#### **Cross-trial prediction of treatment response to transcranial direct current** 1

2

# stimulation in patients with major depressive disorder

- Gerrit Burkhardt, MD<sup>1,2</sup>; Stephan Goerigk, PhD<sup>1,2,3</sup>; Lucia Bulubas, MD, PhD<sup>1,2</sup>; Esther Dechantsreiter, 3
- M.Sc.<sup>1,2</sup>; Daniel Keeser, PhD<sup>1,2</sup>; Ulrike Vogelmann, MD<sup>2,4</sup>; Katharina von Wartensleben<sup>1</sup>; Johannes 4
- Wolf, MD<sup>1,2</sup>; Christian Plewnia, MD<sup>5,6</sup>; Andreas Fallgatter, MD<sup>5,6</sup>; Berthold Langguth, MD<sup>7</sup>; Claus 5
- Normann, MD<sup>8,9</sup>; Lukas Frase, MD<sup>8,10</sup>; Peter Zwanzger, MD<sup>1,11</sup>; Thomas Kammer, MD<sup>12,13</sup>; Carlos 6
- Schönfeldt-Lecuona, MD<sup>12,13</sup>; Daniel Kamp, MD<sup>14</sup>; Malek Bajbouj, MD<sup>15,16</sup>; Nikolaos Koutsouleris, 7
- MD<sup>1,2</sup>; Andre R Brunoni, MD, PhD<sup>17,\*</sup>; Frank Padberg, MD<sup>1,2,\*</sup> 8
- 9
- 10 <sup>1</sup>Department of Psychiatry and Psychotherapy, LMU University Hospital, Munich, Germany
- <sup>2</sup>German Center for Mental Health (DZPG), Site Munich-Augsburg, Germany 11
- 12 <sup>3</sup>Charlotte Fresenius Hochschule, University of Psychology, Munich, Germany
- 13 <sup>4</sup>Department of Psychiatry and Psychotherapy, University Hospital, Technical University of Munich
- 14 (TUM), Munich, Germany
- <sup>5</sup>Tübingen Center for Mental Health, Department of Psychiatry and Psychotherapy, University of 15
- 16 Tübingen, Tübingen, Germany
- <sup>6</sup>German Center for Mental Health (DZPG), Site Tübingen, Germany 17
- 18 <sup>7</sup>Department of Psychiatry and Psychotherapy, University of Regensburg, Regensburg, Germany
- 19 <sup>8</sup>Department of Psychiatry and Psychotherapy, Medical Center, University of Freiburg, Freiburg, 20 Germany
- 21 <sup>9</sup>Center for Basics in Neuromodulation (NeuroModulBasics), University of Freiburg, Freiburg, 22 Germany
- 23 <sup>10</sup>Department of Psychosomatic Medicine and Psychotherapy, Medical Center, University of Freiburg
- 24 <sup>11</sup>kbo-Inn-Salzach-Klinikum, Clinical Center for Psychiatry, Psychotherapy, Psychosomatic Medicine, 25 Geriatrics and Neurology, Gabersee, Wasserburg/Inn, Germany
- 26 <sup>12</sup>Department of Psychiatry and Psychotherapy III, University of Ulm, Ulm, Germany
- 27 <sup>13</sup>German Center for Mental Health (DZPG), Site Mannheim-Heidelberg-Ulm, Germany
- 28 <sup>14</sup>Department of Psychiatry and Psychotherapy, LVR-Klinikum Düsseldorf, Heinrich-Heine-
- 29 Universität Düsseldorf, Medical Faculty, Düsseldorf, Germany
- 30 <sup>15</sup>Department of Psychiatry and Psychotherapy, Charité-Campus Benjamin Franklin, Berlin, Germany
- 31 <sup>16</sup>German Center for Mental Health (DZPG), Site Berlin-Potsdam, Germany
- 32 <sup>17</sup>Department of Psychiatry, University of São Paulo Medical School, São Paulo, Brazil
- 33 \* These authors equally contributed to the manuscript
- 34
- 35 Corresponding author:
- 36 Dr. Gerrit Burkhardt
- Department of Psychiatry and Psychotherapy 37
- 38 LMU University Hospital
- 39 Nußbaumstraße 7
- 40 80336 Munich, Germany
- E-Mail: gerrit.burkhardt@med.uni-muenchen.de 41
- 42 Telephone: +1149 89 4400 53381
- 43
- 44 Running title: Cross-trial prediction of tDCS response in MDD

### 45 Abstract

46 Machine-learning (ML) classification may offer a promising approach for treatment response 47 prediction in patients with major depressive disorder (MDD) undergoing non-invasive brain 48 stimulation. This analysis aims to develop and validate such classification models based on easily 49 attainable sociodemographic and clinical information across two randomized controlled trials on 50 transcranial direct-current stimulation (tDCS) in MDD. Using data from 246 patients with MDD from 51 the randomized-controlled DepressionDC and ELECT-TDCS trials, we employed an ensemble 52 machine learning strategy to predict treatment response to either active tDCS or sham tDCS/placebo, 53 defined as  $\geq$  50% reduction in the Montgomery-Åsberg Depression Rating Scale at 6 weeks. Separate 54 models for active tDCS and sham/placebo were developed in each trial and evaluated for external 55 validity across trials and for treatment specificity across modalities. Additionally, models with above-56 chance detection rates were associated with long-term outcomes to assess their clinical validity. In the 57 DepressionDC trial, models achieved a balanced accuracy of 63.5% for active tDCS and 62.5% for 58 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 59 depression was consistently ranked as the most informative feature. However, response prediction in 60 the ELECT-TDCS trial and across trials was not successful. Our findings indicate that ML-based 61 62 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 63 64 validation in larger samples and with more features. 65

- 66
- 67
- 68
- 69 70
- . .
- 71

## 72 Introduction

73 Major Depressive Disorder (MDD) represents a significant global health challenge, ranking as one of the main causes of disability worldwide<sup>1</sup>. Despite the availability of effective treatments ranging from 74 75 pharmacotherapy and psychotherapy to non-invasive and invasive neurostimulation, many patients do 76 not achieve remission, even after multiple therapeutic attempts<sup>2</sup>. The development of new interventions has proven challenging, possibly due to the heterogeneity of MDD symptoms<sup>3</sup>, its 77 varying time course<sup>4</sup>, and a lack of robust biological correlates<sup>5,6</sup>. While multiple sociodemographic, 78 79 clinical, genetic, and neuroimaging variables have been associated with responses to common 80 treatments like antidepressant medication<sup>7</sup>, these associations have not yet resulted in stratified patient 81 selection algorithms or targeted interventions. Thus, recent research has focused on developing 82 multivariate predictive models that might enable pre-treatment stratification at the individual patient level<sup>8,9</sup> and detect effects beyond the between-group level in randomized controlled trials (RCT)<sup>9,10</sup>. 83 Within this approach, initial machine learning (ML)-based predictive models are typically trained on 84 data from existing RCTs to identify responders to the treatments under investigation<sup>8,11,12</sup>. Models 85 86 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. 87 88 However, efforts to externally validate initial models remain sparse<sup>13</sup>, and consequently, few attempts have been made to validate treatment prediction models in RCTs<sup>14</sup>. 89

90 Predictive approaches are particularly relevant in the field of non-invasive brain stimulation 91 (NIBS), including repetitive transcranial magnetic stimulation (rTMS) and transcranial direct current 92 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<sup>15-17</sup>, but has yielded 93 inconclusive results in recent confirmatory multicenter RCTs<sup>18–20</sup>. Therefore, efforts to optimize 94 95 outcomes on the individual patient level are required to develop the intervention toward clinical 96 applicability<sup>21,22</sup>. A recent study reported a high predictive accuracy of an ML-based prediction model 97 for identifying responders to bifrontal transcranial direct current stimulation (tDCS), yet lacked 98 external validation<sup>23</sup>. For rTMS, ML studies have mainly focused on other neuropsychiatric

conditions, e.g. schizophrenia<sup>9,10</sup>. 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<sup>16</sup> and Germany<sup>18</sup>, 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.

108

112

### 109 Subjects and Methods

### 110 Study design

111 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

113 tDCS or sham/placebo treatment using sociodemographic and clinical baseline variables. Models

114 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

117 were then associated with long-term outcomes at follow-up to explore the clinical validity and utility

118 of the predictions.



121 Figure 1. Model development and validation

122

120

#### 123 **Study population**

We analyzed patients with MDD from two trials: 1.) DepressionDC (trial registration: NCT02530164) 124 was a multicenter RCT investigating the efficacy of 6 weeks of bifrontal tDCS as an additional 125 treatment to selective serotonin reuptake inhibitors (SSRI) in patients with MDD<sup>18</sup>. Between January 126 127 2016 and June 2020, 160 patients were recruited at seven university hospitals and one psychiatric 128 community hospital in Germany. Active tDCS was not superior to sham tDCS in reducing depressive 129 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 130 comparing active tDCS plus placebo medication, escitalopram plus sham tDCS, and sham tDCS plus 131 placebo medication in patients with MDD over 10 weeks<sup>16</sup>. Two hundred forty-five patients were 132 recruited at the University of São Paulo, Brazil, between October 2013 through July 2016. Active 133 tDCS was not non-inferior to escitalopram but superior to placebo treatment. 134

135	Both trials employed rigorous RCT methods with stringent randomization, blinding protocols,
136	and non-active sham conditions. Participants were selected based on the DSM-5 criteria for MDD
137	while excluding patients with bipolar disorder, substance abuse or dependence, dementia, and
138	personality disorders. However, DepressionDC enrolled patients currently receiving SSRIs, whereas
139	ELECT-TDCS required participants to be antidepressant-free. Furthermore, DepressionDC permitted
140	the inclusion of patients with marginally lower symptom severity, as assessed by the Hamilton
141	Depression Rating Scale (HDRS), and restricted participation to a narrower age range. A
142	comprehensive list of eligibility criteria for both trials is provided in the supplement.
143	The intervention followed a basically identical tDCS protocol with 2 mA stimulation of the
144	DLPFC over 30 minutes per session. However, protocols differed in precise electrode placement, i.e.
145	DepressionDC using the F3 (anode) and F4 (cathode) based on the international
146	electroencephalogram 10-20 system, and ELECT-TDCS the Omni-Lateral-Electrode (OLE) system
147	(left DLPFC anode and right DLPFC cathode) with slightly more lateral positions of electrodes <sup>24</sup> .
148	Also, treatment lasted 10 weeks with a total of 22 treatment sessions (15 sessions in the first 3 weeks
149	and 7 weekly sessions for the remaining treatment period) in ELECT-TDCS versus 6 weeks with a
150	total of 24 treatment sessions (20 sessions in the first 4 weeks and 2 sessions per week for 2 weeks) in
151	DepressionDC, with efficacy assessed at these time points using the HDRS and Montgomery-Åsberg
152	Depression Rating Scale (MADRS) as primary outcome measures, respectively. Weekly MADRS
153	scores were additionally collected in ELECT-TDCS as a secondary outcome measure.
154	We used data from all patients with available depression scores on the MADRS at baseline
155	and week 6 after randomization. Thus, the analysis included 136 (active tDCS: 72 patients; sham
156	tDCS: 64 patients) out of 150 patients from the DepressionDC sample and 110 (active tDCS +
157	placebo: 66 patients; sham tDCS + placebo: 44 patients) out of 154 patients from the respective
158	treatment arms of the ELECT-TDCS sample. All participants had provided their written informed
159	consent before inclusion in the respective study. Both studies were approved by the local ethics
160	committees and conducted in accordance with the Declaration of Helsinki.
161	

#### 163 **Prediction target and features**

- 164 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 165 depression)<sup>25</sup>. Pursuing a data-driven approach, we included all variables available across the studies 166 167 at baseline as potential predictors. This amounted to 15 features, including basic sociodemographic 168 information (age, sex, years of education, marriage, unemployment), medical history (body mass 169 index, smoker status, diagnoses of hypertension, diabetes, and/or hypothyroidism), psychiatric history 170 (age of MDD onset, duration of MDD episode, family history of MDD), and baseline depression 171 severity (MADRS and Beck Depression Inventory-II [BDI-II] total scores).
- 172

#### 173 Machine learning analysis

174 All ML analyses were conducted using the in-house, open-source software package

175 NeuroMiner, version 1.05 (https://github.com/neurominer-git/NeuroMiner-1), running on MATLAB

176 (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.

178 In each CV<sub>1</sub> fold, we scaled all features from 0 to 1 and substituted missing values via 7-nearest

179 neighbor-based imputation (Euclidean distance for continuous, Hamming distance for categorical

180 variables)<sup>26</sup>. Following a previous approach<sup>27,28</sup>, each processed  $CV_1$  training sample then entered a

181 greedy stepwise forward search wrapper employing a linear support vector machine algorithm (SVM;

182 LIBSVM 3.12<sup>29</sup>) to iteratively select a subset of 50% features with highest predictive performance

183 (balanced accuracy  $[BAC = \frac{sensitivity + specificity}{2}]^{30}$  on the held-out CV<sub>1</sub> data) across a range of C

184 hyperparameters  $(2^{[-4_{\epsilon}\mathbb{Z}\to+4]})$ . To account for uneven distributions of the outcome labels

185 (response/non-response), optimal C hyperparameters were multiplied with the inverse ratio of the

training group sizes<sup>31</sup>. For each  $CV_1$  permutation, all  $CV_1$  models were retrained with the optimal

187 model hyperparameters, and this ensemble was then applied to the  $CV_2$  test data without modification.

188 Classification probabilities for each CV<sub>2</sub> test subject were retrieved by combining the decisions across

all models. We calculated permutation-based *p*-values to define which models reached above-chance

190	detection rates ( $\alpha$ =0.05; 1000 permutations) <sup>32</sup> . To understand which features were most reliably
191	contributing to the prediction of treatment response, we computed the CV ratio <sup>33</sup> . The feature
192	importance of variables was further estimated using sign-based consistency mapping <sup>34</sup> . For cross-trial
193	and cross-treatment modality validation, we applied $\mathrm{CV}_1$ ensembles with permutation-based above-
194	chance detection rates without modification to the respective other samples. Out-of-sample
195	performance metrics were calculated by comparing the predicted versus the observed outcome labels
196	over all CV <sub>2</sub> predictions.
197	
198	Post-hoc clinical validation
199	Additional validation analyses and visualizations were performed in R, version 4.3.2 <sup>35</sup> . Results were
200	considered significant at $\alpha$ =0.05. To explore the clinical validity of all classifiers with above-chance
201	detection rates, we fit linear mixed models (LMM) using the lme4 package <sup>36</sup> to predict MADRS and
202	GAF changes from baseline until the trials' follow-up appointments based on the models' assigned
203	probability to be a responder (formula: change $\sim$ assigned probability). The model included the
204	treatment site as a random effect (formula: $\sim 1 $ site). The significance of the model factors was
205	determined using omnibus tests (type III ANOVA) with Satterthwaite approximation to degrees of
206	freedom.
207	

# 208 **Results**

### 209 Main classifiers

210 In the DepressionDC trial, 24 patients (33%) had responded to active tDCS treatment at week 6.

211 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];

213 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

215 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%).

217	In the sham group, 32 patients (50%) showed a treatment response. Compared to sham non-
218	responders, these patients had lower clinician-rated (mean [SD] MADRS scores: 22 [5.3] vs. 24.3
219	[4.4]; p=0.017) and self-reported depression severity (mean [SD] BDI scores: 23 [9] vs. 30 [10];
220	p=0.005) at baseline. Sham responders were predicted with an above-chance cross-validated BAC of
221	62.5% (p=0.023; Table 2) and prognostic summary index of 25.1%. In both the tDCS and sham
222	analyses, only baseline BDI scores reliably and significantly contributed to the classifier decisions
223	(Figure 3).
224	In the ELECT trial, 27 patients (41%) had responded to active tDCS treatment at week 6.
225	Compared to tDCS non-responders, they had higher clinician-rated depression at baseline (mean [SD]
226	MADRS scores: 29 [8] vs. 26 [6]; p=0.020). Responders were predicted with a BAC of 61.0% that did
227	not reach our above-chance detection criterion (p=0.071; Table 2). Similarly, our analysis did not
228	yield models with above-chance detection rates for the 14 (32%) responders to sham treatment
229	(BAC=42.9%; p=0.88; Table 2).
230	
231	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.

- 240
- 241
- 242
- 243
- 244

Feature	Dep	pressionDC		ELECT			
	Non-responder, $n = 48^{a}$	Responder, $n = 24^{a}$	p- value <sup>b</sup>	Non-responders, $n = 39^{a}$	Responder, $n = 27^{a}$	p- value <sup>b</sup>	
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 <sup>2</sup>	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	

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

246

<sup>a</sup>Mean (SD); n (%). <sup>b</sup>Pearson's Chi-squared test; Wilcoxon rank sum test. Fisher's exact test.

Model	ТР	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

247 
**Table 2.** Prediction results of MADRS response models.

248 Abbreviations: TP, True Positive; TN, True Negative; FP, False Positive; FN, False Negative; Sens, Sensitivity; Spec, Specificity; PPV, Positive Predictive Value; NPV,

249 Negative Predictive Value; PSI, Prognostic Summary Index; AUC, Area Under the Curve; BAC, Balanced Accuracy. Note: p-values were calculated using permutation 250 analysis ( $\alpha$ =0.05; 1000 permutations).



252 253

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





255

256 Figure 3. Feature importance. A positive cross-validation ratio suggests that a higher feature value predicts 257 MADRS response, whereas a negative ratio implies that lower values do. A higher sign-based consistency 258 suggests that feature weights were more consistently positive or negative across the ensemble. Significance was 259 assessed by defining a hypothesis test for the importance score with a null hypothesis of 0 and an alternative 260 hypothesis of not 0. FDR-corrected p-values were then calculated using a cumulative distribution function of z-261 scores ( $\alpha$ =0.05; red dotted line). 262

#### 263 **Post-hoc clinical validation**

- For the DepressionDC tDCS model, classification probabilities significantly predicted MADRS scores 264
- at the 18-week follow-up visit ( $F_{(1.60)} = 4.53$ ,  $p_{FDR} = .037$ ,  $R^2 = 0.069$ ), but not MADRS scores at 265
- week 30 or GAF scores at weeks 18 and 30 (Supplementary Table S2 and Supplementary Figure S1). 266
- 267 Classification probabilities of the DepressionDC sham model did not predict MADRS or GAF
- 268 outcomes at weeks 18 and 30.

# 270 **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 278 279 attempts to predict responses to antidepressant medication using ML algorithms trained on clinical variables<sup>9,11,12,37</sup>. Although these performances are modest compared to the classification benchmarks 280 set in other medical disciplines, such as neuroradiology<sup>38</sup>, antidepressant response prediction remains 281 282 a challenging task relying on subjective judgment, and thus, even small increases in predictive 283 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 284 to improve treatment outcomes when applied as a decision-making tool<sup>14</sup>, strong indicators are needed 285 to justify further development of classifiers beyond the proof-of-concept stage. In the context of our 286 study, the tDCS response classifier for the DepressionDC sample significantly predicted depression 287 severity at the 18-week follow-up visit, suggesting potential clinical validity. Nonetheless, given the 288 289 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<sup>39</sup>. 290

Recent research, including a validation attempt across trials on antipsychotic medication for schizophrenia<sup>40</sup>, 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

ELECT-TDCS<sup>16,18</sup>. By contrast, baseline depression severity was comparable between trials. Second, 297 298 compared to a previous ML prediction study in the ELECT-TDCS cohort<sup>23</sup>, our models showed considerably lower predictive performance in the same dataset. Since our analysis aimed to develop 299 generalizable prediction models across two RCT cohorts, we took several methodological choices that 300 301 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 generate generalizable models<sup>9,27,28</sup>. We also limited the input variables to features available in both 304 305 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 Magnetic Resonance Imaging (sMRI)<sup>9</sup> by incorporating further data domains (i.e. polygenic risk 311 scores) and multimodal sequential modelling<sup>10</sup>. While not yet validated, this approach improved the 312 prediction performance from 80 % to 94% in the active treatment but not the sham treatment arm. 313 314 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 settings with differing models of reimbursement and access to care. They also subtly differed in 316 317 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

325 scores. This reliance on a single feature presents an interpretative challenge: Without a distinct feature 326 selection profile, it becomes difficult to determine whether the models are tailored specifically to each 327 treatment modality or if they lack generalizability to new patients.

Our analysis has several limitations. Firstly, the relatively small sample size in both datasets 328 might have limited the performance and robustness of our classifiers<sup>41</sup>. Secondly, only a limited set of 329 identical features was available from both trials. For example, negative affect, which was a key 330 predictive feature in the prior ML analysis of the ELECT-TDCS sample<sup>23</sup>, was not collected in 331 332 DepressionDC. This omission might have reduced the predictive accuracy of models in the present 333 analysis. Thirdly, our study was retrospective and served as a proof-of-concept analysis. 334 Consequently, our models have not been validated prospectively, nor have they been benchmarked 335 against clinical judgments. Fourthly, the DepressionDC trial did not demonstrate the efficacy of active 336 tDCS at a group level. This raises the possibility that there may not have been any discernible effects 337 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.<sup>23</sup> evaluated treatment response at 338 339 week 10, we opted to identify responders at week 6 due to the availability of MADRS data at this time 340 point across both trials. Consequently, our study could not explore and compare predictive accuracies 341 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.

# 350 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).

## 358 **Conflict of Interest**

359 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

361 (01KG2003, 01EE1407H, and 01EE1403D) and the Deutsche Forschungsgemeinschaft (PL 525/4-1,

362 PL 525/6-1, and PL 525/6-1) and holds stock options for PsyKit (Tübingen, Germany). BL has

363 received grants from Bayhost, the EU (European School for Interdisciplinary Tinnitus Research

364 [722046] and the Unification of Treatments and Interventions for Tinnitus Patients [848261]), and

365 Neuromod; consulting fees from Neuromod, Decibel Therapeutics, Schwabe, Rovi, Sound

366 Therapeutics, Sonovam, and Sea Pharma; payments from Schwabe, Neuromod, and Desyncra for

367 lectures; payments from Schwabe for expert testimony; has a pending patent for neuronavigated

368 transcranial magnetic stimulation coil positioning for the treatment of tinnitus; participated on a data

369 safety monitoring board or advisory board for the Technical University of Munich (Nicstim) and

370 Neuromod (TENT A2, TENT A3); has a chair on the executive committee of the German Society for

371 Brain Stimulation in Psychiatry and has a fiduciary role in the Tinnitus Research Initiative; has stock

372 or stock options from Sea Pharma; and has received free rental equipment from Magventure, Deymed,

373 and Necstim. LF has received author royalties for book chapters from Wiley-Blackwell, Georg

374 Thieme Verlag, Urban & Fischer Verlag, and Elsevier; and payments for classes from the German

375 Sleep Society, the European Sleep Research Society, and Medical Association Freiburg. DK has

- 376 received a grant from the Manfred-Strohscheer Foundation for the Activity of Cerebral Networks,
- 377 Amyloid and Microglia in Alzheimer's Disease (ActiGliA) project and has received travel and hotel
- 378 expenses for an invited talk from the EU-funded Stimulation in Pediatrics project.MB has received
- 379 consulting fees from Parexel and Bayer and payments for lectures from Johnson & Johnson. LF has
- 380 received author royalties for book chapters from Wiley-Blackwell, Georg Thieme Verlag, Urban &
- 381 Fischer Verlag, and Elsevier; and payments for classes from the German Sleep Society, the European
- 382 Sleep Research Society, and Medical Association Freiburg. FP has received grants from the German
- 383 Research Foundation (BR 4264/6-1) and the German Federal Ministry of Education and Research
- 384 (01EW1903); consulting fees from Brainsway (Jerusalem, Israel) as a member of the European
- 385 Scientific Advisory Board and from Sooma (Helsinki, Finland) as a member of the International
- 386 Scientific Advisory Board; honoraria for workshops from Mag&More (Munich, Germany) and
- 387 honoraria for lectures from the NeuroCare Group and Brainsway; and has received equipment from
- 388 Mag&More, the NeuroCare Group, and Brainsway.

## 389 **References**

- 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.
- 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.
- 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.
- Monroe SM, Harkness KL. Major Depression and Its Recurrences: Life Course Matters. *Annu Rev Clin Psychol* 2022; 18: 329–357.
- 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.
- 403 6 Schmaal L, Pozzi E, C. Ho T, van Velzen LS, Veer IM, Opel N *et al.* ENIGMA MDD: seven
   404 years of global neuroimaging studies of major depression through worldwide data sharing. *Transl* 405 *Psychiatry* 2020; **10**: 172.
- 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.

- 409 Chekroud AM, Zotti RJ, Shehzad Z, Gueorguieva R, Johnson MK, Trivedi MH et al. Cross-trial 8 prediction of treatment outcome in depression: a machine learning approach. The Lancet 410 411 Psychiatry 2016; 3: 243-250.
- 412 9 Koutsouleris N. Wobrock T. Guse B. Langguth B. Landgrebe M. Eichhammer P et al. Predicting 413 Response to Repetitive Transcranial Magnetic Stimulation in Patients With Schizophrenia Using 414 Structural Magnetic Resonance Imaging: A Multisite Machine Learning Analysis. Schizophrenia 415 Bulletin 2018; 44: 1021–1034.
- 416 10 Dong MS, Rokicki J, Dwyer D, Papiol S, Streit F, Rietschel M et al. Multimodal workflows 417 optimally predict response to repetitive transcranial magnetic stimulation in patients with 418 schizophrenia: a multisite machine learning analysis. Transl Psychiatry 2024; 14: 196.
- 419 11 Nie Z, Vairavan S, Narayan VA, Ye J, Li QS. Predictive modeling of treatment resistant 420 depression using data from STAR\*D and an independent clinical study. PLoS One 2018; 13: 421 e0197268.
- 422 12 Nunez J-J, Nguyen TT, Zhou Y, Cao B, Ng RT, Chen J et al. Replication of machine learning 423 methods to predict treatment outcome with antidepressant medications in patients with major 424 depressive disorder from STAR\*D and CAN-BIND-1. PLoS One 2021; 16: e0253023.
- 425 13 Meehan AJ, Lewis SJ, Fazel S, Fusar-Poli P, Steyerberg EW, Stahl D et al. Clinical prediction 426 models in psychiatry: a systematic review of two decades of progress and challenges. Mol 427 Psychiatry 2022; 27: 2700–2708.
- 428 14 Browning M, Bilderbeck AC, Dias R, Dourish CT, Kingslake J, Deckert J et al. The clinical 429 effectiveness of using a predictive algorithm to guide antidepressant treatment in primary care 430 (PReDicT): an open-label, randomised controlled trial. Neuropsychopharmacol 2021; 46: 1307-431 1314.
- 432 15 Brunoni AR, Valiengo L, Baccaro A, Zanão TA, de Oliveira JF, Goulart A et al. The Sertraline vs 433 Electrical Current Therapy for Treating Depression Clinical Study: Results From a Factorial, 434 Randomized, Controlled Trial. JAMA Psychiatry 2013; 70: 383.
- 435 16 Brunoni AR, Moffa AH, Sampaio-Junior B, Borrione L, Moreno ML, Fernandes RA et al. Trial 436 of Electrical Direct-Current Therapy versus Escitalopram for Depression. N Engl J Med 2017; 437 **376**: 2523–2533.
- 438 17 Razza LB, Palumbo P, Moffa AH, Carvalho AF, Solmi M, Loo CK et al. A systematic review 439 and meta-analysis on the effects of transcranial direct current stimulation in depressive episodes. 440 Depress Anxiety 2020; 37: 594–608.
- 441 18 Burkhardt G, Kumpf U, Crispin A, Goerigk S, Andre E, Plewnia C et al. Transcranial direct 442 current stimulation as an additional treatment to selective serotonin reuptake inhibitors in adults 443 with major depressive disorder in Germany (DepressionDC): a triple-blind, randomised, sham-444 controlled, multicentre trial. Lancet
- 2023.https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(23)00640-2/fulltext. 445
- 446 19 Borrione L, Cavendish BA, Aparicio LVM, Luethi MS, Goerigk S, Carneiro AM et al. Home-447 Use Transcranial Direct Current Stimulation for the Treatment of a Major Depressive Episode: A 448 Randomized Clinical Trial. JAMA Psychiatry 2024. doi:10.1001/jamapsychiatry.2023.4948.
- 449 20 Woodham RD, Selvaraj S, Lajmi N, Hobday H, Sheehan G, Ghazi-Noori A-R et al. Home-based 450 transcranial direct current stimulation RCT in major depression. bioRxiv. 2023. 451 doi:10.1101/2023.11.27.23299059.

- 452 21 Padberg F, Bulubas L, Mizutani-Tiebel Y, Burkhardt G, Kranz GS, Koutsouleris N et al. The 453 intervention, the patient and the illness – Personalizing non-invasive brain stimulation in 454 psychiatry. Experimental Neurology 2021; 341: 113713.
- 455 22 Burkhardt G, Goerigk S, Padberg F, Mood Disorders: Predictors of tDCS Response. In: Brunoni 456 AR, Nitsche MA, Loo CK (eds). Transcranial Direct Current Stimulation in Neuropsychiatric 457 Disorders: Clinical Principles and Management. Springer International Publishing: Cham, 2021, 458 pp 481–490.
- 459 23 Kambeitz J, Goerigk S, Gattaz W, Falkai P, Benseñor IM, Lotufo PA et al. Clinical patterns 460 differentially predict response to transcranial direct current stimulation (tDCS) and escitalopram 461 in major depression: A machine learning analysis of the ELECT-TDCS study. Journal of 462 Affective Disorders 2020; 265: 460–467.
- 463 24 Seibt O, Brunoni AR, Huang Y, Bikson M. The Pursuit of DLPFC: Non-neuronavigated Methods 464 to Target the Left Dorsolateral Pre-frontal Cortex With Symmetric Bicephalic Transcranial Direct 465 Current Stimulation (tDCS). Brain Stimul 2015; 8: 590-602.
- 466 25 Montgomery SA, Åsberg M. A New Depression Scale Designed to be Sensitive to Change. Br J Psychiatry 1979; 134: 382-389. 467
- 468 26 Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R et al. Missing value 469 estimation methods for DNA microarrays. Bioinformatics 2001; 17: 520-525.
- 470 27 Burkhardt G, Adorjan K, Kambeitz J, Kambeitz-Ilankovic L, Falkai P, Eyer F et al. A machine 471 learning approach to risk assessment for alcohol withdrawal syndrome. European Neuropsychopharmacology 2020; 35: 61–70. 472
- 473 28 Koutsouleris N, Dwyer DB, Degenhardt F, Maj C, Urquijo-Castro MF, Sanfelici R et al. 474 Multimodal Machine Learning Workflows for Prediction of Psychosis in Patients With Clinical 475 High-Risk Syndromes and Recent-Onset Depression. JAMA Psychiatry 2021; 78: 195-209.
- 476 29 Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM Trans Intell Syst 477 *Technol* 2011; **2**: 1–27.
- 30 Brodersen KH, Ong CS, Stephan KE, Buhmann JM. The Balanced Accuracy and Its Posterior 478 479 Distribution. 2010, pp 3121–3124.
- 31 Huang Y-M, Du S-X. Weighted support vector machine for classification with uneven training 480 481 class sizes. 2005, pp 4365-4369 Vol. 7.
- 482 32 Golland P, Fischl B. Permutation tests for classification: towards statistical significance in image-483 based studies. Inf Process Med Imaging 2003; 18: 330-341.
- 484 33 Koutsouleris N, Kambeitz-Ilankovic L, Ruhrmann S, Rosen M, Ruef A, Dwyer DB et al. 485 Prediction Models of Functional Outcomes for Individuals in the Clinical High-Risk State for 486 Psychosis or With Recent-Onset Depression: A Multimodal, Multisite Machine Learning Analysis. JAMA Psychiatry 2018; 75: 1156–1172. 487
- 488 34 Gómez-Verdejo V, Parrado-Hernández E, Tohka J, Alzheimer's Disease Neuroimaging Initiative. 489 Sign-Consistency Based Variable Importance for Machine Learning in Brain Imaging. 490 Neuroinformatics 2019; 17: 593-609.
- 491 35 R. Team. R: A language and environment for statistical computing. MSOR connections 2014; 492 1.https://apps.dtic.mil/sti/citations/AD1039033.

493 494	36	Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. <i>arXiv</i> preprint arXiv:14065823 2014.http://arxiv.org/abs/1406.5823.
495 496 497	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. <i>European Neuropsychopharmacology</i> 2019; <b>29</b> : 66–75.
498 499 500	38	Wagner DT, Tilmans L, Peng K, Niedermeier M, Rohl M, Ryan S <i>et al.</i> Artificial Intelligence in Neuroradiology: A Review of Current Topics and Competition Challenges. <i>Diagnostics (Basel)</i> 2023; <b>13</b> . doi:10.3390/diagnostics13162670.
501 502	39	Bzdok D, Meyer-Lindenberg A. Machine Learning for Precision Psychiatry: Opportunities and Challenges. <i>Biol Psychiatry Cogn Neurosci Neuroimaging</i> 2018; <b>3</b> : 223–230.
503 504	40	Chekroud AM, Hawrilenko M, Loho H, Bondar J, Gueorguieva R, Hasan A <i>et al.</i> Illusory generalizability of clinical prediction models. <i>Science</i> 2024; <b>383</b> : 164–167.
505 506	41	van der Ploeg T, Austin PC, Steyerberg EW. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. <i>BMC Med Res Methodol</i> 2014; <b>14</b> : 137.
507		
508		
509		
510		
511		
512		
513		
514		
515		
516		
517		
518		
519		
520		
521		
522		
523		
524		
525		
520		
528		
529		
545		