Cross-trial prediction of treatment response to transcranial direct current stimulation in patients with major depressive disorder

Gerrit Burkhardt, MD[1,2]; Stephan Goerigk, PhD[1,2,3]; Lucia Bulubas, MD, PhD[1,2]; Esther Dechantsreiter, M.Sc.[1,2]; Daniel Keeser, PhD[1,2]; Ulrike Vogelmann, MD[2,4]; Katharina von Wartensleben[1]; Johannes Wolf, MD[1,2]; Christian Plewnia, MD[5,6]; Andreas Fallgatter, MD[5,6]; Berthold Langguth, MD[7]; Claus Normann, MD[8,9]; Lukas Frase, MD[8,10]; Peter Zwanzger, MD[1,11]; Thomas Kammer, MD[12,13]; Carlos Schönfeldt-Lecuona, MD[12,13]; Daniel Kamp, MD[14]; Malek Bajbouj, MD[15,16]; Nikolaos Koutsouleris, MD[1,2]; Andre R Brunoni, MD, PhD [17,*]; Frank Padberg, MD[1,2,*]

[1]Department of Psychiatry and Psychotherapy, LMU University Hospital, Munich, Germany
[2]German Center for Mental Health (DZPG), Site Munich-Augsburg, Germany
[3]Charlotte Fresenius Hochschule, University of Psychology, Munich, Germany
[4]Department of Psychiatry and Psychotherapy, University Hospital, Technical University of Munich (TUM), Munich, Germany
[5]Tübingen Center for Mental Health, Department of Psychiatry and Psychotherapy, University of Tübingen, Tübingen, Germany
[6]German Center for Mental Health (DZPG), Site Tübingen, Germany
[7]Department of Psychiatry and Psychotherapy, University of Regensburg, Regensburg, Germany
[8]Department of Psychiatry and Psychotherapy, Medical Center, University of Freiburg, Freiburg, Germany
[9]Center for Basics in Neuromodulation (NeuroModulBasics), University of Freiburg, Freiburg, Germany
[10]Department of Psychosomatic Medicine and Psychotherapy, Medical Center, University of Freiburg
[11]kbo-Inn-Salzach-Klinikum, Clinical Center for Psychiatry, Psychotherapy, Psychosomatic Medicine, Geriatrics and Neurology, Gabersee, Wasserburg/Inn, Germany
[12]Department of Psychiatry and Psychotherapy III, University of Ulm, Ulm, Germany
[13]German Center for Mental Health (DZPG), Site Mannheim-Heidelberg-Ulm, Germany
[14]Department of Psychiatry and Psychotherapy, LVR-Klinikum Düsseldorf, Heinrich-Heine-Universität Düsseldorf, Medical Faculty, Düsseldorf, Germany
[15]Department of Psychiatry and Psychotherapy, Charité-Campus Benjamin Franklin, Berlin, Germany
[16]German Center for Mental Health (DZPG), Site Berlin-Potsdam, Germany
[17]Department of Psychiatry, University of São Paulo Medical School, São Paulo, Brazil
* These authors equally contributed to the manuscript

Corresponding author:

Dr. Gerrit Burkhardt
Department of Psychiatry and Psychotherapy
LMU University Hospital
Nußbaumstraße 7
80336 Munich, Germany
E-Mail: gerrit.burkhardt@med.uni-muenchen.de
Telephone: +1149 89 4400 53381

Running title: Cross-trial prediction of tDCS response in MDD

Abstract

Machine-learning (ML) classification may offer a promising approach for treatment response prediction in patients with major depressive disorder (MDD) undergoing non-invasive brain stimulation. This analysis aims to develop and validate such classification models based on easily attainable sociodemographic and clinical information across two randomized controlled trials on transcranial direct-current stimulation (tDCS) in MDD. Using data from 246 patients with MDD from the randomized-controlled DepressionDC and ELECT-TDCS trials, we employed an ensemble machine learning strategy to predict treatment response to either active tDCS or sham tDCS/placebo, defined as ≥ 50% reduction in the Montgomery-Åsberg Depression Rating Scale at 6 weeks. Separate models for active tDCS and sham/placebo were developed in each trial and evaluated for external validity across trials and for treatment specificity across modalities. Additionally, models with above-chance detection rates were associated with long-term outcomes to assess their clinical validity. In the DepressionDC trial, models achieved a balanced accuracy of 63.5% for active tDCS and 62.5% for sham tDCS in predicting treatment responders. The tDCS model significantly predicted MADRS scores at the 18-week follow-up visit ($F_{(1,60)} = 4.53$, $p_{FDR} = .037$, $R^2 = 0.069$). Baseline self-rated depression was consistently ranked as the most informative feature. However, response prediction in the ELECT-TDCS trial and across trials was not successful. Our findings indicate that ML-based models have the potential to identify responders to active and sham tDCS treatments in patients with MDD. However, to establish their clinical utility, they require further refinement and external validation in larger samples and with more features.

Introduction

Major Depressive Disorder (MDD) represents a significant global health challenge, ranking as one of the main causes of disability worldwide[1]. Despite the availability of effective treatments ranging from pharmacotherapy and psychotherapy to non-invasive and invasive neurostimulation, many patients do not achieve remission, even after multiple therapeutic attempts[2]. The development of new interventions has proven challenging, possibly due to the heterogeneity of MDD symptoms[3], its varying time course[4], and a lack of robust biological correlates[5,6]. While multiple sociodemographic, clinical, genetic, and neuroimaging variables have been associated with responses to common treatments like antidepressant medication[7], these associations have not yet resulted in stratified patient selection algorithms or targeted interventions. Thus, recent research has focused on developing multivariate predictive models that might enable pre-treatment stratification at the individual patient level[8,9] and detect effects beyond the between-group level in randomized controlled trials (RCT)[9,10]. Within this approach, initial machine learning (ML)-based predictive models are typically trained on data from existing RCTs to identify responders to the treatments under investigation[8,11,12]. Models then require testing in independent samples and across diverse populations to ensure their generalizability to unseen patients before they are finally tested prospectively for clinical utility. However, efforts to externally validate initial models remain sparse[13], and consequently, few attempts have been made to validate treatment prediction models in RCTs[14].

Predictive approaches are particularly relevant in the field of non-invasive brain stimulation (NIBS), including repetitive transcranial magnetic stimulation (rTMS) and transcranial direct current stimulation (tDCS), as the clinical application of these interventions is rapidly growing. tDCS is a safe, well-tolerated, and easily applicable treatment option for patients with MDD[15–17], but has yielded inconclusive results in recent confirmatory multicenter RCTs[18–20]. Therefore, efforts to optimize outcomes on the individual patient level are required to develop the intervention toward clinical applicability[21,22]. A recent study reported a high predictive accuracy of an ML-based prediction model for identifying responders to bifrontal transcranial direct current stimulation (tDCS), yet lacked external validation[23]. For rTMS, ML studies have mainly focused on other neuropsychiatric

conditions, e.g. schizophrenia[9,10]. To our knowledge, there are currently no studies available utilizing ML models across RCTs on NIBS interventions.

To investigate whether sociodemographic and clinical data are informative for predicting the individual response to tDCS, we used data from two large RCTs on basically identical tDCS protocols (i.e. bifrontal electrode montage: anode left and cathode right dorsolateral prefrontal cortex [DLPFC], 2 mA intensity and 30 min duration), that were performed in Brazil[16] and Germany[18], evaluating the efficacy of tDCS in patients with MDD. We aimed to develop and externally validate ML-based prediction models to identify patients likely to benefit from tDCS, test those models for treatment specificity, and explore their clinical validity and utility based on long-term outcomes.

Subjects and Methods

Study design

In this secondary analysis of two randomized, blinded, sham-controlled trials, we used an ensemble ML-based strategy with nested cross-validation to identify patients with response to either active tDCS or sham/placebo treatment using sociodemographic and clinical baseline variables. Models were trained separately in each trial for each treatment modality and then applied 1.) across trials to test for external validity and 2.) across treatment modalities to test if predictions were specific to the treatment (see Figure 1). Classification probabilities of models with above-chance detection rates were then associated with long-term outcomes at follow-up to explore the clinical validity and utility of the predictions.

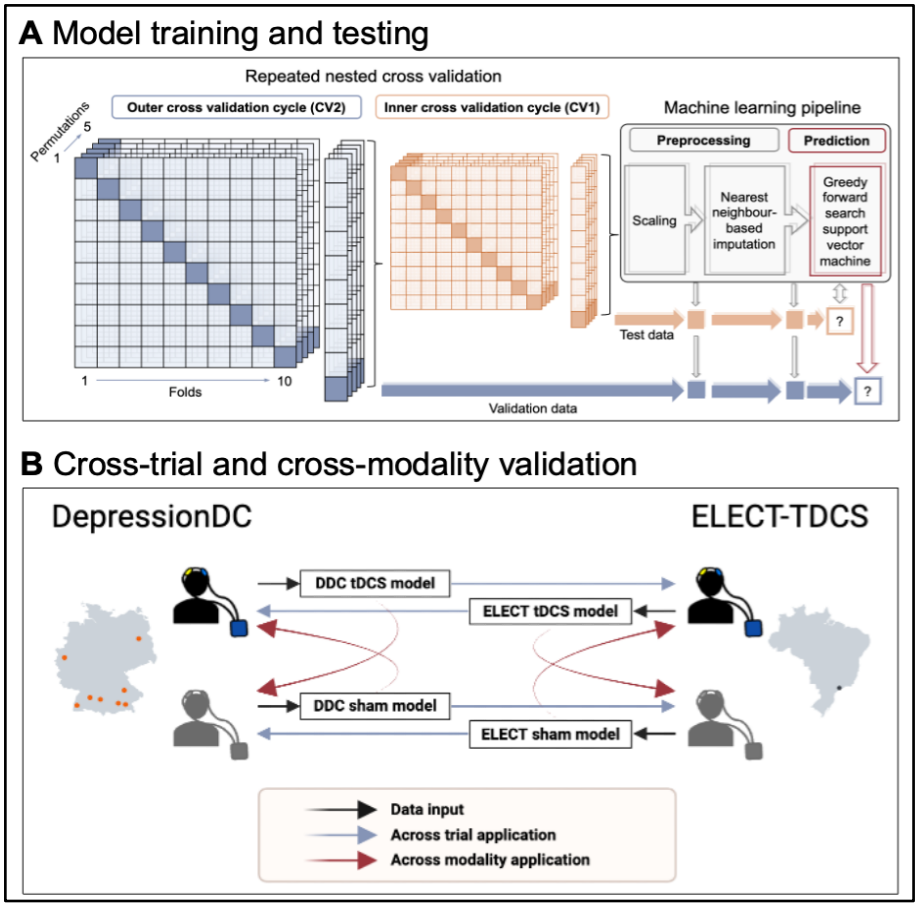

Figure 1. Model development and validation

Study population

We analyzed patients with MDD from two trials: 1.) DepressionDC (trial registration: NCT02530164) was a multicenter RCT investigating the efficacy of 6 weeks of bifrontal tDCS as an additional treatment to selective serotonin reuptake inhibitors (SSRI) in patients with MDD[18]. Between January 2016 and June 2020, 160 patients were recruited at seven university hospitals and one psychiatric community hospital in Germany. Active tDCS was not superior to sham tDCS in reducing depressive symptoms. 2.) The Escitalopram versus Electrical Current Therapy for Treating Depression Clinical Study (ELECT-TDCS; trial registration: NCT01894815) was a single-center, non-inferiority RCT comparing active tDCS plus placebo medication, escitalopram plus sham tDCS, and sham tDCS plus placebo medication in patients with MDD over 10 weeks[16]. Two hundred forty-five patients were recruited at the University of São Paulo, Brazil, between October 2013 through July 2016. Active tDCS was not non-inferior to escitalopram but superior to placebo treatment.

Both trials employed rigorous RCT methods with stringent randomization, blinding protocols, and non-active sham conditions. Participants were selected based on the DSM-5 criteria for MDD while excluding patients with bipolar disorder, substance abuse or dependence, dementia, and personality disorders. However, DepressionDC enrolled patients currently receiving SSRIs, whereas ELECT-TDCS required participants to be antidepressant-free. Furthermore, DepressionDC permitted the inclusion of patients with marginally lower symptom severity, as assessed by the Hamilton Depression Rating Scale (HDRS), and restricted participation to a narrower age range. A comprehensive list of eligibility criteria for both trials is provided in the supplement.

The intervention followed a basically identical tDCS protocol with 2 mA stimulation of the DLPFC over 30 minutes per session. However, protocols differed in precise electrode placement, i.e. DepressionDC using the F3 (anode) and F4 (cathode) based on the international electroencephalogram 10-20 system, and ELECT-TDCS the Omni-Lateral-Electrode (OLE) system (left DLPFC anode and right DLPFC cathode) with slightly more lateral positions of electrodes[24]. Also, treatment lasted 10 weeks with a total of 22 treatment sessions (15 sessions in the first 3 weeks and 7 weekly sessions for the remaining treatment period) in ELECT-TDCS versus 6 weeks with a total of 24 treatment sessions (20 sessions in the first 4 weeks and 2 sessions per week for 2 weeks) in DepressionDC, with efficacy assessed at these time points using the HDRS and Montgomery-Åsberg Depression Rating Scale (MADRS) as primary outcome measures, respectively. Weekly MADRS scores were additionally collected in ELECT-TDCS as a secondary outcome measure.

We used data from all patients with available depression scores on the MADRS at baseline and week 6 after randomization. Thus, the analysis included 136 (active tDCS: 72 patients; sham tDCS: 64 patients) out of 150 patients from the DepressionDC sample and 110 (active tDCS + placebo: 66 patients; sham tDCS + placebo: 44 patients) out of 154 patients from the respective treatment arms of the ELECT-TDCS sample. All participants had provided their written informed consent before inclusion in the respective study. Both studies were approved by the local ethics committees and conducted in accordance with the Declaration of Helsinki.

Prediction target and features

Participants were classified as treatment responders if they achieved a $\geq 50\%$ reduction from baseline to week 6 on the 10-item MADRS (score range 0-60; higher scores indicate more severe depression)[25]. Pursuing a data-driven approach, we included all variables available across the studies at baseline as potential predictors. This amounted to 15 features, including basic sociodemographic information (age, sex, years of education, marriage, unemployment), medical history (body mass index, smoker status, diagnoses of hypertension, diabetes, and/or hypothyroidism), psychiatric history (age of MDD onset, duration of MDD episode, family history of MDD), and baseline depression severity (MADRS and Beck Depression Inventory-II [BDI-II] total scores).

Machine learning analysis

All ML analyses were conducted using the in-house, open-source software package NeuroMiner, version 1.05 (https://github.com/neurominer-git/NeuroMiner-1), running on MATLAB (version R2022a). We used repeated nested cross-validation (CV) with 10 folds and 5 repetitions at both the inner (CV_1) and outer (CV_2) loops to strictly separate the training and testing of the models. In each CV_1 fold, we scaled all features from 0 to 1 and substituted missing values via 7-nearest neighbor-based imputation (Euclidean distance for continuous, Hamming distance for categorical variables)[26]. Following a previous approach[27,28], each processed CV_1 training sample then entered a greedy stepwise forward search wrapper employing a linear support vector machine algorithm (SVM; LIBSVM 3.12[29]) to iteratively select a subset of 50% features with highest predictive performance (balanced accuracy [$BAC = \frac{sensitivity + specificity}{2}$][30] on the held-out CV_1 data) across a range of C hyperparameters ($2^{[-4_{\epsilon}\mathbb{Z}\rightarrow+4]}$). To account for uneven distributions of the outcome labels (response/non-response), optimal C hyperparameters were multiplied with the inverse ratio of the training group sizes[31]. For each CV_1 permutation, all CV_1 models were retrained with the optimal model hyperparameters, and this ensemble was then applied to the CV_2 test data without modification. Classification probabilities for each CV_2 test subject were retrieved by combining the decisions across all models. We calculated permutation-based *p*-values to define which models reached above-chance

detection rates (α=0.05; 1000 permutations)[32]. To understand which features were most reliably contributing to the prediction of treatment response, we computed the CV ratio[33]. The feature importance of variables was further estimated using sign-based consistency mapping[34]. For cross-trial and cross-treatment modality validation, we applied CV_1 ensembles with permutation-based above-chance detection rates without modification to the respective other samples. Out-of-sample performance metrics were calculated by comparing the predicted versus the observed outcome labels over all CV_2 predictions.

Post-hoc clinical validation

Additional validation analyses and visualizations were performed in R, version 4.3.2[35]. Results were considered significant at α=0.05. To explore the clinical validity of all classifiers with above-chance detection rates, we fit linear mixed models (LMM) using the lme4 package[36] to predict MADRS and GAF changes from baseline until the trials' follow-up appointments based on the models' assigned probability to be a responder (formula: change ~ assigned probability). The model included the treatment site as a random effect (formula: ~1| site). The significance of the model factors was determined using omnibus tests (type III ANOVA) with Satterthwaite approximation to degrees of freedom.

Results

Main classifiers

In the DepressionDC trial, 24 patients (33%) had responded to active tDCS treatment at week 6. Compared to tDCS non-responders, these patients were significantly older (mean [SD] age: 37 [13] vs. 30 [13]; p=0.023) and showed lower clinician-rated (mean [SD] MADRS scores: 22 [5] vs. 26 [6]; p=0.049) and self-reported depression severity (mean [SD] BDI scores: 23 [9] vs. 30 [11]; p=0.011) at baseline (other baseline characteristics are shown in Table 1). The classifier ensemble predicted tDCS responders with an above-chance cross-validated BAC of 63.5% (P=0.001; Table 2; Figure 2), increasing the prognostic certainty compared to the base rate (prognostic summary index of 24.1%).

In the sham group, 32 patients (50%) showed a treatment response. Compared to sham non-responders, these patients had lower clinician-rated (mean [SD] MADRS scores: 22 [5.3] vs. 24.3 [4.4]; p=0.017) and self-reported depression severity (mean [SD] BDI scores: 23 [9] vs. 30 [10]; p=0.005) at baseline. Sham responders were predicted with an above-chance cross-validated BAC of 62.5% (p=0.023; Table 2) and prognostic summary index of 25.1%. In both the tDCS and sham analyses, only baseline BDI scores reliably and significantly contributed to the classifier decisions (Figure 3).

In the ELECT trial, 27 patients (41%) had responded to active tDCS treatment at week 6. Compared to tDCS non-responders, they had higher clinician-rated depression at baseline (mean [SD] MADRS scores: 29 [8] vs. 26 [6]; p=0.020). Responders were predicted with a BAC of 61.0% that did not reach our above-chance detection criterion (p=0.071; Table 2). Similarly, our analysis did not yield models with above-chance detection rates for the 14 (32%) responders to sham treatment (BAC=42.9%; p=0.88; Table 2).

External validation and treatment specificity

Models with above-chance detection rates did not generalize across trials. The DepressionDC tDCS classifier showed a BAC of 51.1% in the ELECT tDCS sample, and the DepressionDC sham classifier reached a BAC of 55.5% in the ELECT sham sample (shown in Table 2). Models also did not reach above-chance detection rates across treatment modalities, with the DepressionDC tDCS model showing a BAC of 53.1% when applied to the DepressionDC sham arm and the DepressionDC sham model showing a BAC of 53.1% when applied to the DepressionDC tDCS arm.

245 **Table 1.** Baseline characteristics of patients with MADRS response and non-response to tDCS

Feature	DepressionDC			ELECT		
	Non-responder, n = 48[a]	Responder, n = 24[a]	p-value[b]	Non-responders, n = 39[a]	Responder, n = 27[a]	p-value[b]
Sex			0.9			0.3
Female	29 (60%)	14 (58%)		25 (66%)	21 (78%)	
Male	19 (40%)	10 (42%)		13 (34%)	6 (22%)	
Age at randomization - years	37 (13)	43 (14)	0.1	44 (13)	47 (11)	0.2
Age of depression onset - years	30 (13)	37 (13)	**0.023**	27 (12)	28 (12)	0.8
Duration of episode - weeks	56 (67)	52 (56)	>0.9	28 (72)	26 (33)	0.7
Family history of depression	21 (46%)	15 (65%)	0.1	26 (67%)	19 (70%)	0.8
Years of education	11.91 (2.28)	11.50 (1.65)	0.5	15.8 (4.5)	14.0 (4.7)	0.2
Unemployed	3 (9.1%)	0 (0%)	0.5	13 (35%)	10 (37%)	0.9
Married	5 (15%)	6 (43%)	0.1	19 (49%)	16 (59%)	0.4
Body mass index - kg/m^2	25.7 (5.2)	27.9 (5.9)	0.1	25.8 (4.7)	25.7 (4.6)	>0.9
Smoker	21 (44%)	7 (29%)	0.2	7 (18%)	5 (19%)	>0.9
Hypertension	4 (10%)	4 (24%)	0.2	9 (24%)	5 (19%)	0.7
Diabetes	1 (2.5%)	0 (0%)	>0.9	3 (7.9%)	1 (3.7%)	0.6
Hypothyroidism	2 (5.0%)	4 (24%)	0.1	8 (21%)	5 (19%)	0.8
MADRS score at baseline	26 (6)	22 (5)	**0.049**	26 (6)	29 (8)	**0.020**
BDI score at baseline	30 (11)	23 (9)	**0.011**	30 (10)	30 (11)	0.7

246 [a]Mean (SD); n (%). [b] Pearson's Chi-squared test; Wilcoxon rank sum test. Fisher's exact test.

Table 2. Prediction results of MADRS response models.

Model	TP	TN	FP	FN	Sens	Spec	PPV	NPV	PSI	AUC	BAC	p-value
DDC tDCS	17	27	21	7	70.8	56.2	44.7	79.4	24.1	0.64	63.5	**0.001**
applied to ELECT tDCS	11	24	15	16	40.7	61.5	42.3	60.0	2.3	0.50	51.1	-
applied to DDC sham	17	17	15	15	53.1	53.1	53.1	53.1	6.2	0.62	53.1	-
DDC sham	20	20	12	12	62.5	62.5	62.5	62.5	25.0	0.67	62.5	**0.023**
applied to ELECT sham	9	14	16	5	64.3	46.7	36.0	73.7	9.7	0.59	55.5	-
applied to DDC tDCS	14	23	25	10	58.3	47.9	35.9	69.7	5.6	0.54	53.1	-
ELECT tDCS	17	23	16	10	63.0	59.0	51.5	69.7	21.2	0.60	61.0	0.07
ELECT sham	5	15	15	9	35.7	50.0	25.0	62.5	-12.5	0.44	42.9	0.88

Abbreviations: TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative; Sens, Sensitivity; Spec, Specificity; PPV, Positive Predictive Value; NPV, Negative Predictive Value; PSI, Prognostic Summary Index; AUC, Area Under the Curve; BAC, Balanced Accuracy. Note: p-values were calculated using permutation analysis (α=0.05; 1000 permutations).

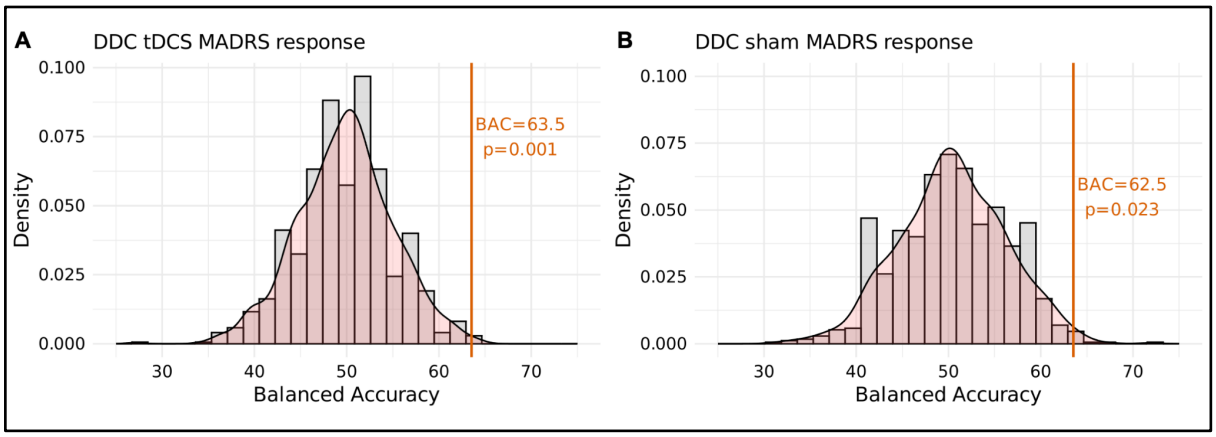

Figure 2. Permutation-based significance. Response was defined as ≥ 50% reduction from baseline.

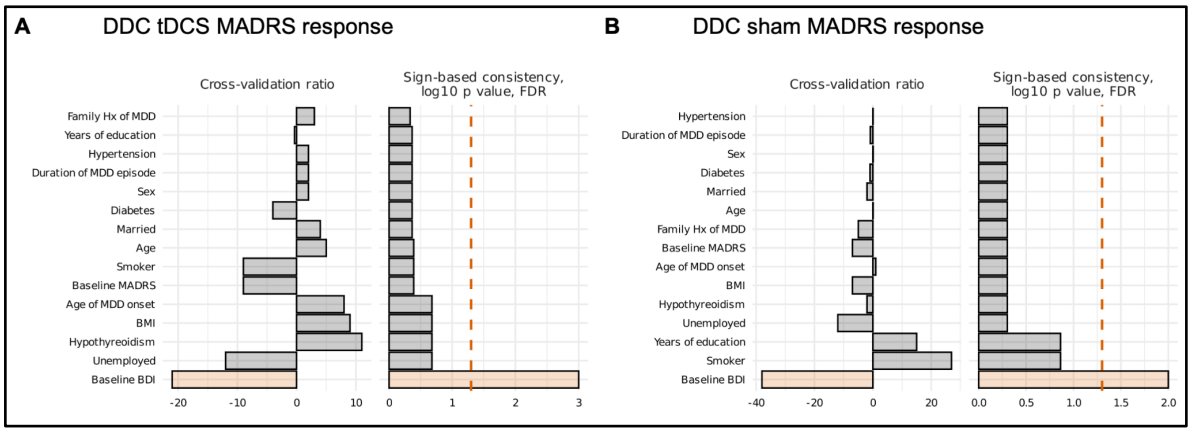

Figure 3. Feature importance. A positive cross-validation ratio suggests that a higher feature value predicts MADRS response, whereas a negative ratio implies that lower values do. A higher sign-based consistency suggests that feature weights were more consistently positive or negative across the ensemble. Significance was assessed by defining a hypothesis test for the importance score with a null hypothesis of 0 and an alternative hypothesis of not 0. FDR-corrected p-values were then calculated using a cumulative distribution function of z-scores (α=0.05; red dotted line).

Post-hoc clinical validation

For the DepressionDC tDCS model, classification probabilities significantly predicted MADRS scores at the 18-week follow-up visit ($F_{(1,60)} = 4.53$, $p_{FDR} = .037$, $R^2 = 0.069$), but not MADRS scores at week 30 or GAF scores at weeks 18 and 30 (Supplementary Table S2 and Supplementary Figure S1). Classification probabilities of the DepressionDC sham model did not predict MADRS or GAF outcomes at weeks 18 and 30.

Discussion

To our knowledge, this is the first study testing the cross-trial validity of ML models trained on easily attainable sociodemographic and clinical baseline variables to predict responders to a NIBS intervention. Whereas ML models increased accuracy in identifying responders to active tDCS and sham tDCS in the 6-week multicenter, randomized-controlled DepressionDC trial, the same variable battery could not be utilized to predict responses in the ELECT-TDCS trial, and models did not generalize across trial datasets and treatment modalities. Our findings underscore existing challenges and limitations inherent in predicting antidepressant responses in RCT populations.

The predictive accuracies of our models with above-chance detection rates align with prior attempts to predict responses to antidepressant medication using ML algorithms trained on clinical variables[9,11,12,37]. Although these performances are modest compared to the classification benchmarks set in other medical disciplines, such as neuroradiology[38], antidepressant response prediction remains a challenging task relying on subjective judgment, and thus, even small increases in predictive accuracy might theoretically inform clinical decisions. However, as exemplified by the only prior attempt to prospectively assess the clinical utility of an antidepressant response classifier, which failed to improve treatment outcomes when applied as a decision-making tool[14], strong indicators are needed to justify further development of classifiers beyond the proof-of-concept stage. In the context of our study, the tDCS response classifier for the DepressionDC sample significantly predicted depression severity at the 18-week follow-up visit, suggesting potential clinical validity. Nonetheless, given the failed external validation and the need for enhanced performance, further refinement and testing of the model would be needed to establish its efficacy and reliability[39].

Recent research, including a validation attempt across trials on antipsychotic medication for schizophrenia[40], suggests that three main reasons might have contributed to our models' failed transferability across trial datasets. First, trial populations might have been too heterogeneous, including patients at different disorder stages or with nuanced differences in psychopathology profiles not captured by the broad DSM-5 inclusion criteria. Indeed, participants in the DepressionDC trial showed numerically later depression onset and longer mean depression episode duration compared to

297 ELECT-TDCS[16,18]. By contrast, baseline depression severity was comparable between trials. Second,
298 compared to a previous ML prediction study in the ELECT-TDCS cohort[23], our models showed
299 considerably lower predictive performance in the same dataset. Since our analysis aimed to develop
300 generalizable prediction models across two RCT cohorts, we took several methodological choices that
301 may partly explain this difference. Instead of an XGBoost classifier, we used a validated ensemble
302 learning strategy applying SVM algorithms within a nested cross-validation framework. This
303 framework was chosen because it has been applied in several multisite analyses and optimized to
304 generate generalizable models[9,27,28]. We also limited the input variables to features available in both
305 datasets. Consequently, this meant omitting data modalities like neuropsychological test results,
306 electrophysiological data, and imaging measurements, which might have been needed to specifically
307 detect patterns of response in the active treatment arm. For example, a recent analysis on data from
308 the RESIS trial, which like the DepressionDC trial was also negative regarding its primary and
309 secondary outcomes, extended a previous attempt to built an active rTMS treatment response
310 classifier for patients with predominant negative-symptom schizophrenia based on structural
311 Magnetic Resonance Imaging (sMRI)[9] by incorporating further data domains (i.e. polygenic risk
312 scores) and multimodal sequential modelling[10]. While not yet validated, this approach improved the
313 prediction performance from 80 % to 94% in the active treatment but not the sham treatment arm.
314 Third, treatment response rates could have been overly influenced by contextual factors that cannot be
315 modeled at the single-subject level. For example, the present trials were conducted in healthcare
316 settings with differing models of reimbursement and access to care. They also subtly differed in
317 eligibility criteria, with participants in the DepressionDC trial kept on a stable SSRI dose while
318 participants in ELECT-TDCS were antidepressant-free. These methodological challenges showcase
319 the current need in ML-based treatment prediction research to systematically assess and compare
320 potential analytic pipelines in larger samples, to include comprehensive phenotyping in RCT
321 protocols, and to harmonize best-practice symptom assessments across brain stimulation trials.
322 Models with above-chance detection rates for active tDCS and sham tDCS in the
323 DepressionDC sample also did not generalize across treatment modalities. Our feature importance
324 analyses indicated that the performance of both models was predominantly driven by baseline BDI

scores. This reliance on a single feature presents an interpretative challenge: Without a distinct feature selection profile, it becomes difficult to determine whether the models are tailored specifically to each treatment modality or if they lack generalizability to new patients.

Our analysis has several limitations. Firstly, the relatively small sample size in both datasets might have limited the performance and robustness of our classifiers[41]. Secondly, only a limited set of identical features was available from both trials. For example, negative affect, which was a key predictive feature in the prior ML analysis of the ELECT-TDCS sample[23], was not collected in DepressionDC. This omission might have reduced the predictive accuracy of models in the present analysis. Thirdly, our study was retrospective and served as a proof-of-concept analysis. Consequently, our models have not been validated prospectively, nor have they been benchmarked against clinical judgments. Fourthly, the DepressionDC trial did not demonstrate the efficacy of active tDCS at a group level. This raises the possibility that there may not have been any discernible effects at the individual subject level either, which would inherently limit the potential of our models to identify specific treatment effects. Lastly, while Kambeitz et al.[23] evaluated treatment response at week 10, we opted to identify responders at week 6 due to the availability of MADRS data at this time point across both trials. Consequently, our study could not explore and compare predictive accuracies at various endpoints.

In conclusion, our findings suggest that readily accessible clinical variables at baseline, particularly self-rated depression severity, have the potential to identify responders to active tDCS and sham tDCS treatments in patients with MDD. However, our findings also caution against the premature dissemination of predictive models that lack external validation. Future research should aim to harmonize and deepen phenotyping efforts in RCT protocols to enable the development of more robust predictive models. Ultimately, such models need to be first externally validated and then prospectively tested for their clinical utility.

Acknowledgements

This research was funded by grant 01EE1403E within the German Center for Brain Stimulation research consortium by the German Federal Ministry of Education and Research and supported within the initial phase of the German Center for Mental Health (Deutsches Zentrum für Psychische Gesundheit [DZPG], grant 01EE2303A). GB was supported by two internal grants for young researchers from the Medical Faculty of LMU Munich (FöFoLe, grant number 1127; FöFoLe+, grant number CS063). JW was supported by an internal grant for young researchers from the Medical Faculty of LMU Munich (FöFoLe, grant number 1150).

Conflict of Interest

GB, SG, LB, ED, DK, UV, KW, JW, AF, CN, PZ, TK, CS, NK and AB have no competing interests to declare. CP has received grants from the German Federal Ministry of Education and Research (01KG2003, 01EE1407H, and 01EE1403D) and the Deutsche Forschungsgemeinschaft (PL 525/4-1, PL 525/6-1, and PL 525/6-1) and holds stock options for PsyKit (Tübingen, Germany). BL has received grants from Bayhost, the EU (European School for Interdisciplinary Tinnitus Research [722046] and the Unification of Treatments and Interventions for Tinnitus Patients [848261]), and Neuromod; consulting fees from Neuromod, Decibel Therapeutics, Schwabe, Rovi, Sound Therapeutics, Sonovam, and Sea Pharma; payments from Schwabe, Neuromod, and Desyncra for lectures; payments from Schwabe for expert testimony; has a pending patent for neuronavigated transcranial magnetic stimulation coil positioning for the treatment of tinnitus; participated on a data safety monitoring board or advisory board for the Technical University of Munich (Nicstim) and Neuromod (TENT A2, TENT A3); has a chair on the executive committee of the German Society for Brain Stimulation in Psychiatry and has a fiduciary role in the Tinnitus Research Initiative; has stock or stock options from Sea Pharma; and has received free rental equipment from Magventure, Deymed, and Necstim. LF has received author royalties for book chapters from Wiley-Blackwell, Georg Thieme Verlag, Urban & Fischer Verlag, and Elsevier; and payments for classes from the German Sleep Society, the European Sleep Research Society, and Medical Association Freiburg. DK has

received a grant from the Manfred-Strohscheer Foundation for the Activity of Cerebral Networks, Amyloid and Microglia in Alzheimer's Disease (ActiGliA) project and has received travel and hotel expenses for an invited talk from the EU-funded Stimulation in Pediatrics project.MB has received consulting fees from Parexel and Bayer and payments for lectures from Johnson & Johnson. LF has received author royalties for book chapters from Wiley-Blackwell, Georg Thieme Verlag, Urban & Fischer Verlag, and Elsevier; and payments for classes from the German Sleep Society, the European Sleep Research Society, and Medical Association Freiburg. FP has received grants from the German Research Foundation (BR 4264/6-1) and the German Federal Ministry of Education and Research (01EW1903); consulting fees from Brainsway (Jerusalem, Israel) as a member of the European Scientific Advisory Board and from Sooma (Helsinki, Finland) as a member of the International Scientific Advisory Board; honoraria for workshops from Mag&More (Munich, Germany) and honoraria for lectures from the NeuroCare Group and Brainsway; and has received equipment from Mag&More, the NeuroCare Group, and Brainsway.

References

1 Institute of Health Metrics and Evaluation. Global Health Data Exchange (GHDx). 2022.http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/d780dffbe8a381b25e1416884959e88b.

2 Rush AJ, Trivedi MH, Wisniewski SR, Stewart JW, Nierenberg AA, Thase ME *et al.* Bupropion-SR, Sertraline, or Venlafaxine-XR after Failure of SSRIs for Depression. *N Engl J Med* 2006; **354**: 1231–1242.

3 Fried EI, Nesse RM. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. *Journal of Affective Disorders* 2015; **172**: 96–102.

4 Monroe SM, Harkness KL. Major Depression and Its Recurrences: Life Course Matters. *Annu Rev Clin Psychol* 2022; **18**: 329–357.

5 Border R, Johnson EC, Evans LM, Smolen A, Berley N, Sullivan PF *et al.* No Support for Historical Candidate Gene or Candidate Gene-by-Interaction Hypotheses for Major Depression Across Multiple Large Samples. *Am J Psychiatry* 2019; **176**: 376–387.

6 Schmaal L, Pozzi E, C. Ho T, van Velzen LS, Veer IM, Opel N *et al.* ENIGMA MDD: seven years of global neuroimaging studies of major depression through worldwide data sharing. *Transl Psychiatry* 2020; **10**: 172.

7 Perlman K, Benrimoh D, Israel S, Rollins C, Brown E, Tunteng J-F *et al.* A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *Journal of Affective Disorders* 2019; **243**: 503–515.

8 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. *The Lancet Psychiatry* 2016; **3**: 243–250.

9 Koutsouleris N, Wobrock T, Guse B, Langguth B, Landgrebe M, Eichhammer P *et al.* Predicting Response to Repetitive Transcranial Magnetic Stimulation in Patients With Schizophrenia Using Structural Magnetic Resonance Imaging: A Multisite Machine Learning Analysis. *Schizophrenia Bulletin* 2018; **44**: 1021–1034.

10 Dong MS, Rokicki J, Dwyer D, Papiol S, Streit F, Rietschel M *et al.* Multimodal workflows optimally predict response to repetitive transcranial magnetic stimulation in patients with schizophrenia: a multisite machine learning analysis. *Transl Psychiatry* 2024; **14**: 196.

11 Nie Z, Vairavan S, Narayan VA, Ye J, Li QS. Predictive modeling of treatment resistant depression using data from STAR*D and an independent clinical study. *PLoS One* 2018; **13**: e0197268.

12 Nunez J-J, Nguyen TT, Zhou Y, Cao B, Ng RT, Chen J *et al.* Replication of machine learning methods to predict treatment outcome with antidepressant medications in patients with major depressive disorder from STAR*D and CAN-BIND-1. *PLoS One* 2021; **16**: e0253023.

13 Meehan AJ, Lewis SJ, Fazel S, Fusar-Poli P, Steyerberg EW, Stahl D *et al.* Clinical prediction models in psychiatry: a systematic review of two decades of progress and challenges. *Mol Psychiatry* 2022; **27**: 2700–2708.

14 Browning M, Bilderbeck AC, Dias R, Dourish CT, Kingslake J, Deckert J *et al.* The clinical effectiveness of using a predictive algorithm to guide antidepressant treatment in primary care (PReDicT): an open-label, randomised controlled trial. *Neuropsychopharmacol* 2021; **46**: 1307–1314.

15 Brunoni AR, Valiengo L, Baccaro A, Zanão TA, de Oliveira JF, Goulart A *et al.* The Sertraline vs Electrical Current Therapy for Treating Depression Clinical Study: Results From a Factorial, Randomized, Controlled Trial. *JAMA Psychiatry* 2013; **70**: 383.

16 Brunoni AR, Moffa AH, Sampaio-Junior B, Borrione L, Moreno ML, Fernandes RA *et al.* Trial of Electrical Direct-Current Therapy versus Escitalopram for Depression. *N Engl J Med* 2017; **376**: 2523–2533.

17 Razza LB, Palumbo P, Moffa AH, Carvalho AF, Solmi M, Loo CK *et al.* A systematic review and meta-analysis on the effects of transcranial direct current stimulation in depressive episodes. *Depress Anxiety* 2020; **37**: 594–608.

18 Burkhardt G, Kumpf U, Crispin A, Goerigk S, Andre E, Plewnia C *et al.* Transcranial direct current stimulation as an additional treatment to selective serotonin reuptake inhibitors in adults with major depressive disorder in Germany (DepressionDC): a triple-blind, randomised, sham-controlled, multicentre trial. *Lancet* 2023.https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(23)00640-2/fulltext.

19 Borrione L, Cavendish BA, Aparicio LVM, Luethi MS, Goerigk S, Carneiro AM *et al.* Home-Use Transcranial Direct Current Stimulation for the Treatment of a Major Depressive Episode: A Randomized Clinical Trial. *JAMA Psychiatry* 2024. doi:10.1001/jamapsychiatry.2023.4948.

20 Woodham RD, Selvaraj S, Lajmi N, Hobday H, Sheehan G, Ghazi-Noori A-R *et al.* Home-based transcranial direct current stimulation RCT in major depression. bioRxiv. 2023. doi:10.1101/2023.11.27.23299059.

21 Padberg F, Bulubas L, Mizutani-Tiebel Y, Burkhardt G, Kranz GS, Koutsouleris N *et al.* The intervention, the patient and the illness – Personalizing non-invasive brain stimulation in psychiatry. *Experimental Neurology* 2021; **341**: 113713.

22 Burkhardt G, Goerigk S, Padberg F. Mood Disorders: Predictors of tDCS Response. In: Brunoni AR, Nitsche MA, Loo CK (eds). *Transcranial Direct Current Stimulation in Neuropsychiatric Disorders: Clinical Principles and Management*. Springer International Publishing: Cham, 2021, pp 481–490.

23 Kambeitz J, Goerigk S, Gattaz W, Falkai P, Benseñor IM, Lotufo PA *et al.* Clinical patterns differentially predict response to transcranial direct current stimulation (tDCS) and escitalopram in major depression: A machine learning analysis of the ELECT-TDCS study. *Journal of Affective Disorders* 2020; **265**: 460–467.

24 Seibt O, Brunoni AR, Huang Y, Bikson M. The Pursuit of DLPFC: Non-neuronavigated Methods to Target the Left Dorsolateral Pre-frontal Cortex With Symmetric Bicephalic Transcranial Direct Current Stimulation (tDCS). *Brain Stimul* 2015; **8**: 590–602.

25 Montgomery SA, Åsberg M. A New Depression Scale Designed to be Sensitive to Change. *Br J Psychiatry* 1979; **134**: 382–389.

26 Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; **17**: 520–525.

27 Burkhardt G, Adorjan K, Kambeitz J, Kambeitz-Ilankovic L, Falkai P, Eyer F *et al.* A machine learning approach to risk assessment for alcohol withdrawal syndrome. *European Neuropsychopharmacology* 2020; **35**: 61–70.

28 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 Psychiatry* 2021; **78**: 195–209.

29 Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. *ACM Trans Intell Syst Technol* 2011; **2**: 1–27.

30 Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior Distribution. 2010, pp 3121–3124.

31 Huang Y-M, Du S-X. Weighted support vector machine for classification with uneven training class sizes. 2005, pp 4365-4369 Vol. 7.

32 Golland P, Fischl B. Permutation tests for classification: towards statistical significance in image-based studies. *Inf Process Med Imaging* 2003; **18**: 330–341.

33 Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer DB *et al.* 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* 2018; **75**: 1156–1172.

34 Gómez-Verdejo V, Parrado-Hernández E, Tohka J, Alzheimer's Disease Neuroimaging Initiative. Sign-Consistency Based Variable Importance for Machine Learning in Brain Imaging. *Neuroinformatics* 2019; **17**: 593–609.

35 R. Team. R: A language and environment for statistical computing. *MSOR connections* 2014; **1**.https://apps.dtic.mil/sti/citations/AD1039033.

36 Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:14065823* 2014.http://arxiv.org/abs/1406.5823.

37 Browning M, Kingslake J, Dourish CT, Goodwin GM, Harmer CJ, Dawson GR. Predicting treatment response to antidepressant medication using early changes in emotional processing. *European Neuropsychopharmacology* 2019; **29**: 66–75.

38 Wagner DT, Tilmans L, Peng K, Niedermeier M, Rohl M, Ryan S *et al.* Artificial Intelligence in Neuroradiology: A Review of Current Topics and Competition Challenges. *Diagnostics (Basel)* 2023; **13**. doi:10.3390/diagnostics13162670.

39 Bzdok D, Meyer-Lindenberg A. Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2018; **3**: 223–230.

40 Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A *et al.* Illusory generalizability of clinical prediction models. *Science* 2024; **383**: 164–167.

41 van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* 2014; **14**: 137.