

ARTICLE

Open Access

# Predicting rehospitalization within 2 years of initial patient admission for a major depressive episode: a multimodal machine learning approach

Micah Cearns<sup>1</sup>, Nils Opel<sup>2,3</sup>, Scott Clark<sup>1</sup>, Claas Kaehler<sup>2</sup>, Anbupalam Thalamuthu<sup>4</sup>, Walter Heindel<sup>5</sup>, Theresa Winter<sup>6,7</sup>, Henning Teismann<sup>8</sup>, Heike Minnerup<sup>8</sup>, Udo Dannlowski<sup>2</sup>, Klaus Berger<sup>8</sup> and Bernhard T. Baune<sup>2,9,10</sup>

## Abstract

Machine learning methods show promise to translate univariate biomarker findings into clinically useful multivariate decision support systems. At current, works in major depressive disorder have predominantly focused on neuroimaging and clinical predictor modalities, with genetic, blood-biomarker, and cardiovascular modalities lacking. In addition, the prediction of rehospitalization after an initial inpatient major depressive episode is yet to be explored, despite its clinical importance. To address this gap in the literature, we have used baseline clinical, structural imaging, blood-biomarker, genetic (polygenic risk scores), bioelectrical impedance and electrocardiography predictors to predict rehospitalization within 2 years of an initial inpatient episode of major depression. Three hundred and eighty patients from the ongoing 12-year Bidirect study were included in the analysis (rehospitalized: yes = 102, no = 278). Inclusion criteria was age  $\geq 35$  and  $< 66$  years, a current or recent hospitalisation for a major depressive episode and complete structural imaging and genetic data. Optimal performance was achieved with a multimodal panel containing structural imaging, blood-biomarker, clinical, medication type, and sleep quality predictors, attaining a test AUC of 67.74 ( $p = 9.99 \times 10^{-5}$ ). This multimodal solution outperformed models based on clinical variables alone, combined biomarkers, and individual data modality prognostication for rehospitalization prediction. This finding points to the potential of predictive models that combine multimodal clinical and biomarker data in the development of clinical decision support systems.

## Introduction

Relapse rates in specialized mental healthcare settings are high in patients with major depressive disorder (MDD) (60% after 5 years, 67% after 10, and 85% after 15)<sup>1</sup>. To better predict relapse, previous research has predominantly focused on the presence and magnitude of clinical symptoms, including residual depressive symptoms<sup>2</sup>, illness severity<sup>3</sup>, number of prior episodes<sup>4</sup>, age of onset<sup>4</sup>, and comorbid personality disorders<sup>5</sup>. Other

studies have explored group level associations between biomarkers and relapse, observing smaller hippocampal volumes<sup>6</sup>, higher levels of post-treatment glucocorticoids<sup>7</sup>, high cortisol response on the combined dexamethasone-CRH test<sup>8</sup>, as well as catecholamine and tryptophan depletion<sup>9</sup>. Such findings have been beneficial in constructing modality specific aetiological hypotheses as well as sign posts for relapse in clinical practice. However, the elucidation of clinically meaningful predictors of relapse is contingent on the construct validity of the prediction outcome and the size of the analysed sample. Some of these studies identifying biomarkers as predictors of relapse have used self-reported relapse into a new depressive episode and samples of less than 50 patients<sup>7,10</sup>. Whether patients have truly relapsed into a

Correspondence: Bernhard T. Baune (bbaune@unimelb.edu.au)

<sup>1</sup>Discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, Australia

<sup>2</sup>Department of Psychiatry, University of Münster, Münster, Germany

Full list of author information is available at the end of the article.

These authors contributed equally: Micah Cearns, Nils Opel

© The Author(s) 2019



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

new depressive episode or merely have a treatment refractory illness is unknown. These observations may help to explain the lack of empirically validated relapse prediction tools and subsequent intervention strategies for relapse prevention<sup>10</sup>.

To overcome this blind spot in patient care, studies of increased size that focus on a well-defined outcome such as rehospitalization are needed to identify clinically robust predictors of this illness trajectory. To facilitate this, a longitudinal, multimodal sample of clinically diagnosed inpatients is required. The BiDirect depression cohort provides such a sample<sup>11</sup>. In addition, multimodal, multivariate modelling techniques that prognosticate individually, rather than at the group level are needed given the heterogeneous nature of MDD and its illness trajectories<sup>12,13</sup>.

Recent work has shown that machine learning (ML) models are well suited to problems of this nature, demonstrating their efficacy in drug response and functional outcome prediction in MDD<sup>14,15</sup>, while other studies using neuroimaging modalities have shown similar success for ML-based MDD diagnostics<sup>16</sup>. Interestingly, no studies have applied ML models to relapse prediction; thus, the utility of differing data modalities for relapse prediction remains poorly understood. Regarding illness trajectory modelling, one previous work from Schmaal et al.<sup>17</sup> combined neuroimaging and clinical data to predict MDD remission trajectories with moderate success (accuracy = 69–73%). Further, Koutsouleris et al.<sup>15</sup> used both neuroimaging and clinical data to predict functional outcomes in recent onset MDD (balanced accuracy = 70.3%). However, studies employing a combination of neuroimaging markers with biomarkers from different modalities as well as clinical data are lacking up until now.

Given the apparent gap in multimodal studies of outcome prediction in MDD, we have combined a range of clinical and biomarker predictors that have shown significant associations to MDD in previous works but are yet to be used for ML-based rehospitalization trajectory modelling. Modalities used included clinical<sup>18</sup>, blood biomarker<sup>19–21</sup>, structural imaging<sup>16</sup>, electrocardiography<sup>22</sup>, genetic<sup>23</sup>, cognitive<sup>24</sup>, nutritional<sup>25</sup>, sleep<sup>26</sup>, and exercise<sup>27</sup>. Using these modalities, we predicted rehospitalization in a cohort of patients within 2 years of initial hospitalization for an acute episode of MDD.

## Materials and methods

### Dataset description

The BiDirect study is an ongoing study of (a) patients, hospitalized for an acute episode of major depression at time of recruitment, (b) population controls randomly drawn from the register of the city of Münster<sup>11</sup>, and (c) patients 3 months after an acute coronary event or myocardial infarction. Examination of all participants

included a computer-assisted face-to-face interview of socio-demographic characteristics and medical history as well as an extensive psychiatric assessment (Supplementary Information 1.1). Only patients in the depression cohort were used in this analysis.

At baseline, 999 MDD patients were recruited and 684 completed their 2-year follow-up and provided their rehospitalization status. Specifically, patients were asked, have you had at least one or more re-admissions to hospital for an acute depressive episode since the initial examination? Patient response was recorded by study assistants. As our aim was to assess both multi and unimodal prediction models of reported rehospitalization, a requirement for inclusion was complete imaging and genetic data. Twenty-nine patients had incomplete genetic data while 294 had incomplete imaging data. In addition, 14 participants were excluded due to poor MRI quality, leaving a final sample of 380 participants (rehospitalized: yes = 102, no = 278). See Table 1 for socio-demographic characteristics. In this final sample, 87.9% (334/380) of patients were taking some form of antidepressant medication at their baseline assessment, 40.3% (153/380) were taking an antipsychotic, while 92.6% (352/380) were taking some form of psychotropic medication (see Table 2).

### Predictor modalities

#### Clinical

Detailed information regarding socio-demographic and socio-economic status, lifetime medical diagnoses, current medication use, healthcare utilization, insurance status, lifestyle and risk behaviour (e.g., diet, physical activity, alcohol consumption, smoking status), and perceived health state was collected via a computer-assisted interview. A combination of individual items as well as total scores were included from the Hamilton Depression Rating Scale (HAM-D), the Hamilton Anxiety Rating Scale (HAM-A), the Center for Epidemiologic Studies Depression Scale (CES-D), the Inventory of Depressive Symptomatology (IDS), the International Physical Activity questionnaire (IPAQ), and the Food Frequency Questionnaire (FFQ) (Supplementary Information 1.1 and 1.2). In addition, we included measures from a cognitive functioning module and several self-report measures (Supplementary Information 1.3). In total, 208 clinical and demographic predictors were included.

#### Structural imaging

We included imaging data derived from structural magnetic imaging sequences (3D-T1). To reduce the size of the predictor space, we a priori selected 15 regions that have been shown to be significantly associated with MDD in previous ENIGMA meta-analyses<sup>6,28</sup>. Selected regions included right and left mean hippocampal volume, cortical

**Table 1 Summary statistics for the final study sample**

	Mean	SD	Min	Max		Mean	SD	Min	Max	P
Rehospitalized? Yes ( <i>n</i> = 102)					Rehospitalized? No ( <i>n</i> = 278)					
Sex (m/f)					Sex (m/f)					
( <i>n</i> = 43/59)					( <i>n</i> = 108/170)					
Age					Age					
( <i>n</i> = 102)	49.03	7.32	34.96	63.96	( <i>n</i> = 278)	49.91	7.38	35.15	65.37	0.3
HAM-D total					HAM-D total					
( <i>n</i> = 101)	15.33	6.59	0.00	27.00	( <i>n</i> = 278)	12.71	6.33	0.00	33.00	<0.01
CES-D total					CES-D total					
( <i>n</i> = 102)	31.30	12.97	1.00	56.00	( <i>n</i> = 276)	25.40	11.50	0.00	48.00	<0.01
Total inpatient episodes					Total inpatient episodes					
( <i>n</i> = 101)	2.06	2.00	0.00	10.00	( <i>n</i> = 274)	1.42	0.90	0.00	6.00	<0.01

Means, standard deviations (SD), minimal (Min), and maximal (Max) values are presented. Significance testing between groups was conducted with independent samples *t*-tests. "Total inpatient episodes" includes the baseline assessment inpatient episode as well as all previous inpatient episodes. *HAM-D total* total score for the first 17 items of the HAM scale; *CES-D total* total score with the inversion of positive items 4, 8, 12, and 16 taken into account

**Table 2 Percentage proportions and total counts for psychotropic medication use in each rehospitalization outcome group**

Medication	Rehospitalized? Yes ( <i>N</i> = 278)	Rehospitalized? No ( <i>N</i> = 102)	<i>P</i>
Selective serotonin reuptake inhibitors	29.50% ( <i>N</i> = 82)	22.55% ( <i>N</i> = 23)	0.18
Beta blocking agents	18.71% ( <i>N</i> = 52)	10.78% ( <i>N</i> = 11)	0.07
Non-selective monoamine reuptake inhibitors	10.43% ( <i>N</i> = 29)	13.73% ( <i>N</i> = 14)	0.37
Other antidepressants	57.91% ( <i>N</i> = 161)	71.57% ( <i>N</i> = 73)	0.06
Benzodiazepines	30.58% ( <i>N</i> = 85)	53.92% ( <i>N</i> = 55)	<0.01
Butyrophenone derivates	3.96% ( <i>N</i> = 11)	8.82% ( <i>N</i> = 9)	0.06
Diazepines, oxazepines, thiazepines, oxepines	24.10% ( <i>N</i> = 67)	41.18% ( <i>N</i> = 42)	<0.01
Lithium	2.52% ( <i>N</i> = 7)	3.92% ( <i>N</i> = 4)	0.47
Other antipsychotics	6.47% ( <i>N</i> = 18)	17.65% ( <i>N</i> = 18)	0.01

Significant differences between groups were assessed using chi-square tests

thickness of the bilateral medial orbitofrontal cortex (OFC), fusiform gyrus, insula, rostral anterior and posterior cingulate cortex and unilaterally in the left middle temporal gyrus, right inferior temporal gyrus and the right caudal anterior cingulate cortex (Supplementary Information 1.4).

#### Serum and genetic markers

Our serum biomarker panel consisted of 10 measures of high sensitive C-reactive protein<sup>19</sup>, free triiodothyronine, thyroxine, thyroid-stimulating hormone<sup>29–31</sup>, 17 beta-estradiol<sup>32,33</sup>, sex hormone-binding globulin, testosterone, the free androgen index<sup>34–36</sup>, total cholesterol<sup>37</sup>, and high-density lipoprotein cholesterol<sup>38</sup>. Due to the correlations between genetic variants shared across psychiatric

traits and common comorbidities between psychiatric disorders<sup>39,40</sup>, we also included seven polygenic risk scores (PGRS) with a *p* value threshold of 0.5, for MDD, anxiety, Alzheimer's, anorexia, autism spectrum disorder, bipolar, and schizophrenia (Supplementary Information 1.5–1.6 and Supplementary Information 2.1–2.6). In addition, we also assessed PGRS with *p* value thresholds of 0.05 and 0.01 (Supplementary Information 3.2).

#### Cardiovascular

We determined cardiovascular and general health status through the assessment of different cardiovascular markers. First, we measured weight (without shoes and heavier clothes), height, and waist circumference. Following,

we used bioelectrical impedance measurements (Body Impedance Analyzer BIA 2000-S, Data Input GmbH) including the determination of body fat and water, extracellular mass, body cell mass, and basic metabolic rate to assess general markers of body composition. In addition, a measurement of a standard three-channel electrocardiogram (ECG) was performed (Supplementary Information 1.7). From these assessments we included the following seven predictors: heart rate (beats per minute), body mass index (BMI), extracellular mass to body cell mass (ECM/BCM ratio), basic metabolic rate, corrected body fat (kg), total body water (kg), and lean body mass (kg). In total, 247 predictors were included in the analysis (see Supplementary Table 1 for all predictors).

### ML pipeline

For our first set of models we entered all predictor modalities into our pipeline for consideration. To ensure the unbiased approximation of the model's generalizability to new patients, we trained and tested all models using repeated nested cross-validation in a pipeline to prevent information leaking between patients used for training and validation. In the inner cross validation loop, we conducted imputation, standardization, feature selection, hyperparameter optimization, and the fitting of a linear support vector machine (SVM).

To begin, we imputed predictors using multivariate imputation of chained equations with the 10 nearest predictors used in the imputation process. No predictors entered into the pipeline had more than 20% missing data. Following, all variables were scaled to have a mean of zero and a standard deviation of one. Next, we used the elastic net, a form of penalized logistic regression to select a final subset of variables for prediction. This approach shrinks the coefficients of highly correlated predictors towards each other while removing irrelevant predictors from the model<sup>41,42</sup>. This process was completed simultaneously with an exhaustive grid search to tune model hyperparameters. For the elastic net, the parameters alpha (the amount of penalization) and the l1 ratio (mixing parameter between the l1 and l2 norms) were tuned for predictor selection. For the l1 ratio parameter we searched the values  $\lambda \in \{0.1, 0.5, 0.7, 0.9, 0.95, 0.99\}$ , with values closer to 1 representing the l1 norm. For alpha, we searched  $a \in \{0.1, 0.2, \dots, 1.0\}$ . For the SVM, we tuned four values of the regularization parameter  $C \in \{0.001, 0.01, 0.1, 1.0\}$ . To accommodate for the class imbalance in the outcome, each  $C$  value was weighted by the inverse percentage proportion of each class label (rehospitalized: yes = 102, no = 278). This approach increases the penalty for misclassifying the minority class (rehospitalized) given its relative scarcity. Finally, we used Platt scaling to calibrate the probability estimates for the SVMs binary predictions<sup>43,44</sup>. The set of predictors and hyperparameters that

maximized area under the curve on the receiver operator characteristic (AUC) were selected in the pipeline. See Supplementary Information 1.8–1.8.7 for further details on the pipeline and Supplementary Information 1.8.5 for the selected hyperparameter values for the elastic net and SVM classifier.

All steps were completed in an inner cross-validation loop with five repeats of 10-fold cross-validation. This method divides the sample into 10 separate subsets, uses nine for training, and then makes predictions on the final set. To avoid favourable splits in the data, this process is repeated five times, initializing splits uniquely for each repeat. For the testing of our final models, we used 10-fold cross validation in the outer cross validation loop, averaging model performance metrics across test folds. To assess the statistical significance of our final best performing model, we used a permutation test ( $m = 10,000$ ) (see Fig. 1).

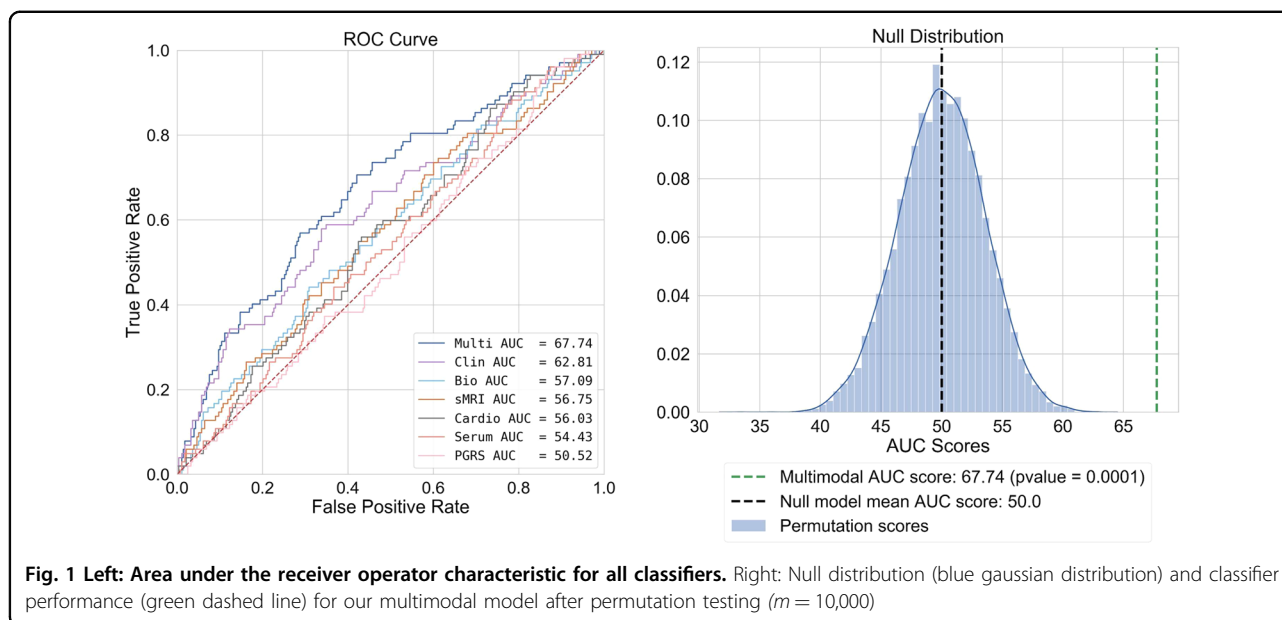
To assess the predictive capacity of each modality, we trained, clinical, sMRI, cardiovascular, and PGRS only models using the same pipeline as above, as well as a biomarker only model that aggregated together all biomarker modalities. To assess for a significant omnibus effect between models, we used a Kruskal–Wallis  $H$ -test. Following, we used Mann–Whitney rank tests to assess for post hoc differences between our multi and unimodal models. All  $p$  values were FDR corrected using the Benjamini and Hochberg method. All ML models were developed using Scikit-learn in Python 3.6.5. All code for analysis is available on request.

### Open sourcing of model

To provide transparency, the use of our model to other research groups, and encourage further external validation of our multimodal model, we uploaded our trained model to the Photon AI online model repository (<https://www.photon-ai.com/repo>). This repository allows our multimodal model to be downloaded and tested by other research groups.

### Analyses of multimodal predictors

To analyse the direction, magnitude, and significance of the models selected predictors after controlling for covariates, we used a non-penalized implementation of logistic regression in Python's statsmodels package. In addition, we included the average SVM weight values from the outer 10-fold cross-validation loop from our ML pipeline (Supplementary Information 3.1 and Supplementary Table 2). Further, we conducted ordinary least-squares regression sub-analyses of variables including age, gender, BMI, overall psychotropic medication load, and individual psychotropic medication use on the biomarkers that were selected in the final multimodal model. This was done to elucidate an understanding of variables that may



be associated with these biomarkers and help explain their inclusion in the multimodal model. In addition, we conducted an exploratory analysis of patient characteristics for those taking a specific type of medication selected in the multimodal model. This was done to determine: (a) whether this medication proxied for the treatment or prevention of a specific comorbid illness, (b) was taken as a polytherapeutic treatment strategy for MDD, or (c) was acting as a proxy for the presence of unmedicated patients that were relapsing at higher rates than the rest of the sample. For these additional analyses and results, see Supplementary Information 3.3–3.5.

## Results

### Multimodal model

Within 2 years of hospitalization for an acute episode of MDD 102 patients (26.8%) were rehospitalized for a depressive episode, while 278 (73.2%) were not. Our best performing model was our multimodal model (test AUC = 67.74). For our multimodal solution, 10 predictors were selected for optimal predictions by the elastic net. Those with a positive association with rehospitalization included the number of previous inpatient depressive episodes, individual CES-D items 5 (Last week I had trouble concentrating) and 3 (In the last week I could not get rid of my mood, although my friends/family tried to cheer me up), Pittsburgh Sleep Quality Index (PSQI) item 7 (During the past month, how often have you had trouble staying awake while driving, eating meals, or engaging in social activity?), the PSQI sleep quality index (global score), taking diazepam, oxazepam, thiazepam and oxepin, and right hippocampal volume. Predictors with a negative association included cholesterol (mmol/l), taking thyroid

medications: yes/no, and how often do you drink alcohol? (At most once a week/more than once a week). See Supplementary Figs. 1–3.

Our final multimodal model provided an increase in prognostic certainty of 23.21% (prognostic summary index = (PPV+NPV)–100). Further, a patient classified as being re-hospitalized by our model was 83% more likely to be rehospitalized than a patient who was not (positive likelihood ratio =  $\frac{\text{sensitivity}}{1-\text{specificity}}$ ). Finally, our multimodal classifier was statistically significant after permutation testing ( $p = 9.99 \times 10^{-5}$ ) (Fig. 1, Table 3). For individual modality models, see Supplementary Information 3.2.

## Discussion

The current study is the first of its kind to investigate the role of multiple predictor modalities for MDD rehospitalization trajectory modelling. Overall, our multimodal model provided a 23.21% increase in prognostic certainty for patient rehospitalization classification while providing an isolated subset of multimodal predictors for analysis. Furthermore, our multimodal model was statistically significant after permutation testing.

Of clinical importance is the positive (PPV) and negative predicted values (NPV) of our multimodal model and how they can be used to inform clinical decision making. As PPV was low (PPV = 41.64), using this model to confirm clinician suspicion of rehospitalization is not supported. On the contrary, NPV was modestly high (NPV = 81.57). Suggesting that the model could be used to confirm a clinician's suspicion of low rehospitalization risk, potentially identifying the patient as suitable for less assertive follow-up.

**Table 3 Performance metrics for all classifiers**

	Train	Test	SVM results										
	AUC	AUC	F1	BAC	Acc	Sens	Spec	PPV	NPV	PSI	PLR	NLR	DOR
Multi	78.86 (2.81)	67.74 (13.86)	67.15	63.05	65.72	57.45	68.65	41.64	81.57	23.21	1.83	0.62	2.96
Clinical	73.59 (1.64)	62.81 (11.14)	64.30	60.10	62.62	54.73	65.44	37.04	79.79	16.83	1.58	0.69	2.29
Bio	63.12 (1.04)	57.09 (11.47)	55.80	51.47	53.64	46.90	56.03	27.44	74.78	2.22	1.07	0.95	1.13
sMRI	64.53 (1.74)	56.75 (10.46)	56.94	52.03	55.31	45.00	59.06	29.16	74.96	4.12	1.10	0.93	1.18
Cardio	61.44 (1.48)	56.03 (13.12)	56.15	54.22	53.65	55.55	52.90	29.62	77.18	6.80	1.18	0.84	1.40
Serum	60.70 (0.75)	54.43 (8.31)	51.81	50.00	63.97	20.00	80.00	5.41	58.57	-36.02	1.00	1.00	1.00
PGRS	59.72 (1.26)	50.52 (13.62)	53.32	50.55	50.85	49.91	51.18	27.88	73.30	1.18	1.02	0.98	1.04

All classifiers used a Linear Support Vector Machine with Platt scaling, only predictor modalities varied across models. Mean (SD) scores from the outer 10-fold cross-validation loops are presented. Model abbreviations: *Multi* our multimodal model (all biomarker modalities, clinical, and demographic variables), *Clinical* clinical and demographic predictors only, *Bio* model with all biomarker modalities (no clinical or demographic data), *sMRI* structural imaging predictor model only, *Cardio* electrocardiography and bioelectrical impedance analysis predictor model only, *Serum* blood biomarkers only, *PGRS* PGRS model only. Metric abbreviations: *AUC* area-under-the-curve, *F1* Harmonic mean of Sens + Spec, *BAC* balanced accuracy, *Acc* accuracy, *Sens* sensitivity, *Spec* specificity, *PPV* positive predicted value, *NPV* negative predicted value, *PSI* prognostic summary index, *PLR* positive likelihood ratio, *NLR* negative likelihood ratio, *DOR* diagnostic odds ratio

As few studies are yet to consider a diverse range of biomarker modalities for rehospitalization trajectory modelling, the discriminative ability of the included modalities, as well as their interaction with clinical markers is of interest. First, our best performing model was multimodal, including a range of clinical, blood-biomarker, and structural imaging predictors. A recent work by Koutsouleris et al.<sup>45</sup> also demonstrated the superiority of multimodal models, finding that a combination of clinical and neuroimaging markers was most predictive of functional outcomes in a cohort of patients at risk for psychosis with recent onset depression. Our findings build on this work as well as provide preliminary evidence for the discriminative ability of blood biomarkers on a well-defined and clinically meaningful dichotomous endpoint.

Given that clinical predictors have consistently demonstrated their discriminative ability in multivariate pattern recognition studies<sup>12,14,46</sup>, it is unsurprising that 8 of 10 predictors in our multimodal solution were clinical. On the contrary, PGRS scores were of no use for rehospitalization classification (test AUC = 50.5) showing the least discriminative ability of all the modelled modalities. Even with lower *p* value thresholds (*p* = 0.05 and 0.01), no changes were seen in our multimodal model. In addition, unimodal PGRS discrimination was still low at these thresholds (AUC = 54.83 and 54.17 respectively, Supplementary Information 3.2). PGRS scores have received strong interest in psychiatric research, hoping that the aggregation of multiple single-nucleotide polymorphisms may illuminate genetic differences in psychiatric traits as well as parse the heterogeneity of outcomes such as medication response<sup>47,48</sup>. Significant associations with outcomes of interest have been consistently demonstrated, however,

commonly explain less than 1–2% of variance, limiting their current clinical use. Given that PGRS scores could not classify rehospitalization any better than chance in the current work, the discovery of more variants as well as non-linear modelling techniques may improve their clinical utility in future works<sup>49</sup>. All other biological modalities performed better, yet still lacked clinically meaningful discriminative ability when modelled without clinical predictors. Given these findings and those of Koutsouleris et al.<sup>45</sup>, it appears that biomarkers may be of prognostic use but likely perform best when modelled with clinical predictors also.

Regarding the selection of diazepam, oxazepam, thiazepam, and oxepines and their positive association with rehospitalization, it is possible that patients on this class of drugs at baseline had a more severe form of illness requiring augmentation with antipsychotic medications, placing them at a greater risk of rehospitalization. In addition, thyroid medication use was negatively associated with rehospitalization. Besides a small handful of patients, those with a past diagnosis of hyper/hypothyroidism who were not currently taking thyroid medications had t3, t4, and TSH levels within healthy reference ranges (Supplementary Information 3.3), suggesting that it was not the presence of unmedicated patients with thyroid disorders relapsing at high rates that were responsible for thyroid medication use and its negative association with rehospitalization. Considering that nearly all patients were taking antidepressant medications at their baseline assessment, it is possible that a synergistic prophylactic relationship between thyroid and antidepressant medication use may exist. Such a polytherapeutic relationship has been robustly demonstrated in the multisite sequenced alternatives to relieve depression trial (STAR\*D)<sup>29</sup>.

However, prophylactic effects against rehospitalization should be explored in future works.

Of interest, right hippocampal volume showed a positive association with rehospitalization. This was surprising, given the well-established effect of smaller hippocampal volumes in MDD patients compared to healthy controls<sup>6</sup>. To better understand this observation, we conducted a range of sub-analysis (Supplementary Information 3.4). In accordance with previous works<sup>50,51</sup>, we showed that those taking any form of antipsychotic at baseline had larger right hippocampal volumes than those who were not, albeit, this effect was not significant after controlling for covariates. In addition, we showed that there was a significant gender/diazepine, oxazepine, thiazepine, and oxepine use interaction effect on right hippocampal volumes, with women currently taking medications from this class having significantly larger right hippocampal volumes than those who were not. Forty-four percent of these women ended up being rehospitalized between baseline and their 2-year follow-up assessment, compared to only 22% of women who were not. While we had binary usage data for medication, dosage data were not available. Dosage effects were likely prevalent, but not quantifiable in the current work. Overall, it is plausible that changes in right hippocampal volume proxied for a complex gender/medication/dosage-specific aetiology not fully represented in our clinical data.

The first limitation of the current work is the models' scope. As our model was trained on a middle-aged European cohort, it is plausible that some of the selected predictors in our pipeline were unique to this demographic. Predictors such as the number of previous inpatient episodes will likely be larger on an older cohort, offering more discriminative ability. Therefore, the clinical utility of the model needs to be considered within this scope. To overcome this limitation, we have provided our trained model online through the Photon AI model repository.

Further limitations include the reporting of nominal significance in our logistic regression model (Supplementary Information 3.1). It is important to consider that the nominally significant effects that were found in our non-penalized logistic regression model did not survive FDR corrections (Supplementary Table 2). However, our primary interest was the emergent multivariate pattern that demonstrated statistically significant class separation between cases and controls at  $p = 9.99 \times 10^{-5}$ . Given the importance and statistical significance of this aggregated multivariate pattern, we deemed it necessary to conduct an exploratory analysis to illuminate the contribution of each individual predictor after rigorously controlling for known covariates age, gender, smoker status, BMI, and intracranial volume. Given this, we believe the nominal

significance of predictors in our logistic regression model to be of use for hypothesis generation in future works.

Regarding clinical use, while our multimodal model was statistically significant after permutation testing, its balanced accuracy was relatively low (BA = 63.05). Whether this level of performance is sufficient for clinical use is unknown; however, this BA is similar to that attained in a previous work that is now deployed clinically (BA = 59.60–64.6)<sup>14</sup>. Further, we argue that it is not the absolute accuracy of a model that should dictate whether it is of clinical use, but whether or not it outperforms current clinical best practice. For example, as rehospitalization risk is not formally and routinely quantified by clinicians to inform their clinical decision making, we would not expect clinician prognostication of rehospitalization to be any better than chance. Therefore, even a modestly performing model would theoretically confer clinical advantage when deployed at scale. In addition, such a model would be free of well-documented decision-making biases that are known to affect clinician prognostication<sup>52</sup> (e.g., anchoring effects), potentially providing further clinical benefit under high workloads commonly seen in inpatient care. Nonetheless, for these questions to be answered, future works that benchmark clinicians' prognostic abilities are needed before such incremental utility can be quantified.

Finally, it is important to note that while our multimodal model provided the highest degree of class separation, after testing for post hoc differences between all models and FDR correcting  $p$  values, our multimodal model did not significantly outperform our model containing only clinical predictors ( $p = 0.15$ ) (Supplementary Information 3.2). Future works of increased size may be required to elucidate a statistically significant difference given the small predictive discrepancy between the multimodal and clinical model. On the contrary, clinical data of sufficient depth (for example, that captures the previously discussed gender/medication/dosage effects and their association with right hippocampal volumes) may even mitigate the contribution of biomarkers in future works. If so, these models will bestow both ease of use and economic advantage, rendering the inclusion of costly and harder to attain biomarkers ineffectual. To answer these questions a greater dearth of clinical data, discovery samples of increased size, as well as validation samples that are phenotypically and geographically diverse will be required.

In conclusion, the presented findings suggest that the combination of ML techniques with multimodal clinical and biomarker data may lead to an increase in prognostic certainty compared to chance level. Continued research is required, but ML may be of use to derive models for clinicians to make personalized predictions regarding

## rehospitalization risk as well as better inform prophylactic treatment strategies.

### Acknowledgements

The BiDirect Study is funded by the German Federal Ministry of Education and Research (BMBF (01ER0816 and 01ER1506, and 01ER1205)). Additional funding for the analysis was provided by the German Research Foundation (DFG, grant FOR2107 DA1151/5-1 and DA1151/5-2 to U.D.; SB-TRR58, Projects C09 and Z02), the Interdisciplinary Center for Clinical Research (IZKF) of the medical faculty of Münster (grant Dan3/012/17 to U.D., SEED 11/18 to NO) and the Deanery of the Medical Faculty of the University of Münster. In addition, this work was supported with supercomputing resources provided by the Phoenix HPC service at the University of Adelaide. The funding sources had no role in study design, collection, analysis, or interpretation of the data, in the writing of the report, or in the decision to submit the article for publication.

### Author details

<sup>1</sup>Discipline of Psychiatry, School of Medicine, University of Adelaide, Adelaide, Australia. <sup>2</sup>Department of Psychiatry, University of Münster, Münster, Germany. <sup>3</sup>Interdisciplinary Centre for Clinical Research (IZKF), University of Münster, Münster, Germany. <sup>4</sup>Centre for Healthy Brain Ageing, UNSW, Sydney, Australia. <sup>5</sup>Institute of Clinical Radiology, University of Münster, Münster, Germany. <sup>6</sup>Institute of Clinical Chemistry and Laboratory Medicine, University Medicine Greifswald, Greifswald, Germany. <sup>7</sup>Integrated Research Biobank, University Medicine Greifswald, Greifswald, Germany. <sup>8</sup>Institute of Epidemiology and Social Medicine, University of Münster, Münster, Germany. <sup>9</sup>Department of Psychiatry, Melbourne Medical School, The University of Melbourne, Melbourne, Australia. <sup>10</sup>The Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Melbourne, Australia

### Conflict of interest

The authors declare that they have no conflict of interest.

### Publisher's note

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

**Supplementary Information** accompanies this paper at (<https://doi.org/10.1038/s41398-019-0615-2>).

Received: 30 April 2019 Revised: 8 October 2019 Accepted: 20 October 2019

Published online: 11 November 2019

### References

- Hardevelde, F., Spijker, J., De Graaf, R., Nolen, W. A. & Beekman, A. T. Prevalence and predictors of recurrence of major depressive disorder in the adult population. *Acta Psychiatr. Scand.* **122**, 184–191 (2010).
- Paykel, E. S. et al. Residual symptoms after partial remission: an important outcome in depression. *Psychol. Med.* **25**, 1171–1180 (1995).
- Ramana, R. et al. Remission and relapse in major depression: a two-year prospective follow-up study. *Psychol. Med.* **25**, 1161–1170 (1995).
- Keller, M. B., Lavori, P. W., Lewis, C. E. & Klerman, G. L. Predictors of relapse in major depressive disorder. *JAMA* **250**, 3299–3304 (1983).
- Grilo, C. M. et al. Personality disorders predict relapse after remission from an episode of major depressive disorder: a 6-year prospective study. *J. Clin. Psychiatry* **71**, 1629–1635 (2010).
- Schmaal, L. et al. Subcortical brain alterations in major depressive disorder: findings from the ENIGMA Major Depressive Disorder working group. *Mol. Psychiatry* **21**, 806–812 (2016).
- Appelhof, B. C. et al. Glucocorticoids and relapse of major depression (dexamethasone/corticotropin-releasing hormone test in relation to relapse of major depression). *Biol. Psychiatry* **59**, 696–701 (2006).
- Zobel, A. W., Yassouridis, A., Frieboes, R. M. & Holsboer, F. Prediction of medium-term outcome by cortisol response to the combined dexamethasone-CRH test in patients with remitted depression. *Am. J. Psychiatry* **156**, 949–951 (1999).
- Berman, R. M. et al. Transient depressive relapse induced by catecholamine depletion: potential phenotypic vulnerability marker? *Arch. Gen. Psychiatry* **56**, 395–403 (1999).
- Gili, M., Vicens, C., Roca, M., Andersen, P. & McMillan, D. Interventions for preventing relapse or recurrence of depression in primary health care settings: a systematic review. *Prev. Med.* **76**, S16–S21 (2015).
- Teismann, H., et al. Establishing the bidirectional relationship between depression and subclinical arteriosclerosis—rationale, design, and characteristics of the BiDirect Study. *BMC Psychiatry* **14**, 3–4. <https://doi.org/10.1186/1471-244x-14-174> (2014).
- Chekroud, A. M. et al. Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. *JAMA Psychiatry* **74**, 370–378 (2017).
- Drysdale, A. T. et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nat. Med.* **23**, 28–38 (2017).
- Chekroud, A. M. et al. Cross-trial prediction of treatment outcome in depression: a machine learning approach. *Lancet Psychiatry* **3**, 243–250 (2016).
- Koutsouleris, N. 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*. <https://doi.org/10.1001/jamapsychiatry.2018.2165> (2018).
- Kambeitz, J. et al. Detecting neuroimaging biomarkers for depression: a meta-analysis of multivariate pattern recognition studies. *Biol. Psychiatry* **82**, 330–338 (2017).
- Schmaal, L. et al. Predicting the naturalistic course of major depressive disorder using clinical and multimodal neuroimaging information: a multivariate pattern recognition study. *Biol. Psychiatry* **78**, 278–286 (2015).
- Moayedoddin, B. et al. Prevalence and clinical characteristics of the DSM IV major depression among general internal medicine patients. *Eur. J. Intern. Med.* **24**, 763–766 (2013).
- Lopresti, A. L., Maker, G. L., Hood, S. D. & Drummond, P. D. A review of peripheral biomarkers in major depression: the potential of inflammatory and oxidative stress biomarkers. *Prog. Neuropsychopharmacol. Biol. Psychiatry* **48**, 102–111 (2014).
- Weber, B., Lewicka, S., Deuschle, M., Colla, M. & Heuser, I. Testosterone, androstenedione and dihydrotestosterone concentrations are elevated in female patients with major depression. *Psychoneuroendocrinology* **25**, 765–771 (2000).
- Schweiger, U. et al. Testosterone, gonadotropin, and cortisol secretion in male patients with major depression. *Psychosom. Med.* **61**, 292–296 (1999).
- Musselman, D. L., Evans, D. L. & Nemeroff, C. B. The relationship of depression to cardiovascular disease: epidemiology, biology, and treatment. *Arch. Gen. Psychiatry* **55**, 580–592 (1998).
- Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
- Knight, M. J. & Baune, B. T. Cognitive dysfunction in major depressive disorder. *Curr. Opin. Psychiatry* **31**, 26–31 (2018).
- Lopresti, A. L., Hood, S. D. & Drummond, P. D. A review of lifestyle factors that contribute to important pathways associated with major depression: diet, sleep and exercise. *J. Affect. Disord.* **148**, 12–27 (2013).
- Murphy, M. J. & Peterson, M. J. Sleep disturbances in depression. *Sleep Med. Clin.* **10**, 17–23 (2015).
- Chekroud, S. R. & Chekroud, A. M. Efficacy of resistance exercise training with depressive symptoms. *JAMA Psychiatry* **75**, 1091–1092 (2018).
- Schmaal, L. et al. Cortical abnormalities in adults and adolescents with major depression based on brain scans from 20 cohorts worldwide in the ENIGMA Major Depressive Disorder Working Group. *Mol. Psychiatry* **22**, 900–909 (2017).
- Nierenberg, A. A. et al. A comparison of lithium and T(3) augmentation following two failed medication treatments for depression: a STAR\*D report. *Am. J. Psychiatry* **163**, 1519–1530 (2006). quiz 665.
- Joffe, R. T. & Marriott, M. Thyroid hormone levels and recurrence of major depression. *Am. J. Psychiatry* **157**, 1689–1691 (2000).
- Joffe, R., Segal, Z. & Singer, W. Change in thyroid hormone levels following response to cognitive therapy for major depression. *Am. J. Psychiatry* **153**, 411–413 (1996).
- Sassarini, D. J. Depression in midlife women. *Maturitas* **94**, 149–154.
- Schiller, C. E., Meltzer-Brody, S. & Rubinow, D. R. The role of reproductive hormones in postpartum depression. *CNS Spectr.* **20**, 48–59 (2015).
- Rodgers, S. et al. Serum testosterone levels and symptom-based depression subtypes in men. *Front. Psychiatry* **6**, 61 (2015).



35. Almeida, O. P., Yeap, B. B., Hankey, G. J., Jamrozik, K. & Flicker, L. Low free testosterone concentration as a potentially treatable cause of depressive symptoms in older men. *Arch. Gen. Psychiatry* **65**, 283–289 (2008).
36. Kische, H. et al. Associations of androgens with depressive symptoms and cognitive status in the general population. *PLoS One* **12**, e0177272 (2017).
37. Parekh, A., Smeeth, D., Milner, Y. & Thure, S. The role of lipid biomarkers in major depression. *Healthcare (Basel)*. **5**, <https://doi.org/10.3390/healthcare5010005> (2017).
38. Maes, M. et al. Lower serum high-density lipoprotein cholesterol (HDL-C) in major depression and in depressed men with serious suicidal attempts: relationship with immune-inflammatory markers. *Acta Psychiatr. Scand.* **95**, 212–221 (1997).
39. Lee, S. H. et al. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45**, 984–994 (2013).
40. Brainstorm C. et al. Analysis of shared heritability in common disorders of the brain. *Science* **360**, <https://doi.org/10.1126/science.aap8757> (2018).
41. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **67**, 301–320 (2005).
42. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1 (2010).
43. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A. J., Barlett, P. L., Scholkopf, B. & Schuurmans, D. (eds.) *Advances in Large Margin Classifiers*. Vol. 10, 61–74 (The MIT press, Cambridge, Massachusetts, London, England, 1999).
44. Niculescu-Mizil A., Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning*. 625–632 (ACM, 2005).
45. Koutsouleris, N. 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* **75**, 1156–1172 (2018).
46. Koutsouleris, N., Kahn, R. S. & Chekroud, A. M. Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: a machine learning approach (vol 3, pg 935, 2016). *Lancet Psychiatry* **4**, 95 (2017).
47. Amare, A. T. et al. Association of polygenic score for schizophrenia and HLA antigen and inflammation genes with response to lithium in bipolar affective disorder: a genome-wide association study. *JAMA Psychiatry* **75**, 65–74 (2018).
48. Amare, A. T. et al. Association of the polygenic scores for personality traits and response to selective serotonin reuptake inhibitors in patients with major depressive disorder. *Front. Psychiatry* **9**, 65 (2018).
49. Zheutlin, A. B. et al. Multivariate pattern analysis of genotype–phenotype relationships in schizophrenia. *Schizophrenia Bull.* **44**, 1045–1052 (2018).
50. Bodnar, M. et al. The effect of second-generation antipsychotics on hippocampal volume in first episode of psychosis: longitudinal study. *BJPsych Open*. **2**, 139–146 (2016).
51. Chakos, M. H. et al. Duration of illness and treatment effects on hippocampal volume in male patients with schizophrenia. *Br. J. Psychiatry* **186**, 26–31 (2005).
52. Saposnik, G., Redelmeier, D., Ruff, C. C. & Tobler, P. N. Cognitive biases associated with medical decisions: a systematic review. *BMC Med Inf. Decis. Mak* **16**, 138 (2016).