American Journal of Medical Genetics Part B: Neuropsychiatric Genetics

WILEY

RESEARCH ARTICLE OPEN ACCESS

Optimizing the Prediction of Depression Remission: A Longitudinal Machine Learning Approach

Ewan Carr[1] | Marcella Rietschel[2] | Ole Mors[3] | Neven Henigsberg[4] | Katherine J. Aitchison[5,6,7,8] | Wolfgang Maier[9,10] | Rudolf Uher[11] | Anne Farmer[12] | Peter McGuffin[12] | Raquel Iniesta[1,13]

[1]Department of Biostatistics & Health Informatics, Institute of Psychiatry, Psychology and Neuroscience (IoPPN), London, UK | [2]Department of Genetic Epidemiology in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany | [3]Psychosis Research Unit, Aarhus University Hospital—Psychiatry, Aarhus, Denmark | [4]Croatian Institute for Brain Research, Medical School, University of Zagreb, Zagreb, Croatia | [5]College of Health Sciences, Department of Psychiatry and Medical Genetics, University of Alberta, Edmonton, Alberta, Canada | [6]Neuroscience and Mental Health Institute, University of Alberta, Edmonton, Alberta, Canada | [7]Women and Children's Health Research Institute, University of Alberta, Edmonton, Alberta, Canada | [8]Northern Ontario School of Medicine, Thunder Bay, Ontario, Canada | [9]Department of Neurodegenerative Diseases and Geriatric Psychiatry, University Hospital Bonn, Bonn, Germany | [10]German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany | [11]Department of Psychiatry, Dalhousie University, Halifax, Nova Scotia, Canada | [12]Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK | [13]King's Institute for Artificial Intelligence, King's College London, London, UK

Correspondence: Raquel Iniesta (raquel.iniesta@kcl.ac.uk)

Received: 10 October 2023 | **Revised:** 5 August 2024 | **Accepted:** 1 October 2024

Funding: This work has been supported by the Brain and Behaviour Foundation awarded to Raquel Iniesta (Award number 26338). The funding body played no role in the design of the study, the collection, analysis, interpretation of data, or in writing the manuscript. This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Keywords: depression remission | machine learning | repeated measures | topological data analysis

ABSTRACT

Decisions about when to change antidepressant treatment are complex and benefit from accurate prediction of treatment outcome. Prognostic accuracy can be enhanced by incorporating repeated assessments of symptom severity collected during treatment. Participants ($n = 714$) from the Genome-Based Therapeutic Drugs for Depression study received escitalopram or nortriptyline over 12 weeks. Remission was defined as scoring ≤ 7 on the Hamilton Rating Scale. Predictors included demographic, clinical, and genetic variables (at 0 weeks) and measures of symptom severity (at 0, 2, 4, and 6 weeks). Longitudinal descriptors extracted with growth curves and topological data analysis were used to inform prediction of remission. Repeated assessments produced gradual and drug-specific improvements in predictive performance. By Week 4, models' discrimination in all samples reached levels that might usefully inform treatment decisions (area under the receiver operating curve (AUC) = 0.777 for nortriptyline; AUC = 0.807 for escitalopram; AUC = 0.794 for combined sample). Decisions around switching or modifying treatments for depression can be informed by repeated symptom assessments collected during treatment, but not until 4 weeks after the start of treatment.

This paper is dedicated to the memory of Professor Peter McGuffin.

1 | Introduction

Past studies have highlighted the need to consider individual trajectories of antidepressant treatment response (Uher et al. 2011) by incorporating repeated assessments of symptom severity collected (e.g., weekly) during treatment. Whereas some patients may experience early improvements within 2 or 3 weeks, others may not see a benefit until later. This raises important questions about modifying or switching treatments. Since delayed treatment response is common, there can be uncertainty around how long to wait before considering alternative treatments. When predicting treatment response, a balance is needed between earlier but potentially inaccurate predictions versus delayed predictions with increased accuracy. The former risks miss-classifying patients with delayed improvements and curtailing treatment that could have led to remission; the latter risks patients staying on ineffective treatment longer than necessary.

Previous investigations into the timing of antidepressant treatment response have been inconsistent. Some suggest that differences between control and treatment arms can be identified within 1 week of treatment (Browning et al. 2019). A review of 10 placebo-controlled trials found improvements within 2 weeks (≥ 20% reduction in symptom severity) to be predictive of treatment response, but these improvements did not reliably predict eventual remission (Kemp et al. 2011). Other studies suggest waiting 3 weeks to identify a therapeutic response (Szegedi et al. 2009), that patients not improving within 4 weeks should have their treatment altered (Quitkin et al. 1996), or that accurate outcome prediction is only possible after 8 weeks (Uher et al. 2011). Several clinical guidelines support altering antidepressant treatment at 4 weeks (Gelenberg et al. 2010; NICE 2022).

In addition to questions about the timing of treatment response and modification, incorporating weekly assessments raises methodological challenges (Bull et al. 2020).

Several approaches have been used to incorporate repeated measures when predicting antidepressant response, each with varying strengths and limitations. Uher et al. (2011) used growth mixture models to show individual differences in trajectories of symptom response in the GENDEP study (Genome-Based Therapeutic Drugs for Depression), a partially randomized multicentre pharmacogenetic study that compared two antidepressants for patients with moderate-to-severe unipolar depression (Uher et al. 2009, 2010). Athreya et al. (2021) used probabilistic graph models and unsupervised machine learning (ML) to predict remission and treatment response by 8 weeks. Several studies have shown ML techniques to outperform regression-based methods for predictors measured at baseline (Hilbert et al. 2020; Iniesta et al. 2016, 2018; Kessler et al. 2016; Koutsouleris et al. 2016; Pearson et al. 2019; Webb et al. 2020); several have used ML techniques to predict response trajectories (Chekroud et al. 2021). Some have combined unsupervised and supervised ML first to identify potential subtypes of responders (based on baseline symptoms and/or symptoms trajectories) and then develop classifiers to predict specific response trajectory and/or total severity outcomes (Bondar et al. 2020; Chekroud et al. 2017, 2021; Dinga et al. 2019; Drysdale et al. 2017; Gueorguieva, Chekroud, and Krystal 2017; Paul et al. 2019).

We have previously reported that the combination of baseline demographic, clinical, and genetic variables can help select between escitalopram and nortriptyline treatments (Iniesta et al. 2018). Recent research has suggested that a Polygenic Risk Score (PRS) could allow MDD patients to be stratified for antidepressant response by showing that a greater genetic loading for MDD is associated with a less favorable response (Ward et al. 2018).

In the present study, we evaluate the extent to which the combination of baseline demographic and clinical variables, a PRS for MDD, and repeated measures of depression severity can predict depression remission with specific treatments at the individual level. In particular, we aim to (1) use ML to optimize the prediction of depression remission in GENDEP by incorporating repeated measures of depression severity into a multivariable prediction model of demographic, clinical and genetic variables and (2) evaluate how discrimination of remission improves by adding severity measurements collected at successive weeks during treatment. For this aim, we extracted longitudinal features using remission curve modeling and topological data analysis, a recently developed method for handling large longitudinal data (Chazal and Michel 2017; Iniesta et al. 2022; Riihimäki et al. 2020). To the best of our knowledge, no other studies have compared multiple methods to handle repeated measurements when predicting depression remission.

2 | Methods

2.1 | Editorial Policies and Ethical Considerations

The GENDEP study was approved by the ethics boards of participating centres. All participants provided written informed consent.

2.2 | Data

GENDEP included 868 European adult patients with major depressive disorder, recruited in nine European clinical centres between 2004 and 2008. Eligibility and study procedures are described elsewhere (Uher et al. 2009, 2010). Participants received nortriptyline (a norepinephrine reuptake inhibitor) or escitalopram (a serotonin reuptake inhibitor) for 12 weeks. Patients were randomly allocated to receive nortriptyline or escitalopram unless they had a contraindication for one of the drugs, in which case they were allocated to the other antidepressant. We restricted our analysis to participants who provided information on at least four occasions as this is a minimum needed to establish an initial trend in clinical response, including the baseline assessment and the two-week follow-up as the minimum requirement based on maximizing completeness of data (see Figure S1). We considered three analytical samples: (A) Participants randomly allocated to receive nortriptyline ($n = 196$); (B) Participants randomly allocated to receive escitalopram ($n = 214$); (C) participants receiving either drug including both random and non-random allocations ($n = 714$).

2.3 | Measures

Our outcome was depression remission, defined as a score of ≤7 on the Hamilton Rating Scale for Depression (HRSD-17; Hamilton 1967) at the latest available assessment after 4–12 weeks of treatment. As per existing evidence in the literature, to meaningfully predict who will respond to which antidepressant, it may be necessary to combine biological and clinical variables. We considered predictors measured at baseline and weekly during treatment. Baseline predictors included 140 demographic and clinical variables and a PRS for MDD computed from GENDEP SNPs data (see Table S1 for a complete list of predictor variables). We created a PRS for MDD for 714 GENDEP participants of European ancestry included in this study, with 6,737,401 autosomes-only variants. The PRS was calculated from the last available Psychiatric Genomics Consortium (PGC) MDD summary statistics in European ancestry. Further details on genetic data and the PRS derivation are described in Supporting Information. For models developed in the 'combined' sample, we included a dummy variable indicating the treatment drug (0 = *nortriptyline*; 1 = *escitalopram*). Longitudinal predictors included individual items and total scores for the HRSD-17, the Beck Depression Inventory (BDI; Beck, Steer, and Carbin 1988), the Montgomery-Åsberg Depression Rating Scale (MADRS; Montgomery and Åsberg 1979) and total scores for three factors (observed mood, cognitive symptoms, neurovegetative symptoms) and six dimensions (mood, anxiety, pessimism, interest-activity, sleep, and appetite) from a published factor analysis (Uher et al. 2008, 2012).

2.4 | Derivation of Longitudinal Features

We considered two approaches to incorporating predictors measured repeatedly during treatment in a prediction model. First, we derived (GC) growth curve parameters by separately modeling each longitudinal feature using a linear mixed model that included linear and quadratic terms for time (weeks since baseline) with participant-specific random intercepts and slopes. We extracted the best linear unbiased predictions of the three random effects (intercept and linear and quadratic terms for time) and included these as predictors in subsequent models. Second, a simpler approach (LW) consisted of entering the latest weekly (LW) measurement into the considered prediction model as a predictor (e.g., the "Week 4" model included the baseline predictors and Week 4 longitudinal measurements only).

We supplemented both approaches (GC and LW) using topological data analysis (TDA). TDA is a set of mathematical tools to extract topological and geometric information from data to provide new features and descriptors of the data. They can be combined with other features for further analysis and machine-learning tasks (Chazal and Michel 2017; Iniesta et al. 2022). For each participant, we summarized the topological shape characteristics of their longitudinal features using the persistence diagram. The information encoded in the persistence diagram was vectorised using persistence landscapes (Bubenik 2018; see Figure S2) to generate features for subsequent models. To assess the incremental benefit of increasing weeks of feature information for each approach, we estimated models with baseline features and repeated measures up to 2, 4, and 6 weeks after baseline.

2.5 | Prediction of Remission Outcome by 12 Weeks

We used binary logistic regression with elastic net regularization to predict HRSD remission by 12 weeks. Regularized regression models are general linear models with penalties to avoid extreme parameters that could cause overfitting. The elastic net is an application of regularized regression that provides efficient feature selection from a large set of variables. Elastic net allows estimation of the combined predictive ability of a high number of variables and maximises model transparency, which is key for the explainable and ethical development of AI applications in medicine (https://github.com/ewancarr/tda_prediction) while reducing overfitting. All models included baseline features and one of four sets of longitudinal features defined above: (i) GC, (ii) GC supplemented with landscape features, (iii) LW, and (iv) LW supplemented with landscape features. Each model was estimated with and without Polygenic Risk Scores.

Predictive performance for each feature set was evaluated using repeated, nested k-fold cross-validation (10-fold repeated 20 times). All modeling steps were repeated within each fold to avoid data leakage, including pre-processing (e.g., standardization, removal of features with zero variance), hyperparameter tuning (e.g., alpha), feature selection, derivation of growth curve parameters, and missing value imputation. Missing values were imputed using *k*-nearest neighbors ($k = 5$) using the KNNImputer class from scikit-learn (Troyanskaya et al. 2001). Predictive performance was described using the area under the receiver operating curve (AUC) as a value of discrimination. We interpreted the AUC values under the benchmark proposed by Hosmer and Lemeshow (2000, 156–164). Based on this benchmark, a model with an AUC higher or equal to 0.7 and lower than 0.8 is considered to have an "acceptable discrimination," with an AUC higher or equal to 0.8 and lower than 0.9 has "excellent discrimination" and with an AUC equal or higher to 0.9 is considered to have an "outstanding discrimination." We also reported sensitivity, specificity, and negative predictive value (NPV). To balance sensitivity and specificity, we chose classification thresholds of 40% for the escitalopram and combined samples and 30% for nortriptyline. Cross-validation metrics were summarized by averaging each repetition (i.e., the mean over 10 folds) and reporting the median of the resulting values. Uncertainty intervals were computed based on the outer loop of the repeated, nested cross-validation. We reported NPV because we were more concerned with false positives than negatives when making decisions around treatment adaptations. A false positive would indicate that a patient's treatment was left unchanged, and the patient failed to achieve remission; a false negative would indicate that treatment was changed unnecessarily (i.e., the patient would have achieved remission on their original treatment). Since all changes to treatment would be reviewed and decided by a clinician, we considered the risk of not changing treatment to be greater than that of potentially unnecessary changes.

All analyses were conducted in Python 3.9.2 (Van Rossum and Drake 2009). Growth curve models were estimated using StatsModels (Seabold and Perktold 2010); topological summaries were computed using GUDHI (The GUDHI Project 2020);

elastic net logistic regression models were estimated using the LogitNet function from the glmnet for Python package (Python GLMNET 2021) and Scikit-Learn (Pedregosa et al. 2011). All analytical code used in our analyses is available at https://github.com/ewancarr/tda_prediction.

3 | Results

From 868 participants in GENDEP, we excluded 75 participants with < 4 assessments and 79 participants missing information at baseline or Week 2. The analytical samples comprised $n = 196$ (nortriptyline, randomized), $n = 214$ (escitalopram, randomized), and $n = 714$ (combined sample with both drugs, partially randomized) participants. Compared to the combined sample, excluded participants were similar in terms of age, years of education and BMI but reported lower symptom severity based on the HDRS (19.9 vs. 22.1; p-value = 0.002) and MADRS (26.9 vs. 29.0; p-value = 0.018). Table S2 summarizes the demographic and clinical characteristics of the three samples. Participants had a mean age of 42.8 years and were majority female gender (62%). In the combined sample of 714 participants, 285 (40%) achieved remission by the last observation.

3.1 | Prediction of Remission Incorporating Repeated Measures

Figure 1 and Table S3 present model performance for increasing weeks of repeated measures (0, 2, 4 or 6). For each sample and week, we present results for the best-performing methodological approach (i.e., GC, LW or these methods supplemented with persistence landscape variables). For the combined sample ($n = 714$), we observed a gradual increase in predictive performance for each additional fortnight, from an AUC of 0.703 (baseline) to 0.754 (Week 2), 0.794 (Week 4), and 0.844 (Week 6). NPV, sensitivity, and specificity show similar incremental improvements over time. At Week 4, the AUC was approaching 0.80 in both treatment arms ("Acceptable discrimination" (2000, 156–164)) and above 0.80 at Week 6 ("Excellent discrimination").

We saw differing results for the drug-specific samples. For nortriptyline ($n = 196$), initial performance was lower, but the improvement by Week 2 was much larger (from AUC of 0.618 to 0.763 for nortriptyline). In the combined sample, the increase in AUC from baseline by Week 2 was more moderate (0.703–0.754). Conversely, the model for escitalopram ($n = 214$) had higher initial performance, compared to other samples, but saw no improvement by Week 2 (e.g., from AUC at baseline of 0.718 to 0.717 at Week 2). From Weeks 2 to 4, increases in model performance were similar for all three samples. Table S3 summarizes the predictive performance with vs. without the baseline PRS for MDD. Improvement in performance due to the addition of the PRS was nearly negligible for all samples and weeks. The overlap observed across uncertainty intervals, as shown in Figure 1, suggested similar trends in models' performance across different treatment allocations.

3.2 | Predictive Performance of Each Modeling Approach

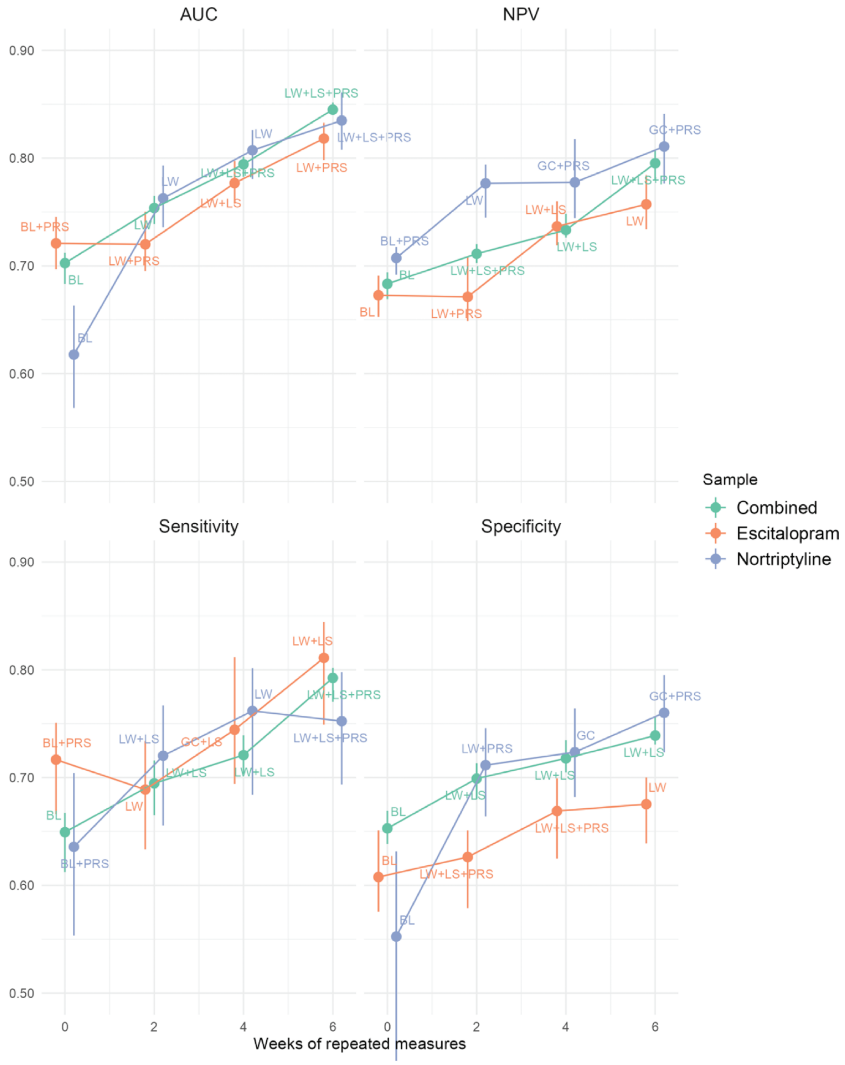
Table S4 summarizes the predictive performance of each method (LW, GC, and each supplemented with TDA variables). At most time points, entering the LW measure directly into the model (LW) tended to outperform the corresponding growth curve. Performance for growth curves was notably lower at Week 2 (e.g., AUC in the combined sample of 0.698 for GC vs. 0.754 for LW), which is expected given the limited number of time points available to construct the curve. Supplementing the LW with landscape parameters (TDA) produced small performance increases, and only for Weeks 4–6. For example, for the combined sample at Week 6, supplementing the LW model increased AUC from 0.824 (0.817, 0.835) to 0.844 (0.839, 0.853), with uncertainty intervals not overlapping, this suggesting a statistically significant improvement in models' performance because of adding TDA information.

4 | Discussion

In a large study of antidepressant treatment, we showed how repeated symptom measurements collected during treatment can improve prediction of remission by 12 weeks. Supplementing baseline measures with repeated assessments produced an incremental benefit for each additional fortnight. We reported important novel information: (1) a combination of the mentioned predictor variables was predictive of antidepressant treatment outcome, suggesting a potential for personalized antidepressant treatment, and particularly, (2) that models were able to predict remission by week four with adequate discrimination. Our results varied by treatment drug. Models for nortriptyline had worse initial performance but saw larger increases over the first 2 weeks; models for escitalopram had better initial performance but saw no improvement by Week 2. This suggests drug-specific differences in the minimum follow-up period required for a clinically meaningful prediction of antidepressant treatment response.

Our team had previously investigated if antidepressant treatment outcomes definition could be improved by using symptoms trajectories and established different classes of responders, the rapid improvement trajectory and the gradual improvement trajectory (Uher et al. 2010). The current work aimed to answer whether a combination of clinical, demographic, genetic, baseline, and longitudinal symptoms measurements can significantly predict remission at the individual level, if the addition of weekly measurements is relevant for the prediction and investigating how many weeks of symptoms measurements is optimal to consider for a clinically significant prediction of depression remission. Given the nature of the present work was predictive and not an association study, we selected the machine learning methodology as the basis of our predictive models.

Our analysis was motivated by the need for accurate prediction of treatment outcome to inform decisions about the timing of treatment adaptations. While adding post-baseline information



BL = Baseline only.
LW = Latest weekly measures only*.
GC = Growth curves only*.
+LS = Supplemented with topological landscape variables
+PRS = Supplemented with polygenic risk score
*All models included baseline features.

		0 weeks	2 weeks	4 weeks	6 weeks
Escitalopram	AUC	0.718	0.717	0.777	0.818
	NPV	0.673	0.668	0.737	0.757
	Sensitivity	0.700	0.689	0.744	0.811
	Specificity	0.608	0.625	0.668	0.675
Nortriptyline	AUC	0.618	0.763	0.807	0.834
	NPV	0.706	0.777	0.777	0.809
	Sensitivity	0.630	0.720	0.762	0.747
	Specificity	0.552	0.711	0.724	0.756
Combined	AUC	0.703	0.754	0.794	0.844
	NPV	0.683	0.711	0.733	0.792
	Sensitivity	0.649	0.695	0.721	0.791
	Specificity	0.653	0.699	0.718	0.739

Notes. The classification threshold was set at 40% for the Escitalopram and 'Combined' samples and 30% for Nortriptyline.

FIGURE 1 | Area under the receiver operating curve (AUC), negative predictive value (NPV), sensitivity and specificity of the best-performing model for increasing weeks of data (0, 2, 4, 6) stratified by sample (nortriptyline, escitalopram, combined) with uncertainty intervals based on the outer loop of the repeated, nested cross-validation.

improved discrimination of depressive remission, resulting in a 20% increase in AUC for the combined sample, it was only by the fourth or sixth week that accuracy reached a point where they could offer meaningful insights for treatment decisions. Improvements over the first 2 weeks were generally not informative. Larger improvements by week two were observed for nortriptyline, although from a low starting point. A key concern during depression therapy is deciding if and when treatment adaptations are needed. Patients responding poorly to one antidepressant may need to switch to an alternative. Repeated assessments collected during treatment have previously been used to highlight differing response trajectories (Skelton et al. 2022), but the question of how long to wait before considering alternative treatments remains unanswered. A balance is needed between earlier but less accurate predictions versus later predictions that can be enhanced by longitudinal assessments. Overall, we saw a gradual improvement in predictive performance over successive weeks; there was no single point where accuracy markedly increased.

Past studies are inconclusive on the timing of therapeutic response. We found that assessments from the first 2 weeks of treatment were insufficient to predict eventual response reliably. This is consistent with earlier analyses of GENDEP (Uher et al. 2011), where patients showing no improvement within 2 weeks still had a 40%–50% chance of treatment response by 12 weeks. Others have concluded that eventual outcome can only be accurately predicted after 8 weeks of treatment. Our findings are partially consistent with Kemp et al. (2011), who found that while initial improvements predicted later response, there was a high rate of false positives, such that many patients showing initial improvements did not subsequently respond to treatment. By contrast, a large meta-analysis (Szegedi et al. 2009) found improvements within the first 2 weeks of treatment could predict treatment outcome with high sensitivity (> 80%). However, the meta-analysis included studies that had a large dropout rate and performed last observation carried forward (LOCF), which can inflate prediction accuracy.

We found poorer initial performance from models for nortriptyline compared to models for escitalopram. This finding aligns with our previous GENDEP analyses (Iniesta et al. 2016), which, using baseline information only, reported poorer discrimination of remission for nortriptyline versus escitalopram. One possibility is that individuals receiving nortriptyline were more likely to experience adverse effects and end treatment early. It is possible that those continuing beyond 2 weeks represent a pharmacogenetically distinct subset for whom subsequent model prediction is then stronger. Few studies have considered drug-specific predictive accuracy of treatment response using repeated assessments. Athreya et al. (2021) considered the prediction of treatment response for patients receiving a range of antidepressants, including citalopram and escitalopram. For an outcome of remission measured at 8 weeks, they found comparable predictive performance between different drugs (e.g., accuracy of 77% and 72% in training and test cohorts, respectively).

We compared several approaches to incorporating repeated measures into our models. In most samples, the simplest approach of including the LW measures tended to outperform other approaches. This is notable, given the significant methodological differences, and suggests that improvements in accuracy are driven by the new information gained at successive weeks rather than contrasting algorithms. Interestingly, the inclusion of topological features did improve performance at Week 6 for the nortriptyline sample and Weeks 4 and 6 for the combined sample. Uncertainty intervals did not overlap for Week 6 in the combined sample, this suggesting a statistically significant potential for TDA to enhance the prediction of depression remission when using repeated assessments. The performance of growth curves at 2 weeks was particularly poor. The inclusion of topological parameters at 2 weeks similarly had no benefit on performance. These results are likely to reflect the sparseness of our data. More sophisticated techniques, such as growth curve modeling and especially topological data analysis, benefit from larger, high-dimensional datasets. Deriving topological parameters for each individual from a sparse grid is particularly challenging. These techniques performed better at Week 6 when more longitudinal information was available. This finding raises a broader point about competing methods' data requirements and suitability. Topological data analysis is better suited to large, multivariate, and multidimensional data; growth curves require repeated assessments, ideally more than three, but can be univariate. Including the latest measure (LW) is simplest regarding data requirements and analysis but omits potentially important information about the trajectory and intercorrelations of repeated assessments. Further interpretations in terms of confidence or statistical significance of estimations based on the computed intervals should be made cautiously given that the hold-out data in the nested-cross validation are not independent of the test data and can cause some dependency, a fact that has been shown to potentially bias the intervals coverage (Bates, Hastie, and Tibshirani 2023).

4.1 | Strengths and Limitations

Our study is among the first to contrast different approaches to incorporating repeated measurements into a multivariable prediction model for depression remission. We benefited from a large combined sample and were able to consider drug-specific predictive performance. We considered wide-ranging baseline and longitudinal predictors, including item-level data and total scores. We conducted a robust internal validation using repeated, nested cross-validation where all modeling steps were repeated within each fold to avoid data leakage.

However, several limitations must be acknowledged. First, our drug-specific samples were small, and, as a secondary analysis of the GENDEP cohort, no a priori sample size calculation was conducted. Second, while we undertook robust internal validation, our findings were not externally validated. Third, several steps in our pipeline were computationally demanding, in particular, the derivation of landscape parameters for each individual and the generation of growth curve parameters at every cross-validation fold. Simpler methods, such as directly entering the latest measure into the model, would be more computationally feasible in practice. Fourth, to minimize the computational burden of our models, we only used fortnightly measurements (at 0, 2, 4, and 6 weeks), whereas it has been shown that symptoms can change weekly (Rapaport and Maddux 2002). Fifth,

we did not consider dropout information as we were aiming to produce models that could prospectively predict remission and inform a personalized treatment approach. Hence, we could not use the time in the study as a predictor in our models, as this measurement would not be available for new starting treatment patients. In previous research in GENDEP (Uher et al. 2010) dropouts were only found to be related with reduced drug effects in the analyses with remission in the combined sample (patients treated with either nortriptyline or escitalopram). To lower this effect, we stratified our analysis by drug, and considered the prediction of remission separately within the two drug groups. Finally, our models considered symptom level information, as they included the individual items and total scores for the different symptoms considered as predictors. We were aware that symptom-level assessments of our models might provide specific and valuable insights into the predictive ability of the models for particular groups of patients. However, we did not consider including such analysis, as stratifying our sample by symptoms would reduce even further the available sample size, which was already rather small, and would preclude even further the interpretation of our results.

5 | Conclusions

Our analysis highlights the potential benefits of incorporating repeated measurements collected during treatment to predict eventual treatment response. In a large sample of patients receiving antidepressant treatment, adding post-baseline weeks of information suggested incremental benefits in performance over successive weeks, with no single point showing a marked increase in performance. Our machine learning models predicted depression remission by week four with adequate discrimination, suggesting a potential to inform treatment decisions. In summary, this work provides further evidence that models to aid antidepressant decisions can potentially benefit from longitudinal symptom assessments. Future works, including bigger samples and independent validation, are required.

Author Contributions

Ewan Carr: conceptualization and design of the study, formal analysis, writing of the paper, review, and editing. **Marcella Rietschel:** review and editing. **Ole Mors:** review and editing. **Neven Henigsberg:** review and editing. **Katherine J. Aitchison:** review and editing. **Peter McGuffin:** conceptualization of the study, review and editing. **Wolfgang Maier:** review and editing. **Rudolf Uher:** conceptualization of the study, review and editing. **Anne Farmer:** conceptualization of the study, review and editing. **Raquel Iniesta:** conceptualization and design of the study, writing of the paper, review, and editing.

Acknowledgments

This work has been supported by the Brain and Behaviour Foundation awarded to Raquel Iniesta (Award number 26338). The funding body played no role in the design of the study, the collection, analysis, interpretation of data, or in writing the manuscript. This paper represents independent research part-funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

We thank Joanna Hauser, Mojca Zvezdana Dernovsek, Daniel Souery, and Cathryn Lewis for their contribution to the research. This paper is dedicated to the memory of Professor Peter McGuffin (1949–2024), who passed away during the review stage of this work and whose brilliant ideas first brought this project to light. Thank you for being an extraordinary visionary in your field, a generous and kind colleague, and a compassionate, influential, and inspirational figure. We are grateful for your enormous contributions to science and the profound impact you made on all those you encountered. We have nothing but gratitude for you.

Conflicts of Interest

K.J.A. is a member of the Clinical Pharmacogenetics Implementation Consortium, the Pharmacogene Variation Consortium, and the ISPG Genetic Testing Committee, and the founder of DigHap Ltd., a commercial entity that did not make any contribution to the work described herein.

Data Availability Statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

Athreya, A. P., T. Brückl, E. B. Binder, et al. 2021. "Prediction of Short-Term Antidepressant Response Using Probabilistic Graphical Models With Replication Across Multiple Drugs and Treatment Settings." *Neuropsychopharmacology* 46, no. 7: 1272–1282. https://doi.org/10.1038/s41386-020-00943-x.

Bates, S., T. Hastie, and R. Tibshirani. 2023. "Cross-Validation: What Does It Estimate and How Well Does It Do It?" *Journal of the American Statistical Association* 119: 1434–1445. https://doi.org/10.1080/01621459.2023.2197686.

Beck, A. T., R. A. Steer, and M. G. Carbin. 1988. "Psychometric Properties of the Beck Depression Inventory: Twenty-Five Years of Evaluation." *Clinical Psychology Review* 8, no. 1: 77–100. https://doi.org/10.1016/0272-7358(88)90050-5.

Bondar, J., A. Caye, A. M. Chekroud, and C. Kieling. 2020. "Symptom Clusters in Adolescent Depression and Differential Response to Treatment: A Secondary Analysis of the Treatment for Adolescents With Depression Study Randomised Trial." *Lancet Psychiatry* 7, no. 4: 337–343. https://doi.org/10.1016/S2215-0366(20)30060-2.

Browning, M., J. Kingslake, C. T. Dourish, G. M. Goodwin, C. J. Harmer, and G. R. Dawson. 2019. "Predicting Treatment Response to Antidepressant Medication Using Early Changes in Emotional Processing." *European Neuropsychopharmacology* 29, no. 1: 66–75. https://doi.org/10.1016/j.euroneuro.2018.11.1102.

Bubenik, P. 2018. "The Persistence Landscape and Some of Its Properties," arXiv:1810.04963 [Math, Stat]. http://arxiv.org/abs/1810.04963.

Bull, L. M., M. Lunt, G. P. Martin, K. Hyrich, and J. C. Sergeant. 2020. "Harnessing Repeated Measurements of Predictor Variables for Clinical Risk Prediction: A Review of Existing Methods." *Diagnostic and Prognostic Research* 4, no. 1: 9. https://doi.org/10.1186/s41512-020-00078-z.

Chazal, F., and B. Michel. 2017. "An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists," arXiv:1710.04019 [Cs, Math, Stat]. http://arxiv.org/abs/1710.04019.

Chekroud, A. M., J. Bondar, J. Delgadillo, et al. 2021. "The Promise of Machine Learning in Predicting Treatment Outcomes in Psychiatry." *World Psychiatry* 20, no. 2: 154–170. https://doi.org/10.1002/wps.20882.

Chekroud, A. M., R. Gueorguieva, H. M. Krumholz, M. H. Trivedi, J. H. Krystal, and G. McCarthy. 2017. "Reevaluating the Efficacy and Predictability of Antidepressant Treatments: A Symptom Clustering Approach." *JAMA Psychiatry* 74, no. 4: 370–378. https://doi.org/10.1001/jamapsychiatry.2017.0025.

Dinga, R., L. Schmaal, B. W. J. H. Penninx, et al. 2019. "Evaluating the Evidence for Biotypes of Depression: Methodological Replication and Extension of." *NeuroImage. Clinical* 22, no. 101: 796. https://doi.org/10.1016/j.nicl.2019.101796.

Drysdale, A. T., L. Grosenick, J. Downar, et al. 2017. "Resting-State Connectivity Biomarkers Define Neurophysiological Subtypes of Depression." *Nature Medicine* 23, no. 1: 28–38. https://doi.org/10.1038/nm.4246.

Gelenberg, A. J., M. P. Freeman, J. C. Markowitz, et al. 2010. *Practice Guideline for the Treatment of Patients With Major Depressive Disorder*. Virginia: American Psychiatric Association.

Gueorguieva, R., A. M. Chekroud, and J. H. Krystal. 2017. "Trajectories of Relapse in Randomised, Placebo-Controlled Trials of Treatment Discontinuation in Major Depressive Disorder: An Individual Patient-Level Data Meta-Analysis." *Lancet Psychiatry* 4, no. 3: 230–237. https://doi.org/10.1016/S2215-0366(17)30038-X.

Hamilton, M. 1967. "Development of a Rating Scale for Primary Depressive Illness." *British Journal of Social & Clinical Psychology* 6, no. 4: 278–296. https://doi.org/10.1111/j.2044-8260.1967.tb00530.x.

Hilbert, K., S. L. Kunas, U. Lueken, N. Kathmann, T. Fydrich, and L. Fehm. 2020. "Predicting Cognitive Behavioral Therapy Outcome in the Outpatient Sector Based on Clinical Routine Data: A Machine Learning Approach." *Behaviour Research and Therapy* 124, no. 103: 530. https://doi.org/10.1016/j.brat.2019.103530.

Hosmer, D. W., and S. Lemeshow. 2000. *Applied Logistic Regression*. Hoboken, NJ: John Wiley & Sons.

Iniesta, R., E. Carr, M. Carriere, N. Yerolemou, B. Michel, and F. Chazal. 2022. "Topological Data Analysis and Its Usefulness for Precision Medicine Studies." *Statistics and Operations Research Transactions* 46, no. 1: 22.

Iniesta, R., K. Hodgson, D. Stahl, et al. 2018. "Antidepressant Drug-Specific Prediction of Depression Treatment Outcomes From Genetic and Clinical Variables." *Scientific Reports* 8, no. 1: 5530. https://doi.org/10.1038/s41598-018-23584-z.

Iniesta, R., K. Malki, W. Maier, et al. 2016. "Combining Clinical Variables to Optimize Prediction of Antidepressant Treatment Outcomes." *Journal of Psychiatric Research* 78: 94–102. https://doi.org/10.1016/j.jpsychires.2016.03.016.

Kemp, D. E., S. J. Ganocy, M. Brecher, et al. 2011. "Clinical Value of Early Partial Symptomatic Improvement in the Prediction of Response and Remission During Short-Term Treatment Trials in 3369 Subjects With Bipolar I or II Depression." *Journal of Affective Disorders* 130, no. 1–2: 171–179. https://doi.org/10.1016/j.jad.2010.10.026.

Kessler, R. C., H. M. van Loo, K. J. Wardenaar, et al. 2016. "Testing a Machine-Learning Algorithm to Predict the Persistence and Severity of Major Depressive Disorder From Baseline Self-Reports." *Molecular Psychiatry* 21, no. 10: 1366–1371. https://doi.org/10.1038/mp.2015.198.

Koutsouleris, N., R. S. Kahn, A. M. Chekroud, et al. 2016. "Multisite Prediction of 4-Week and 52-Week Treatment Outcomes in Patients With First-Episode Psychosis: A Machine Learning Approach." *Lancet Psychiatry* 3, no. 10: 935–946. https://doi.org/10.1016/S2215-0366(16)30171-7.

Montgomery, S. A., and M. Åsberg. 1979. "A New Depression Scale Designed to Be Sensitive to Change." *British Journal of Psychiatry* 134: 382–389.

NICE. 2022. "Depression in Adults: Treatment and Management," National Institute for Health and Care Excellence. https://www.nice.org.uk/guidance/ng222.

Paul, R., T. F. M. Andlauer, D. Czamara, et al. 2019. "Treatment Response Classes in Major Depressive Disorder Identified by Model-Based Clustering and Validated by Clinical Prediction Models." *Translational Psychiatry* 9, no. 1: 187. https://doi.org/10.1038/s41398-019-0524-4.

Pearson, R., D. Pisner, B. Meyer, J. Shumake, and C. G. Beevers. 2019. "A Machine Learning Ensemble to Predict Treatment Outcomes Following an Internet Intervention for Depression." *Psychological Medicine* 49, no. 14: 2330–2341. https://doi.org/10.1017/S003329171800315X.

Pedregosa, F., G. Varoquaux, A. Gramfort, et al. 2011. "Scikit-Learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.

Python GLMNET. 2021. "Python Civis Analytics,".

Quitkin, F. M., P. J. McGrath, J. W. Stewart, et al. 1996. "Chronological Milestones to Guide Drug Change. When Should Clinicians Switch Antidepressants?" *Archives of General Psychiatry* 53, no. 9: 785–792. https://doi.org/10.1001/archpsyc.1996.01830090031005.

Rapaport, M. H., and R. E. Maddux. 2002. "Challenges in the Development of Clinical Trials for Major Depressive Disorder: Lessons Learned From Trials in Minor Depression." *Dialogues in Clinical Neuroscience* 4, no. 4: 402–407.

Riihimäki, H., W. Chachólski, J. Theorell, J. Hillert, and R. Ramanujam. 2020. "A Topological Data Analysis Based Classification Method for Multiple Measurements." *BMC Bioinformatics* 21, no. 1: 336. https://doi.org/10.1186/s12859-020-03659-3.

Seabold, S., and J. Perktold. 2010. "Statsmodels: Econometric and statistical modeling with python," 9th Python in Science Conference.

Skelton, M., E. Carr, J. E. J. Buckman, et al. 2022. "Trajectories of Depression and Anxiety Symptom Severity During Psychological Therapy for Common Mental Health Problems." *Psychological Medicine* 53, no. 13: 6183–6193. https://doi.org/10.1017/S0033291722003403.

Szegedi, A., W. T. Jaansen, A. P. P. van Willigenburg, E. van der Meulen, H. H. Stassen, and M. E. Thase. 2009. "Early Improvement in the First 2 Weeks as a Predictor of Treatment Outcome in Patients With Major Depressive Disorder: A Meta-Analysis Including 6562 Patients." *Journal of Clinical Psychiatry* 70, no. 3: 344–353.

The GUDHI Project. 2020. "GUDHI user and reference manual (3.1.1). GUDHI Editorial Board," https://gudhi.inria.fr/doc/3.1.1/.

Troyanskaya, O., M. Cantor, G. Sherlock, et al. 2001. "Missing Value Estimation Methods for DNA Microarrays." *Bioinformatics* 17, no. 6: 520–525. https://doi.org/10.1093/bioinformatics/17.6.520.

Uher, R., A. Farmer, W. Maier, et al. 2008. "Measuring Depression: Comparison and Integration of Three Scales in the GENDEP Study." *Psychological Medicine* 38, no. 2: 289–300. https://doi.org/10.1017/S0033291707001730.

Uher, R., P. Huezo-Diaz, N. Perroud, et al. 2009. "Genetic Predictors of Response to Antidepressants in the GENDEP Project." *Pharmacogenomics Journal* 9, no. 4: 225–233. https://doi.org/10.1038/tpj.2009.12.

Uher, R., O. Mors, M. Rietschel, et al. 2011. "Early and Delayed Onset of Response to Antidepressants in Individual Trajectories of Change During Treatment of Major Depression: A Secondary Analysis of Data From the Genome-Based Therapeutic Drugs for Depression (GENDEP) Study." *Journal of Clinical Psychiatry* 72, no. 11: 1478–1484. https://doi.org/10.4088/JCP.10m06419.

Uher, R., R. H. Perlis, N. Henigsberg, et al. 2012. "Depression Symptom Dimensions as Predictors of Antidepressant Treatment Outcome: Replicable Evidence for Interest-Activity Symptoms." *Psychological Medicine* 42, no. 5: 967–980. https://doi.org/10.1017/S0033291711001905.

Uher, R., N. Perroud, M. Y. M. Ng, et al. 2010. "Genome-Wide Pharmacogenetics of Antidepressant Response in the GENDEP Project." *American Journal of Psychiatry* 167, no. 5: 555–564. https://doi.org/10.1176/appi.ajp.2009.09070932.

Van Rossum, G., and F. L. Drake. 2009. "Python 3 Reference Manual," CreateSpace.

Ward, J., N. Graham, R. J. Strawbridge, et al. 2018. "Polygenic Risk Scores for Major Depressive Disorder and Neuroticism as Predictors of Antidepressant Response: Meta-Analysis of Three Treatment Cohorts." *PLoS One* 13, no. 9: e0203896. https://doi.org/10.1371/journal.pone.0203896.

Webb, C. A., Z. D. Cohen, C. Beard, M. Forgeard, A. Peckham, and T. Björgvinsson. 2020. "Personalized Prognostic Prediction of Treatment Outcome for Depressed Patients in a Naturalistic Psychiatric Hospital Setting: A Comparison of Machine Learning Approaches." *Journal of Consulting and Clinical Psychology* 88, no. 1: 25–38. https://doi.org/10.1037/ccp0000451.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.