Developing, validating and comparing competing models of individual symptoms to make prognostic predictions for adults with depression in primary care.

Authors

Buckman, J. E. J.,*[1,2] Cohen, Z. D.,[3] O'Driscoll, C.,[1] Fried, E. I.,[4] Saunders, R.,[1] Ambler, G.,[5] DeRubeis, R. J.,[6] Gilbody, S.,[7] Hollon, S. D.,[8] Kendrick, T.,[9] Watkins, E.,[10] Eley, T.C.,[11], Peel, A. J.,[11] Rayner, C.,[11] Kessler, D.,[12] Wiles, N.,[13] Lewis, G.,[14] & Pilling, S.[1,15]

Contact

* Joshua.buckman@ucl.ac.uk, 0207 679 1785

1 – Centre for Outcomes Research and Effectiveness (CORE), Research Department of Clinical, Educational & Health Psychology, University College London, 1-19 Torrington Place, London WC1E 7HB, UK

2- iCope – Camden & Islington Psychological Therapies Services – Camden & Islington NHS Foundation Trust, St Pancras Hospital, London NW1 0PE, UK.

3 – Department of Psychiatry, University of California, Los Angeles, Los Angeles, CA, USA

4 – Department of Clinical Psychology, Leiden University, Leiden, The Netherlands.

5 – Statistical Science, University College London, 1-19 Torrington Place, London WC1E 7HB

6 – School of Arts and Sciences, Department of Psychology, 425 S. University Avenue, Philadelphia PA, 19104-60185, USA

7 – Department of Health Sciences, University of York, Seebohm Rowntree Building, Heslington, York YO10 5DD, UK

8 –Department of Psychology, Vanderbilt University, Nashville, TN, USA

9 – Primary Care & Population Sciences, Faculty of Medicine, University of Southampton, Aldermoor Health Centre, Southampton SO16 5ST.

10- Department of Psychology, University of Exeter, Sir Henry Wellcome Building for Mood Disorders Research, Perry Road, Exeter EX4 4QG

11 – Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology & Neuroscience, King's College London; London SE5 8AF, UK

12 – Centre for Academic Primary Care, Population Health Sciences, Bristol Medical School, University of Bristol, Canynge Hall, Bristol, UK

13 – Centre for Academic Mental Health, Population Health Sciences, Bristol Medical School, University of Bristol, Oakfield House, Bristol, UK

14 – Division of Psychiatry, University College London, Maple House, London W1T 7NF, UK

15 – Camden & Islington NHS Foundation Trust, St Pancras Hospital, London NW1 0PE, UK.

Abstract

Aims

To develop, validate, and compare the performance of nine models predicting post-treatment outcomes for depressed adults based on pre-treatment data.

Methods

Individual patient data from all six eligible RCTs were used to develop (k=3, n=1722) and test (k=3, n=1136) nine models. Predictors included depressive and anxiety symptoms, social support, life events and alcohol use. Weighted sum-scores were developed using coefficient weights derived from network centrality statistics (Models 1-3) and factor loadings from a confirmatory factor analysis (Model 4). Unweighted sum-score models were tested using Elastic Net Regularized (ENR) and ordinary least squares (OLS) regression (Models 5-6). Individual items were then included in ENR and OLS (Models 7-8). All models were compared to one another and to a null model using the mean post-baseline BDI-II score in the training data (Model 9). Primary outcome: BDI-II scores at 3-4 months.

Results

Models 1-7 all outperformed the null model. Individual-item models (particularly Model 8) explained less variance. Model performance was very similar across models 1-6, meaning that differential weights applied to the baseline sum-scores had little impact.

Conclusions

Any of the modelling techniques (1-7) could be used to inform prognostic predictions for depressed adults with differences in the proportions of patients reaching remission based on the predicted severity of depressive symptoms post-treatment. However, the majority of variance in prognosis remained unexplained. It may be necessary to include a broader range of biopsychosocial variables to better adjudicate between competing models, and to derive models with greater clinical utility for treatment-seeking adults with depression.

Keywords: depressive symptoms; major depression; network analysis; prognosis; prediction modelling

Introduction

Depression affects approximately 320 million people worldwide every year [1,2]. Despite the existence of many effective treatments, roughly half of depressed patients fail to recover with the first treatment they are given. This can lead patients to disengage from treatment or to poor long term prognoses [3,4]. Providing accurate predictions about the likelihood of treatment response for patients would be of great value, informing clinical management and providing patients and clinicians with information they want to know [5,6]. However, despite advances in both the methods of predictive modelling and availability of computational power and resources necessary to build and test complex models, there is a lack of accurate, validated prognostic models for adults in treatment for depression [7]. Central to this vacancy in the literature are methodological inconsistencies, debates about how best to develop predictive models, and what variables to include in such models. Only relatively recently has the field begun to reach consensus on how best to test the utility of such models robustly [8–12].

One factor consistently found to be associated with prognosis of depression is the severity of depressive symptoms pre-treatment [13–16]. This is often captured with sum-scores on depressive symptom scales. However, depression is a heterogeneous disorder [17] so it might be the case that utilizing information at the symptom level could provide more nuanced information on patient's experience of depression, and consequently improve the accuracy of prognostic predictions [18–21]. Network theory [22,23] has given rise to an approach that can capture the relationships between individual symptoms. These relationships could reflect potential causal pathways, thereby elucidating maintenance mechanisms that could be targeted with treatment, and might therefore inform prognosis [24]. The arrangement and inter-relationships of symptoms within networks have most often been captured with one or more measures of centrality—i.e. the interconnectedness of each symptom with other symptoms in the network [25,26].

A recent study used centrality metrics to weight individual items of a depressive symptom questionnaire, which when summed together created a new, or weighted, sum-score. A regression model using this weighted sum-score was found to outperform a model containing the original sum-score in an exploratory analysis predicting first onsets of depression in adults [21]. Other studies have utilized the centrality of symptoms within networks to generate predictions about changes in particular symptoms over time [21,27–29], or to generate predictions of post-treatment outcomes [30,31]. However, such studies have not tested the developed models against simpler comparative models. Further, the predictive utility of models developed based on network centrality statistics has not been tested in data separate from those used to develop the models [8,9,32], and past studies have not adhered to recent conventions for the transparency of conducting such research by following pre-registered analysis plans or protocols [33]. Therefore, the extent to which the use of centrality metrics can add incremental value in prognostic models remains unclear. The present paper aims to fill this gap and further the consideration of the development of models that can be translated into clinical settings.

There are several potentially equally valid ways to estimate item centrality in network models. We will therefore investigate several methods that have been used in the recent network modeling literature. One method uses the estimated arrangement of items into communities of highly partially correlated items, we will compare this to a model from a factor analysis in which it is assumed that there is a single latent factor. We will use these methods to investigate the benefit of using item centrality scores and factor loadings to create weighted sum-scores, and compare these to an unweighted regression model, and to a penalized regression model. We will then compare these methods against models that use all the individual items rather than sum-scores, and to a simple null model [21]. In this way, this study aims to develop, validate, and compare the predictive

performance of prognostic models for depressed adults in primary care, based on pre-treatment data including individual symptoms of depression.

Methods

The methods for the present study were pre-registered (https://osf.io/vzk65/). We have reported the details in accordance with TRIPOD, brief details are given below, and further information including a TRIPOD checklist is available in the Supplementary materials.

Participants

Individual patient data (IPD) were drawn from a subset of the Depression in General Practice (Dep-GP) IPD dataset [34]. Studies were included here if they were Randomized Clinical Trials (RCTs) that recruited adults with depression in primary care centres, used the Revised Clinical Interview Schedule (CIS-R) [35] to collect depressive and anxiety symptom data and determine diagnoses. This was in order to bring uniformity to the items available to use in the predictive models across the studies. From our previous work we have found that the CIS-R is the most commonly used comprehensive measure of this kind in studies of depression in primary care [36]. Studies also had to use the Beck Depression Inventory Second Edition (BDI-II) [37] to collect individual symptoms of depression. Six RCTs met inclusion criteria and were split such that half (k=3, n=1722) would form a dataset to develop the predictive models (the 'training set') and half (k=3, n=1136) would form a separate dataset to test the models (the 'test set'). See Supplementary Table 1 and Supplementary Figure 1, for details of each study. It was decided that studies with similar types of treatment would be split across the training and test sets, and where this was the case, those with the larger sample sizes would go into the training data.

Predictors and Measures

Predictors varied depending on the model used, as detailed below (Table 1). Models either included total scores (with items either weighted or unweighted) or individual items from the BDI-II. All models used total scores for eight anxiety subscales from CIS-R (generalized anxiety, worry, compulsions, obsessions, phobic anxiety, health anxiety, somatic concerns, and panic; with items either weighted or unweighted), and total scores for alcohol use, social support, and life events. However, the Null models used the BDI-II total score only. See supplementary Table 2 for details of the measures.

Outcomes

The primary outcome was the BDI-II score at 3-4 months post-baseline. The secondary outcome was remission at 3-4 months post-baseline, defined as a score of 10 or less on the BDI-II. In all but one of the six studies, assessors and analysts were blind to treatment allocation when collecting these data.

Data Analysis

For details of the pre-processing stages and handling of missing data see Supplementary Materials. In brief, missing data were imputed in the training set for all variables with less than 30% missing data, using the "missForest" package in R [38]. In the test set the same approach was used but outcome data were not imputed. The maximum amount of missing data of any of the variables used in the predictive models here, at baseline in any of the six studies was 0.83%. For one study whose data were included in the training set, COBALT, BDI-IIs at 3-4 months were not collected. These scores were imputed using the methods above based on all available variables in that study including baseline BDI-II scores and PHQ-9 scores, three-month PHQ-9 scores, six-month BDI-II and PHQ-9 scores, and 12-month BDI-II and PHQ-9 scores.

Model Building

For both primary and secondary outcomes, nine models were constructed in the training set (Table 1).

For the first four models we developed separate weighted sum-scores for the CIS-R anxiety subscales by summing together coefficient weights for each of the eight subscales, and for the BDI-II by summing together coefficient weights for each of the 21 BDI-II items. Weighted sum-scores for the CIS-R anxiety subscales and BDI-II, and coefficient weights for the total scores for social support, life events, and alcohol were used as predictors by entering them into regression models (ordinary least squares for the primary outcome and logistic regression for the secondary outcome). This follows a method used by others to develop predictive models from networks [21]. As described below, Models 5 and 7 were based on a method that develops model weights internally (Elastic Net Regularized Regression: ENR). Models 6 and 8 used the original, unweighted scores as a means of comparison. Model 9 was a Null model, detailed further below.

Network Analyses

For models 1-3, Gaussian Graphical 'network' Models (GGM) were estimated using item-level data from CIS-R anxiety subscales and the BDI-II. In order to estimate networks in the training dataset, multiple GGMs were jointly estimated implementing a penalty based on the density of the network and edge weight differences between samples, with tuning parameters selected through 10-fold cross validation [39,40]. Centrality metrics derived from the GGM were used to construct weights after re-scaling these to be between 0-1. The three methods for determining coefficient weights from the estimated networks were: Model 1) 1-step expected influence (EI: sum of all edges connected to the focal node); Model 2) 2-step expected influence (sum of all edges connected to either the focal node or any other node directly connected to the focal node) [41]; and Model 3) the geometric mean of the participation coefficient (PC) and participation ratio (PR) [42]. For details of these methods see Supplementary Materials. The EI metrics are widely used and have recently been proposed to be informative for predicting treatment outcomes [30,31]. PC/PR is a newer approach which is thought to be more sensitive to the use of different scale measures within the same network, as it takes the community structure (multidimensionality) into account [42]. This is important here as we used measures of severity beyond depressive symptoms, given their importance for prognosis [36,43].

Confirmatory Factor Analyses

Model 4 was a unidimensional confirmatory factor analytic (CFA) model that assumes the data come from a single dimensional latent construct (in contrast to Model 3, which is based on a Walktrap algorithm that identifies densely connected communities of items via random walks). Factor loadings were rescaled to be between 0-1 (as with the weights for Models 1-3) and summed to develop the weighted total scores.

Penalized Regression Analyses

Model 5 was an ENR model built using the unweighted total scores on the same scales that were used for models 1-4. In ENR, variables are selected and model weights are assigned through the use of LASSO (least absolute shrinkage and selection operator) and ridge penalizations. A parameter space search was conducted using 10-fold cross-validation to identify the optimal settings for these parameters before building the final ENR model [32,44]. Model 7 was an ENR model using all of the individual items from the BDI-II and the CIS-R anxiety subscales, and total scores for life events, social support and alcohol use.

Non-Penalized Regression Analyses

Two simple comparison models were constructed using non-penalized regression (OLS regression for continuous outcomes and logistic regression for binary outcomes). Model 6 used the unweighted total scores on the five baseline measures, and Model 8 used the same items as Model 7.

Null Models

A null model was built for each outcome for the purpose of comparison. For the primary outcome this used the mean 3-4 month BDI-II score in the training set as the prediction for all patients in the test set, and for the secondary outcome the proportion of participants in remission in the training set was used as the prediction for all patients in the test set.

Sensitivity Analyses

In order to assess the impact of having to impute the 3-4 month BDI-II outcomes for the COBALT study, we conducted two sensitivity analyses. All analyses using BDI-II as the outcome were re-done excluding COBALT from the training dataset. Then, a different way of capturing depressive symptoms at 3-4 months post-baseline was calculated and used as an outcome variable. This was based on a method of converting scores from different depressive symptom measures to a single comparable score; the PROMIS T-score [45]. In order to achieve this we used a multidimensional item-response theory (IRT) based conversion tool [46], see Supplement for further details.

Model Evaluation

Models were first evaluated in the full test set comprising three studies (TREAD, IPCRESS, and MIR), and then separately in each of the three study samples. They were also evaluated in a 10-fold internal cross-validation of the training data.

For the continuous outcomes the amount of variance explained (R^2), the root mean squared error (RMSE) and the mean absolute error (MAE) in the predictions were used to compare the predictive accuracy of the models. For the binary outcome the area under the receiver operating characteristic curve (AUC) and Brier scores were used to compare the models. Since the R^2 in this study is a comparison of the predicted BDI-II score values to the mean BDI-II score at 3-4 months in the test set, and the training and test set BDI-II score means at 3-4 months differed, it was expected that some models might have R^2 values less than zero. There are limits to the inferences that can be drawn from the above metrics due to the variability in the modelling schemes that were applied, which differed in a variety of ways, including: which variables were made available; the number of variables made available; whether or not network analysis or factor analysis was used to create weighted sum-scores; and whether or not penalized regression was applied to the variables that were made available. To make these performance metrics more accessible, we have provided three visualizations that demonstrate the potential clinical relevance of each model. For each of the eight models the predicted BDI-II scores at 3-4 months were arrayed from lowest to highest, then: 1) we plotted the observed BDI-II score at 3-4 months against the predicted score in groups (“bins”) of n=50; 2) predicted scores were split into categories of severity in line with delineations made by the originators of the scale [37] (i.e. scores between 0-13 were considered minimal, 14-19 mild, 20-28 moderate, and 29-63 severe), and the rate of remission observed in the test set samples was calculated for each category; and 3) to provide a more granular visualization of remission we plotted the observed percentage of participants in remission against BDI-II predicted scores at 3-4 months, again in bins of n=50.

Results

Characteristics of the included studies

Six RCTs were identified as meeting inclusion criteria, see Supplementary Figure 1 for flow of studies and Supplementary Table 1 for details of each study.

Descriptive Statistics

Descriptive statistics and comparisons of the distributions of socio-demographics and markers of severity across the training set and test set samples are provided in Table 2. There were some differences between the training and test datasets: fewer people of non-white ethnicities were in the test set, more of the training sample were unemployed, and the mean score on the AUDIT-PC was higher in the test set. In addition, the mean BDI-II scores were higher in the test set (by 2.47 points at baseline and 3.53 points at 3-4 months). This corresponded with a large difference in the proportions of each sample reaching remission: 48.83% in the training set and 32.53% in the test set.

Formation of the Models

The weights given to the individual items for models 1-4 are shown in Supplementary Table 6. Final model coefficients are presented in Supplementary Tables 7 and 8.

Comparison of Model Performance

After the models were developed they were evaluated using the test dataset. Despite slight differences in the formation of some of the models, they made very similar predictions of who would get better (remit) and by what magnitude (BDI-II score) at 3-4 months. To illustrate this the predictions produced for the primary outcome by the models were highly correlated (all correlation coefficients above r=0.90 for models 1-6 and above r=0.75 for models 7-8) see Supplementary Figure 2.

For the primary outcome (BDI-II score at 3-4 months post-baseline) in the combined test sets, the RMSE was similar for models 1-6 (the largest difference was between Model 2 which had the lowest RMSE and Model 4, =0.057) with slightly higher RMSE for the OLS individual-item model (Model 8) (difference between Model 2 and Model 8 =0.214). Models 1-8 made similar predictions for those with BDI-II scores at 3-4 months that were below 18 or above 25, but diverged more in the predictions for those with scores between 18-24, see Figure 1 (for ease of presentation, results are displayed for groups of 50 participants, each point shows the mean predicted and observed score for the 50 participants closest to that point on the graph). All models (1-8) had lower RMSE scores than the Null model (ranging between 0.944 for the difference between Model 8 and 9, to 1.158 for the difference between Model 2 and 9), see Table 3. The amount of variance explained by models 1-7 was again very similar with R^2 values between 0.157 and 0.169. Model 8 (R^2=0.109) explained less variance, but all models had R^2 values well above the Null model (R^2= -0.01). MAE values were similar for Models 1-7 (ranging between 9.089 for Model 5, and 9.173 for Model 7). MAE was slightly higher in Model 8 (=9.279) and higher again in the Null model (9.935), see Table 3. For the secondary outcome there was a similar pattern to the results, although the Null model (9) had a similar Brier score to models 1-7 and this was slightly lower than that of Model 8 (=0.246), see Supplementary Table 3. There were greater variations between the models in the separate test set studies than in the overall test set and for all models (1-9) the RMSE and MAE scores were lower, and R^2s were higher, in the internal cross-validation than in the external test set data.

In order to evaluate the potential clinical relevance of the models we determined the observed proportion of participants in remission at 3-4 months based on the predicted score made by each model (Supplementary Figure 3), and the same based on categories of severity of symptoms taken from the predicted scores (see Figure 2). From these figures we can see that when the models predicted high BDI-II scores at 3-4 month the chances of being in remission were very low. Models 7

and 8 predicted more participants would have severe depression at 3-4 months than the other models. When the models predicted minimal symptoms (BDI-II scores less than 10) the observed rate of remission was around 50%. There were few differences between the models overall, although greater variations in the observed rates of remission between the models for patients predicted to have mild to moderate BDI-II scores at 3-4 months.

Sensitivity analyses did not lead to any substantive differences in our findings, see Supplementary Tables 2-3.

Discussion

There were few differences in the performance of the majority of the predictive models: the first seven models all outperformed the null models on all metrics for primary and secondary outcomes, and those using weighted or unweighted sum-scores (the first six models) performed better in the held-out test data than the individual item models did, particularly Model 8 (the ordinary least squares regression model using all of the individual BDI-II score items and eight CIS-R anxiety subscale scores instead of the sum-scores for each). Any of the eight models could be used to predict the severity of depressive symptoms at 3-4 months after starting treatment based on pre-treatment data. The large difference in observed remission rates between those predicted to have high compared to low BDI-II scores at 3-4 months informs the potential clinical relevance of these models.

Strengths and limitations

This study was the first to provide robust tests of the ability of centrality statistics from GGM networks and factor loadings from a factor-analytic model to develop weighted total scale scores to inform predictive models of treatment outcomes. This is something that has been proposed as a promising method for using individual symptom data to build informative predictive models [21]. We tested these methods against *bone fide* predictive models and simple comparison models, and in entirely held-out (test) data, and found there to be little evidence of any advantage to the above approaches. We used a large individual patient data dataset comprising six RCTs with a variety of widely available treatments for depression, all of the RCTs were situated in primary care, and five were pragmatic trials, increasing the generalizability of these results [47]. We included a range of psychopathology measures at baseline, not just depression symptoms from a single measure, as there is good evidence that such factors are associated with prognosis for depressed adults [36]. We also used the most commonly utilized comprehensive measure of depressive and anxiety symptoms and diagnoses from RCTs of depression in primary care, to minimize bias in harmonizing data, and ensure a broad range of depressive and anxiety based symptoms could be included in the models we developed.

However, there were a number of limitations. Not all important covariates were controlled for: we did not include data on durations of depression or anxiety despite their associations with prognosis for adults with depression [36,43]. Including such data would have led to problems of multi-collinearity with the symptoms of the individual comorbid anxiety disorders experienced by each participant, and across durations of anxiety disorders and depression, biasing centrality estimates and factor loadings for Models 1-4. The intercepts and coefficient weights provided in the supplementary materials could be used to derive prognostic predictions for future depressed patients using models developed here. However, all of the models had large amounts of variance in the outcome that could not be explained. Whilst some proportion is likely due to measurement error, for the majority of patients it is likely that other factors including those that better capture the biopsychosocial complexity of depression would need to be included before the predictive models could more accurately predict prognosis for any individual patient [48]. Crucially, for this study, such

improvements in accuracy may also have been required for us to find any differences in the performance of the modelling schemes.

The present study used prognostic outcomes including depressive symptom severity at 3-4 months and remission, but both of these relied on sum-scores from the BDI-II. As the BDI-II items or sum-score were used in the development of the predictive models it might have been informative to consider model performance with an entirely separate but clinically meaningful outcome such as functioning, quality of life, or mental pain [49]; data on such outcomes were not available here. In addition, models here used IPD but the networks were estimated based on aggregated data, a number of studies have shown the potential utility of using idiographic networks to predict outcomes for individual patients [50–52], this may yet prove the most fruitful avenue for using networks to inform prognostic models which are able to outperform classic regression models of the same factors.

Implications and Conclusions

Prognoses generated by the models developed here could be informative for depressed patients seeking treatment in primary care. However, there were few differences between the models, with no clear advantage in using individual items over sum-scores, or in using network models or factor analytic models to weight individual items, in order to derive prognostic predictions. In all of the models the degree of inaccuracy in their predictions might be unacceptable to any individual patient, although for those predicted to have particularly low or high scores, there were clear differences in the number of people reaching remission. It is noteworthy that all of the models utilised both depressive and anxiety symptom data, and all but one included the total score from the life events scale, and six of the eight included the Social Support Scale score. It might therefore be informative for prognosis to assess for these factors routinely in clinic. The individual item models outperformed the others in the internal cross-validation data suggesting that narrow constructs (e.g. anhedonia) might be more informative for prognosis than broad constructs (e.g. depression), but issues of measurement error arise, particularly with the validity of the single items to measure each narrow construct. The findings presented here also highlight the importance of external validation in accounting for issues of overfitting.

Authors contributions

JEJB, CO'D, ZDC, and EF conceived of the original project, JEJB along with SP, GL, RJD, SDH, SG, TK, EW, and GA applied for and received funding to support this work. All ten of the above, and RS wrote the initial protocol document and plan for the current study. ZDC, NW, DK, TK, SG, and GL provided data and liaison to resolve issues and discrepancies between received datasets and publications about those studies. JEJB, RS, GL and SP were responsible for the screening of studies, data extraction, and additional data cleaning. JB, CO'D, and ZDC conducted the data analyses with support from EF, GA, RS, GL and SP, and consultation from all other authors. JEJB wrote the original manuscript will support from ZDC, RS, CO'D, EF, GA, GL and SP. All authors contributed to consecutive drafts and approved the final manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted.

Conflict of interest statements

The authors have no conflicts of interest to declare.

Funding Sources

This work was supported by the Wellcome Trust through a Clinical Research Fellowship to JEJB (201292/Z/16/Z), MQ Foundation (for ZDC: MQDS16/72), the Higher Education Funding Council for England, the National Institute of Health Research (NIHR), NIHR University College London Hospitals Biomedical Research Centre (CO'D, RS, GL and SP), NIHR Biomedical Research Centre at the University Hospitals Bristol and Weston NHS Foundation Trust and the University of Bristol (NW and DK), University College London (GA, GL), University of Pennsylvania (RJD), Vanderbilt University (SDH), University of Southampton (TK), University of Exeter (EW), and University of York (SG), National Institute for Health Research (NIHR) Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London (TE, AP, and CR). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

The included studies that make up the Dep-GP IPD database were funded by:

1. COBALT: The National Institute for Health Research Health Technology Assessment (NIHR HTA) programme (project number 06/404/02).
2. GENPOD: Medical Research Council and supported by the Mental Health Research Network.
3. IPCRESS: BUPA Foundation
4. MIR: National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme (project 11/129/76) and supported by the NIHR Biomedical Research Centre at University Hospitals Bristol NHS Foundation Trust and the University of Bristol
5. PANDA: NIHR Programme Grant for Applied Research (RP-PG-0610-10048).
6. TREAD: National Institute for Health Research (NIHR) Health Technology Assessment (HTA) programme.

The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. All authors were fully independent of their respective funders and had responsibility for the decision to submit for this manuscript for publication. The guarantor accepts full responsibility for the work and the conduct of the study, had access to the data, and controlled the decision to publish.

References

1 Thornicroft G, Chatterji S, Evans-Lacko S, Gruber M, Sampson N, Aguilar-Gaxiola S, et al. Undertreatment of people with major depressive disorder in 21 countries. Br J Psychiatry. 2017;210(2):119–24.

2 Vos T, Allen C, Arora M, Barber RM, Bhutta ZA, Brown A, et al. Global , regional , and national incidence , prevalence , and years lived with disability for 310 diseases and injuries , 1990 – 2015 : a systematic analysis for the Global Burden of Disease Study 2015. Lancet. 2016;388:1545–602.

3 Buckman JEJ, Underwood A, Clarke K, Saunders R, Hollon SD, Fearon P, et al. Risk factors for relapse and recurrence of depression in adults and how they operate: A four-phase systematic review and meta-synthesis. Clin Psychol Rev. 2018 Aug;64(7):13–38.

4 Judd LL, Akiskal HS, Maser JD, Zeller PJ, Endicott J, Coryell W, et al. Major depressive disorder: A prospective study of residual subthreshold depressive symptoms as predictor of rapid relapse. J Affect Disord. 1998;50(2–3):97–108.

5 Hayden JA, Windt DA Van Der, Cartwright JL, Côté P, Bombardier C. Assessing Bias in Studies of Prognostic Factors. Ann Intern Med. 2013;158:280–6.

6 Morgan AJ, Reavley NJ, Jorm AF. Beliefs about mental disorder treatment and prognosis: Comparison of health professionals with the Australian public. Aust N Z J Psychiatry. 2014;48(5):442–51.

7 Cohen ZD, DeRubeis RJ. Treatment Selection in Depression. Annu Rev Clin Psychol. 2018 May;14(1):209–36.

8 Harrell FE, Lee KL, Mark DB. Prognostic/Clinical Prediction Models: Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors. Tutorials Biostat Stat Methods Clin Stud. 2005;1:223–49.

9 Dwyer DB, Falkai P, Koutsouleris N. Machine Learning Approaches for Clinical Psychology and Psychiatry. Annu Rev Clin Psychol. 2018 May;14(1):91–118.

10 Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the Performance of Prediction Models. Epidemiology. 2010 Jan;21(1):128–38.

11 Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. Ann Intern Med. 2015 Jan;162(1):W1.

12 Adibi A, Sadatsafavi M, Ioannidis JPA. Validation and Utility Testing of Clinical Prediction Models. JAMA. 2020 Jul;324(3):235.

13 Bower P, Kontopantelis E, Sutton A, Kendrick T, Richards DA, Gilbody S, et al. Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. BMJ. 2013 Feb;346(feb26 2):f540–f540.

14 Driessen E, Cuijpers P, Hollon SD, Dekker JJM. Does Pretreatment Severity Moderate the Efficacy of Psychological Treatment of Adult Outpatient Depression ? A Meta-Analysis. J Consult Clin Psychol. 2010;78(5):668–80.

15 Fournier JC, Derubeis RJ, Hollon SD, Shelton RC, Fawcett J. Antidepressant Drug Effects and Depression Severity. J Am Med Assoc. 2010;303(1):47–53.

16 Weitz ES, Hollon SD, Twisk J, Van Straten A, Huibers MJH, David D, et al. Baseline depression severity as moderator of depression outcomes between cognitive behavioral therapy vs pharmacotherapy: An individual patient data meta-analysis. JAMA Psychiatry. 2015;72(11):1102–9.

17 Fried EI, Nesse RM. Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study. J Affect Disord. 2015 Feb;172:96–102.

18 Fava GA, Ruini C, Belaise C. The concept of recovery in major depression. Psychol Med. 2007;37(3):307–17.

19 Fried EI, Nesse RM. The impact of individual depressive symptoms on impairment of psychosocial functioning. PLoS One. 2014;9(2). DOI: 10.1371/journal.pone.0090311

20 Fried EI, Nesse RM. Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. BMC Med. 2015;13(1):1–11.
21 Boschloo L, van Borkulo CD, Borsboom D, Schoevers RA. A Prospective Study on How Symptoms in a Network Predict the Onset of Depression. Psychother Psychosom. 2016;85(3):183–4.
22 Fried EI, Cramer AOJ. Moving Forward: Challenges and Directions for Psychopathological Network Theory and Methodology. Perspect Psychol Sci. 2017 Nov;12(6):999–1020.
23 Borsboom D, Cramer AOJ. Network Analysis: An Integrative Approach to the Structure of Psychopathology. Annu Rev Clin Psychol. 2013 Mar;9(1):91–121.
24 Borsboom D. A network theory of mental disorders. World Psychiatry. 2017 Feb;16(1):5–13.
25 Bringmann LF, Elmer T, Epskamp S, Krause RW, Schoch D, Wichers M, et al. What do centrality measures measure in psychological networks? J Abnorm Psychol. 2019 Nov;128(8):892–903.
26 Fried EI, Epskamp S, Nesse RM, Tuerlinckx F, Borsboom D. What are "good" depression symptoms? Comparing the centrality of DSM and non-DSM symptoms of depression in a network analysis. J Affect Disord. 2016;189:314–20.
27 van Borkulo C, Boschloo L, Borsboom D, Penninx BWJH, Waldorp LJ, Schoevers RA. Association of Symptom Network Structure With the Course of Depression. JAMA Psychiatry. 2015 Dec;72(12):1219.
28 Wichers M, Groot PC. Critical Slowing Down as a Personalized Early Warning Signal for Depression. Psychother Psychosom. 2016;85(2):114–6.
29 Koenders MA, De Kleijn R, Giltay EJ, Elzinga BM, Spinhoven P, Spijker AT. A network approach to bipolar symptomatology in patients with different course types. PLoS One. 2015;10(10):1–16.
30 Elliott H, Jones PJ, Schmidt U. Central Symptoms Predict Posttreatment Outcomes and Clinical Impairment in Anorexia Nervosa: A Network Analysis. Clin Psychol Sci. 2020;8(1):139–54.
31 Berlim MT, Richard-Devantoy S, Dos Santos NR, Turecki G. The network structure of core depressive symptom-domains in major depressive disorder following antidepressant treatment: A randomized clinical trial. Psychol Med. 2020;(May). DOI: 10.1017/S0033291720001002
32 Webb CA, Cohen ZD, Beard C, Forgeard M, Peckham AD, Björgvinsson T. Personalized prognostic prediction of treatment outcome for depressed patients in a naturalistic psychiatric hospital setting: A comparison of machine learning approaches. J Consult Clin Psychol. 2020 Jan;88(1):25–38.
33 Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. BMC Med. 2015;13(1):1–10.
34 Buckman JEJ, Saunders R, Cohen ZD, Clarke K, Ambler G, DeRubeis RJ, et al. What factors indicate prognosis for adults with depression in primary care? A protocol for meta-analyses of individual patient data using the Dep-GP database. Wellcome Open Res. 2020 Apr;4:69.
35 Lewis G, Pelosi AJ, Araya R, Dunn G. Measuring psychiatric disorder in the community : a standardized assessment for use by lay interviewers. Psychol Med. 1992;22:465–86.
36 Buckman JEJ, Saunders R, Cohen ZD, Barnett P, Clarke K, Ambler G, et al. Indicators of Prognosis Independent of Treatment for Adults with Depression in Primary Care, Going Beyond Baseline Symptom-Severity: A Systematic Review and Individual Patient Data Meta-Analysis. SSRN Electron J. 2020 DOI: 10.2139/ssrn.3520082
37 Beck AT, Steer RA, Brown GK. Manual for the Beck Depression Inventory-II. Man Beck Depress Invent. 1996
38 Stekhoven ADJ, Stekhoven MDJ. Package ' missForest .' 2011
39 Costantini G, Epskamp S. Package "EstimateGroupNetwork." 2017;1–10.
40 Danaher P, Wang P, Witten DM. for Inverse Covariance Estimation Across Multiple Classes. J

R Stat Soc. 2014;76(2):373–97.
41 Robinaugh DJ, Millner AJ, McNally RJ. Identifying Highly Influential Nodes in the Complicated Grief Network. J Abnorm Psychol. 2016;125(6):747–57.
42 Letina S, Blanken TF, Deserno MK, Borsboom D. Expanding Network Analysis Tools in Psychological Networks: Minimal Spanning Trees, Participation Coefficients, and Motif Analysis Applied to a Network of 26 Psychological Attributes. Complexity. 2019;2019. DOI: 10.1155/2019/9424605
43 Lorenzo-Luaces L, Rodriguez-Quintana N, Bailey AJ. Double trouble: Do depression severity and duration interact to predicting treatment outcomes in adolescent depression? Behav Res Ther. 2020;103637.
44 Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
45 Choi SW, Schalet BD, Cook KF, Cella D. Establishing a common metric for depressive symptoms: Linking the BDI-II , CES-D, and PHQ-9 to PROMIS Depression. Psychol Assess. 2014;26(2):513–27.
46 Fischer HF, Rose M. www.common-metrics.org: a web application to estimate scores from different patient-reported outcome measures on a common scale. BMC Med Res Methodol. 2016;16(142):1–5.
47 Rothwell PM. Subgroup analysis in randomised controlled trials : importance, indications, and interpretation. Lancet. 2005;365:176–86.
48 Fried EI, Robinaugh DJ. Systems all the way down: Embracing complexity in mental health research. BMC Med. 2020;18(1):4–7.
49 Fava GA, Tomba E, Brakemeier EL, Carrozzino D, Cosci F, Eöry A, et al. Mental Pain as a Transdiagnostic Patient-Reported Outcome Measure. Psychother Psychosom. 2019;88(6):341–9.
50 Fisher AJ, Boswell JF. Enhancing the Personalization of Psychotherapy With Dynamic Assessment and Modeling. Assessment. 2016 Aug;23(4):496–506.
51 Fisher AJ, Reeves JW, Lawyer G, Medaglia JD, Rubel JA. Exploring the idiographic dynamics of mood and anxiety via network analysis. J Abnorm Psychol. 2017;126(8):1044–56.
52 Fisher AJ, Medaglia JD, Jeronimus BF. Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci U S A. 2018;115(27):E6106–15.

Tables and Figures

Table 1. Description of the modelling approaches for the primary outcome

Type of approach	Weighting Approach	Model Number	Method	Predictors Included	Description
Weighted sum-scores	1-step EI (GGM)	1	OLS	CIS-R weighted sum-score for anxiety subscales, BDI-II weighted score, SSS score, LE score, AUDIT-PC score	Sum of all edges connected to the focal node used to weight items to construct weighted sum-scores
	2-step EI (GGM)	2	OLS	CIS-R weighted sum-score for anxiety subscales, BDI-II weighted score, SSS score, LE score, AUDIT-PC score	Sum of all edges connected to either the focal node or any other node directly connected to the focal node
	PC/PR (GGM)	3	OLS	CIS-R weighted sum-score for anxiety subscales, BDI-II weighted score, SSS score, LE score, AUDIT-PC score	the geometric mean between the participation coefficient (PC) and participation ratio (PR)
	CFA	4	OLS	CIS-R weighted sum-score for anxiety subscales, BDI-II weighted score, SSS score, LE score, AUDIT-PC score	Factor loadings from CFA were used as weights to develop the weighted total scores.
Unweighted sum-scores	Shrinkage	5	ENR	CIS-R unweighted sum-score for anxiety subscales, BDI-II score, SSS score, LE score, AUDIT-PC score	ENR, built using the unweighted total scores.
	None	6	OLS	CIS-R unweighted sum-score for anxiety subscales, BDI-II score, SSS score, LE score, AUDIT-PC score	OLS model with unweighted total scores on the baseline measures
Individual symptoms	Shrinkage	7	ENR	CIS-R anxiety subscale items, BDI-II individual items, SSS score, LE score, AUDIT-PC score	ENR model using all of the individual items of BDI-II, Anxiety sub-scores of CIS-R and total scores of other measures
	None	8	OLS	CIS-R anxiety subscale items, BDI-II individual items, SSS score, LE score, AUDIT-PC score	OLS regression model with items assessing the same symptoms included in weighted models.
Null model	None	9	OLS	Mean BDI-II sum-score	Mean BDI-II score in training set studies used as prediction for all cases in test set

Abbreviations: BDI-II – Beck Depression Inventory Second Edition; CFA – Confirmatory Factor Analysis; CIS-R – Revised Clinical Interview Schedule; EI – Expected Influence; ENR – Elastic Net Regularized Regression; GGM – Graphical Gaussian Model; LE – Life Events; OLS – Ordinary Least Squares Regression; PC/PR – Geometric Mean between the Participation Ratio and Participation Coefficient; SSS: Social Support Scale.

Table 2. Descriptive statistics for training and test set samples, and comparison of the two datasets.

		Train Set	Test Set	t-test or χ^2
Self-reported Baseline Characteristics	**Factor**	**N(%) or Mean(SD)**	**N(%) or Mean(SD)**	**p-value**
	Sample Size	*1772*	*1136*	
Age in years	Mean(SD)	42.1(14.0)	43.2(14.3)	.051
Gender	Female	1131(65.7)	769(67.8)	.237
	Male	59(34.3)	365(32.2)	
Ethnicity	White	1613(93.7)	1085(95.6)	.028
	Non-White	109(6.33)	50(4.41)	
Employment status	Employed	996(57.8)	643(56.7)	.002
	Not seeking employment	379(22.0)	306(27.0)	
	Unemployed	347(20.2)	185(16.3)	
Marital Status	Married/cohabiting	819(47.6)	560(49.3)	.608
	Single	560(32.5)	351(30.9)	
	No longer married	343(19.9)	225(19.8)	
Number of recent life events	Mean(SD)	1.39(1.26)	1.28(1.20)	.021
Social Support Total	Median (IQR)	21(18 to 24)	22(18 to 24)	.752
AUDIT-PC score	Mean(SD)	2.57(2.87)	3.13(3.26)	<.001
Past Antidepressant use	No	537(31.2)	371(32.7)	.408
	Yes	1185(68.8)	765(67.3)	
CIS-R Sum of Anxiety Subscales score	Mean(SD)	13.7(6.85)	13.9(6.31)	.437
CIS-R durations	Depression	3.38(1.44)	3.48(1.25)	.056
	Average Anxiety Duration	2.14(1.00)	2.13(0.97)	.780
Baseline BDI-II score	Mean(SD)	29.5(11.1)	31.9(9.45)	<.001
3-4 month BDI-II score	Mean(SD)	14.4(11.4)	17.9(12.4)	<.001
Remission 3-4 months	No	742(51.2)	621(67.7)	<.001
	Yes	708(48.8)	297(32.4)	
Baseline PROMIS score	Mean(SD)	70.3(8.38)	73.3(6.36)	<.001
3-4 month PROMIS score	Mean(SD)	60.1(11.5)	60.4(12.5)	.499

Table 3. Performance of the models predicting BDI-II scores at 3-4 months post-baseline in the test datasets individually and combined.

Type of approach	Model	**All studies combined (n=918)**			**IPCRESS (n=206)**			**MIR (n=424)**			**TREAD (n=288)**			Internal Cross-validation		
		RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE
Weighted Sum-scores	1. EI 1-step	11.285	0.168	9.122	11.642	0.171	9.572	11.216	0.174	8.993	11.127	0.137	8.990	9.995	0.216	7.991
	2. EI 2-Step	11.281	0.169	9.119	11.646	0.170	9.575	11.209	0.175	8.987	11.122	0.137	8.985	9.992	0.216	7.989
	3. PR/PC	11.326	0.162	9.097	11.626	0.173	9.526	11.226	0.177	9.053	11.253	0.117	8.856	9.941	0.223	7.940
	4. CFA	11.338	0.160	9.100	11.655	0.169	9.548	11.219	0.175	9.041	11.284	0.112	8.865	9.953	0.221	7.946
Unweighted Sum-scores	5. ENR	11.311	0.165	9.089	11.638	0.171	9.541	11.232	0.173	9.046	11.189	0.127	8.827	9.946	0.223	7.950
	6. OLS	11.319	0.163	9.091	11.631	0.172	9.544	11.220	0.175	9.045	11.237	0.119	8.836	9.947	0.222	7.944
Individual Symptoms	7. ENR	11.359	0.157	9.173	11.869	0.138	9.798	11.201	0.178	9.075	11.216	0.123	8.871	9.881	0.233	7.886
	8. OLS	11.495	0.137	9.279	12.192	0.090	10.084	11.225	0.174	9.094	11.375	0.098	8.976	9.904	0.230	7.881
Null	9. Null	12.439	-0.010	9.935	12.852	-0.011	10.396	12.544	-0.031	9.993	11.975	0.000	9.521	11.270	-0.001	9.026

Abbreviations: CFA - Confirmatory Factor Analysis; EI - Expected Influence; ENR - Elastic Net Regularized Regression; MAE - Mean Absolute Error; OLS - Ordinary Least Squares; PC - Participation Coefficient; PR - Participation Ratio; RMSE - Root Mean-Squared Error. Note there is no calculation of r^2 for the Null model as all there was no variability in prediction.

Figure 1. Predicted and observed BDI-II score at 3-4 months in combined Test set data (n=918) by the eight models (excluding the null model) built in the Training set data.

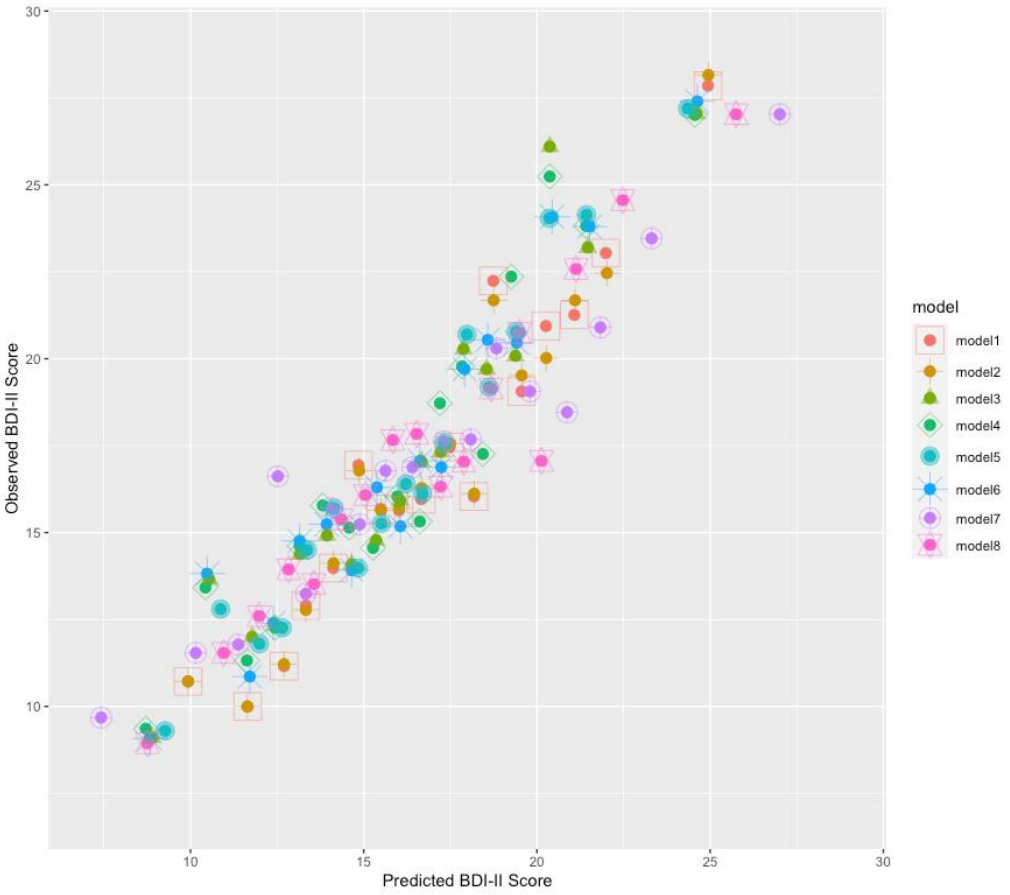
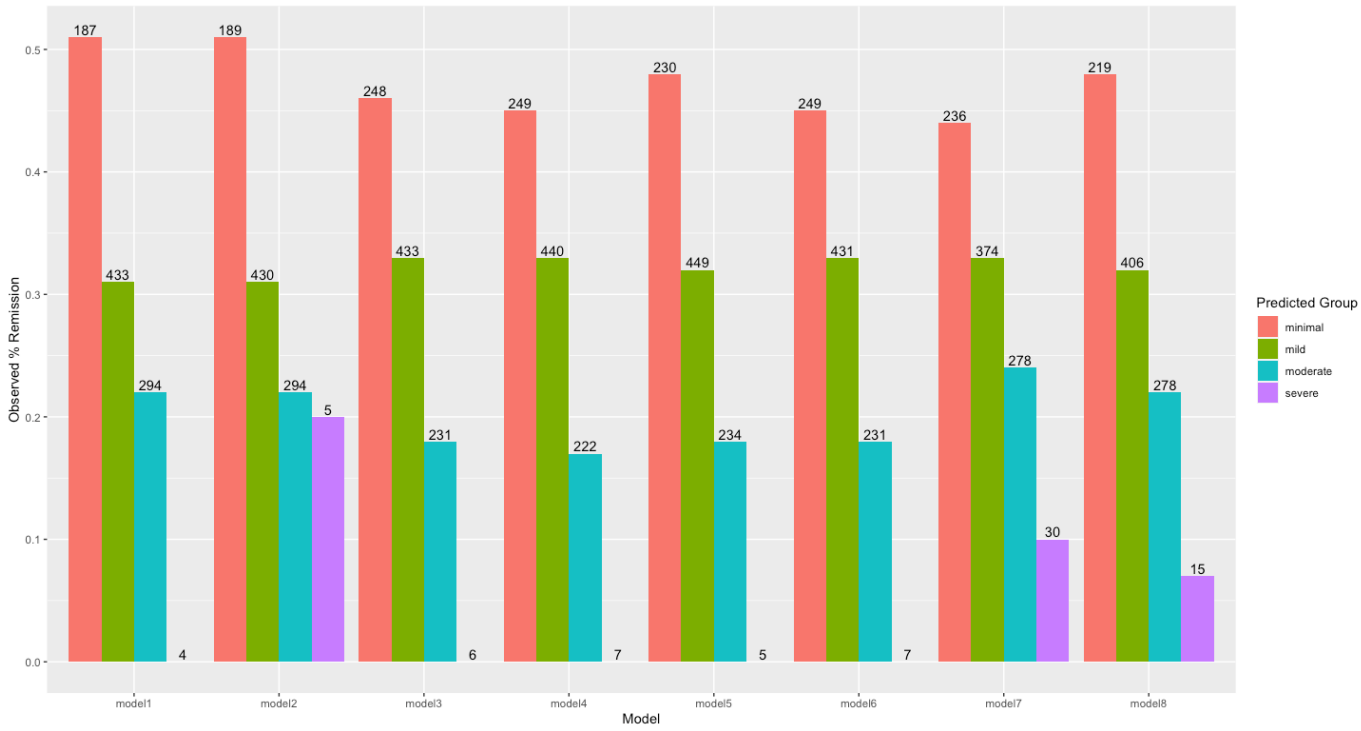


Figure 2. Proportion of participants in remission at 3-4 months post-baseline in the test set studies (n=918) by predicted category of depressive severity at 3-4 months, for each of the eight models.



Supplementary Materials

Further details of data analyses

Pre-processing

Due to overlap between symptoms measured by the CIS-R and BDI-II, pre-processing will identify multicollinearity. Items to be removed will be chosen based on having lower variability in the pair. Additionally, assumptions relating to near zero variance, approximately equal variance of nodes, asymmetrical distributions will be assessed. Items will removed if they clearly violate assumptions across 2 or more training studies.

The pre-processing will apply to the three network informed weighted models and the unidimensional confirmatory factor analytic (CFA) model, any item removed from the network at the pre-processing stage will be weighted to zero.

For the ENR, 'dummy coded' variables will be created for each category of any non-continuous/ordinal variables.

Potential Deviations from Pre-processing Procedure

If we find that the pre-processing stages above related to removing variables due to multi-collinearity or assumption violations for the network modelling result in numerous variables being removed from one set of models (the network models) but not from others, we will consider using less-conservative means of pre-processing so as to not invalidate model comparisons. For example, we might suggest that only those variables that violate network-modelling assumptions in all of the training set studies be removed rather than those that do so in just two of the studies.

Missing Data

Missing data will be imputed using the "missForest" package [63] in R Studio (R Core Team, 2013). This uses a random forest model to impute missing data on all types of variables (continuous, categorical and binary) generating a single dataset with imputed values taken by averaging across a large number of regression trees. The imputation model will be run separately in each of the six RCTs, the results from which will be merged to form the two datasets (train and test). For studies in the train-set, all individuals with ≥30% missingness on baseline variables and all variables with ≥30% missingness across all participants will be excluded. Missing baseline data and outcomes for the remaining cases will be imputed. The primary analyses in the test-set studies will be completers only (those without outcomes will be excluded) although the systematically missing BDI-II scores at 3-4 months post-baseline in COBALT will be interpolated using "missForest" using data at baseline, 6-8 months and the PHQ-9 scores at 3-4 months. As with the training data, cases in the test sample with ≥30% missingness on baseline variables will be excluded and imputation will be performed via random forest. However, outcome data will not be used to inform imputation of missing baseline data for the primary analyses in the test sample.

Consistency Checks and Additional Model Evaluation

In addition to considering the performance of the models using the metrics specified in the 'Model Evaluation' section above the predictions of outcome for each model were compared in a correlation matrix, the weights applied to the predictor variables were also compared.

Software & Packages

Data handling and cleaning prior to the development of the Dep-GP database was performed in Stata 15.0 [65]. All data pre-processing, imputation and analyses for the outlined study will be performed in R (R Core Team, 2019).

The R packages to be used are:

- Bootnet [66]

- Caret [67]
- EGAnet [68]
- EstimateGroupNetwork [39]
- glmnet [69]
- missForest [63]
- mgm [70]
- networktools [71]
- qgraph [72]
- mirt [73]
- lavaan (Rosseel, 2012)

Supplementary Table 1. Description of included studies

Study	Sample and Recruitment	Interventions (N)	Outcome Measure (N for analysis)	Dataset
COBALT [57]	Adults aged 18-75 with treatment resistant depression, scoring ≥14 BDI-II, recruited between November 2008 and September 2010 from 73 general practices in urban and rural settings in three UK centres: Bristol, Exeter, and Glasgow	TAU (n=235) vs CBT+TAU (n=234)	BDI-II also PHQ-9 (n=469)	Train
GENPOD [58]	Adults aged 18-74 with depressive episode, recruited by GPs in three UK centres: Bristol, Birmingham and Newcastle between October 2005 and February 2008.	Citalopram (n=298) vs Reboxetine (n=303)	BDI-II (n=601)	Train
PANDA [59]	Adults presenting with low mood or depression to GP in last 2 years, free of ADM for 8 weeks up to baseline. Recruited between January 2015 and August 2018 from 179 primary care surgeries in four UK cities (Bristol, Liverpool, London, and York)	Sertraline (n=323) vs Placebo (n=329)	PHQ-9 also BDI-II (n=652	Train
TREAD [60]	Adults aged 18-69 who met diagnostic criteria for MDD and scored ≥14 on BDI-II. Recruited from 65 primary care centres in Bristol and Exeter, UK, from August 2007 to October 2009.	TAU (n=179) vs Physical Activity + TAU (n=182)	BDI-II (n=288)	Test
IPCRESS [61]	Adults scoring ≥14 BDI-II and GP confirmed diagnosis of depression. Recruited from 55 general practices in Bristol, London, and Warwickshire, between October 2005 and February 2008	iCBT (n=148) vs TAU (n=147)	BDI-II (n=206)	Test
MIR [62]	Adults ≥18 taking SSRIs or SNRIs at adequate dose for≥ 6 weeks, and scored ≥14 on BDI-II. Recruited from general practices surrounding four centres in Bristol, Exeter, Hull, and Keele/North Staffordshire, UK, between August 2013 and October 2015.	Mirtazapine (n=241) vs Placebo (n=239)	BDI-II also PHQ-9 (n=424)	Test

Abbreviations: ADM – antidepressant medication; BDI-II – Beck Depression Inventory; GP – General Practitioner; iCBT (internet based therapist delivered cognitive behavioural therapy); MDD – Major Depressive Disorder; PHQ-9 – Patient Health Questionnaire 9-item version; SNRI – Serotonin-Norepinephrine Reuptake Inhibitor; SSRI – Selective Serotonin Reuptake Inhibitor; TAU – treatment as usual

Supplementary Table 2. Measures used across the studies of the Dep-GP IPD database

Measure	Details	Scores and Cut-offs for Remission
The CIS-R [35]	Consists of 14 symptom subsections scored 0-4, five of which measure depressive symptoms: core features of depression, depressive thoughts (scored 0-5), fatigue, concentration/forgetfulness, and sleep. Nine sections measure anxiety symptoms: generalized anxiety, worry, irritability, obsessions, compulsions, health anxiety, somatic concerns, phobic anxiety (split into agoraphobia, social phobia, and specific phobia), and panic. A final section measures general health, impairment and weight change. Here only eight anxiety subscales were used, irritability was not used given the similarity between this and the agitation item of the BDI-II.	The total score ranges from 0-57 with a cut-off of ≥12 used to indicate likely common mental disorder, primary and secondary diagnoses using ICD-10 criteria are given as are binary indictors of diagnosis for all the disorders assessed. Scores of <12 among those that were previously depressed can be used to indicate remission.
Beck Depression Inventory 2nd Edition (BDI-II) [37]	Consists of 21 items to assess depressive symptoms, each item is scored 0-3.	There is a maximum score obtainable of 63, and a cut-off of ≥10 is used indicate significant symptoms of depression, scores of <10 are therefore used to indicate remission in those that were previously depressed/scored ≥10.
Patient Health Questionnaire 9-item version (PHQ-9) [74]	This is a depression screening measure, with respondents asked to rate how often they have been bothered by each of the nine symptom items over the preceding two weeks. Each item is scored 0-3	There is a maximum score of 27 with a cut-off of ≥10 is used to indicate "caseness" for depression, a score of 9 or below for those that were previously depressed is therefore considered to indicate remission
Social Support Scale - adapted by authors of RCTs [61] included in this IPD by adding one item to the Health and Lifestyles Survey Social Support Measure [75]	An 8-item instrument (the first seven of which are from the Health and Lifestyles Survey) assessing the degree to which participants rated the social support of their friends and family in each of the following domains: 1) being accepted for who one is; 2) feeling cared about; 3) feeling loved; 4) feeling important to them; 5) being able to rely on them; 6) feeling well supported and encouraged by them; 7) being made to feel happy by them; and 8) feeling able to talk to them whenever one might like. Items are scored 1-3, with total scores ranging from 8-24; higher scores indicate higher levels of perceived social support. The authors of the Health and Lifestyles Survey suggested the maximum score for social support (which was 21 on that scale) indicated 'no lack of social support', scores between 18-20 indicated a 'moderate lack of social support', and scores of 17 or below indicated a 'severe lack of social support'.	N/A
Life events: adapted by the authors of the Adult Psychiatric Morbidity Surveys [76] based on the Social Readjustment Rating Scale [77]	Participants are asked to respond yes/no to whether they have suffered any of eight events within the last six months e.g. a death/bereavement; being physically attacked/injured; or going through a divorce/separation. Each item is scored yes (1) or no (0) and the total score is the sum of all the items.	N/A
Alcohol use: the alcohol use disorder identification test primary care version (AUDIT-PC) [78].	Used to assess alcohol misuse, this includes five items scored 0-4. A cut-off of ≥5 indicates hazardous alcohol use that may be harmful to one's health	N/A

All measures apart from the PHQ-9 were used in all six studies, PHQ-9 was used in three studies (COBALT, MIR, & PANDA), here it was only used for imputation and in the formation of the PROMIS T-Score in sensitivity analyses.

Supplementary Table 3. Performance of the models predicting remission at 3-4 months post-baseline in the test datasets individually and combined.

		All studies combined (n=918)		**IPCRESS (n=206)**		**MIR (n=424)**		**TREAD (n=288)**		Internal Cross-validation	
Type of approach	Model	AUC	Brier	AUC	Brier	AUC	Brier	AUC	Brier	AUC	Brier
Weighted Sum-scores	1. EI 1-step	0.628	0.227	0.670	0.211	0.645	0.228	0.591	0.237	0.731	0.208
	2. EI 2-Step	0.628	0.227	0.669	0.211	0.645	0.228	0.591	0.237	0.731	0.208
	3. PR/PC	0.633	0.234	0.681	0.209	0.649	0.230	0.593	0.259	0.737	0.206
	4. CFA	0.631	0.237	0.672	0.211	0.652	0.229	0.588	0.267	0.735	0.207
Unweighted Sum-scores	5. ENR	0.626	0.234	0.653	0.235	0.648	0.232	0.592	0.237	0.724	0.243
	6. Logistic Regression	0.632	0.236	0.675	0.210	0.647	0.229	0.593	0.264	0.737	0.206
Individual Symptoms	7. ENR	0.618	0.233	0.668	0.233	0.625	0.232	0.590	0.236	0.716	0.242
	8. OLS	0.599	0.246	0.642	0.219	0.608	0.236	0.585	0.278	0.738	0.207
Null	9. Null	N/A	0.237	N/A	0.239	N/A	0.235	N/A	0.239	N/A	0.237

Abbreviations: AUC – Area Under the receiver operating characteristic Curve; CFA - Confirmatory Factor Analysis; EI - Expected Influence; ENR - Elastic Net Regularized Regression; OLS - Ordinary Least Squares; PC - Participation Coefficient; PR -Participation Ratio;

Supplementary Table 4. Performance of the models predicting PROMIS T-score scores at 3-4 months post-baseline in the test datasets individually and combined.

		Test set (n with complete data at 3-4 months post-baseline)														
		All studies combined (n=918)			**IPCRESS (n=206)**			**MIR (n=424)**			**TREAD (n=288)**			Internal Cross-validation		
Type of approach	**Model**	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE
Weighted Sum-scores	1. EI 1-step	11.855	0.103	9.332	13.624	0.039	10.644	11.077	0.143	8.920	11.514	0.081	8.940	10.169	0.152	7.999
	2. EI 2-Step	11.853	0.104	9.329	13.628	0.039	10.647	11.072	0.143	8.914	11.509	0.081	8.937	10.167	0.152	7.998
	3. Geometric-mean PR/PC	11.843	0.105	9.447	13.575	0.046	10.593	11.081	0.142	8.999	11.563	0.081	9.327	10.117	0.161	7.961
	4. CFA	11.863	0.102	9.477	13.613	0.041	10.639	11.067	0.144	8.978	11.563	0.073	9.327	10.122	0.160	7.978
Unweighted Sum-scores	5. ENR	11.859	0.103	9.473	13.599	0.043	10.635	11.085	0.141	8.991	11.539	0.077	9.299	10.114	0.161	7.961
	6. OLS	11.851	0.104	9.462	13.599	0.043	10.625	11.076	0.143	8.980	11.524	0.079	9.287	10.115	0.161	7.961
Individual Symptoms	7. ENR	12.082	0.069	9.538	14.379	-0.070	10.949	11.128	0.135	8.939	11.503	0.082	9.348	10.025	0.176	7.869
	8. OLS	12.232	0.046	9.643	14.706	-0.120	11.107	11.185	0.126	8.969	11.622	0.063	9.525	10.035	0.174	7.867
Null	9. Null	12.522	0.000	10.007	13.995	-0.014	11.360	12.022	0.000	9.719	12.045	-0.006	9.399	11.045	0.000	8.775

Abbreviations: CFA - Confirmatory Factor Analysis; EI - Expected Influence; ENR - Elastic Net Regularized Regression; MAE - Mean Absolute Error; OLS - Ordinary Least Squares; PC - Participation Coefficient; PR -Participation Ratio; RMSE - Root Mean-Squared Error. Note there is no calculation of r^2 for the Null model as all there was no variability in prediction.

Supplementary Table 5. Performance of the models predicting BDI-II scores at 3-4 months post-baseline in the test datasets individually and combined. Models were only developed in two training set studies, excluding COBALT.

		Test set (n with complete data at 3-4 months post-baseline)													
		All studies combined (n=918)			**IPCRESS (n=206)**			**MIR (n=424)**			**TREAD (n=288)**			Internal Cross-validation		
Type of approach	**Model**	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE
Weighted Sum-scores	1. EI 1-step	11.441	0.145	9.173	11.603	0.176	9.424	11.578	0.121	9.241	11.115	0.138	8.895	9.906	0.191	7.869
	2. EI 2-Step	11.435	0.146	9.170	11.607	0.176	9.427	11.569	0.123	9.234	11.111	0.139	8.891	9.903	0.191	7.868
	3. Geometric-mean PR/PC	11.550	0.129	9.202	11.615	0.174	9.442	11.542	0.127	9.270	11.515	0.075	8.930	9.848	0.200	7.832
	4. CFA	11.566	0.126	9.217	11.604	0.176	9.404	11.619	0.115	9.313	11.458	0.084	8.944	9.848	0.200	7.817
Unweighted Sum-scores	5. ENR	11.595	0.126	9.225	11.625	0.173	9.424	11.613	0.116	9.320	11.462	0.084	8.944	9.849	0.200	7.821
	6. OLS	11.540	0.130	9.207	11.587	0.178	9.392	11.596	0.119	9.307	11.423	0.090	8.926	9.850	0.200	7.818
Individual Symptoms	7. ENR	11.552	0.129	9.262	11.896	0.134	9.782	11.463	0.139	9.229	11.431	0.089	8.939	9.865	0.198	7.836
	8. OLS	11.683	0.109	9.373	12.186	0.091	10.047	11.489	0.135	9.256	11.597	0.062	9.064	9.899	0.192	7.834
Null	9. Null	12.664	-0.047	9.918	13.076	-0.046	10.414	12.871	-0.086	10.081	12.044	-0.012	9.322	11.021	-0.001	8.740

Abbreviations: CFA - Confirmatory Factor Analysis; EI - Expected Influence; ENR - Elastic Net Regularized Regression; MAE - Mean Absolute Error; OLS - Ordinary Least Squares; PC - Participation Coefficient; PR -Participation Ratio; RMSE - Root Mean-Squared Error. Note there is no calculation of r^2 for the Null model as all there was no variability in prediction.

Supplementary Table 6. Item weights from the three ways of determining item centrality from the GGM network and factor loadings from the CFA model.

Variable	Model			
	EI 1-Step (1)	EI 2-Step (2)	PC_PR (3)	CFA (4)
Anxiety (cisr)	0.773	0.757	0.724	0.165
Compulsions (cisr)	0.638	0.626	0.716	0.172
Health anxiety (cisr)	0.656	0.620	0.812	0.173
Obsessions (cisr)	0.547	0.543	0.598	0.181
Panic (cisr)	0.895	0.874	0.891	0.188
Phobia (cisr)	0.810	0.803	0.797	0.168
Somatic (cisr)	0.439	0.437	0.602	0.186
Worry (cisr)	0.719	0.704	0.696	0.176
Sadness (BDI)	0.955	0.943	0.978	0.182
Pessimism (BDI)	0.736	0.792	0.801	0.181
Failure (BDI)	0.814	0.872	0.718	0.172
Loss of pleasure (BDI)	0.785	0.834	0.844	0.180
Guilt (BDI)	0.861	0.881	0.851	0.194
Punishment (BDI)	0.813	0.825	0.860	0.176
Self dislike (BDI)	0.696	0.732	0.515	0.182
Self criticism (BDI)	0.853	0.866	0.763	0.156
Suicidal thoughts (BDI)	0.741	0.750	0.811	0.173
Crying (BDI)	0.578	0.586	0.733	0.163
Agitation (BDI)	0.698	0.686	0.709	0.184
Loss of Interest (BDI)	0.800	0.820	0.857	0.184
Indecisiveness (BDI)	0.736	0.765	0.828	0.186
Worthlessness (BDI)	1.000	1.000	0.736	0.169
Loss of energy (BDI)	0.900	0.916	0.888	0.156
Sleep (BDI)	0.492	0.486	0.547	0.177
Irritability (BDI)	0.825	0.817	0.830	0.177
Concentration (BDI)	0.912	0.921	0.823	0.170
Fatigue (BDI)	0.837	0.853	0.665	0.164
Libido (BDI)	0.452	0.451	0.510	0.000
Social Support	0.000	0.000	0.437	0.157
Recent life events	0.469	0.447	0.508	0.106
Alcohol use (AUDIT)	0.300	0.294	0.000	0.251

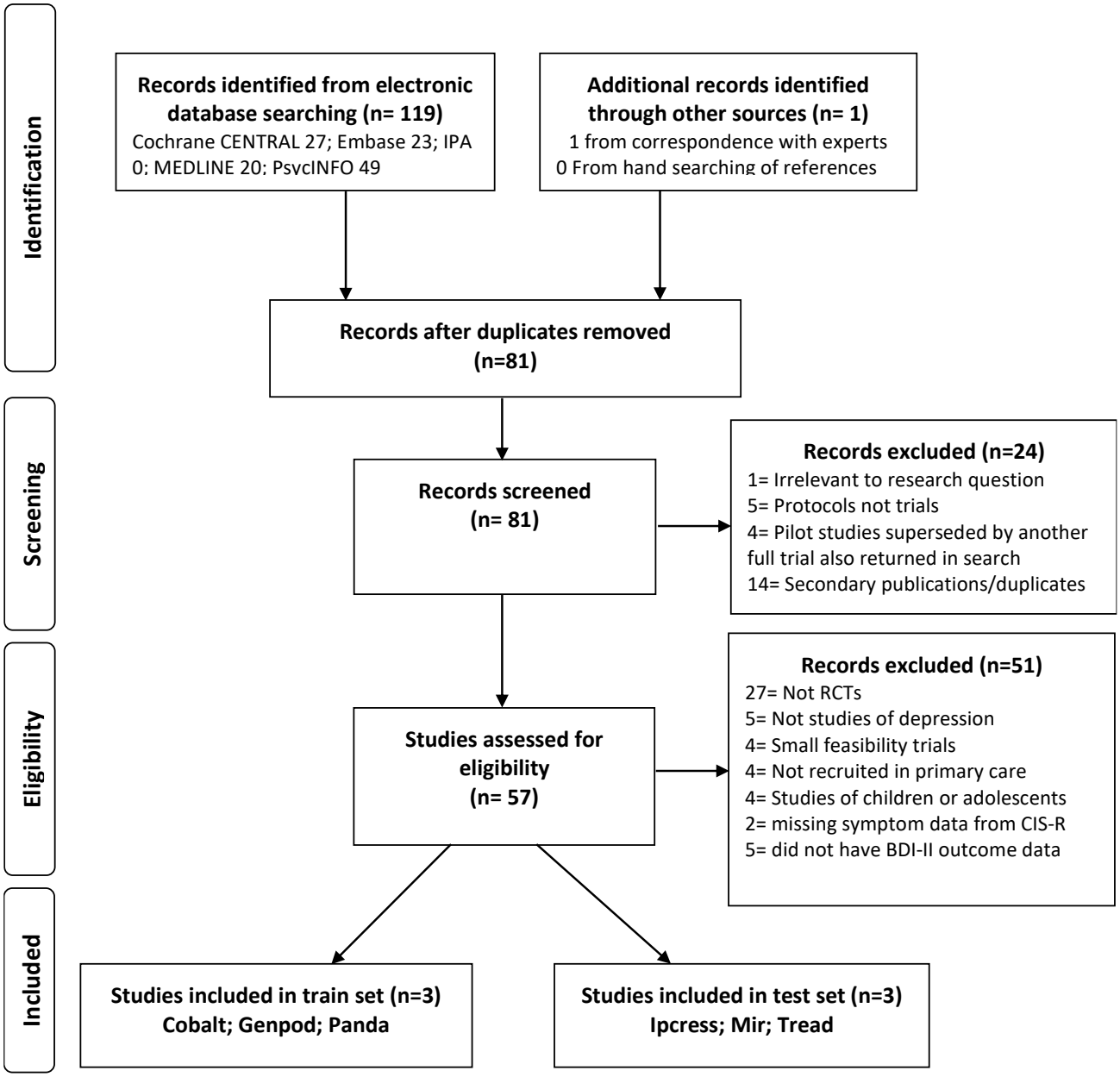
Supplementary Table 7. Coefficients and Intercepts from Models 1-6 for BDI-II score at 3-4 month outcome.

Model	Intercept	Coefficients				
		CIS-R Anxiety Sum-score	BDI-II Sum-score	Social Support Sum-score	Life Events Sum-score	AUDIT-PC Sum-score
Model 1 (EI 1-Step)	2.883	0.300	0.501	.	0.200	-0.072
Model 2 (EI 2-Step)	2.858	0.304	0.494	.	0.215	-0.074
Model 3 (PC/PR)	7.251	0.324	0.454	-0.642	0.038	.
Model 4 (CFA)	7.293	1.323	2.092	-1.796	0.179	-0.143
Model 5 (ENR)	16.641	1.329	3.867	-0.931	.	.
Model 6 (OLS)	7.010	0.230	0.355	-0.280	0.034	-0.018

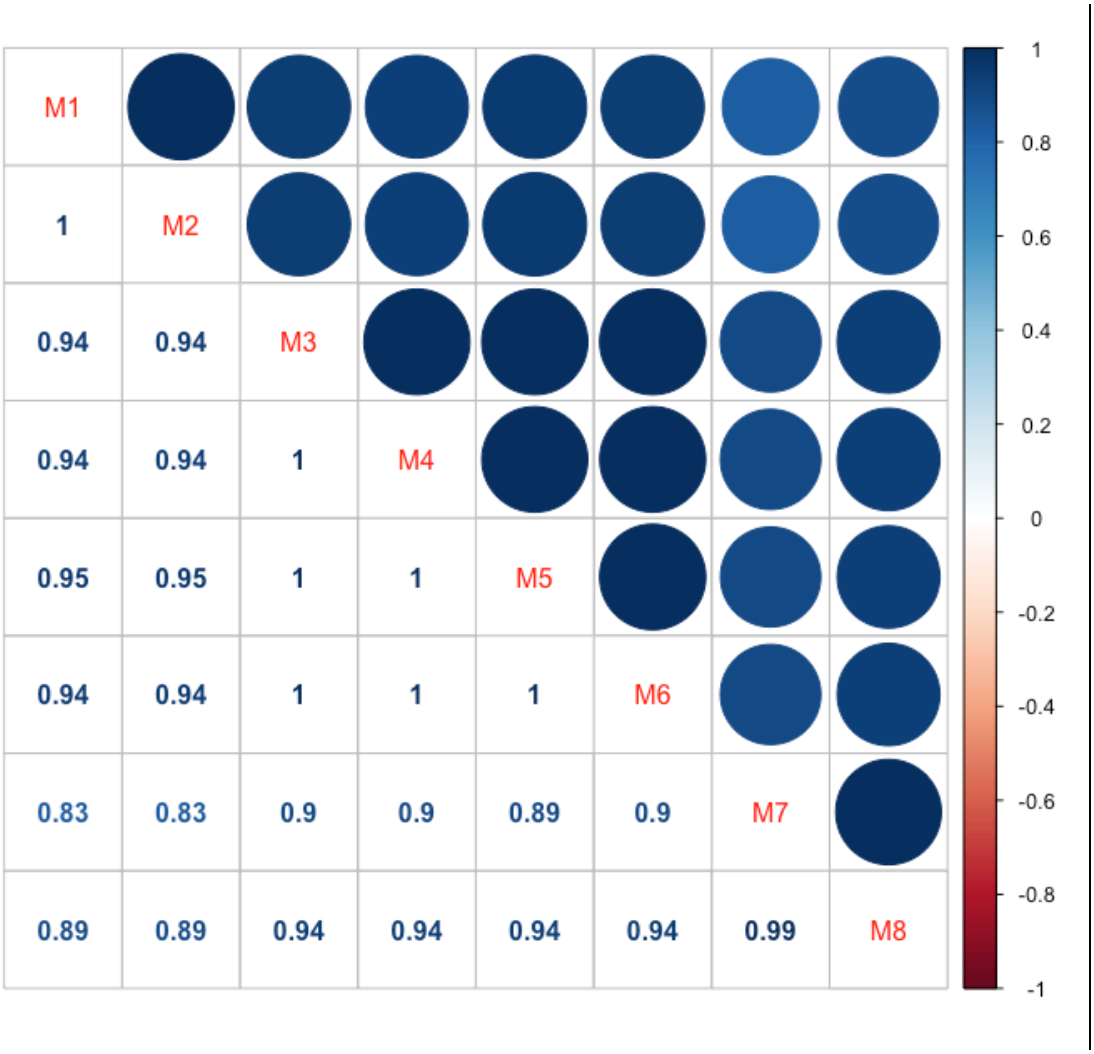
Supplementary Table 8. Coefficient weights and intercepts from individual item models for BDI-II score at 3-4 months.

Item	Model	
	Model 7 (ENR)	**Model 8 (OLS)**
Intercept	16.641	6.714
Anxiety (cisr)	.	0.015
Compulsions (cisr)	0.228	0.203
Health anxiety (cisr)	0.768	0.747
Obsessions (cisr)	-0.077	-0.169
Panic (cisr)	0.330	0.273
Phobia (cisr)	0.604	0.559
Somatic (cisr)	0.582	0.501
Worry (cisr)	.	-0.061
Sadness (BDI)	-0.033	-0.585
Pessimism (BDI)	0.439	0.539
Failure (BDI)	0.412	0.479
Loss of pleasure (BDI)	0.539	0.856
Guilt (BDI)	0.555	0.823
Punishment (BDI)	0.108	0.071
Self dislike (BDI)	0.437	0.569
Self criticism (BDI)	0.086	0.018
Suicidal thoughts (BDI)	0.944	1.933
Crying (BDI)	0.383	0.481
Agitation (BDI)	0.273	0.582
Loss of Interest (BDI)	0.208	0.189
Indecisiveness (BDI)	0.336	0.395
Worthlessness (BDI)	0.265	0.230
Loss of energy (BDI)	0.788	1.293
Sleep (BDI)	.	-0.096
Irritability (BDI)	-0.278	-0.810
Appetite (BDI)	0.069	0.075
Concentration (BDI)	0.316	0.486
Fatigue (BDI)	0.407	0.569
Libido (BDI)	0.222	0.236
Social Support	-0.805	-0.256
Recent life events	0.019	0.052
Alcohol use (AUDIT)	.	0.002

Supplementary Figure 1. Flow diagram of study selection.



Supplementary Figure 2. Correlation of predictions by the six models in the Test set data.



Supplementary Figure 3. Proportions of participants in remission at 3-4 months post-baseline in the test set (n=918) based on predicted 3-4 month BDI-II scores by each of the eight models.