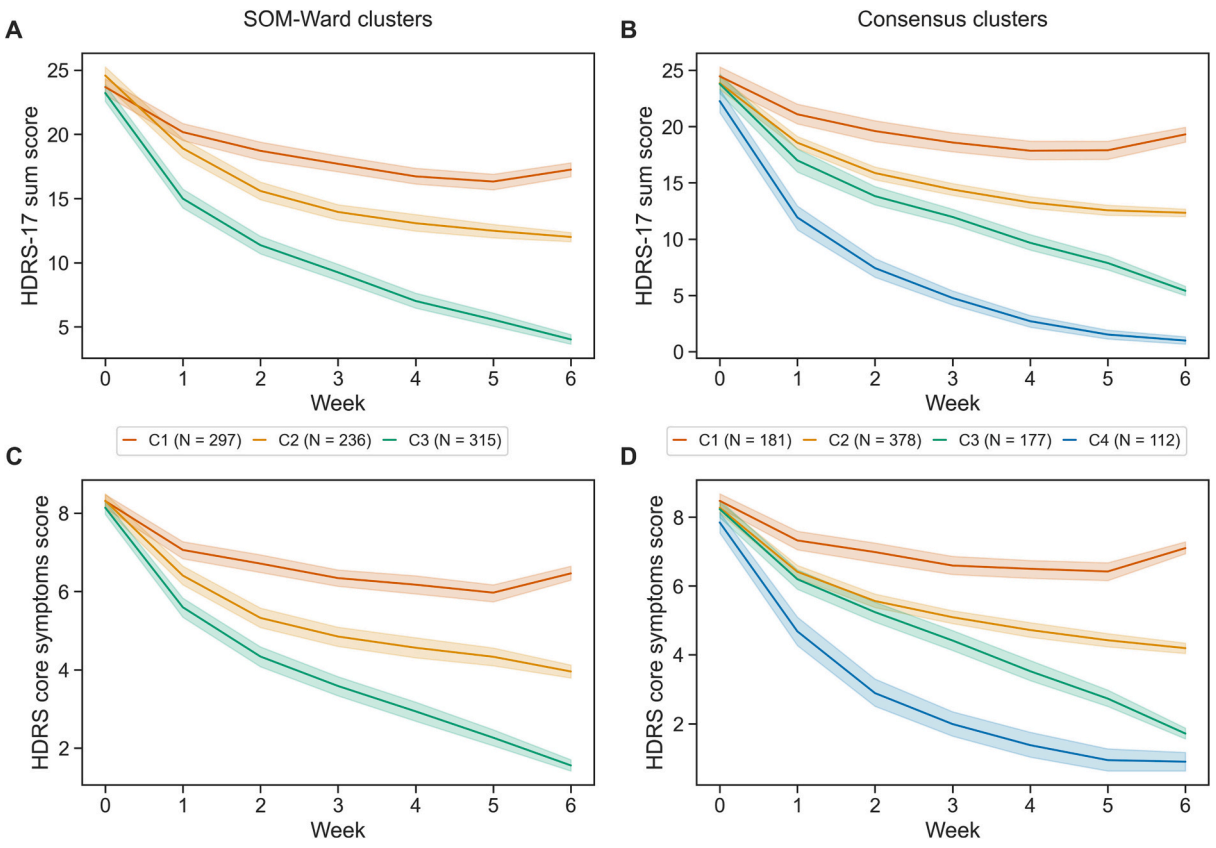


Fig. 1. Trajectories of HDRS-17 sum scores and core symptom scores (sum of scores on items 1, 7, and 13) for both cluster solutions over the observed duration of treatment (baseline until week 6) in the training sample. The lines represent the group means, the shaded areas the respective 95 % confidence intervals. A: HDRS-17 sum scores for the SOM-Ward clusters, B: HDRS-17 sum scores for the consensus clusters, C: core symptom scores for the SOM-Ward clusters, D: core symptom scores for the consensus clusters. HDRS-17, Hamilton Rating Scale for Depression (17-item version); SOM, self-organizing map.

Table 3
Contingency tables between both clustering methods in the training sample.

		Consensus cluster				
		C1	C2	C3	C4	Total
SOM-Ward cluster	C1	181 (60.9)	116 (39.1)	0 (0)	0 (0)	297
	C2	0 (0)	232 (98.3)	4 (1.7)	0 (0)	236
	C3	0 (0)	30 (9.5)	173 (54.9)	112 (35.6)	315
	Total	181	378	177	112	

Note: The numbers in parentheses represent row-wise percentages. SOM, self-organizing map.

Table 4
Contingency tables between both clustering methods and response and remission in the training sample.

		Response		Remission	
		Yes $N = 463$	No $N = 385$	Yes $N = 268$	No $N = 580$
SOM-Ward clusters	C1	12 (4.0)	285 (96.0)	2 (0.7)	295 (99.3)
	C2	142 (60.2)	94 (39.8)	10 (4.2)	226 (95.8)
	C3	309 (98.1)	6 (1.9)	256 (81.3)	59 (18.7)
Consensus clusters	C1	4 (2.2)	177 (97.8)	0 (0)	181 (100)
	C2	173 (45.8)	205 (54.2)	20 (5.3)	358 (94.7)
	C3	174 (98.3)	3 (1.7)	136 (76.8)	41 (23.2)
	C4	112 (100)	0 (0)	112 (100)	0 (0)

Note: The numbers in parentheses represent row-wise percentages. SOM, self-organizing map.

predictions based on the non-biological (0.08–0.28) and the complete feature set (0.06–0.25). Differences between the non-biological and the complete set were comparatively small, indicating an overall negligible additional predictive value of biological on top of non-biological features.

Key predictor variables (mean permutation importance > 0) ranked by their permutation importance for the model with the best performance on the test sample (non-biological feature set predicting response) are depicted in Fig. 3. Information on symptom profiles (i.e., items from the HDRS, the Panic and Agoraphobia Scale (Bandelow, 1995), and the Symptom Check-List-90-R (Derogatis and Savitz, 1999)), medical history and social demographics were most influential. The top predictors for the best model from the biological features and for the complete feature set are shown in Supplementary Figs. 6 and 7, respectively. The top predictors from the complete feature set were all non-biological variables, underscoring the restricted predictive value of the available biological features.

3.3. Biological completer analysis

Since most biological baseline measures could not be included in the predictive analysis due to many missing values, we wanted to ensure that feature selection and feature importance were not exclusively driven by differences in the amount of missing data. Thus, we retrained the best-performing model from the biological and non-biological feature set on a subsample ($N = 142$) including only patients with complete sMRI, methylation and PRS data (for details, see Supplementary Methods). Response classification on the resulting 160 non-biological features led to a BAC of 54.9 % on the test sample while the model based on the 101 biological features performed worse than

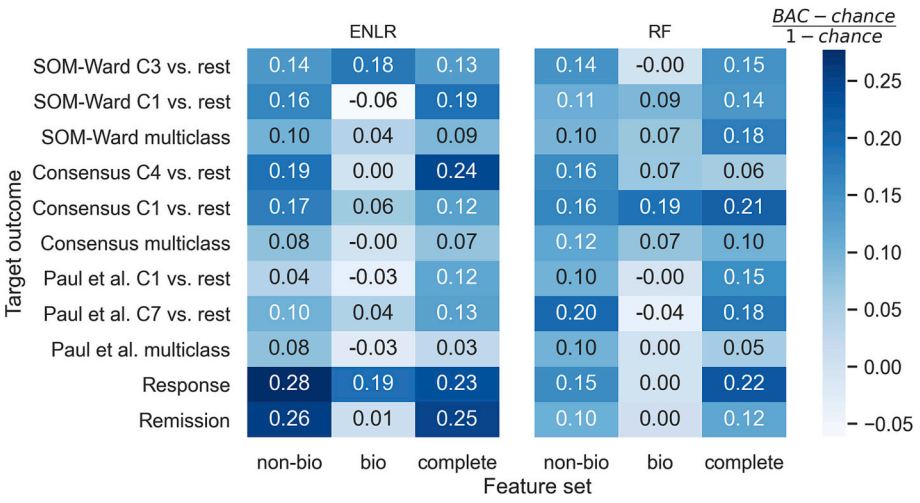


Fig. 2. Overview over model performances on the test sample across all prediction targets, feature sets and classifiers. BAC, Balanced accuracy; bio, biological feature set; ENLR, elastic net regularized logistic regression; non-bio, non-biological feature set; RF, random forests.

chance with a BAC of 37.7 %.

3.4. Expertise and ratings by clinicians

53 mental health professionals (20 resident physicians, 12 senior physicians, 14 psychologists (in psychotherapy training), 6 certified psychotherapists) completed the online survey and had an average clinical experience of 6.9 years (see Supplementary Fig. 8). Ratings on the predictive value of the 11 included predictor categories regarding treatment outcome were highest for treatment history (M = 3.31, SD = 0.83), personality traits (M = 3.12, SD = 1.08) and course of the disorders (M = 3.06, SD = 0.83). The categories ranked as least important were vital parameters (M = 1.35, SD = 0.99), sociodemographic (M = 1.48, SD = 0.98), and clinical characteristics (M = 1.98, SD = 1.15; Fig. 4 A). Free text answers were also grouped into categories (see Supplementary Fig. 9). The most frequently mentioned indicators of non-response were treatment history (e.g., non-response to antidepressants in prior episodes), patient attitude towards treatment and disorder (e.g., lacking trust into pharmacotherapy, lacking openness to different forms of therapy), personality factors, course of the disorder (e.g., number of prior episodes, duration of the disorder) and environmental factors (e.g., stressful life events, lack of social support). Of the 20 top predictors from the prediction model (Fig. 3), we were able to match ten with ratings from the online survey (Fig. 4 B). Among those, duration of the current depressive episode had consistently high rankings according to both the data and the clinical survey. Compared with modeling results, clinicians appeared to underestimate the influence of anxiety and suicidality as well as of sociodemographic information.

4. Discussion

Using two clustering methods, we were able to identify three and four treatment outcome classes in a clinical cohort of 1060 MDD inpatients. Subsequent classification models, however, showed that outcome classes were not easier to predict based on baseline measurements than standard outcome definitions, that is, response (≥ 50 % reduction of HDRS-17 sum score) and remission (HDRS-17 sum score < 8). A possible explanation for this could be that the (bi)weekly HDRS assessments were the only measure of treatment outcome in the MARS cohort and all outcome definitions used here were based on this scale. The use of unsupervised ML to create outcome classes might be of greater benefit when different outcome measures have been collected and need to be combined (Rost et al., 2022a). As such, they could help to identify more broadly characterized critical pathways of the disorder and to find corresponding predictors, beyond response and remission.

Classification performances for our test sample and both validation samples were largely within the range of previous studies using similar approaches to predict treatment response and remission (Lee et al., 2018; Sajjadian et al., 2021). Compared to the work by Paul et al. (2019), our models showed worse performance in predicting their outcome classes. This was not surprising as we opted for a more rigorous prospective approach and included only information measured at baseline as predictors. However, recent findings suggest that more accurate predictions are needed for successful implementation into clinical practice (Browning et al., 2021).

We additionally assessed the predictive value of biological markers by splitting the baseline feature set into biological measures (PRS, blood levels, and vital parameters), non-biological measures (mainly sociodemographic and clinical measures) and the combination of both. Overall, models that used only biological information showed the worst performance. Models built on the remaining two feature sets in general performed similarly well. While the highest observed BAC of 70.4 % was achieved by a model that included biological information (combined feature set predicting remission in the internal validation data), the overall results underscored the low incremental value of biological over clinical and sociodemographic characteristics. Our results replicated prior findings that clinical features, particularly symptom profiles and baseline depression severity, as well as sociodemographic information were most predictive of outcome classification (e.g., Chekroud et al., 2016; Iniesta et al., 2016). While it has been suggested (in psychosis research, for instance (Koutsouleris et al., 2021, 2018)) that multimodal modeling might outperform unimodal prediction models, the influence of biological markers remained limited in our analysis, whether they were included in the modeling alone or as part of the full feature set. This is consistent with previous results from MDD research showing that adding biological features to psychometric data leads to only minor improvements in performance (Dinga et al., 2018; Iniesta et al., 2018). Since biological characteristics often require a lot of effort in measurement and preprocessing, e.g., brain imaging or omics data, they should be included in predictive models only if their inclusion leads to a substantial increase in predictive accuracy. Self-reports and clinician-based ratings, on the other hand, remain the most robust predictors identified so far, are easy to assess in clinical settings and therefore may be most valuable for clinical application (Chekroud et al., 2021). Nevertheless, deeper multimodal characterizations including novel objective measures, such as sensor-based data ('digital phenotyping'), have led to promising results and might be of greater relevance to predictive models in the future (Bzdok and Meyer-Lindenberg, 2018; Durstewitz et al.,

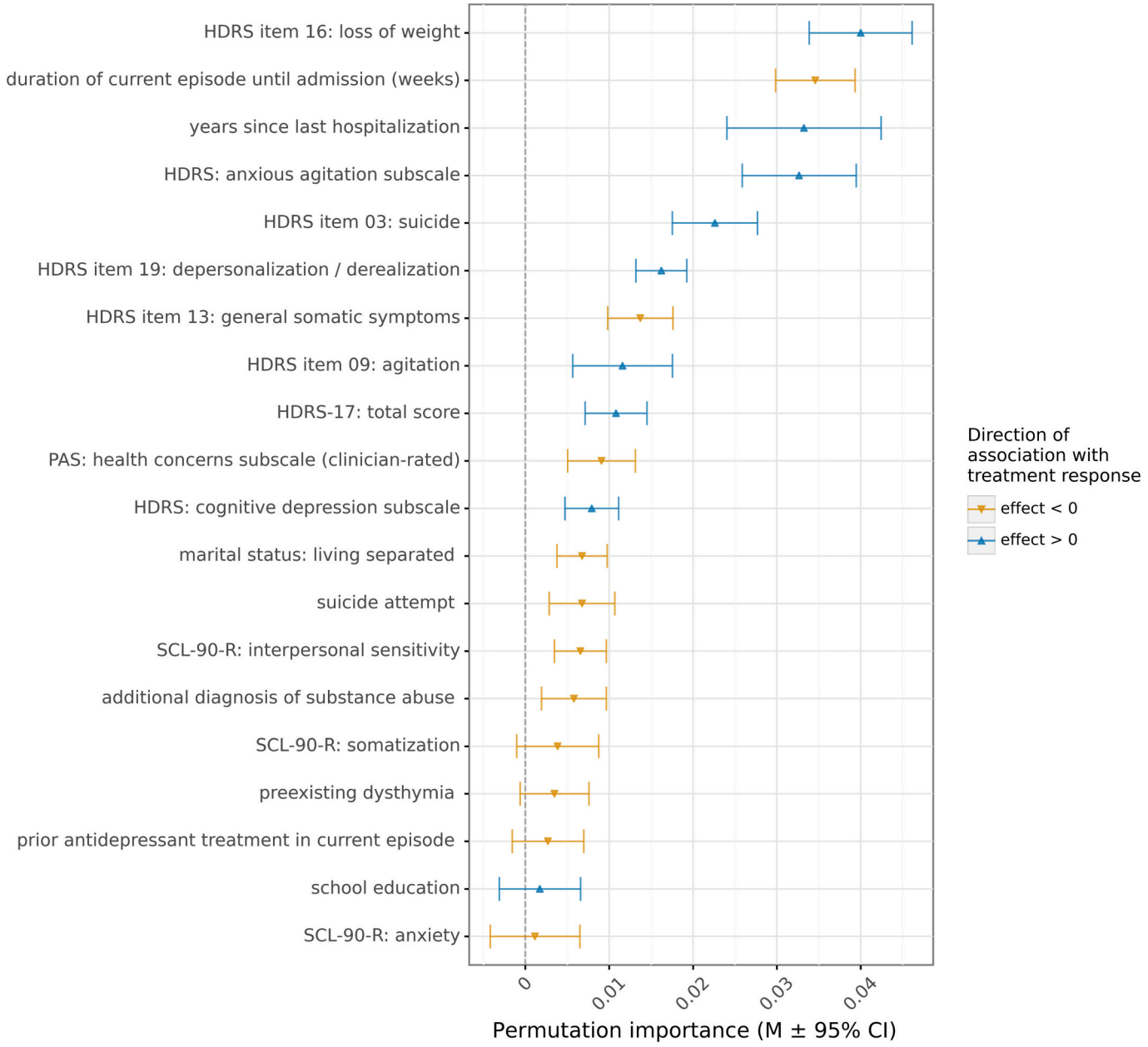


Fig. 3. Most important features for the best-performing model (regularized logistic regression on treatment response using the non-biological feature set) indicted by permutation importance. Color coding represents the direction of the effect, i.e., of the association of each feature with treatment response. HDRS, Hamilton Rating Scale for Depression; PAS, Panic and Agoraphobia Scale; SCL-90-R, Symptom Check-List-90-R.

2019; Huckvale et al., 2019). Future discovery of biomarkers that may relate to true biological differences will likely improve their usefulness in prediction algorithms.

Consistent with the modeling results, mental health professionals indicated that the course of the disorder and previous treatments had a large impact on treatment outcome. Participants further mentioned patient attitude (e.g., lacking trust into pharmacotherapy, fixation on one specific type of therapy) as an important prognostic factor. While attitude and beliefs of patients are rarely measured in clinical trials, it might be worthwhile to include corresponding assessments to empirically evaluate their predictive value. Recent studies on patients' treatment preferences and treatment efficacy, however, did not show any benefits of matching patients with their preferred form of therapy (Kuzminskaite et al., 2021; Windle et al., 2020). Another highly rated category from the clinical survey were personality factors. In MARS, data from personality questionnaires were assessed at discharge and therefore not included in the baseline feature sets. Nevertheless, univariate analyses revealed statistically significant differences between all outcome groups on several personality traits, e.g., extraversion, neuroticism, and harm avoidance (see Supplementary Tables 7–10), suggesting that including these variables could have further improved model performance. Since previous studies (Paul et al., 2019; Takahashi et al., 2013) have also suggested that personality factors are predictive of treatment outcome, these measures should be included in future prediction models. The same could apply to information on environmental factors, such as stressful life events, social support, and childhood trauma, which also proved comparatively important in the clinical survey and showed associations with treatment outcome in our statistical tests and in prior publications (Nanni et al., 2012; Nelson et al., 2017; Williams et al., 2016). Clinical characteristics, on the other hand, received rather low scores from clinicians although they contributed substantially to the prediction model. These factors, e.g., symptom

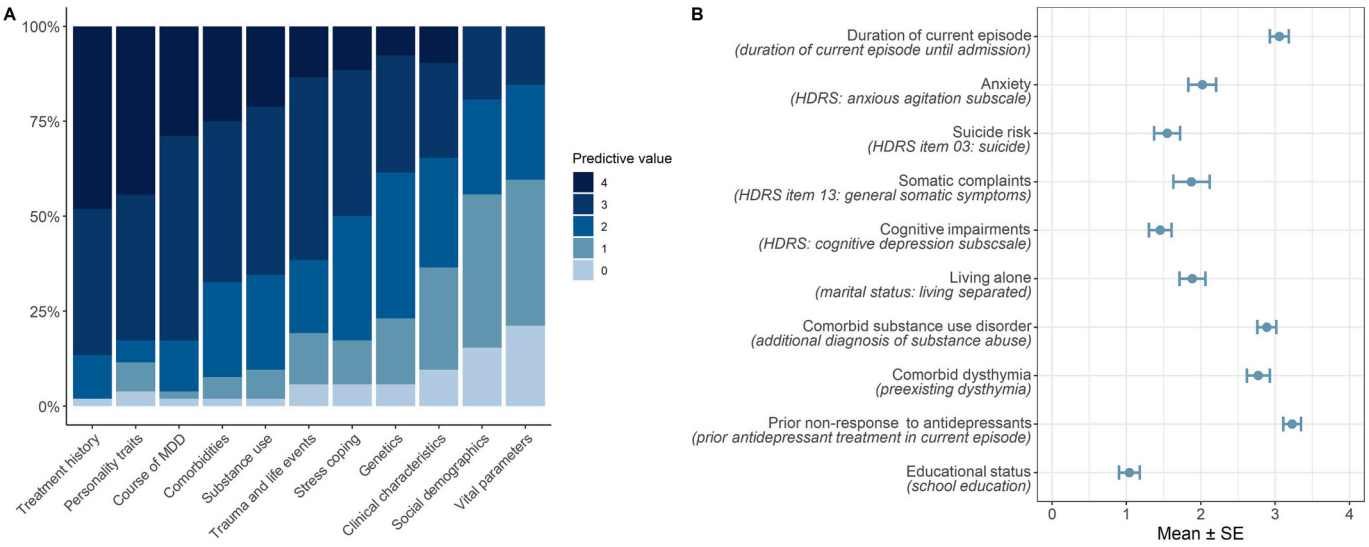


Fig. 4. Results from the online survey among clinicians. A: categories of predictors, rated by their predictive value for MDD treatment outcome and ordered by their corresponding mean ratings, B: clinician ratings for data-driven top predictors from the prediction model that could be matched to items in the survey. The parentheses behind each item represent the matching feature from Fig. 3. Ratings ranged from 0 (no predictive value) to 4 (very high predictive value).

profiles and initial symptom severity, have been identified as robust and generalizable predictors in predictive modeling studies (Chekroud et al., 2021; De Carlo et al., 2016; DeRubeis et al., 2014; Fournier et al., 2010), which is why it might be beneficial for clinicians to consider this information when making treatment decisions.

Our study comes with certain limitations. First, the MARS sample is very heterogeneous in terms of age, symptom profiles, and medication. Consequently, no differential predictions on the outcomes to specific drugs could be concluded. The goal of personalized matching of individual patients to specific treatment options thus remains unmet in our investigation. Nonetheless, our sample represents the broad natural spectrum of MDD patients, speaking for high robustness of the resulting models. On the other hand, the use of a naturalistic inpatient cohort might limit comparability and generalizability of our findings to other studies that primarily used outpatient data from randomized controlled trials and often focused on specific antidepressants (e.g., Athreya et al., 2019; Iniesta et al., 2018; Sajjadian et al., 2022). A second issue is the amount of missing values in the data. Although we aimed at including as many baseline features as possible in the predictive modeling and wanted to make use of the extensive biological characterization of the sample, we could not include several modalities in the main analyses as they were available only for a specific subsample. This was the case for sMRI data, DNA methylation, gene expression, and neuropsychological testing. We attempted to approximate their predictive value using univariate analyses and by performing a predictive 'biological completer analysis', which did not lead to any different conclusions compared to the main findings. However, our biological predictions most likely lack power, which may have negatively affected the identification of biological markers and their predictive value. The lacking relevance of PRSs in our analyses was not unexpected given that no strong biomarkers for MDD have been discovered to date and corresponding PRSs explain less than 5 % variance (Howard et al., 2019). Finally, the questions and ratings from the clinical survey were not fully comparable to the variables in the data, and not all features could be captured in the survey. Hence, not all attributes in Fig. 4 B are completely congruent with the corresponding model predictors in Fig. 3. While we aimed at assessing as many characteristics from the data set as possible in the survey, we had to make restrictions in terms of the scope and length of the survey. All higher-level predictor categories, however, were covered in the survey.

In conclusion, reliable and robust markers are still needed to predict treatment outcome in MDD and to advance personalized psychiatry, especially with respect to biological measures. Whereas data-driven treatment outcome classes do not seem to facilitate this task compared to standard binary outcome measures based on a symptom scale sum score, clinician expertise should be considered when planning design and data collection in future clinical trials.

Contributors

N.R., D.M., E.B.B., and T.M.B. developed the study design. N.R. conducted the statistical analyses and wrote the initial draft of the manuscript. D.B.D., S.G., S.R., and D.M. contributed to the statistical analyses. E.B.B. and T.M.B. curated data, critically contributed to the writing of the manuscript, and supervised the project. All authors contributed to and have approved the final manuscript.

Funding sources

N.R. received funding from the Bavarian Ministry of Economic Affairs, Regional Development and Energy (BayMED, PBN_MED-1711-0003).

Conflict of interest

None.

Data availability

Original data from the Munich Antidepressant Response Signature (MARS) study, the venlafaxin XR augmentation trial and from the online survey (in German) can be requested by contacting Dr. Tanja Brückl (brueckl@psych.mpg.de). Analysis scripts, the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) checklist for prediction model development, and the online survey are available under https://osf.io/fdy8k/.

Acknowledgements

We thank Dr. Darina Czamara, Dr. Marcus Ising, Dr. Nils Rek, Dr. Janine Knauer-Arloth, Dr. Riya Paul and Dr. Philipp Sämann for data curation, provision, and preprocessing.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jad.2023.02.007.

References

American Psychiatric Association, 1994. Diagnostic and Statistical Manual of Mental Disorders, 4th ed. American Psychiatric Association, Washington, DC.

Athreya, A.P., Neavin, D., Carrillo-Roa, T., Skime, M., Biernacka, J., Frye, M.A., Rush, A. J., Wang, L., Binder, E.B., Iyer, R.K., Weinshilboum, R.M., Bobo, W.V., 2019. Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine-learning approach with multi-trial replication. Clin. Pharmacol. Ther. 106, 855–865. https://doi.org/10.1002/cpt.1482.

Athreya, A.P., Brückl, T., Binder, E.B., Rush, A.J., Biernacka, J., Frye, M.A., Neavin, D., Skime, M., Monrad, D., Iyer, R.K., Mayes, T., Trivedi, M., Carter, R.E., Wang, L., Weinshilboum, R.M., Croarkin, P.E., Bobo, W.V., 2021. Prediction of short-term antidepressant response using probabilistic graphical models with replication across multiple drugs and treatment settings. Neuropsychopharmacology. https://doi.org/10.1038/s41386-020-00943-x.

Bandelow, B., 1995. Assessing the efficacy of treatments for panic disorder and agoraphobia: II. The Panic and Agoraphobia Scale. Int. Clin. Psychopharmacol. 10, 73–81.

Brodersen, K.H., Ong, C.S., Stephan, K.E., Buhmann, J.M., 2010. The balanced accuracy and its posterior distribution. Proc. Int. Conf. Pattern Recognit. 3121–3124. https://doi.org/10.1109/ICPR.2010.764.

Browning, M., Bilderbeck, A.C., Dias, R., Dourish, C.T., Kingslake, J., Deckert, J., Goodwin, G.M., Gorwood, P., Guo, B., Harmer, C.J., Morriss, R., Reif, A., Ruhe, H.G., van Schaik, A., Simon, J., Sola, V.P., Veltman, D.J., Elices, M., Lever, A.G., Menke, A., Scanferla, E., Stäblein, M., Dawson, G.R., 2021. The clinical effectiveness of using a predictive algorithm to guide antidepressant treatment in primary care (PReDicT): an open-label, randomised controlled trial. Neuropsychopharmacology 46, 1307–1314. https://doi.org/10.1038/s41386-021-00981-z.

Bzdok, D., Meyer-Lindenberg, A., 2018. Machine learning for precision psychiatry: opportunities and challenges. Biol. Psychiatry Cogn. Neurosci. Neuroimaging 3, 223–230. https://doi.org/10.1016/j.bpsc.2017.11.007.

Bzdok, D., Varoquaux, G., Steyerberg, E.W., 2021. Prediction, not association, paves the road to precision medicine. JAMA Psychiatry 78, 127–128. https://doi.org/10.1001/jamapsychiatry.2020.2549.

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, 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X.

Chekroud, A.M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., Choi, K., 2021. The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 20, 154–170. https://doi.org/10.1002/wps.20882.

De Carlo, V., Calati, R., Serretti, A., 2016. Socio-demographic and clinical predictors of non-response/non-remission in treatment resistant depressed patients: a systematic review. Psychiatry Res. 240, 421–430. https://doi.org/10.1016/j.psychres.2016.04.034.

Derogatis, L.R., Savitz, K.L., 1999. The SCL-90-R, brief symptom inventory, and matching clinical rating scales. In: The Use of Psychological Testing for Treatment Planning and Outcomes Assessment, 2nd ed. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, pp. 679–724.

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 recommendationsA demonstration. PLoS One 9, 1–8. https://doi.org/10.1371/journal.pone.0083875.

Dinga, R., Marquand, A.F., Veltman, D.J., Beekman, A.T.F., Schoevers, R.A., van Hemert, A.M., Penninx, B.W.J.H., Schmaal, L., 2018. Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. Transl. Psychiatry 8, 241. https://doi.org/10.1038/s41398-018-0289-1.

Durstewitz, D., Koppe, G., Meyer-Lindenberg, A., 2019. Deep neural networks in psychiatry. Mol. Psychiatry 24, 1583–1598. https://doi.org/10.1038/s41380-019-0365-9.

Dwyer, D.B., Falkai, P., Koutsouleris, N., 2018. Machine learning approaches for clinical psychology and psychiatry. Annu. Rev. Clin. Psychol. 14, 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037.

Fournier, J.C., DeRubeis, R.J., Hollon, S.D., Dimidjian, S., Amsterdam, J.D., Shelton, R. C., Fawcett, J., 2010. Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA 303, 175–177. https://doi.org/10.1001/jama.2009.1943.

Frässle, S., Marquand, A.F., Schmaal, L., Dinga, R., Veltman, D.J., van der Wee, N.J.A., van Tol, M.J., Schöbi, D., Penninx, B.W.J.H., Stephan, K.E., 2020. Predicting individual clinical trajectories of depression with generative embedding. NeuroImage Clin. 26, 102213 https://doi.org/10.1016/j.nicl.2020.102213.

García-González, J., Tansey, K.E., Hauser, J., Henigsberg, N., Maier, W., Mors, O., Placentino, A., Rietschel, M., Souery, D., Žagar, T., Czerski, P.M., Jerman, B., Buttenschøn, H.N., Schulze, T.G., Zobel, A., Farmer, A., Aitchison, K.J., Craig, I., McGuffin, P., Giupponi, M., Perroud, N., Bondolfi, G., Evans, D., O'Donovan, M., Peters, T.J., Wendland, J.R., Lewis, G., Kapur, S., Perlis, R.H., Arolt, V., Domschke, K., Breen, G., Curtis, C., Sang-Hyuk, L., Kan, C., Newhouse, S., Patel, H., Baune, B.T., Uher, R., Lewis, C.M., Fabbri, C., 2017. Pharmacogenetics of antidepressant response: a polygenic approach. Prog. Neuro-Psychopharmacology Biol. Psychiatry 75, 128–134. https://doi.org/10.1016/j.pnpbp.2017.01.011.

GENDEP Investigators, MARS Investigators, STAR*D Investigators, 2013. Common genetic variation and antidepressant efficacy in major depressive disorder: A meta-analysis of three genome-wide pharmacogenetic studies. Am. J. Psychiatry 170, 207–217. https://doi.org/10.1176/appi.ajp.2012.12020237.

Greenberg, P.E., Fournier, A.A., Sisitsky, T., Pike, C.T., Kessler, R.C., 2015. The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J. Clin. Psychiatry 76, 155–162. https://doi.org/10.4088/JCP.14m09298.

Hamilton, M., 1960. A rating scale for depression. J. Neurol. Neurosurg. Psychiatry 23, 56–62.

Hennings, J.M., Owashi, T., Binder, E.B., Horstmann, S., Menke, A., Kloiber, S., Dose, T., Wollweber, B., Spieler, D., Messer, T., Lutz, R., Künzel, H., Bierner, T., Pollmächer, T., Pfister, H., Nickel, T., Sonntag, A., Uhr, M., Ising, M., Holsboer, F., Lucae, S., 2009. Clinical characteristics and treatment outcome in a representative sample of depressed inpatients - findings from the Munich antidepressant response signature (MARS) project. J. Psychiatr. Res. 43, 215–229. https://doi.org/10.1016/j.jpsychires.2008.05.002.

Howard, D.M., Adams, M.J., Clarke, T.K., Hafferty, J.D., Gibson, J., Shirali, M., Coleman, J.R.I., Hagenaars, S.P., Ward, J., Wigmore, E.M., Alloza, C., Shen, X., Barbu, M.C., Xu, E.Y., Whalley, H.C., Marioni, R.E., Porteous, D.J., Davies, G., Deary, I.J., Hemani, G., Berger, K., Teismann, H., Rawal, R., Arolt, V., Baune, B.T., Dannlowski, U., Domschke, K., Tian, C., Hinds, D.A., Trzaskowski, M., Byrne, E.M., Ripke, S., Smith, D.J., Sullivan, P.F., Wray, N.R., Breen, G., Lewis, C.M., McIntosh, A. M., 2019. Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. Nat. Neurosci. 22, 343–352. https://doi.org/10.1038/s41593-018-0326-7.

Huckvale, K., Venkatesh, S., Christensen, H., 2019. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. npj Digit. Med. 2 https://doi.org/10.1038/s41746-019-0166-1.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M.Z., Souery, D., Stahl, D., Dobson, R., Aitchison, K.J., Farmer, A., Lewis, C.M., McGuffin, P., Uher, R., 2016. Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J. Psychiatr. Res. 78, 94–102. https://doi.org/10.1016/j.jpsychires.2016.03.016.

Iniesta, R., Hodgson, K., Stahl, D., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M.Z., Souery, D., Dobson, R., Aitchison, K.J., Farmer, A., McGuffin, P., Lewis, C.M., Uher, R., 2018. Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci. Rep. 8, 1–9. https://doi.org/10.1038/s41598-018-23584-z.

Jacobs, M., Pradier, M.F., McCoy, T.H., Perlis, R.H., Doshi-Velez, F., Gajos, K.Z., 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. Transl. Psychiatry 11. https://doi.org/10.1038/s41398-021-01224-x.

Kang, S.-G., Cho, S.-E., 2020. Neuroimaging biomarkers for predicting treatment response and recurrence of major depressive disorder. Int. J. Mol. Sci. 21, 2148. https://doi.org/10.3390/ijms21062148.

Khan, A., Fahl Mar, K., Faucett, J., Khan Schilling, S., Brown, W.A., 2017. Has the rising placebo response impacted antidepressant clinical trial outcome? Data from the US Food and Drug Administration 1987–2013. World Psychiatry 16, 181–192. https://doi.org/10.1002/wps.20421.

Kilsdonk, E., Peute, L.W., Jaspers, M.W.M., 2017. Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis. Int. J. Med. Inform. 98, 56–64. https://doi.org/10.1016/j.ijmedinf.2016.12.001.

Kohonen, T., 1982. Self-organized formation of topologically correct feature maps. Biol. Cybern. 43, 59–69. https://doi.org/10.1007/bf00337288.

Koutsouleris, N., Kambeitz-Ilankovic, L., Ruhrmann, S., Rosen, M., Ruef, A., Dwyer, D.B., Paolini, M., Chisholm, K., Kambeitz, J., Haidl, T., Schmidt, A., Gillam, J., Schultze-Lutter, F., Falkai, P., Reiser, M., Riecher-Rössler, A., Upthegrove, R., Hietala, J., Salokangas, R.K.R., Pantelis, C., Meisenzahl, E., Wood, S.J., Beque, D., Brambilla, P., Borgwardt, S., 2018. Prediction models of functional outcomes for individuals in the clinical high-risk state for psychosis or with recent-onset depression: a multimodal, multisite machine learning analysis. JAMA Psychiatry 75, 1156–1172. https://doi.org/10.1001/jamapsychiatry.2018.2165.

Koutsouleris, N., Dwyer, D.B., Degenhardt, F., Maj, C., Urquijo-Castro, M.F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S.S., Weiske, J., Ruef, A., Kambeitz-Ilankovic, L., Antonucci, L.A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T.K., Rosen, M., Chisholm, K., Riecher-Rössler, A., Egloff, L., Schmidt, A., Andreou, C., Hietala, J., Schirmer, T., Romer, G., Walger, P., Franscini, M., Traber-Walker, N., Schimmelmann, B.G., Flückiger, R., Michel, C., Rössler, W., Borisov, O., Krawitz, P.M., Heekeren, K., Buechler, R., Pantelis, C., Falkai, P., Salokangas, R.K.R., Lencer, R., Bertolino, A., Borgwardt, S., Noethen, M., Brambilla, P., Wood, S.J., Upthegrove, R., Schultze-Lutter, F., Theodoridou, A., Meisenzahl, E., 2021. Multimodal machine learning workflows for prediction of psychosis in patients with clinical high-risk syndromes and recent-onset depression. JAMA Psychiatry 78, 195–209. https://doi.org/10.1001/jamapsychiatry.2020.3604.

Kuzminskaite, E., Lemmens, L.H.J.M., van Bronswijk, S.C., Peeters, F., Huibers, M.J.H., 2021. Patient choice in depression psychotherapy: outcomes of patient-preferred therapy versus randomly allocated therapy. Am. J. Psychother. 74, 103–111.

Lee, Y., Ragguett, R.M., Mansur, R.B., Boutilier, J.J., Rosenblat, J.D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T.C.Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V.C.H., Ho, R., Rong, C., McIntyre, R.S., 2018. Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J. Affect. Disord. 241, 519–532. https://doi.org/10.1016/j.jad.2018.08.073.

Lépine, J.P., Briley, M., 2011. The increasing burden of depression. Neuropsychiatr. Dis. Treat. 7, 3–7. https://doi.org/10.2147/NDT.S19617.
Nanni, V., Uher, R., Danese, A., 2012. Childhood maltreatment predicts unfavorable course of illness and treatment outcome in depression: a meta-analysis. Am. J. Psychiatry 169, 141–151. https://doi.org/10.1176/appi.ajp.2011.11020335.
Nelson, J., Klumparendt, A., Doebler, P., Ehring, T., 2017. Childhood maltreatment and characteristics of adult depression: meta-analysis. Br. J. Psychiatry 210, 96–104. https://doi.org/10.1192/bjp.bp.115.180752.
Nie, Z., Vairavan, S., Narayan, V.A., Ye, J., Li, Q.S., 2018. Predictive modeling of treatment resistant depression using data from STARD and an independent clinical study. PLoS One 13, 1–18. https://doi.org/10.1371/journal.pone.0197268.
Oluboka, O.J., Katzman, M.A., Habert, J., McIntosh, D., MacQueen, G.M., Milev, R.V., McIntyre, R.S., Blier, P., 2018. Functional recovery in major depressive disorder: providing early optimal treatment for the individual patient. Int. J. Neuropsychopharmacol. 21, 128–144. https://doi.org/10.1093/ijnp/pyx081.
Paul, R., Andlauer, T.F.M., Czamara, D., Hoehn, D., Lucae, S., Pütz, B., Lewis, C.M., Uher, R., Müller-Myhsok, B., Ising, M., Sämann, P.G., 2019. Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Transl. Psychiatry 9. https://doi.org/10.1038/s41398-019-0524-4.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.
Perlis, R.H., 2013. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biol. Psychiatry 74, 7–14. https://doi.org/10.1088/1367-2630/15/1/015008.Fluid.
Perlman, K., Benrimoh, D., Israel, S., Rollins, C., Brown, E., Tunteng, J.F., You, R., You, E., Tanguay-Sela, M., Snook, E., Miresco, M., Berlim, M.T., 2019. A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. J. Affect. Disord. 243, 503–515. https://doi.org/10.1016/j.jad.2018.09.067.
Rost, N., Binder, E.B., Brückl, T.M., 2022a. Predicting treatment outcome in depression: an introduction into current concepts and challenges. Eur. Arch. Psychiatry Clin. Neurosci. https://doi.org/10.1007/s00406-022-01418-4.
Rost, N., Brückl, T.M., Koutsouleris, N., Binder, E.B., Müller-Myhsok, B., 2022b. Creating sparser prediction models of treatment outcome in depression: a proof-of-concept study using simultaneous feature selection and hyperparameter tuning. BMC Med. Inform. Decis. Mak. 22, 181. https://doi.org/10.1186/s12911-022-01926-2.
Sajjadian, M., Lam, R.W., Milev, R., Rotzinger, S., Frey, B.N., Soares, C.N., Parikh, S.V., Foster, J.A., Turecki, G., Müller, D.J., Strother, S.C., Farzan, F., Kennedy, S.H., Uher, R., 2021. Machine learning in the prediction of depression treatment outcomes: a systematic review and meta-analysis. Psychol. Med. 51, 2742–2751. https://doi.org/10.1017/S0033291721003871.
Sajjadian, M., Uher, R., Ho, K., Hassel, S., Milev, R., Frey, B.N., Farzan, F., Blier, P., Foster, J.A., Parikh, S.V., Müller, D.J., Rotzinger, S., Soares, C.N., Turecki, G., Taylor, V.H., Lam, R.W., Strother, S.C., Kennedy, S.H., 2022. Prediction of depression treatment outcome from multimodal data : a CAN-BIND-1 report. Psychol. Med. 1–11 https://doi.org/10.1017/S0033291722002124.
Sämann, P.G., Höhn, D., Chechko, N., Kloiber, S., Lucae, S., Ising, M., Holsboer, F., Czisch, M., 2013. Prediction of antidepressant treatment response from gray matter volume across diagnostic categories. Eur. Neuropsychopharmacol. 23, 1503–1515. https://doi.org/10.1016/j.euroneuro.2013.07.004.
Takahashi, M., Shirayama, Y., Muneoka, K., Suzuki, M., Sato, K., Hashimoto, K., 2013. Personality traits as risk factors for treatment-resistant depression. PLoS One 8, 1–7. https://doi.org/10.1371/journal.pone.0063756.
Thomas, L., Kessler, D., Campbell, J., Morrison, J., Peters, T.J., Williams, C., Lewis, G., Wiles, N., 2013. Prevalence of treatment-resistant depression in primary care: cross-sectional data. Br. J. Gen. Pract. 63, 852–858. https://doi.org/10.3399/bjgp13X675430.
Viscovery Software GmbH, 2021. Viscovery SOMine.
Ward Jr., J.H., 1963. Hierarchical grouping to optimize an objective funcction. J. Am. Stat. Assoc. 58, 236–244.
Williams, L.M., Debattista, C., Duchemin, A.M., Schatzberg, A.F., Nemeroff, C.B., 2016. Childhood trauma predicts antidepressant response in adults with major depression: Data from the randomized international study to predict optimized treatment for depression. Transl. Psychiatry 6, e799-7. https://doi.org/10.1038/tp.2016.61.
Windle, E., Tee, H., Sabitova, A., Jovanovic, N., Priebe, S., Carr, C., 2020. Association of patient treatment preference with dropout and clinical outcomes in adult psychosocial mental health interventions: a systematic review and meta-analysis. JAMA Psychiatry 77, 294–302. https://doi.org/10.1001/jamapsychiatry.2019.3750.
World Health Organization, 1992. The ICD-10 Classification of Mental and Behavioural Disorders: Clinical Descriptions and Diagnostic Guidelines. World Health Organisation, Geneva.

References

Altman, D. G., & Royston, P. (2006). The cost of dichotomising continuous variables. *British Medical Journal*, *332*(7549), 1080. https://doi.org/10.1136/bmj.332.7549.1080

Bagby, R. M., Ryder, A. G., Schuller, D. R., & Marshall, M. B. (2004). The Hamilton Depression Rating Scale: Has the gold standard become a lead weight? *American Journal of Psychiatry*, *161*(12), 2163–2177. https://doi.org/10.1176/appi.ajp.161.12.2163

Bennabi, D., Aouizerate, B., El-Hage, W., Doumy, O., Moliere, F., Courtet, P., Nieto, I., Bellivier, F., Bubrovsky, M., Vaiva, G., Holztmann, J., Bougerol, T., Richieri, R., Lancon, C., Camus, V., Saba, G., Haesbaert, F., D'Amato, T., Charpeaud, T., … Haffen, E. (2015). Risk factors for treatment resistance in unipolar depression: A systematic review. *Journal of Affective Disorders*, *171*, 137–141. https://doi.org/10.1016/j.jad.2014.09.020

Bowling, N. A., Gibson, A. M., Houpt, J. W., & Brower, C. K. (2021). Will the Questions Ever End? Person-Level Increases in Careless Responding During Questionnaire Completion. *Organizational Research Methods*, *24*(4), 718–738. https://doi.org/10.1177/1094428120947794

Browning, M., Bilderbeck, A. C., Dias, R., Dourish, C. T., Kingslake, J., Deckert, J., Goodwin, G. M., Gorwood, P., Guo, B., Harmer, C. J., Morriss, R., Reif, A., Ruhe, H. G., van Schaik, A., Simon, J., Sola, V. P., Veltman, D. J., Elices, M., Lever, A. G., … Dawson, G. R. (2021). The clinical effectiveness of using a predictive algorithm to guide antidepressant treatment in primary care (PReDicT): an open-label, randomised controlled trial. *Neuropsychopharmacology*, *46*(7), 1307–1314. https://doi.org/10.1038/s41386-021-00981-z

Brückl, T. M., Spoormaker, V. I., Sämann, P. G., Brem, A. K., Henco, L., Czamara, D., Elbau, I., Grandi, N. C., Jollans, L., Kühnel, A., Leuchs, L., Pöhlchen, D., Schneider, M., Tontsch, A., Keck, M. E., Schilbach, L., Czisch, M., Lucae, S., Erhardt, A., & Binder, E. B. (2020). The biological classification of mental disorders (BeCOME) study: A protocol for an observational deep-phenotyping study for the identification of biological subtypes. *BMC Psychiatry*, *20*(1), 1–25. https://doi.org/10.1186/s12888-020-02541-z

Bzdok, D., Varoquaux, G., & Steyerberg, E. W. (2021). Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry*, *78*(2), 127–128. https://doi.org/10.1001/jamapsychiatry.2020.2549

Chekroud, A. M., Bondar, J., Delgadillo, J., Doherty, G., Wasil, A., Fokkema, M., Cohen, Z., Belgrave, D., DeRubeis, R., Iniesta, R., Dwyer, D., & Choi, K. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*, *20*(2), 154–170. https://doi.org/10.1002/wps.20882

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. *The Lancet Psychiatry*, *3*(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

Chevance, A. M., Ravaud, P., Tomlinson, A., Le Berre, C., Teufer, B., Touboul, S., Fried, E. I., Gartlehner, G., Cipriani, A., & Tran, V. T. (2020). Identifying outcomes for depression that matter to patients, informal caregivers and healthcare professionals:

qualitative content analysis of a large international online survey. *Lancet Psychiatry*, *7*, 692–702. https://doi.org/10.1016/S2215-0366(20)30191-7

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annual Review of Clinical Psychology*, *14*, 209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746

Dawson, N. V., & Weiss, R. (2012). Dichotomizing continuous variables in statistical analysis: A practice to avoid. *Medical Decision Making*, *32*(2), 225–226. https://doi.org/10.1177/0272989X12437605

Dinga, R., Marquand, A. F., Veltman, D. J., Beekman, A. T. F., Schoevers, R. A., van Hemert, A. M., Penninx, B. W. J. H., & Schmaal, L. (2018). Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. *Translational Psychiatry*, *8*(1), 241. https://doi.org/10.1038/s41398-018-0289-1

Durstewitz, D., Koppe, G., & Meyer-Lindenberg, A. (2019). Deep neural networks in psychiatry. *Molecular Psychiatry*, *24*(11), 1583–1598. https://doi.org/10.1038/s41380-019-0365-9

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine Learning Approaches for Clinical Psychology and Psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Fried, E. I. (2017). The 52 symptoms of major depression: Lack of content overlap among seven common depression scales. *Journal of Affective Disorders*, *208*, 191–197. https://doi.org/10.1016/j.jad.2016.10.019

Fried, E. I., Flake, J. K., & Robinaugh, D. J. (2022). Revisiting the theoretical and methodological foundations of depression measurement. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-022-00050-2

Fried, E. I., & Nesse, R. M. (2015). Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR∗D study. *Journal of Affective Disorders*, *172*, 96–102. https://doi.org/10.1016/j.jad.2014.10.010

Gennatas, E. D., Friedman, J. H., Ungar, L. H., Pirracchio, R., Eaton, E., Reichmann, L. G., Interian, Y., Luna, J. M., Simone, C. B., Auerbach, A., Delgado, E., van der Laan, M. J., Solberg, T. D., & Valdes, G. (2020). Expert-augmented machine learning. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(9), 4571–4577. https://doi.org/10.1073/pnas.1906831117

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857–871.

Greenberg, P. E., Fournier, A. A., Sisitsky, T., Pike, C. T., & Kessler, R. C. (2015). The economic burden of adults with major depressive disorder in the United States (2005 and 2010). *Journal of Clinical Psychiatry*, *76*(2), 155–162. https://doi.org/10.4088/JCP.14m09298

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery, and Psychiatry*, *23*(1), 56–62.

Hartmann, A., von Wietersheim, J., Weiss, H., & Zeeck, A. (2018). Patterns of symptom change in major depression: Classification and clustering of long term courses. *Psychiatry Research*, *267*, 480–489. https://doi.org/10.1016/j.psychres.2018.03.086

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning:*

Data Mining, Inference, and Prediction (Second Edi). Springer. https://doi.org/10.1111/j.1467-985X.2004.298_11.x

Hennings, J. M., Owashi, T., Binder, E. B., Horstmann, S., Menke, A., Kloiber, S., Dose, T., Wollweber, B., Spieler, D., Messer, T., Lutz, R., Künzel, H., Bierner, T., Pollmächer, T., Pfister, H., Nickel, T., Sonntag, A., Uhr, M., Ising, M., … Lucae, S. (2009). Clinical characteristics and treatment outcome in a representative sample of depressed inpatients - Findings from the Munich Antidepressant Response Signature (MARS) project. *Journal of Psychiatric Research*, *43*(3), 215–229. https://doi.org/10.1016/j.jpsychires.2008.05.002

Herzog, D. P., Wagner, S., Ruckes, C., Tadic, A., Roll, S. C., Härter, M., & Lieb, K. (2017). Guideline adherence of antidepressant treatment in outpatients with major depressive disorder: a naturalistic study. *European Archives of Psychiatry and Clinical Neuroscience*, *267*(8), 711–721. https://doi.org/10.1007/s00406-017-0798-6

Hollon, S. D., Aréan, P. A., Craske, M. G., Crawford, K. A., Kivlahan, D. R., Magnavita, J. J., Ollendick, T. H., Sexton, T. L., Spring, B., Bufka, L. F., Galper, D. I., & Kurtzman, H. (2014). Development of clinical practice guidelines. *Annual Review of Clinical Psychology*, *10*, 213–241. https://doi.org/10.1146/annurev-clinpsy-050212-185529

Iniesta, R., Hodgson, K., Stahl, D., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M. Z., Souery, D., Dobson, R., Aitchison, K. J., Farmer, A., McGuffin, P., Lewis, C. M., & Uher, R. (2018). Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. *Scientific Reports*, *8*(1), 1–9. https://doi.org/10.1038/s41598-018-23584-z

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J., Henigsberg, N., Dernovsek, M. Z., Souery, D., Stahl, D., Dobson, R., Aitchison, K. J., Farmer, A., Lewis, C. M., McGuffin, P., & Uher, R. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. *Journal of Psychiatric Research*, *78*(6), 94–102. https://doi.org/10.1016/j.jpsychires.2016.03.016

Jacobson, N. C., Weingarden, H., & Wilhelm, S. (2019). Digital biomarkers of mood disorders and symptom change. *Npj Digital Medicine*, *2*(1), 88–90. https://doi.org/10.1038/s41746-019-0078-0

James, S. L., Abate, D., Abate, K. H., Abay, S. M., Abbafati, C., Abbasi, N., Abbastabar, H., Abd-Allah, F., Abdela, J., Abdelalim, A., Abdollahpour, I., Abdulkader, R. S., Abebe, Z., Abera, S. F., Abil, O. Z., Abraha, H. N., Abu-Raddad, L. J., Abu-Rmeileh, N. M. E., Accrombessi, M. M. K., … Murray, C. J. L. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1789–1858. https://doi.org/https://doi.org/10.1016/S0140-6736(18)32279-7

Kelley, M. E., Dunlop, B. W., Nemeroff, C. B., Lori, A., Carrillo-Roa, T., Binder, E. B., Kutner, M. H., Rivera, V. A., Craighead, W. E., & Mayberg, H. S. (2018). Response rate profiles for major depressive disorder: Characterizing early response and longitudinal nonresponse. *Depression and Anxiety*, *35*(10), 992–1000. https://doi.org/10.1002/da.22832

Kilsdonk, E., Peute, L. W., & Jaspers, M. W. M. (2017). Factors influencing

implementation success of guideline-based clinical decision support systems: A systematic review and gaps analysis. *International Journal of Medical Informatics*, *98*, 56–64. https://doi.org/10.1016/j.ijmedinf.2016.12.001

König, H., König, H. H., & Konnopka, A. (2019). The excess costs of depression: A systematic review and meta-analysis. *Epidemiology and Psychiatric Sciences*. https://doi.org/10.1017/S2045796019000180

Koutsouleris, N., Dwyer, D. B., Degenhardt, F., Maj, C., Urquijo-Castro, M. F., Sanfelici, R., Popovic, D., Oeztuerk, O., Haas, S. S., Weiske, J., Ruef, A., Kambeitz-Ilankovic, L., Antonucci, L. A., Neufang, S., Schmidt-Kraepelin, C., Ruhrmann, S., Penzel, N., Kambeitz, J., Haidl, T. K., … Meisenzahl, E. (2021). Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients with Clinical High-Risk Syndromes and Recent-Onset Depression. *JAMA Psychiatry*, *78*(2), 195–209. https://doi.org/10.1001/jamapsychiatry.2020.3604

Kraus, C., Kadriu, B., Lanzenberger, R., Zarate, C. A., & Kasper, S. (2019). Prognosis and improved outcomes in major depression: a review. *Translational Psychiatry*, *9*(1). https://doi.org/10.1038/s41398-019-0460-3

Kubat, M. (2017). An Introduction to Machine Learning. In *An Introduction to Machine Learning*. https://doi.org/10.1007/978-3-319-63913-0

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., Subramaniapillai, M., Chan, T. C. Y., Fus, D., Park, C., Musial, N., Zuckerman, H., Chen, V. C. H., Ho, R., Rong, C., & McIntyre, R. S. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: A meta-analysis and systematic review. *Journal of Affective Disorders*, *241*, 519–532. https://doi.org/10.1016/j.jad.2018.08.073

Maj, M., Stein, D. J., Parker, G., Zimmerman, M., Fava, G. A., De Hert, M., Demyttenaere, K., McIntyre, R. S., Widiger, T., & Wittchen, H. U. (2020). The clinical characterization of the adult patient with depression aimed at personalization of management. *World Psychiatry*, *19*(3), 269–293. https://doi.org/10.1002/wps.20771

Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, *52*(1–2), 91–118. https://doi.org/10.1023/A:1023949509487

Nanni, V., Uher, R., & Danese, A. (2012). Childhood Maltreatment Predicts Unfavorable Course of Illness and Treatment Outcome in Depression: A Meta-Analysis. *American Journal of Psychiatry*, *169*(2), 141–151. https://doi.org/10.1176/appi.ajp.2011.11020335

Nelson, J., Klumparendt, A., Doebler, P., & Ehring, T. (2017). Childhood maltreatment and characteristics of adult depression: Meta-analysis. *British Journal of Psychiatry*, *210*(2), 96–104. https://doi.org/10.1192/bjp.bp.115.180752

Olbert, C. M., Gala, G. J., & Tupler, L. A. (2014). Quantifying heterogeneity attributable to polythetic diagnostic criteria: Theoretical framework and empirical application. *Journal of Abnormal Psychology*, *123*(2), 452–462. https://doi.org/10.1037/a0036068

Paul, R., Andlauer, T. F. M., Czamara, D., Hoehn, D., Lucae, S., Pütz, B., Lewis, C. M., Uher, R., Müller-Myhsok, B., Ising, M., & Sämann, P. G. (2019). Treatment

response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. *Translational Psychiatry*, *9*(1). https://doi.org/10.1038/s41398-019-0524-4

Perlman, K., Benrimoh, D., Israel, S., Rollins, C., Brown, E., Tunteng, J. F., You, R., You, E., Tanguay-Sela, M., Snook, E., Miresco, M., & 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. https://doi.org/10.1016/j.jad.2018.09.067

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: Is shorter better? A review and meta-analysis. *Value in Health*, *14*(8), 1101–1108. https://doi.org/10.1016/j.jval.2011.06.003

Rost, N., Binder, E. B., & Brückl, T. M. (2022). Predicting treatment outcome in depression: an introduction into current concepts and challenges. *European Archives of Psychiatry and Clinical Neuroscience*. https://doi.org/10.1007/s00406-022-01418-4

Rost, N., Brückl, T. M., Koutsouleris, N., Binder, E. B., & Müller-Myhsok, B. (2022). Creating sparser prediction models of treatment outcome in depression: a proof-of-concept study using simultaneous feature selection and hyperparameter tuning. *BMC Medical Informatics and Decision Making*, *22*(1), 181. https://doi.org/10.1186/s12911-022-01926-2

Rost, N., Dwyer, D. B., Gaffron, S., Rechberger, S., Maier, D., Binder, E. B., & Brückl, T. M. (2023). Multimodal predictions of treatment outcome in major depression: A comparison of data-driven predictors with importance ratings by clinicians. *Journal of Affective Disorders*, *327*, 330–339. https://doi.org/10.1016/j.jad.2023.02.007

Rush, A. J., Trivedi, M. H., Wisniewski, S. R., Nierenberg, A. A., Stewart, J. W., Warden, D., Niederehe, G., Thase, M. E., Lavori, P. W., Lebowitz, B. D., McGrath, P. J., Rosenbaum, J. F., Sackeim, H. A., Kupfer, D. J., Luther, J., & Fava, M. (2006). Acute and Longer-Term Outcomes in Depressed Outpatients Requiring One or Seve ... *American Journal of Psychiatry*, *163*(11), 1905–1917. https://doi.org/10.1176/appi.ajp.163.11.1905

Sajjadian, M., Lam, R. W., Milev, R., Rotzinger, S., Frey, B. N., Soares, C. N., Parikh, S. V., Foster, J. A., Turecki, G., Müller, D. J., Strother, S. C., Farzan, F., Kennedy, S. H., & Uher, R. (2021). Machine learning in the prediction of depression treatment outcomes: A systematic review and meta-analysis. *Psychological Medicine*, *51*(16), 2742–2751. https://doi.org/10.1017/S0033291721003871

Sajjadian, M., Uher, R., Ho, K., Hassel, S., Milev, R., Frey, B. N., Farzan, F., Blier, P., Foster, J. A., Parikh, S. V, Müller, D. J., Rotzinger, S., Soares, C. N., Turecki, G., Taylor, V. H., Lam, R. W., Strother, S. C., & Kennedy, S. H. (2022). Prediction of depression treatment outcome from multimodal data : a CAN-BIND-1 report. *Psychological Medicine*, 1–11. https://doi.org/10.1017/S0033291722002124

Sanchez-Pinto, L. N., Venable, L. R., Fahrenbach, J., & Churpek, M. M. (2018). Comparison of variable selection methods for clinical predictive modeling. *International Journal of Medical Informatics*, *116*(February), 10–17. https://doi.org/10.1016/j.ijmedinf.2018.05.006

Schultebraucks, K., Choi, K. W., Galatzer-Levy, I. R., & Bonanno, G. A. (2021). Discriminating Heterogeneous Trajectories of Resilience and Depression after Major Life Stressors Using Polygenic Scores. *JAMA Psychiatry*, *78*(7), 744–752.

https://doi.org/10.1001/jamapsychiatry.2021.0228

Skryabin, V., Rozochkin, I., Zastrozhin, M., Lauschke, V., Franck, J., Bryun, E., & Sychev, D. (2022). Meta-analysis of pharmacogenetic clinical decision support systems for the treatment of major depressive disorder. *The Pharmacogenomics Journal*, 1–5. https://doi.org/10.1038/s41397-022-00295-3

Souery, D., Oswald, P., Massat, I., Bailer, U., Bollen, J., Demyttenaere, K., Kasper, S., Lecrubier, Y., Montgomery, S., Serretti, A., Zohar, J., Mendlewicz, J., & Group for the Study of Resistant Depression (GSRD). (2007). Clinical factors associated with treatment resistance in major depressive disorder: Results from a European multicenter study. *Journal of Clinical Psychiatry*, *68*(7), 1062–1070. https://doi.org/10.4088/JCP.v68n0713

Takahashi, M., Shirayama, Y., Muneoka, K., Suzuki, M., Sato, K., & Hashimoto, K. (2013). Personality Traits as Risk Factors for Treatment-Resistant Depression. *PLoS ONE*, *8*(5), 1–7. https://doi.org/10.1371/journal.pone.0063756

Thomas, L., Kessler, D., Campbell, J., Morrison, J., Peters, T. J., Williams, C., Lewis, G., & Wiles, N. (2013). Prevalence of treatment-resistant depression in primary care: Cross-sectional data. *British Journal of General Practice*, *63*(617), 852–858. https://doi.org/10.3399/bjgp13X675430

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, *25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Torous, J., Kiang, M. V., Lorme, J., & Onnela, J. P. (2016). New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health*, *3*(2). https://doi.org/10.2196/mental.5165

Trajković, G., Starčević, V., Latas, M., Leštarević, M., Ille, T., Bukumirić, Z., & Marinković, J. (2011). Reliability of the Hamilton Rating Scale for Depression: A meta-analysis over a period of 49 years. *Psychiatry Research*, *189*(1), 1–9. https://doi.org/10.1016/j.psychres.2010.12.007

Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., Mors, O., Elkin, A., Williamson, R. J., Schmael, C., Henigsberg, N., Perez, J., Mendlewicz, J., Janzing, J. G. E., Zobel, A., Skibinska, M., Kozel, D., Stamp, A. S., Bajs, M., … Aitchison, K. J. (2008). Measuring depression: Comparison and integration of three scales in the GENDEP study. *Psychological Medicine*, *38*(2), 289–300. https://doi.org/10.1017/S0033291707001730

Uher, R., Mors, O., Rietschel, M., Rajewska-Rager, A., Petrovic, A., Zobel, A., Henigsberg, N., Mendlewicz, J., Aitchison, K. J., Farmer, A., & McGuffin, P. (2011). Early and delayed onset of response to antidepressants in individual trajectories of change during treatment of major depression: A secondary analysis of data from the genome-based therapeutic drugs for depression (GENDEP) study. *Journal of Clinical Psychiatry*, *72*(11), 1478–1484. https://doi.org/10.4088/JCP.10m06419

Uher, R., Perlis, R. H., Placentino, A., Dernovšek, M. Z., Henigsberg, N., Mors, O., Maier, W., McGuffin, P., & Farmer, A. (2012). Self-report and clinician-rated measures of depression severity: Can one replace the other? *Depression and Anxiety*, *29*(12), 1043–1049. https://doi.org/10.1002/da.21993

Viscovery Software GmbH. (2021). *Viscovery SOMine* (7.2).

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective funccction. *Journal of the American Statistical Association*, *58*(301), 236–244.

Williams, L. M., Debattista, C., Duchemin, A. M., Schatzberg, A. F., & Nemeroff, C. B. (2016). Childhood trauma predicts antidepressant response in adults with major depression: Data from the randomized international study to predict optimized treatment for depression. *Translational Psychiatry*, *6*(5), e799-7. https://doi.org/10.1038/tp.2016.61

Wolpert, D. (1992). Stacked Generalization. *Neural Networks*, *5*, 241–259.

World Health Organization. (1992). *The ICD-10 classification of mental and behavioural disorders: Clinical descriptions and diagnostic guidelines*. World Health Organisation.

World Health Organization. (2017). *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organisation.

Zarate, D., Stavropoulos, V., Ball, M., de Sena Collier, G., & Jacobson, N. C. (2022). Exploring the digital footprint of depression: a PRISMA systematic literature review of the empirical evidence. *BMC Psychiatry*, *22*(1), 1–24. https://doi.org/10.1186/s12888-022-04013-y

Appendix A: Paper III

Rost, N., Binder, E. B., & Brückl, T. M. (2022). Predicting treatment outcome in depression: an introduction into current concepts and challenges. European Archives of Psychiatry and Clinical Neuroscience. https://doi.org/10.1007/s00406-022-01418-4

European Archives of Psychiatry and Clinical Neuroscience
https://doi.org/10.1007/s00406-022-01418-4

ORIGINAL PAPER

Predicting treatment outcome in depression: an introduction into current concepts and challenges

Nicolas Rost[1,2] · **Elisabeth B. Binder**[1] · **Tanja M. Brückl**[1]

Received: 26 November 2021 / Accepted: 11 April 2022

Abstract
Improving response and remission rates in major depressive disorder (MDD) remains an important challenge. Matching patients to the treatment they will most likely respond to should be the ultimate goal. Even though numerous studies have investigated patient-specific indicators of treatment efficacy, no (bio)markers or empirical tests for use in clinical practice have resulted as of now. Therefore, clinical decisions regarding the treatment of MDD still have to be made on the basis of questionnaire- or interview-based assessments and general guidelines without the support of a (laboratory) test. We conducted a narrative review of current approaches to characterize and predict outcome to pharmacological treatments in MDD. We particularly focused on findings from newer computational studies using machine learning and on the resulting implementation into clinical decision support systems. The main issues seem to rest upon the unavailability of robust predictive variables and the lacking application of empirical findings and predictive models in clinical practice. We outline several challenges that need to be tackled on different stages of the translational process, from current concepts and definitions to generalizable prediction models and their successful implementation into digital support systems. By bridging the addressed gaps in translational psychiatric research, advances in data quantity and new technologies may enable the next steps toward precision psychiatry.

Keywords Major depressive disorder · Treatment outcome · Predictive modeling · Clinical decision support system · Precision psychiatry

Introduction

With over 300 million affected people worldwide, depressive disorders have become one of the main causes of disability [1, 2]. Even though there has been an increasing number of studies investigating the optimization of treatment for major depressive disorder (MDD), response rates in patients remain unsatisfactory [3, 4]. In fact, rates have not much improved since the Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study reported in 2006 that only 30% of patients reach the clinical goal of remission, i.e., the absence of symptoms, after the first trial of medication [5]. These numbers need to be taken seriously given the high level of suffering during depressive episodes, the high risk for suicide and comorbidities, and the huge social and economic impact [6, 7]. The question of *what* constitutes the best treatment option for a *specific* patient with a depressive episode under certain individual circumstances is still difficult to answer. Approaches that allow the matching of patients with personalized treatments, often termed 'precision medicine', are widely called for in psychiatry [8, 9]. Particularly in early stages of MDD treatment, it is often unclear whether an individual patient will profit most from pharmacotherapy or if other approaches, such as psychotherapy, brain stimulation, or a combination of treatments, might be more beneficial [10]. Models predicting treatment outcome on the basis of individual baseline characteristics can inform the stratification of patients according to their response chances and consequently, the physician's choice of individualized treatment strategies. In oncology, for example, molecular approaches for tumor characterization have led to the discovery of important subtypes and greatly

✉ Nicolas Rost
nicolas_rost@psych.mpg.de

[1] Department of Translational Research in Psychiatry, Max Planck Institute of Psychiatry, Kraepelinstraße 2-10, 80804 Munich, Germany

[2] International Max Planck Research School for Translational Psychiatry, Munich, Germany

Published online: 19 May 2022

improved individualized treatments [11, 12]. However, in psychiatry, prediction models have not yielded any reliable and valid (bio)markers that are ready for incorporation into clinical tools to support diagnoses or guide treatment decisions (for a review, see [13]). For the treatment of specific psychiatric disorders, such as MDD, mental health professionals can refer to evidence-based, mostly country-specific, guidelines that have been formulated by a committee of experts, such as the American Psychiatric Association [14] or corresponding organization in other countries (e.g., Germany; [15]). These guidelines typically recommend, depending on depression severity, different initial treatment trials as well as a stepwise increase in treatment intensity if initial treatments fail. To some extent, they also take individual patient characteristics into account by adapting treatment recommendations to specific comorbidity or symptom patterns and the patient's prior subjective experience with tolerability and efficacy of certain antidepressants. Standardized approaches in the treatment of MDD, such as guideline- and measurement-based [16] treatments, can help to improve treatment success rates [17]. However, treatment guidelines for MDD are also limited by the non-availability of accurate and validated makers of treatment outcome that are needed for the personalization of treatment. Therefore, treatment administration in MDD is often based on the physician's individual experiences and the patient's personal preferences [18], potentially adding to the low success rates of MDD treatment [19]. With the current lack of personalized treatment, it is more likely that a chosen treatment will be inefficient than efficient for a certain patient [20].

Thus, a better understanding of individual factors contributing to treatment outcome in MDD continues to be a major topic in psychiatry. The present review summarizes definitions of and issues with the current concepts of treatment outcome and provides an introduction into approaches to study and predict antidepressant outcome in MDD. It focuses on clinical implications from these approaches and on implementations into clinical decision support systems.

How is treatment outcome in MDD defined?

In the absence of measurable biological indicators of depression severity, it is important to understand how treatment outcome in MDD is commonly defined and how patients are evaluated based on their rate of recovery.

Changes in symptom severity

In clinical studies, the efficacy of any kind of treatment in MDD as in other psychiatric disorders is routinely assessed with symptom questionnaires, including both clinician-based ratings as well as patient self-ratings. Table 1 summarizes the most typical definitions of treatment outcome based on these ratings. Among the most commonly used scales are the Hamilton Rating Scale for Depression (HDRS; [21]), the Montgomery–Åsberg Depression Rating Scale (MADRS; [22]), the Quick Inventory of Depressive Symptomatology (QIDS; [23]), and the Beck Depression Inventory (BDI; [24]). While the HDRS and MADRS are both clinician-based ratings and require a certain amount of clinical training from the rater [25, 26], the QIDS and BDI are scales based on self-assessments. Even though all these scales were initially created to measure the same construct, i.e., MDD symptom severity, studies have shown that they are not entirely congruent but should rather be used as complementary measures, irrespective of their assessment method [27, 28].

Symptom questionnaires are commonly analyzed by adding up their single items into a sum score. Treatment outcome can then be evaluated by simply interpreting this sum score after a certain length of treatment or by comparing it to a baseline score. However, even though the scales are semiquantitative, binary outcome definitions are widely used, the most common ones being 'response' and 'remission'. Treatment response implies a reduction of symptom severity compared to baseline severity by a certain amount (usually by at least 50%), whereas remission requires symptom scores to drop below a certain threshold (e.g., ≤ 7 on the 17-item HDRS; [29]). Since the concept of response relies on the percentage change in symptom severity, it strongly depends on the baseline score. Remission, on the contrary, does not rely on baseline symptom severity at all. From a clinical perspective, remission is the more desired outcome as remitted patients are generally considered symptom-free and, for the time being, fully recovered. Compared to patients who report residual symptoms after treatment (e.g., response without remission), remitters show a reduced risk of subsequent relapse [30, 31].

If depressive symptoms are continuously measured over time, outcome definitions are not restricted to absolute or relative measures, such as response or remission. Instead, trajectories of symptom development over time can be considered to evaluate treatment success. Many longitudinal studies and clinical trials collect data by applying symptom scales on a weekly basis, which allows outcome definitions built on data from more than one or two timepoints. With this information, more refined interpretations of treatment effects can be made for individual patients. Furthermore, symptom trajectories can be used to identify subgroups of patients with similar outcome patterns but different dynamics in change. With increases in computing power, advances in statistical methods and sufficient sample sizes, such

Table 1 Definitions of treatment outcome in MDD

Concept	Operationalization in studies	Evaluation
Symptom severity	Raw (sum) score derived from a depression severity questionnaire	Continuous measurement Independent from baseline severity As a stand-alone measure (without reference to baseline measure) no clear clinical interpretation
Change in symptom severity	Percentage change or difference of (sum) scores derived from depression severity questionnaire between two time points	Continuous measurement Highly dependent on baseline value
Partial response Small amount of symptom reduction	Reduction of (sum) score on a depression severity questionnaire usually by 25–49% between treatment start and a specific treatment week	Dichotomous outcome measure (yes/no) Dichotomization by arbitrary threshold Early indicator for stable response later Highly dependent on baseline value
Response Considerable amount of symptom reduction	Reduction of (sum) score on a depression severity questionnaire usually by at least 50% between treatment start and a specific treatment week	Dichotomous outcome measure (yes/no) Dichotomization by arbitrary threshold Highly dependent on baseline value Does not necessarily imply that core symptoms have improved
Remission Absence of clinically significant amount of symptoms	Depression severity score below a certain threshold. Cut-off values vary by scale: HDRS-17: ≤ 8 / ≤ 7 (e.g., [5, 15]) HDRS-21: ≤ 8 [118] MADRS: ≤ 6 [15] QIDS-SR: ≤ 5 [5] BDI-II: ≤ 9 [5]	Dichotomous outcome measure (yes/no) Dichotomization by arbitrary threshold Reflects the treatment goal Compares an almost symptom-free group to the remainder group
Symptom trajectories Pattern of symptom severity changes over time	Identification of patient subgroups with distinct patterns of change in symptom severity over duration of treatment; patients with similar pattern are assigned to the same group; methods often based on unsupervised machine learning	Categorical outcome measure Data-driven method of outcome definition No dichotomization or cut-off value needed Dependent on selected variables and scales Heterogeneous methods and algorithms for subgroup identification
Treatment resistance Persistent lack of considerable symptom reduction	Often defined as no significant symptom reduction after at least two adequate antidepressant trials coming from different pharmacological classes; staging approaches	Heterogeneous definitions and staging models (categorical vs. dimensional approaches) Potentially stigmatizing terminology
Functional recovery Recovery in daily functioning beyond symptom reduction	Changes in quality of life and disability measures	Important additional measures that imply a broader understanding of recovery (beyond symptom reduction) Often not assessed in clinical studies Subscales partly not applicable in inpatient settings

HDRS Hamilton Rating Scale for Depression, *MADRS* Montgomery–Åsberg Depression Rating Scale, *QIDS-SR* Quick Inventory of Depressive Symptomatology (Self-Report), *BDI-II* Beck Depression Inventory-II

approaches are becoming more and more prevalent [32–36].

Treatment resistance

In contrast to response and remission, non-response and non-remission can be precursors of so-called 'treatment-resistant depression' (TRD). Definitions of TRD also depend primarily on scores from symptom questionnaires and are mainly focusing on pharmacotherapy. Even though there is no unique definition [37], TRD is most commonly described as a major depressive episode with no response after two or more trials of adequate antidepressant medication coming from different pharmacological classes [38–40]. Still, although this definition seems to be the most prevalent and a useful common ground, many different definitions exist. Some of them vary fundamentally in their criteria, making them difficult to compare [38, 41].

Recovery of cognition and daily functioning

Apart from reduction of symptom severity and failed treatment trials, the desired outcome after a depressive episode also includes other aspects of the patient's recovery. Ideally, patients return to the same (or even a higher) level of well-being as well as to their way of living from before the disorder, including their daily functioning, i.e., their work, social contacts, and general quality of life [42, 43]. This overarching goal of MDD treatment, helping patients to achieve all aspects of recovery, seems to be a stepwise process. For patients with acute moderate or severe episodes, a reduction of symptoms is naturally the first target. Hence, in clinical studies, especially in inpatient settings, symptom severity is more commonly measured than levels of functioning and positive affect [44], the assessments of which are not necessarily well-suited for routine use [19].

Nevertheless, restoration of daily functioning and positive affect are important factors of a holistic picture of recovery. Any potentially impaired cognitive abilities, such as attention, learning, memory, and executive functions [45], should improve, as should components of positive affect, such as optimism and self-confidence [46]. Whereas cognition is routinely assessed using different neurocognitive tests or batteries [47], functional aspects are less well defined [19]. Still, numerous scales and questionnaires with varying foci exist, including the Global Assessment of Functioning [48], the Quality of Life Enjoyment and Satisfaction Questionnaire [49], and the World Health Organization Disability Assessment Schedule [50].

Prediction models of treatment outcome in MDD

The endeavor of finding indicators of treatment efficacy in MDD has led to a remarkable amount of publications from different psychiatric subfields. A large subset of these have looked at associations of preselected psychological and biological factors with treatment outcome. The main aim hereby was the identification of new (bio)markers using classical statistical approaches, such as regression models with null hypothesis significance testing based on p-values of the investigated predictors. The results from these association studies have been summarized in several systematic reviews and meta-analyses, often focusing on selected data modalities (but see [51, 52]), such as sociodemographic and clinical measures [53], cognitive functioning [54], or blood biomarkers [55]. Table 2 provides a list of these publications grouped by data modality and by their ease of access in clinical practice. Overall, the most consistently identified and most predictive factors were derived from sociodemographic and clinical characteristics [19]. Information on a patient's social support, their baseline symptom severity, psychiatric comorbidities (e.g., anxiety disorders), or chronicity of the disorder, for instance, have repeatedly been associated with MDD treatment outcome [51–53]. However, an important shortcoming of these results is that none of the identified measures has been proven informative enough to sufficiently predict treatment outcome on their own.

This issue has led to a "new generation" of studies which aim at creating prediction models based on a multitude of variables. These models use machine learning (ML) methods, mainly supervised learning with classification algorithms such as regularized logistic regression or tree-based methods [56], to combine the effects of many variables and to increase predictive accuracy. Hence, they do not necessarily focus on the identification of new predictors of treatment outcome but rather try to find the best combination of variables to maximize their predictive power. A clear and comprehensive review on ML models and their value for predicting treatment outcome in psychiatry was recently published [57], as well as a systematic review and meta-analysis of these approaches in MDD specifically [58]. Crucially, the development of such models needs to include some kind of validation in order assure that predictions are not specific to the data they were created from but also generalize to new data. Validation is often performed by dividing the initial data set into subsamples (e.g., training sample and validation sample) or by testing the model's performance on a completely independent sample [59]. Furthermore, sufficiently large data sets in terms of sample size are required

Table 2 Different measurement techniques used in psychiatric research and corresponding examples of derived factors associated with antidepressant treatment effects

Measurement technique	Requirements	Example indicators of antidepressant treatment outcome
Easily accessible and usable		
Questionnaires and clinician-based ratings or interviews	Manuals (Clinical training)	Social demographics [53, 119, 120] Symptom profiles [53, 119, 120] Comorbidities [121, 122] Personality traits [123] Exposure to environmental risk factors, e.g., childhood abuse [124–126]
Tests and tasks	Manuals Technical devices for digital implementations	Cognitive functioning [54] Emotional processing [127, 128]
Technically feasible but additional efforts and expenses needed		
Blood draw or saliva sampling for established parameters	Medical training and equipment Laboratory capacities	Immune parameters, e.g., cytokines [129–131] Metabolites [132] Pharmacogenomic testing [133–135]
Dynamic function tests	Medical training and equipment Laboratory capacities	HPA-axis regulation [136, 137]
Technically feasible but high complexity and expenses		
Genotyping pipelines (based on blood draw or saliva sampling or other biospecimen)	Medical training and equipment Laboratory capacities Computational expertise and resources	Candidate genes without established testing [138, 139] Genome-wide associations [140–145] Polygenic risk scores [146] Epigenetic, transcriptomic and metabolomic markers [147–151]
Technical recording devices	Special equipment Technical training and expertise	Neuroimaging [152–155] Electroencephalographic markers [156, 157] Peripheral physiological markers [158]

Measurements are grouped by their accessibility and usability for routine clinical practice and licensed physicians. Note that this table is neither exhaustive nor based on a systematic literature search but meant to show exemplary indicators and their translational value

to guarantee robustness and generalizability of the predictions. The majority of predictive ML models of MDD treatment outcome have thus been created on data from large patient cohorts coming either from clinical trials (such as STAR*D [60, 61], Genome-based Therapeutic Drugs for Depression [62, 63], or Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care in Depression [64]), or from observational studies (such as the Munich Antidepressant Response Signature project [32] or the Netherlands Study of Depression and Anxiety (NESDA) [65, 66]). Since clinical trials usually compare different treatment arms (or treatment against placebo), the resulting predictions are likely be treatment-specific and may not be readily applied to other treatments [60, 62]. Observational studies, on the other hand, follow a more naturalistic approach by observing patients who are treated based on routine clinical decisions, which might lead to more heterogeneity in the data [67, 68]. In general, prediction models of MDD treatment outcome based on sample sizes of at least several hundred patients (e.g., [60–63]) can predict treatment outcome (most often response vs. non-response or remission vs. non-remission) with moderate to good accuracies of 65%–75% [58]. This means that up to three quarters of 'true' responders/remitters are recognized as such by these prediction models. Most models that have been published so far have confirmed that the most reliable predictors of MDD treatment outcome come from established clinical and sociodemographic factors that had already been identified in earlier studies, such as initial symptom severity (e.g., [32, 36, 60, 62]), number and duration of depressive episodes (e.g., [32, 60]), personality traits (e.g., [32, 66]), as well as employment status and education (e.g., [61, 66]). However, only few studies exist that have assessed the additional value of other data modalities by comparing the performance of a multimodal model to a model using sociodemographic or clinical variables only. We here provide two examples of studies that have followed this approach using large sample sizes (at least several hundred samples) and ML methods. Iniesta et al. [63] showed that a prediction model combining demographic and clinical variables (e.g., depressive symptom scores, medication status, and stressful life events) with over 500,000 genetic markers (single nucleotide polymorphisms and copy number variants) led to slightly more accurate predictions (area under the receiver operating characteristic curve (AUC)

of 0.77) than a model trained on the non-genetic variables only (AUC of 0.74; [62]). Similarly, Dinga et al. [66] compared a prediction model combining clinical and biological data (primarily somatic health measures, inflammatory and metabolic markers) to models including only one of the available predictor domains. Across all comparisons, the full model containing all variables performed better than the alternative models. The largest differences occurred when the alternative model was based on biological measures only, the smallest differences when it was based on depressive symptom severity scores (differences in AUC of 0.01–0.05). These results suggest that even though adding biological markers to prediction models can lead to increases in performance, their additional value on top of clinical data still remains small.

Clinical decision support systems in psychiatry

A suitable instrument to transfer predictive models from research into clinical practice is a Clinical Decision Support System (CDSS). CDSSs are any kinds of computer systems that work with clinical data or knowledge and are set up to assist healthcare professionals in decision processes [69]. These decisions can refer to both diagnosing a patient and selecting the best treatment [70]. Concretely, a patient's characteristics enter a CDSS to be evaluated based on implemented clinical knowledge in order to return recommendations to the clinicians [71]. Hence, these systems can improve clinical processes and help healthcare professionals benefit from scientific findings [72].

CDSSs have been used successfully in many medical disciplines (for a review, see [73]), but use in psychiatry or mental health is lagging behind. However, some systems have been developed for the diagnoses of mental disorders, e.g., for attention deficit hyperactivity disorder [74], MDD and anxiety disorders [75], subtypes of schizophrenia [76], or a broader range of disorders [77]. Other systems were designed more specifically and can also be of value for MDD, such as the NetDSS [78], a web-based CDSS with various functions, from patient registry to clinical outcome monitoring. An elegant tool for physicians and patients was set up by Henshall et al. [79]. They developed a recommendation system and tested it on a focus group comprising physicians, caregivers, and patients with several mental disorders, including MDD. By entering basic sociodemographic and clinical variables as well as by setting preferences for potential side effects, the software returned a graphical illustration of recommended interventions and their corresponding probabilities of effectiveness. A benefit of such a tool is that it uses individual data to tailor a treatment to each patient. Similarly, a few commercial tools have been developed lately, promoting improvements of treatment efficacy for mental disorders using individual patient data and predictive models [80–82].

Ultimately, such predictive systems can enhance personalized treatment, e.g., by indicating from the beginning which medication has the highest probability to lead to a beneficial response. Moreover, these tools can save physicians time and increase preciseness of clinical judgements [83, 84].

Current challenges and unmet needs

With the increasing interest in precision psychiatry and outcome prognosis, many efforts have been invested in this field of research. Nonetheless, the core problem in translational psychiatry remains: translations of research findings into daily clinical work, in such a way that patients and clinicians could directly benefit from them, are practically non-existent. Due to the lack of validated tests as guidance for personalized medication, treatment administration still has to rely on generic guidelines and physicians' personal judgements. The potential solution appears to be twofold: first, robust (bio)markers of treatment efficacy need to be identified and built into prognostic models. Subsequently, if models are proven useful, the second step will be their translation into new tools for clinicians. The main issues and current challenges in this translational process as well as potential solution approaches are outlined below. Additionally, they are illustrated in Fig. 1.

Challenges in concepts and definitions

Up to 16,400 potential symptom combinations can lead to a diagnosis of MDD [85], which might essentially be a conglomerate of many different pathophysiologies [86]. Moreover, MDD shows a high degree of comorbidity with other mental disorders, both cross-sectionally [87–89] and over time [90]. Longitudinal studies, especially using registry data [91], have shown large variability of diagnoses across lifetime which is why a cross-sectional focus on MDD diagnosis might miss relevant longitudinal information that discriminates among disorder subtypes. Hence, transdiagnostic and longitudinal approaches (e.g., assessing lifetime disorders in diagnostic interviews) should be considered in clinical studies.

A second challenge is posed by the measurements and definitions of antidepressant outcome (see Table 1). Unlike other medical disciplines, which provide objective biological measures of disease severity or treatment success, psychiatry defines clinical outcomes on subjective ratings (self-reported or clinician-rated). However, some of the

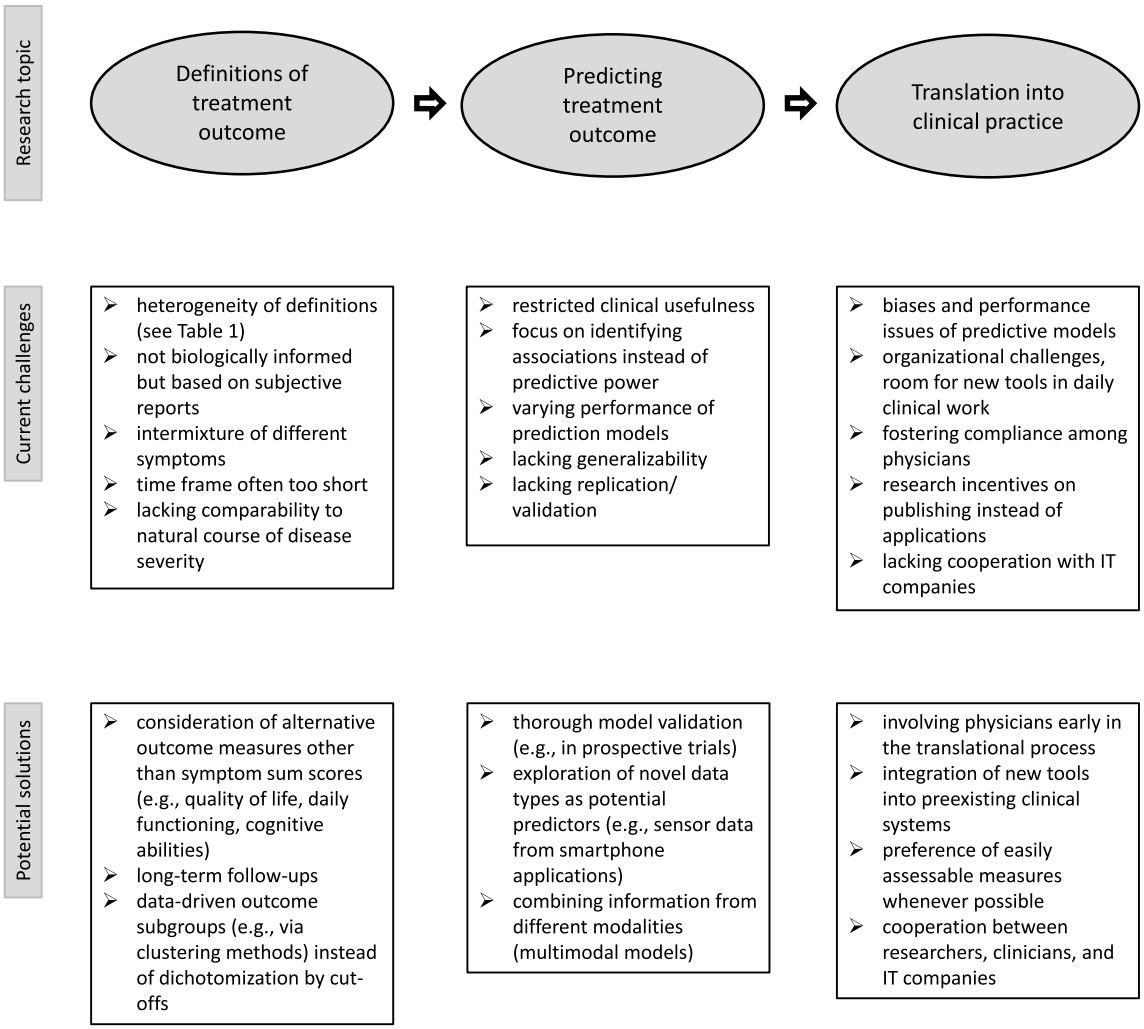


Fig. 1 Current challenges with respect to different stages of research on treatment outcome in MDD patients and its translation into clinical practice

most common ratings were shown to lack reliability [27, 92, 93] and to be incongruent among themselves, meaning that they do not measure exactly the same construct and are thus not fully comparable [28]. These issues limit the validity of findings and the generalizability from one outcome scale to others. Moreover, ratings of depressive symptom severity, such as the HDRS, the QIDS, or the BDI, evaluate many different symptoms and aspects of MDD, all influencing the respective sum score. It is possible for patients to show a 50% reduction of the sum score and be classified as responders, even when none of the core symptoms of MDD (depressed mood or reduced interest/pleasure in activities) have improved. Furthermore, patients with the same overall severity score can show very different symptom profiles, and have thus very different subjective experiences of their disorder. This important information gets lost when sum score data are used [94]. Explicitly differentiating between symptoms instead of using sum scores could help to identify indicators of specific symptoms and could thus lead toward more targeted treatments [95].

Moreover, antidepressant outcome is often defined as (partial) response or remission (see Table 1). Both terms represent artificially dichotomized variables, created based on more or less arbitrary cut-off values on a continuous scale, that is, the respective sum score (for remission) or the difference in sum scores (for response) on a symptom scale. Dichotomizing continuous variables always brings certain risks and comes with loss of information [96]. Consider two patients with very similar symptom scores during the course of treatment, e.g., symptom reductions of 55% and 45%, respectively. According to the common definition of treatment response, the first patient would be classified as a 'responder' whereas the second patient would be treated as a 'non-responder'. In fact, the second patient would be categorized together with patients who do not show any symptom reduction at all. Classifying patients in a data-driven manner,

e.g., using clustering techniques to create more homogenous outcome classes, might be a promising alternative that has already been implemented in several studies [32–35]. Still, the resulting outcome groups strongly depended on the selected variables and the chosen clustering method. Hence, the number of identified groups varied, e.g., from five [33] to seven [32, 34] up to nine [35]. These discrepancies challenge their clinical usefulness as the obtained classes are likely not generalizable to most other settings. Nevertheless, especially if more than one type of outcome measure is available, clustering methods might be a good way to combine information and identify subgroups.

Another issue with common measurements of treatment efficacy is the time frame. Patients in clinical trials are often measured over a few weeks only. Especially in disorders such as MDD, which can appear recurrently and show a risk of chronification [97], it is important to follow up on patients after a longer period of time. This could help differentiate between temporary improvements and long-term recovery. In the NESDA sample, 22% of initially remitted patients developed a recurrent episode within the following 2 years [98]. Identifying these at-risk patients early on might help to prevent subsequent episodes by scheduling regular checkups and implementing prevention strategies [99].

Even in the absence of reliable (biological) alternatives, sum scores on symptom questionnaires alone do not seem to be the most specific and clinically meaningful measures [95, 100]. In a recent online survey, MDD patients, informal caregivers, and healthcare professionals were asked to indicate outcome domains that matter most in their opinion. They identified not only depressive symptoms but also domains of functioning, healthcare organization, and social representation, many of which are not measured in most clinical studies, let alone included in depression rating scales [44, 101], highlighting the importance of including patient centered outcomes. Another research team explicitly differentiated between opinions from doctors and patients [102]. Their survey revealed that physicians mainly considered alleviation of depressive symptoms to be most important for relief and cure from MDD whereas patients rather focused on rehabilitation of positive affect. These results suggest that definitions and measures of treatment outcome should go beyond plain ratings of symptom changes and need to be broadened and potentially lengthened [42]. Relevant assessment instruments for many different domains of MDD characterization, including neurocognition, functioning and quality of life, as well as their suitability for routine clinical use have recently been reviewed [19] and should be considered when measuring treatment outcome in future studies.

Finally, novel objective measures that do not rely on subjective self- or external reports, such as behavioral and functional data generated by smartphones, wearables or other digital devices, could be of further value [103]. As long as no direct biological measure of treatment outcome exists, personal data collected from mobile devices, i.e., 'digital phenotyping', might become a promising alternative [104]. Ecological momentary assessments, actimetry, speech characteristics, or movement patterns, for instance, can be continuously and mainly passively collected in large amounts and in high temporal resolution. Sensor data and other information from wearable devices like smartphones have already been successfully applied in psychiatric research, especially in combination with ML and deep learning [103]. Future studies will need to prove if they can contribute to a deeper and broader characterization of treatment outcome and MDD.

Challenges for prediction models

Except for a few psychometric and sociodemographic factors, there are still no robust or well replicated predictors of treatment outcome. Apart from a few promising pharmacogenetic tests [81, 105], no biological measures qualify as stable biomarkers nor are they used in clinical practice. Associations between specific measurements and treatment outcome are often of limited prognostic value as statistically significant associations do not guarantee accurate and robust predictions. Therefore, the focus has started to shift from testing associations to improving predictions in order to forecast what is most beneficial for an individual patient and to personalize clinical decision-making [106].

Predictive ML models tackle this issue as they are built to be as accurate and robust as possible. The robustness of a model should be assessed by validating it on an independent data set [57], ideally by testing its performance and safety on new patients in a prospective clinical study. Nonetheless, several prediction models were not validated on external data sets at all (e.g., [32, 64, 65]). Others were less predictive when they were applied to other classes of antidepressants, suggesting that the identified predictors of treatment outcome might be agent-specific [60, 62]. In addition, the main target variables in studies using ML were response and remission in their binary form [58], the downsides of which have already been discussed. Furthermore, psychiatric data often face the problem of high dimensionality while samples sizes remain relatively small [107]. This is often referred to as the 'curse of dimensionality': the more variables a data set contains, the more the sample size needs to increase (per variable) to allow reliable results [59]. Otherwise, resulting prediction models are likely to be biased and therefore need to be carefully validated on independent data to ensure their reliability. Moreover, prediction models based on biological data often only show restricted translatability into clinical practice as they require precisely preprocessed data from time-consuming and expensive measurements. A prerequisite for a successful translation of a predictive model into

clinical practice is that it consists of parameters that can be routinely accessed by a licensed physician without producing a lot of extra costs. Psychological and clinical features as well as sociodemographic information can be evaluated easily by any trained clinician or via self-ratings. On the other hand, as indicated in Table 2, many biological measures, i.e., potential biomarkers, are comparatively expensive or hard to assess for physicians in common clinical settings. This is especially the case for neuroimaging, omics data, and endocrinological markers derived from a challenge test, for instance. Such parameters should only be preferred over less costly data modalities, e.g., questionnaire data, if their predictive performance is notably higher and thus justifies the additional expenses. Making use of other objective measures, such as data collected from smartphones and other wearable devices, might become a promising alternative [103]. Their collection would be economical and profitable for researchers as well as less time-consuming and free of stress for patients.

In summary, well-performing and externally validated ML models are promising tools for future psychiatric practice [59], including the prognosis of treatment outcome in MDD.

Challenges for CDSSs

In order to translate predictive models into digital tools for everyday clinical use, CDSSs could be of help. Iniesta et al. [108] sketched a concise outline of the workflow for designing and choosing predictive models and, crucially, explained how to bring them into CDSSs. Still, as appealing as the idea of such publicly used tools might sound, they have not yet become prevalent in healthcare institutions.

The main challenge in MDD outcome prediction seems to be the lack of powerful models and established predictive patient characteristics. As outlined above, predictive models are still not robust and generalizable enough to guide daily clinical decisions. Only if additional value coming from a predictive model is proven, will an implementation into a CDSSs lead to a successful supporting device. Biases in such systems, for instance, were shown to lead to underestimations of their effectiveness [109], high non-compliance rates among users [73], and even to wrong diagnoses by physicians [110]. This is particularly concerning given that working with a CDSS might influence clinicians in their decisions later on even when they are not explicitly using the system anymore [111].

Furthermore, before CDSSs can be fully implemented into clinical workflows, substantial ethical challenges need to be considered. Apart from data protection, which needs to be assured, questions regarding liability and responsibility for treatment decisions have to be addressed, especially when it comes to disagreement between physicians and support systems. Also, human interactions, conversations and relations between patients and mental health professionals play an important role, not only in psychiatric care [112, 113]. Further necessary ethical considerations have been summarized by Chekroud et al. [57].

Due to these problems, a number of factors needed to sustainably establish CDSSs in clinical settings should be considered [73]: First, apart from having appealing visual designs and being user-friendly, the system should implement personalized, transparent, and reliable recommendations as well as comprehensive overviews for each patient. Second, physicians should keep the authority over treatment decisions and should still oversee algorithmic outputs [114]. They should be involved in the development of the system, receive training and not have to make adaptations in their daily working processes in order to use the application. Third, to circumvent organizational obstacles, CDSSs should be integrated into preexisting clinical computerized systems, such as electronic medical records or physician order entries [73].

Ultimately, however, the main incentive in research seems to remain the publication of novel findings, indeed funding for the translation of existing findings into applications and technical devices is often more difficult to obtain [115, 116]. Therefore, interdisciplinary work is needed, bringing together scientists, clinicians and, e.g., information technologists for successful development of CDSSs.

Conclusion

Tackling the medical treatment of MDD and increasing treatment efficacy have always been major challenges in psychiatric research. In this narrative review, we summarized current approaches to operationalize and predict treatment outcome in MDD. We highlighted findings from ML approaches and discussed their implementation into CDSSs. To date, numerous studies have investigated and discovered associations between biological and phenotypic patient characteristics and treatment outcome, producing growing evidence for potential underlying mechanisms. Large patient cohort data and ML methods have additionally produced predictive models with promising accuracies (e.g., [32, 36, 60, 62, 64, 65]). Nevertheless, psychiatry has made comparatively little progress in applying the acquired knowledge into daily clinical work and in personalizing decisions based on empirically derived patient characteristics.

The main issue of this lacking translation seems to be the absence of robust and generalizable predictors of treatment outcome, especially of biological and other objectively measurable markers. Further quantitative characterizations of patients might help to identify more robust predictors and could provide support in medical decisions, such as choosing

the most beneficial treatment for individual patients or subgroups of patients [117]. Once reliable indicators and prognostic models are established, the next challenge will be their implementation into clinical practice. Efficient systems with clear interpretation of results need to be introduced and made available for healthcare professionals. CDSSs can be useful tools to implement tests and predictive models to guarantee benefits for physicians and patients. To make this happen, research funding needs to put more emphasis on translational systems, i.e., the development of target-oriented and clinically useful applications. Cooperation with companies specialized in health information technologies might be of particular use for this endeavor. Finally, there needs to be a shift in psychiatry toward a data-driven stratification of patients as well as more precise, personalized treatments based on individual patient data.

Acknowledgements We thank Dr. Jessica Keverne for proof-reading the article.

Author contributions NR wrote the initial draft of the manuscript. EBB and TMB critically contributed to the writing of the manuscript and its revisions. All authors contributed to and have approved the final manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. This publication is funded by the Max Planck Institute of Psychiatry. NR is supported by the International Max Planck Research School of Translational Psychiatry (IMPRS-TP) and received funding from the Bavarian Ministry of Economic Affairs, Regional Development and Energy (BayMED, PBN_MED-1711–0003).

Declarations

Conflict of interest EBB is an editor of the journal *European Archives of Psychiatry and Clinical Neuroscience*. Otherwise, the authors have no financial or non-financial competing interests to declare that are relevant to the content of this article.

Employment All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

References

1. World Health Organization (2017) Depression and other common mental disorders: global health estimates. World Health Organisation, Geneva
2. GBD (2016) Disease and Injury Incidence and Prevalence Collaborators (2017) Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 390:1211–1259. https://doi.org/10.1016/S0140-6736(17)32154-2
3. Thomas L, Kessler D, Campbell J et al (2013) Prevalence of treatment-resistant depression in primary care: cross-sectional data. Br J Gen Pract 63:852–858. https://doi.org/10.3399/bjgp13X675430
4. Khan A, Fahl Mar K, Faucett J et al (2017) Has the rising placebo response impacted antidepressant clinical trial outcome? Data from the US food and drug administration 1987–2013. World Psychiatry 16:181–192. https://doi.org/10.1002/wps.20421
5. Trivedi MH, Rush AJ, Wisniewski SR et al (2006) Evaluation of outcomes with citalopram for depression using measurement-based care in STAR*D: implications for clinical practice. Am J Psychiatry 163:28–40. https://doi.org/10.1176/appi.ajp.163.1.28
6. Bingham KS, Rothschild AJ, Mulsant BH, et al (2017) The Association of Baseline Suicidality With Treatment Outcome in Psychotic Depression. J Clin Psychiatry 78:1149–1154. https://doi.org/10.4088/JCP.14m09658
7. Greenberg PE, Fournier AA, Sisitsky T et al (2015) The economic burden of adults with major depressive disorder in the United States (2005 and 2010). J Clin Psychiatry 76:155–162. https://doi.org/10.4088/JCP.14m09298
8. Cohen ZD, DeRubeis RJ (2018) Treatment selection in depression. Annu Rev Clin Psychol 14:209–236. https://doi.org/10.1146/annurev-clinpsy-050817-084746
9. Friston KJ, Redish AD, Gordon JA (2017) Computational nosology and precision psychiatry. Comput psychiatry (Cambridge, Mass) 1:2–23. https://doi.org/10.1162/CPSY_a_00001
10. DeRubeis RJ, Siegle GJ, Hollon SD (2008) Cognitive therapy versus medication for depression: treatment outcomes and neural mechanisms. Nat Rev Neurosci 9:788–796. https://doi.org/10.1038/nrn2345
11. Collins FS, Varmus H (2015) A new initiative on precision medicine. N Engl J Med 372:793–795. https://doi.org/10.1056/NEJMp1002530
12. National Research Council (2011) Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. The National Academies Press, Washington, D.C.
13. Kraus C, Kadriu B, Lanzenberger R, et al (2019) Prognosis and improved outcomes in major depression: a review. Transl Psychiatry 9:. https://doi.org/10.1038/s41398-019-0460-3
14. American Psychiatric Association (2000) Practice guideline for the treatment of patients with major depressive disorder (revision). Am J Psychiatry 157:1–45
15. DGPPN, BÄK, KBV, et al (2015) S3-Leitlinie/Nationale VersorgungsLeitlinie Unipolare Depression–Langfassung, 1. Auflage. Version 5
16. Hong RH, Murphy JK, Michalak EE et al (2021) Implementing measurement-based care for depression: practical solutions for psychiatrists and primary care physicians. Neuropsychiatr Dis Treat 17:79–90. https://doi.org/10.2147/NDT.S283731
17. Härter M, Watzke B, Daubmann A et al (2018) Guideline-based stepped and collaborative care for patients with depression in a cluster-randomised trial. Sci Rep 8:1–9. https://doi.org/10.1038/s41598-018-27470-6

18. McHugh RK, Whitton SW, Peckham AD et al (2013) Patient preference for psychological vs. pharmacological treatment of psychiatric disorders: a meta-analytic review. J Clin Psychiatry 74:595–602. https://doi.org/10.4088/JCP.12r07757.Patient
19. Maj M, Stein DJ, Parker G et al (2020) The clinical characterization of the adult patient with depression aimed at personalization of management. World Psychiatry 19:269–293. https://doi.org/10.1002/wps.20771
20. Malhi GS, Das P, Mannie Z, Irwin L (2019) Treatment-resistant depression: problematic illness or a problem in our approach? Br J Psychiatry 214:1–3. https://doi.org/10.1192/bjp.2018.246
21. Hamilton M (1960) A rating scale for depression. J Neurol Neurosurg Psychiatry 23:56–62
22. Montgomery SA, Åsberg M (1979) A new depression scale designed to be sensitive to change. Br J Psychiatry 134:382–389. https://doi.org/10.1192/bjp.134.4.382
23. Rush AJ, Trivedi MH, Ibrahim HM et al (2003) The 16-item Quick inventory of depressive symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 54:573–583. https://doi.org/10.1016/S0006-3223(02)01866-8
24. Beck AT, Ward C, Mendelson M et al (1961) Beck depression inventory (BDI). Arch Gen Psychiatry 4:561–571. https://doi.org/10.1093/ndt/gfr086
25. Williams JBW, Kobak KA (2008) Development and reliability of a structured interview guide for the Montgomery-Åsberg Depression Rating Scale (SIGMA). Br J Psychiatry 192:52–58. https://doi.org/10.1192/bjp.bp.106.032532
26. Hooijer C, Zitman FG, Griez E et al (1991) The Hamilton Depression Rating Scale (HDRS): changes in scores as a function of training and version used. J Affect Disord 22:21–29. https://doi.org/10.1016/0165-0327(91)90079-8
27. Uher R, Farmer A, Maier W et al (2008) Measuring depression: comparison and integration of three scales in the GENDEP study. Psychol Med 38:289–300. https://doi.org/10.1017/S0033291707001730
28. Uher R, Perlis RH, Placentino A et al (2012) Self-report and clinician-rated measures of depression severity: Can one replace the other? Depress Anxiety 29:1043–1049. https://doi.org/10.1002/da.21993
29. Rush AJ, Kraemer HC, Sackeim HA et al (2006) Report by the ACNP Task Force on response and remission in major depressive disorder. Neuropsychopharmacology 31:1841–1853. https://doi.org/10.1038/sj.npp.1301131
30. Paykel ES, Ramana R, Cooper Z et al (1995) Residual symptoms after partial remission: an important outcome in depression. Psychol Med 25:1171–1180. https://doi.org/10.1017/S0033291700033146
31. Thase ME (2003) Achieving remission and managing relapse in depression. J Clin Psychiatry 64(Suppl 1):3–7
32. Paul R, Andlauer TFM, Czamara D, et al (2019) Treatment response classes in major depressive disorder identified by model-based clustering and validated by clinical prediction models. Transl Psychiatry 9:. https://doi.org/10.1038/s41398-019-0524-4
33. Kelley ME, Dunlop BW, Nemeroff CB et al (2018) Response rate profiles for major depressive disorder: characterizing early response and longitudinal nonresponse. Depress Anxiety 35:992–1000. https://doi.org/10.1002/da.22832
34. Hartmann A, von Wietersheim J, Weiss H, Zeeck A (2018) Patterns of symptom change in major depression: classification and clustering of long term courses. Psychiatry Res 267:480–489. https://doi.org/10.1016/j.psychres.2018.03.086
35. Uher R, Mors O, Rietschel M et al (2011) Early and delayed onset of response to antidepressants in individual trajectories of change during treatment of major depression: a secondary analysis of data from the genome-based therapeutic drugs for depression (GENDEP) study. J Clin Psychiatry 72:1478–1484. https://doi.org/10.4088/JCP.10m06419
36. Athreya AP, Brückl T, Binder EB et al (2021) Prediction of short-term antidepressant response using probabilistic graphical models with replication across multiple drugs and treatment settings. Neuropsychopharmacology. https://doi.org/10.1038/s41386-020-00943-x
37. McIntyre RS, Filteau MJ, Martin L et al (2014) Treatment-resistant depression: definitions, review of the evidence, and algorithmic approach. J Affect Disord 156:1–7. https://doi.org/10.1016/j.jad.2013.10.043
38. Berlim MT, Turecki G (2007) Definition, assessment, and staging of treatment-resistant refractory major depression: a review of current concepts and methods. Can J Psychiatry 52:46–54
39. Souery D, Amsterdam J, De Montigny C et al (1999) Treatment resistant depression: methodological overview and operational criteria. Eur Neuropsychopharmacol 9:83–91. https://doi.org/10.1016/S0924-977X(98)00004-2
40. Anderson IM (2018) We all know what we mean by treatment-resistant depression—Don't we? Br J Psychiatry 212:259–261. https://doi.org/10.1192/bjp.2018.56
41. Conway CR, George MS, Sackeim HA (2017) Toward an evidence-based, operational definition of treatment-resistant depression: When enough is enough. JAMA Psychiat 74:9–10. https://doi.org/10.1001/jamapsychiatry.2016.2586
42. Slofstra C, Booij SH, Rogier Hoenders HJ, Castelein S (2019) Redefining therapeutic outcomes of depression treatment. J Pers Res 5:115–122. https://doi.org/10.17505/jpor.2019.10
43. Ishak WW, Greenberg JM, Balayan K et al (2011) Quality of life: the ultimate outcome measure of interventions in major depressive disorder. Harv Rev Psychiatry 19:229–239. https://doi.org/10.3109/10673229.2011.614099
44. McKnight PE, Kashdan TB (2009) The importance of functional impairment to mental health outcomes: a case for reassessing our goals in depression treatment research. Clin Psychol Rev 29:243–259. https://doi.org/10.1016/j.cpr.2009.01.005
45. Lee RSC, Hermens DF, Porter MA, Redoblado-Hodge MA (2012) A meta-analysis of cognitive deficits in first-episode Major Depressive Disorder. J Affect Disord 140:113–124. https://doi.org/10.1016/j.jad.2011.10.023
46. Zimmerman M, McGlinchey JB, Posternak MA et al (2006) How should remission from depression be defined? The depressed patient's perspective. Am J Psychiatry 163:148–150
47. McIntyre RS, Cha DS, Soczynska JK et al (2013) Cognitive deficits and functional outcomes in major depressive disorder: determinants, substrates, and treatment interventions. Depress Anxiety 30:515–527. https://doi.org/10.1002/da.22063
48. American Psychiatric Association (1994) Diagnostic and statistical manual of mental disorders, 4th edn. American Psychiatric Association, Washington, DC
49. Endicott J, Nee J, Harrison W, Blumenthal R (1993) Quality of life enjoyment and satisfaction questionnaire: a new measure. Psychopharmacol Bull 29:321–326
50. Üstün TB, Chatterji S, Kostanjsek N et al (2010) Developing the World Health Organization disability assessment schedule 2.0. Bull World Health Organ 88:815–823. https://doi.org/10.2471/BLT.09.067231
51. Bennabi D, Aouizerate B, El-Hage W et al (2015) Risk factors for treatment resistance in unipolar depression: a systematic review. J Affect Disord 171:137–141. https://doi.org/10.1016/j.jad.2014.09.020

52. Perlman K, Benrimoh D, Israel S et al (2019) A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. J Affect Disord 243:503–515. https://doi.org/10.1016/j.jad.2018.09.067
53. De Carlo V, Calati R, Serretti A (2016) Socio-demographic and clinical predictors of non-response/non-remission in treatment resistant depressed patients: a systematic review. Psychiatry Res 240:421–430. https://doi.org/10.1016/j.psychres.2016.04.034
54. Pimontel MA, Rindskopf D, Rutherford BR et al (2016) A meta-analysis of executive dysfunction and antidepressant treatment response in late-life depression. Am J Geriatr Psychiatry 24:31–41. https://doi.org/10.1161/CIRCULATIONAHA.114.010270. Hospital
55. Mora C, Zonca V, Riva MA, Cattaneo A (2018) Blood biomarkers and treatment response in major depression. Expert Rev Mol Diagn 18:513–529. https://doi.org/10.1080/14737159.2018.1470927
56. Hastie T, Tibshirani R, Friedman J (2009) The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edi. Springer, New York, NY
57. Chekroud AM, Bondar J, Delgadillo J et al (2021) The promise of machine learning in predicting treatment outcomes in psychiatry. World Psychiatry 20:154–170. https://doi.org/10.1002/wps.20882
58. Lee Y, Ragguett RM, Mansur RB et al (2018) Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. J Affect Disord 241:519–532. https://doi.org/10.1016/j.jad.2018.08.073
59. Dwyer DB, Falkai P, Koutsouleris N (2018) Machine learning approaches for clinical psychology and psychiatry. Annu Rev Clin Psychol 14:91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037
60. Chekroud AM, Zotti RJ, Shehzad Z et al (2016) Cross-trial prediction of treatment outcome in depression: a machine learning approach. The Lancet Psychiatry 3:243–250. https://doi.org/10.1016/S2215-0366(15)00471-X
61. Jain FA, Hunter AM, Brooks JO, Leuchter AF (2013) Predictive socioeconomic and clinical profiles of antidepressant response and remission. Depress Anxiety 30:624–630. https://doi.org/10.1002/da.22045
62. Iniesta R, Malki K, Maier W et al (2016) Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J Psychiatr Res 78:94–102. https://doi.org/10.1016/j.jpsychires.2016.03.016
63. Iniesta R, Hodgson K, Stahl D et al (2018) Antidepressant drug-specific prediction of depression treatment outcomes from genetic and clinical variables. Sci Rep 8:1–9. https://doi.org/10.1038/s41598-018-23584-z
64. Wu W, Zhang Y, Jiang J et al (2020) An electroencephalographic signature predicts antidepressant response in major depression. Nat Biotechnol 38:439–447. https://doi.org/10.1038/s41587-019-0397-3
65. Frässle S, Marquand AF, Schmaal L et al (2020) Predicting individual clinical trajectories of depression with generative embedding. NeuroImage Clin 26:102213. https://doi.org/10.1016/j.nicl.2020.102213
66. Dinga R, Marquand AF, Veltman DJ et al (2018) Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. Transl Psychiatry 8:241. https://doi.org/10.1038/s41398-018-0289-1
67. Ross JS (2014) Randomized clinical trials and observational studies are more often alike than unlike. JAMA Intern Med 174:1557. https://doi.org/10.1001/jamainternmed.2014.3366
68. Webb CA, Cohen ZD, Beard C et al (2020) Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: a comparison of machine learning approaches. J Consult Clin Psychol 88:25–38. https://doi.org/10.1037/ccp0000451
69. Kemppinen J, Korpela J, Elfvengren K, et al (2014) Decision Support in Evaluating the Impacts of Mental Disorders on Work Ability. 2014 47th Hawaii Int Conf Syst Sci 2958–2966. https://doi.org/10.1109/HICSS.2014.368
70. Musen MA, Middleton B, Greenes RA (2014) Clinical decision-support systems. In: Shortliffe E, Cimino J (eds) Biomedical informatics. Springer, London, pp 643–674
71. Sim I, Gorman P, Greenes RA et al (2001) Clinical decision support systems for the practice of evidence-based medicine. J Am Med Assoc 8:527–534
72. Bright TJ, Wong A, Dhurjati R et al (2012) Effect of clinical decision-support systems: a systematic review. Ann Intern Med 157:29–43. https://doi.org/10.7326/0003-4819-157-1-201207030-00450
73. Kilsdonk E, Peute LW, Jaspers MWM (2017) Factors influencing implementation success of guideline-based clinical decision support systems: a systematic review and gaps analysis. Int J Med Inform 98:56–64. https://doi.org/10.1016/j.ijmedinf.2016.12.001
74. Kemppinen J, Korpela J, Elfvengren K, et al (2013) A Clinical Decision Support System for adult ADHD diagnostics process. Proc Annu Hawaii Int Conf Syst Sci 2616–2625. https://doi.org/10.1109/HICSS.2013.30
75. Suhasini A, Palanivel S, Ramalingam V (2011) Multimodel decision support system for psychiatry problem. Expert Syst Appl 38:4990–4997. https://doi.org/10.1016/j.eswa.2010.09.152
76. Razzouk D, Mari JJ, Shirakawa I, et al (2006) Decision support system for the diagnosis of schizophrenia disorders. Brazilian J Med Biol Res = Rev Bras Pesqui medicas e Biol 39:119–128. S0100-879X2006000100014
77. Bergman LG, Fors UGH (2008) Decision support in psychiatry—a comparison between the diagnostic outcomes using a computerized decision support system versus manual diagnosis. BMC Med Inform Decis Mak 8:9. https://doi.org/10.1186/1472-6947-8-9
78. Fortney JC, Pyne JM, Steven CA et al (2010) A web-based clinical decision support system for depression care management. Am J Manag Care 16:849–954. https://doi.org/10.1016/j.atherosclerosis.2009.05.009.Effect
79. Henshall C, Marzano L, Smith K et al (2017) A web-based clinical decision tool to support treatment decision-making in psychiatry: a pilot focus group study with clinicians, patients and carers. BMC Psychiatry 17:265. https://doi.org/10.1186/s12888-017-1406-z
80. aifred health (2020) Aifred Health. https://aifredhealth.com/
81. Assurex Health Inc. (2020) GeneSight. Changing lives through genetic insight. https://genesight.com/
82. Spring Care Inc. (2020) Spring Health. https://www.springhealth.com/
83. Dawes RM, Faust D, Meehl PE (1989) Clinical versus actuarial assessments. Science (80-) 243:1668–1674
84. Trivedi MH, Rush AJ, Crismon ML et al (2004) Clinical results for patients with major depressive disorder in the Texas medication algorithm project. Arch Gen Psychiatry 61:669–680. https://doi.org/10.1001/archpsyc.61.7.669
85. Fried EI, Nesse RM (2015) Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR∗D study. J Affect Disord 172:96–102. https://doi.org/10.1016/j.jad.2014.10.010
86. Olbert CM, Gala GJ, Tupler LA (2014) Quantifying heterogeneity attributable to polythetic diagnostic criteria: theoretical framework and empirical application. J Abnorm Psychol 123:452–462. https://doi.org/10.1037/a0036068

87. Hasin DS, Sarvet AL, Meyers JL et al (2018) Epidemiology of adult DSM-5 major depressive disorder and its specifiers in the United States. JAMA Psychiat 75:336–346. https://doi.org/10.1001/jamapsychiatry.2017.4602
88. Lamers F, Van Oppen P, Comijs HC et al (2011) Comorbidity patterns of anxiety and depressive disorders in a large cohort study: the Netherlands Study of Depression and Anxiety (NESDA). J Clin Psychiatry 72:341–348. https://doi.org/10.4088/JCP.10m06176blu
89. Jacobi F, Wittchen H-U, Hölting C et al (2004) Prevalence, comorbidity and correlates of mental disorders in the general population: results from the German Health Interview and Examination Survey (GHS). Psychol Med 34:597–611. https://doi.org/10.1017/S0033291703001399
90. Caspi A, Houts RM, Ambler A et al (2020) Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the dunedin birth cohort study. JAMA Netw open 3:e203221. https://doi.org/10.1001/jamanetworkopen.2020.3221
91. Plana-Ripoll O, Pedersen CB, Holtz Y et al (2019) Exploring comorbidity within mental disorders among a danish national population. JAMA Psychiat 76:259–270. https://doi.org/10.1001/jamapsychiatry.2018.3658
92. Trajković G, Starčević V, Latas M et al (2011) Reliability of the Hamilton rating scale for depression: a meta-analysis over a period of 49 years. Psychiatry Res 189:1–9. https://doi.org/10.1016/j.psychres.2010.12.007
93. Bagby RM, Ryder AG, Schuller DR, Marshall MB (2004) The Hamilton depression rating scale: Has the gold standard become a lead weight? Am J Psychiatry 161:2163–2177. https://doi.org/10.1176/appi.ajp.161.12.2163
94. McNeish D, Wolf MG (2020) Thinking twice about sum scores. Behav Res Methods 52:2287–2305. https://doi.org/10.3758/s13428-020-01398-0
95. Fried EI, Nesse RM (2015) Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. BMC Med 13:1–11. https://doi.org/10.1186/s12916-015-0325-4
96. Altman DG, Royston P (2006) The cost of dichotomising continuous variables. Br Med J 332:1080. https://doi.org/10.1136/bmj.332.7549.1080
97. Hardeveld F, Spijker J, De Graaf R et al (2010) Prevalence and predictors of recurrence of major depressive disorder in the adult population. Acta Psychiatr Scand 122:184–191. https://doi.org/10.1111/j.1600-0447.2009.01519.x
98. Penninx BWJH, Nolen WA, Lamers F et al (2011) Two-year course of depressive and anxiety disorders: results from the Netherlands study of depression and anxiety (NESDA). J Affect Disord 133:76–85. https://doi.org/10.1016/j.jad.2011.03.027
99. Otte C, Gold SM, Penninx BW et al (2016) Major depressive disorder. Nat Rev Dis Prim 2:1–21. https://doi.org/10.1038/nrdp.2016.65
100. Fried EI (2017) The 52 symptoms of major depression: lack of content overlap among seven common depression scales. J Affect Disord 208:191–197. https://doi.org/10.1016/j.jad.2016.10.019
101. Chevance AM, Ravaud P, Tomlinson A et al (2020) Identifying outcomes for depression that matter to patients, informal caregivers and healthcare professionals: qualitative content analysis of a large international online survey. Lancet Psychiatry 7:692–702. https://doi.org/10.1016/S2215-0366(20)30191-7
102. Demyttenaere K, Donneau AF, Albert A et al (2015) What is important in being cured from depression? Discordance between physicians and patients (1). J Affect Disord 174:390–396. https://doi.org/10.1016/j.jad.2014.12.004
103. Durstewitz D, Koppe G, Meyer-Lindenberg A (2019) Deep neural networks in psychiatry. Mol Psychiatry 24:1583–1598. https://doi.org/10.1038/s41380-019-0365-9
104. Huckvale K, Venkatesh S, Christensen H (2019) Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. npj Digit Med 2:. https://doi.org/10.1038/s41746-019-0166-1
105. Zeier Z, Carpenter LL, Kalin NH et al (2018) Clinical implementation of pharmacogenetic decision support tools for antidepressant drug prescribing. Am J Psychiatry 175:873–886. https://doi.org/10.1176/appi.ajp.2018.17111282.Clinical
106. Bzdok D, Varoquaux G, Steyerberg EW (2021) Prediction, not association, paves the road to precision medicine. JAMA Psychiat 78:127–128. https://doi.org/10.1001/jamapsychiatry.2020.2549
107. Rutledge RB, Chekroud AM, Huys QJ (2019) Machine learning and big data in psychiatry: toward clinical applications. Curr Opin Neurobiol 55:152–159. https://doi.org/10.1016/j.conb.2019.02.006
108. Iniesta R, Stahl D, McGuffin P (2016) Machine learning, statistical learning and the future of biological research in psychiatry. Psychol Med 46:2455–2465. https://doi.org/10.1017/S0033291716001367
109. Tsai C-Y, Wang S-H, Hsu M-H, Li Y-C (2016) Do false positive alerts in naïve clinical decision support system lead to false adoption by physicians? A randomized controlled trial. Comput Methods Programs Biomed 132:83–91. https://doi.org/10.1016/j.cmpb.2016.04.011
110. Dreiseitl S, Binder M (2005) Do physicians value decision support? A look at the effect of decision support systems on physician opinion. Artif Intell Med 33:25–30. https://doi.org/10.1016/j.artmed.2004.07.007
111. Browning M, Bilderbeck AC, Dias R et al (2021) The clinical effectiveness of using a predictive algorithm to guide antidepressant treatment in primary care (PReDicT): an open-label, randomised controlled trial. Neuropsychopharmacology 46:1307–1314. https://doi.org/10.1038/s41386-021-00981-z
112. Kelley JM, Kraft-Todd G, Schapira L, et al (2014) The influence of the patient-clinician relationship on healthcare outcomes: A systematic review and meta-analysis of randomized controlled trials. PLoS One 9:. https://doi.org/10.1371/journal.pone.0094207
113. Thompson L, McCabe R (2012) The effect of clinician-patient alliance and communication on treatment adherence in mental health care: A systematic review. BMC Psychiatry 12:. https://doi.org/10.1186/1471-244X-12-87
114. Topol EJ (2019) High-performance medicine: the convergence of human and artificial intelligence. Nat Med 25:44–56. https://doi.org/10.1038/s41591-018-0300-7
115. Grimes DR, Bauch CT, Ioannidis JPA (2018) Modelling science trustworthiness under publish or perish pressure. R Soc Open Sci 5:. https://doi.org/10.1098/rsos.171511
116. Fanelli D (2010) Do pressures to publish increase scientists' bias? An empirical support from US states data. PLoS One 5:. https://doi.org/10.1371/journal.pone.0010271
117. Akil H, Gordon J, Hen R et al (2018) Treatment resistant depression: A multi-scale, systems biology approach. Neurosci Biobehav Rev 84:272–288. https://doi.org/10.1016/j.neubiorev.2017.08.019
118. Houston JP, Gatz JL, Degenhardt EK, Jamal HH (2010) Symptoms predicting remission after divalproex augmentation with olanzapine in partially nonresponsive patients experiencing mixed bipolar i episode: a post-hoc analysis of a randomized controlled study. BMC Res Notes 3:1–6. https://doi.org/10.1186/1756-0500-3-276
119. DeRubeis RJ, Cohen ZD, Forand NR et al (2014) The personalized advantage index: translating research on prediction into individualized treatment recommendations. A demonstration PLoS One 9:1–8. https://doi.org/10.1371/journal.pone.0083875

120. Fournier JC, DeRubeis RJ, Hollon SD et al (2010) Antidepressant drug effects and depression severity: a patient-level meta-analysis. JAMA 303:175–177. https://doi.org/10.1001/jama.2009.1943
121. Souery D, Oswald P, Massat I et al (2007) Clinical factors associated with treatment resistance in major depressive disorder: results from a European multicenter study. J Clin Psychiatry 68:1062–1070. https://doi.org/10.4088/JCP.v68n0713
122. Howland RH, Rush AJ, Wisniewski SR et al (2009) Concurrent anxiety and substance use disorders among outpatients with major depression: Clinical features and effect on treatment outcome. Drug Alcohol Depend 99:248–260. https://doi.org/10.1016/j.drugalcdep.2008.08.010
123. Takahashi M, Shirayama Y, Muneoka K et al (2013) Personality traits as risk factors for treatment-resistant depression. PLoS One 8:1–7. https://doi.org/10.1371/journal.pone.0063756
124. Nanni V, Uher R, Danese A (2012) Childhood maltreatment predicts unfavorable course of illness and treatment outcome in depression: a meta-analysis. Am J Psychiatry 169:141–151. https://doi.org/10.1176/appi.ajp.2011.11020335
125. Nelson J, Klumparendt A, Doebler P, Ehring T (2017) Childhood maltreatment and characteristics of adult depression: meta-analysis. Br J Psychiatry 210:96–104. https://doi.org/10.1192/bjp.bp.115.180752
126. Williams LM, Debattista C, Duchemin AM et al (2016) Childhood trauma predicts antidepressant response in adults with major depression: data from the randomized international study to predict optimized treatment for depression. Transl Psychiatry 6:e799–e807. https://doi.org/10.1038/tp.2016.61
127. Godlewska BR, Browning M, Norbury R et al (2016) Early changes in emotional processing as a marker of clinical response to SSRI treatment in depression. Transl Psychiatry 6:e957–e967. https://doi.org/10.1038/tp.2016.130
128. Browning M, Kingslake J, Dourish CT et al (2019) Predicting treatment response to antidepressant medication using early changes in emotional processing. Eur Neuropsychopharmacol 29:66–75. https://doi.org/10.1016/j.euroneuro.2018.11.1102
129. Haroon E, Daguanno AW, Woolwine BJ et al (2018) Antidepressant treatment resistance is associated with increased inflammatory markers in patients with major depressive disorder. Psychoneuroendocrinology 95:43–49. https://doi.org/10.1016/j.psyneuen.2018.05.026
130. Liu JJ, Bin WY, Strawbridge R et al (2020) Peripheral cytokine levels and response to antidepressant treatment in depression: a systematic review and meta-analysis. Mol Psychiatry 25:339–350. https://doi.org/10.1038/s41380-019-0474-5
131. Zhou C, Zhong J, Zou B et al (2017) Meta-analyses of comparative efficacy of antidepressant medications on peripheral BDNF concentration in patients with depression. PLoS One 12:1–18. https://doi.org/10.1371/journal.pone.0172270
132. Kaddurah-Daouk R, Boyle SH, Matson W et al (2011) Pretreatment metabotype as a predictor of response to sertraline or placebo in depressed outpatients: a proof of concept. Transl Psychiatry 1:1–7. https://doi.org/10.1038/tp.2011.22
133. Altar CA, Carhart JM, Allen JD et al (2015) Clinical validity: combinatorial pharmacogenomics predicts antidepressant responses and healthcare utilizations better than single gene phenotypes. Pharmacogenomics J 15:443–451. https://doi.org/10.1038/tpj.2014.85
134. Bousman CA, Arandjelovic K, Mancuso SG et al (2019) Pharmacogenetic tests and depressive symptom remission: a meta-analysis of randomized controlled trials. Pharmacogenomics 20:37–47. https://doi.org/10.2217/pgs-2018-0142
135. Brown L, Vranjkovic O, Li J et al (2020) The clinical utility of combinatorial pharmacogenomic testing for patients with depression: a meta-analysis. Pharmacogenomics 21:559–569. https://doi.org/10.2217/pgs-2019-0157
136. Binder EB, Künzel HE, Nickel T et al (2009) HPA-axis regulation at in-patient admission is associated with antidepressant therapy outcome in male but not in female depressed patients. Psychoneuroendocrinology 34:99–109. https://doi.org/10.1016/j.psyneuen.2008.08.018
137. Fischer S, Macare C, Cleare AJ (2017) Hypothalamic-pituitary-adrenal (HPA) axis functioning as predictor of antidepressant response–meta-analysis. Neurosci Biobehav Rev 83:200–211. https://doi.org/10.1016/j.neubiorev.2017.10.012
138. Fabbri C, Corponi F, Souery D et al (2019) The genetics of treatment-resistant depression: a critical review and future perspectives. Int J Neuropsychopharmacol 22:93–104. https://doi.org/10.1093/ijnp/pyy024
139. Uher R, Perroud N, Ng MYM et al (2010) Genome-wide pharmacogenetics of antidepressant response in the GENDEP Project. Am J Psychiatry 167:1–10. https://doi.org/10.1176/appi.ajp.2009.09070932
140. Adkins DE, Åberg K, McClay JL et al (2010) A genomewide association study of citalopram response in major depressive disorder—a psychometric approach. Biol Psychiatry 68:e25–e27. https://doi.org/10.1016/j.biopsych.2010.05.018
141. Biernacka JM, Sangkuhl K, Jenkins G et al (2015) The International SSRI Pharmacogenomics Consortium (ISPC): A genome-wide association study of antidepressant treatment response. Transl Psychiatry 5:1–9. https://doi.org/10.1038/tp.2015.47
142. Garriock HA, Kraft JB, Shyn SI et al (2010) A Genomewide Association Study of Citalopram Response in Major Depressive Disorder. Biol Psychiatry 67:133–138. https://doi.org/10.1016/j.biopsych.2009.08.029
143. GENDEP Investigators, MARS Investigators, STAR*D Investigators (2013) Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of three genome-wide pharmacogenetic studies. Am J Psychiatry 170:207–217. https://doi.org/10.1176/appi.ajp.2012.12020237
144. Ising M, Lucae S, Binder EB et al (2009) A genomewide association study points to multiple loci that predict antidepressant drug treatment outcome in depression. Arch Gen Psychiatry 66:966–975
145. Tansey KE, Guipponi M, Hu X et al (2013) Contribution of common genetic variants to antidepressant response. Biol Psychiatry 73:679–682. https://doi.org/10.1016/j.biopsych.2012.10.030
146. García-González J, Tansey KE, Hauser J et al (2017) Pharmacogenetics of antidepressant response: a polygenic approach. Prog Neuro-Psychopharmacology Biol Psychiatry 75:128–134. https://doi.org/10.1016/j.pnpbp.2017.01.011
147. Domschke K, Tidow N, Schwarte K et al (2014) Serotonin transporter gene hypomethylation predicts impaired antidepressant treatment response. Int J Neuropsychopharmacol 17:1167–1176. https://doi.org/10.1017/S146114571400039X
148. Lisoway AJ, Zai CC, Tiwari AK, Kennedy JL (2018) DNA methylation and clinical response to antidepressant medication in major depressive disorder: a review and recommendations. Neurosci Lett 669:14–23. https://doi.org/10.1016/j.neulet.2016.12.071
149. Belzeaux R, Lin R, Ju C et al (2018) Transcriptomic and epigenomic biomarkers of antidepressant response. J Affect Disord 233:36–44. https://doi.org/10.1016/j.jad.2017.08.087
150. Caspani G, Turecki G, Lam RW et al (2021) Metabolomic signatures associated with depression and predictors of antidepressant response in humans: A CAN-BIND-1 report. Commun Biol 4:. https://doi.org/10.1038/s42003-021-02421-6
151. Alshehri T, Mook- Kanamori DO, Willems Van Dijk K et al (2021) Metabolomics dissection of depression heterogeneity and related cardiometabolic risk. Psychol Med. https://doi.org/10.1017/S0033291721001471

152. Fu CHY, Steiner H, Costafreda SG (2013) Predictive neural biomarkers of clinical response in depression: a meta-analysis of functional and structural neuroimaging studies of pharmacological and psychological therapies. Neurobiol Dis 52:75–83. https://doi.org/10.1016/j.nbd.2012.05.008
153. Lener MS, Iosifescu DV (2015) In pursuit of neuroimaging biomarkers to guide treatment selection in major depressive disorder: a review of the literature. Ann NY Acad Sci 1344:50–65. https://doi.org/10.1111/nyas.12759
154. Enneking V, Leehr EJ, Dannlowski U, Redlich R (2019) Brain structural effects of treatments for depression and biomarkers of response: a systematic review of neuroimaging studies. Psychol Med 50:187–209. https://doi.org/10.1017/S0033291719003660
155. Kang S-G, Cho S-E (2020) Neuroimaging biomarkers for predicting treatment response and recurrence of major depressive disorder. Int J Mol Sci 21:2148. https://doi.org/10.3390/ijms21062148
156. Baskaran A, Milev R, McIntyre RS (2012) The neurobiology of the EEG biomarker as a predictor of treatment response in depression. Neuropharmacology 63:507–513. https://doi.org/10.1016/j.neuropharm.2012.04.021
157. Widge AS, Bilge MT, Montana R et al (2019) Electroencephalographic biomarkers for treatment response prediction in major depressive illness: a meta-analysis. Am J Psychiatry 176:44–56. https://doi.org/10.1176/appi.ajp.2018.17121358
158. Kircanski K, Williams LM, Gotlib IH (2019) Heart rate variability as a biomarker of anxious depression response to antidepressant medication. Depress Anxiety 36:63–71. https://doi.org/10.1002/da.22843

Acknowledgements

I am extremely grateful to all the people who have made this work and my PhD journey possible.

First, I would like to thank my supervisors and thesis advisory committee (TAC). I want to thank my first TAC member and supervisor, Elisabeth Binder, for all the support, guidance and reliability over the past years. Her detailed and overarching knowledge as well as her quick thinking, excellent memory, and decision-making abilities have been invaluable to my research. In the same way, I want to thank Tanja Brückl for her close and continuous supervision. I could always rely on her clinical expertise, her eagerness to help and her open-mindedness during our weekly meetings. As my second TAC member, I want to thank Nikolaos Koutsouleris for his profound methodological expertise as well as his highly valuable contributions and new ideas to my projects. As my third TAC member, I want to thank Bertram Müller-Myhsok for his kind support during all our meetings and his excellent scientific supervision, especially with paper I.

Second, I would like to thank my colleagues at the institute for their scientific help and contributions. My special thanks go to Darina Czamara, Dominic Dwyer, Marcus Ising, Lee Jollans, Janine Knauer-Arloth, Nils Rek, Philipp Sämann, and Natan Yusupov. Additionally, I have benefited from a great institutional support by several people whom I would like to thank at this point: my graduate school coordinators Isabel Oßwald, Bettina Schönherr, and André Vogel; Nicole Grodzycki and Heike Junkert for their steady help with administrative matters; and the information technology department for solving (almost) all technical problems that occurred. I also want to thank our collaboration partners from Biomax Informatics AG and the Stack Overflow community. Furthermore, I am very grateful to the psychophysiology group for regularly inviting me to their socials, lunches, and coffee breaks, which has always enriched my daily work life.

Last but not least, I would like to thank my friends and family for their enduring support. I am very grateful to my parents and my brother for their unconditional encouragement and care as well as for their interest in my work and their advice. I am especially grateful to my partner for all her wonderful support during the ups and downs of my PhD, her empathy, patience, humor, and ability to take my mind off things when I worry too much about work. Finally, I want to thank our cat for soundtracking the many hours I spent working from home with his constant purrs.