Journal of Psychiatric Research 78 (2016) 94–102

Contents lists available at ScienceDirect

Journal of Psychiatric Research

journal homepage: www.elsevier.com/locate/psychires

Combining clinical variables to optimize prediction of antidepressant treatment outcomes

Raquel Iniesta [a, *], Karim Malki [a], Wolfgang Maier [b], Marcella Rietschel [c], Ole Mors [d], Joanna Hauser [e], Neven Henigsberg [f], Mojca Zvezdana Dernovsek [g], Daniel Souery [h], Daniel Stahl [i], Richard Dobson [a], Katherine J. Aitchison [a, j], Anne Farmer [a], Cathryn M. Lewis [a], Peter McGuffin [a], Rudolf Uher [a, k]

[a] Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, 16 De Crespigny Park, Denmark Hill, London, SE5 8AF, UK
[b] Department of Psychiatry, University of Bonn, Regina-Pacis-Weg 3, 53113 Bonn, Germany
[c] Central Institute of Mental Health, Division of Genetic Epidemiology in Psychiatry, Square J5, 68159, Mannheim, Germany
[d] Research Department P, Aarhus University Hospital, Norrebrogade 44, Aarhus C, DK-8000, Risskov, Denmark
[e] Laboratory of Psychiatric Genetics, Department of Psychiatry, Poznan University of Medical Sciences, Collegium Maius, Fredry 10, 61-701, Poznań, Poland
[f] Croatian Institute for Brain Research, Medical School, University of Zagreb, Salata 3, 10 000, Zagreb, Croatia
[g] University Psychiatric Clinic and the Medical Faculty, University of Ljubljana, Kongresni trg 12, 1000, Ljubljana, Slovenia
[h] Laboratoire de Psychologie Médicale, Université Libre de Bruxelles and Psy Pluriel – Centre Européen de Psychologie Médicale, Av Jack Pastur 47a, 1180, Uccle, Belgium
[i] Institute of Psychiatry, Psychology and Neuroscience, Kings College London, 16 De Crespigny Park, London, SE5 8AF, UK
[j] Department of Psychiatry and Medical Genetics, University of Alberta, 116 St and 85 Ave, Edmonton, AB, T6G 2R3, Canada
[k] Dalhousie University Department of Psychiatry, 5909 Veterans' Memorial Drive, Halifax, B3H 2E2, Nova Scotia, Canada

ARTICLE INFO

Article history:
Received 22 October 2015
Received in revised form
12 March 2016
Accepted 30 March 2016

Keywords:
Depression
Outcome
Antidepressant
Prediction
Machine learning
Statistical learning

ABSTRACT

The outcome of treatment with antidepressants varies markedly across people with the same diagnosis. A clinically significant prediction of outcomes could spare the frustration of trial and error approach and improve the outcomes of major depressive disorder through individualized treatment selection. It is likely that a combination of multiple predictors is needed to achieve such prediction. We used elastic net regularized regression to optimize prediction of symptom improvement and remission during treatment with escitalopram or nortriptyline and to identify contributing predictors from a range of demographic and clinical variables in 793 adults with major depressive disorder. A combination of demographic and clinical variables, with strong contributions from symptoms of depressed mood, reduced interest, decreased activity, indecisiveness, pessimism and anxiety significantly predicted treatment outcomes, explaining 5–10% of variance in symptom improvement with escitalopram. Similar combinations of variables predicted remission with area under the curve 0.72, explaining approximately 15% of variance (pseudo R^2) in who achieves remission, with strong contributions from body mass index, appetite, interest-activity symptom dimension and anxious-somatizing depression subtype. Escitalopram-specific outcome prediction was more accurate than generic outcome prediction, and reached effect sizes that were near or above a previously established benchmark for clinical significance. Outcome prediction on the nortriptyline arm did not significantly differ from chance. These results suggest that easily obtained demographic and clinical variables can predict therapeutic response to escitalopram with clinically meaningful accuracy, suggesting a potential for individualized prescription of this antidepressant drug.

1. Introduction

Major depressive disorder is a common condition, responsible for a substantial proportion of disability world-wide (Whiteford et al., 2013). Although a number of pharmacological and

* Corresponding author.
E-mail address: raquel.iniesta@kcl.ac.uk (R. Iniesta).

http://dx.doi.org/10.1016/j.jpsychires.2016.03.016

psychological treatment options are available, the outcomes are unsatisfactory. While some individuals experience dramatic improvements, most do not benefit sufficiently from the first treatment and have to undergo multiple treatment trials. Each trial takes weeks, with delays causing frustration, prolonging disability and risking adverse outcomes, including suicide. The unsatisfactory state of depression therapeutics has led to the consensus that diagnosis of depression may not be sufficient for treatment selection and additional information needs to be considered to estimate which treatment is likely to work for whom (Kupfer et al., 2012).

There is little evidence to guide clinicians in selecting a treatment for a given individual (Simon and Perlis, 2010). A single piece of information is unlikely to predict treatment outcome with an accuracy that is meaningful in clinical practice. Therefore, multiple factors may have to be considered to make the best prediction of outcomes at the individual level. The need for prediction at individual level has prompted the use of new methods, such as machine learning and statistical learning (Hastie et al., 2009). Unlike traditional statistics that focus on testing whether a single variable makes a statistically significant contribution, learning methods consider all available information across a number of variables to make the best prediction for an individual. The accuracy of prediction can then be compared to a standard benchmark to evaluate whether it is likely to be clinically significant (Uher et al., 2012d), i.e. whether it makes a meaningful difference to a particular individual.

Individualized treatment selection could be useful if it is based on predictors that are easily obtained (e.g. questionnaires and rating scales) and if it can differentially predict outcomes with alternative treatments. Two prior studies suggest that meaningful prediction of treatment outcomes from easy-to-obtain variables is achievable. A study of the STAR*D cohort found that 48 demographic and clinical variables robustly predicted treatment success with a clinically significant effect size (area under the curve 0.71, 11.4% variance explained) (Perlis, 2013). The prediction was robust in stringent validation test. A second study found that the relative benefits of cognitive-behavioural therapy and antidepressant medication can be predicted from eight demographic and clinical variables in a way that makes a meaningful difference in outcomes for 60% of 154 participants (DeRubeis et al., 2014). While both studies show promising results, they also leave caveats. The STAR*D study predicted overall outcome rather than outcomes of specific treatments. The strongest predictor was race, raising questions about how the findings generalize to populations with different ethnic composition. The study of cognitive-behavioural therapy and antidepressants established differential prediction, but due to a limited sample size, it had to derive a small number of predictors based on results obtained in the same sample and relied on a less stringent leave-one-out cross-validation.

Therefore, in the present study we evaluate to what extend can demographic and clinical variables predict outcomes with specific treatments at the level of individual. We have applied statistical learning to a study comparing treatment with two different antidepressants in an ethnically homogeneous sample large enough to allow robust 10-fold-split-sample cross-validation and permutations (Kohavi, 1995; Perez-Guaita et al., 2015).

2. Materials and methods

2.1. Study design and sample

The Genome-based Therapeutic Drugs for Depression (GENDEP) is a 12-week comparative study that aims to personalize treatment choice in major depressive disorder using clinical and genetic predictors of response to a serotonin-reuptake-inhibiting antidepressant escitalopram and a norepinephrine-reuptake-inhibiting antidepressant nortriptyline (Uher et al., 2009a, 2010). GENDEP included 868 treatment-seeking adults of White-European ethnicity from nine centers, diagnosed with ICD-10/DSM-IV major depressive disorder and a current depressive episode of at least moderate severity established with the Schedules for Clinical Assessment in Neuropsychiatry (SCAN) interview (Wing et al., 1990). Exclusion criteria were personal or family history of bipolar disorder or schizophrenia and active substance dependence. Eligible patients with no contraindications were randomly allocated to receive treatment with one of the two antidepressants for 12 weeks. Escitalopram is a selective serotonin reuptake inhibitor (SSRI) and has no effect on norepinephrine reuptake. Nortriptyline is a second-generation tricyclic antidepressant (TCA) with a much higher affinity for the norepinephrine transporter than for the serotonin transporter. A protocol guided treatment with escitalopram 10–30 mg daily and nortriptyline 50–150 mg daily, adjusted according to therapeutic effect and tolerability (Uher et al., 2009a). Participants with contraindications or history of intolerance of one of the drugs were offered treatment with the other drug non-randomly (Uher et al., 2009a). Seventy-six percent of GENDEP participants remained on the allocated antidepressant for 8 weeks or longer. In the present study, we include 793 participants (328 on nortriptyline and 465 on escitalopram), who had four or more depression severity measurements, a minimum needed to establish at least an initial trend in clinical response. Since participants non-randomly allocated to escitalopram and nortriptyline differed on some clinical characteristics (Uher et al., 2009a), we also repeated analyses restricting the sample to randomly allocated participants ($n = 450$) to provide drug-specific estimates in comparable samples. The ethics boards of all centers approved the protocol and all participants signed an informed consent.

2.2. Outcomes

The clinician-rated Montgomery–Åsberg Depression Rating Scale (MADRS) (Montgomery and Asberg, 1979), the 17-item Hamilton Rating Scale for Depression (HRSD) (Hamilton, 1967) and the self-report Beck Depression Inventory (BDI) (Beck et al., 1961) were administered at baseline and then weekly for 12 weeks with high inter-rater reliability (Uher et al., 2008, 2012c). Following a consensus reached in a meta-analysis (Investigators et al., 2013), we considered one primary continuous outcome and one primary categorical outcome. The primary continuous outcome was the percentage of improvement in MADRS score (the primary GENDEP outcome measure) over the twelve weeks, based on week twelve measurement if available and on the mixed effects model best unbiased linear estimate from earlier measurements if the week twelve measurement was missing, adjusted for center of recruitment, age and sex (Uher et al., 2010). On average, GENDEP participants improved by 56.2%, from a mean initial MADRS score of 29.0 to a mean end-of-treatment MADRS score of 12.7 (Uher et al., 2009a). The primary categorical outcome was remission, defined as a HRSD score of 7 or less on the last available measurement without imputation (we have selected the HRSD since this is the most established definition of remission; there is less agreement about which cut-off on the MADRS should be used as a threshold for defining remission). Secondary continuous and categorical outcomes included completion of an adequate treatment trial (six weeks or more on allocated antidepressant) and treatment resistant depression (TRD; lack of response to two adequate antidepressant treatment trials, including the GENDEP treatment and previous treatment trials). Of the analyzed sample, 326 (41.1%) participants achieved remission on HRSD-17, 710 (89.5%) completed an adequate treatment trial and 105 (13.3%) had TRD.

2.3. Predictors

We combined multiple predictors that were previously tested one at a time (Supplementary Table S1) (Keers et al., 2010; Uher et al., 2009b, 2011, 2012b, 2009a, 2012c). Demographics data included current age, age at onset of depression, sex, smoking status, body mass index (BMI), occupation, marital status, years of education and number of children (Uher et al., 2009b). We included items and total scores on baseline severity measures (MADRS, HRSD, BDI) (Uher et al., 2012c), individual depressive symptoms from SCAN and depression subtypes (atypical, melancholic, anxious) (Uher et al., 2011). We included observed mood, cognitive and neurovegetative symptom factors and six dimensions (mood, anxiety, pessimism, interest-activity, sleep, appetite) from a published factor analysis (Uher et al., 2008; 2012b). We further included stressful life events (SLEs) experienced during the six months prior to the baseline assessment (Keers et al., 2010), measured with the List of Threatening Experiences Questionnaire (LTE-Q) (Brugha et al., 1985). Medication history included the use of antidepressant at the time of recruitment, any prior antidepressant treatment, number and types of antidepressants tried (SSRI, tricyclic, dual, monoamine oxidase inhibitor, other) established with Medication History Form (MHF) (Uher et al., 2012a). Four combinations of variables were tested:

Combination I: Demographic data and baseline severity (60 predictors).
Combination II: Combination I plus depression subtypes, symptoms, and dimensions (107 predictors).
Combination III: Combination II plus stressful life events (121 predictors).
Combination IV: Combination II plus medication history (125 predictors).

The combinations allowed us comparing the addition of predictors to a minimal set of variables just including demographic and baseline severity, as well as quantifying the improvement in prediction when adding stressful life events and medication history to a set of known strong predictors such as depression symptoms and dimensions (Uher et al., 2008; 2012b). Baseline characteristics of the whole sample and the drug subgroups are shown in Supplementary Tables S2 and S3.

2.4. Statistical learning

Machine and statistical learning methods are designed to provide the best estimate of an outcome in a given individual from multiple predictors. These methods use a 'learning' dataset to learn from relationships between predictors and outcomes to derive the best model for predicting an outcome. The prediction model is then tested in a separate group of individuals, 'testing' dataset, to establish how well it predicts an unknown outcome in an independent sample. Machine learning methods are power tools for prediction at individual level (Hastie et al., 2009), but have been criticized as a "black box" that leaves the mechanism of the prediction unexplained (Wall et al., 2003). Statistical learning is a set of tools that integrates machine learning with additional statistical methods for understanding complex data sets. We use the statistical learning method of elastic net regularized regression (ENRR) (Zou and Hastie, 2005) to predict continuous and categorical outcomes and map the contribution of predictors. 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 an efficient internal method of search and selection of predictors from a large set of variables. Elastic net allows to estimate the combined predictive ability of a high number of variables whilst preventing the models from overfitting. Final model coefficients are interpreted as in a usual regression output and allow ranking of predictors by the magnitude of their contribution to predicting the outcome. We used CARET (Kuhn, 2008) and GLMNET (Friedman et al., 2010) R packages to implement a series of linear and logistic ENRR and select predictors leading to optimized final model for each outcome. Following recommendations for optimized balance between variance explained and minimum bias, we tested the model for each combination across a range of model parameters in 10-fold cross-validation with resampling (Kohavi, 1995). In each step, we split data randomly into 10 subsets, use 9 subsets as the training dataset and the remaining subset as the testing dataset, so that we use each part of the dataset once for testing. To minimize variation across testing datasets, we repeated the 10-fold cross-validation 100 times with independent random dataset partitions, a procedure that optimizes the stability of results (Kim, 2009). We tested drug-specificity by comparing same-drug prediction (training and testing datasets treated with the same drug) with a cross-drug analysis (training and testing datasets treated with a different drug). For continuous outcomes, we indexed the accuracy of prediction with the coefficient of determination R^2, which quantifies the proportion of variance in outcome explained by the predictive model, averaged across the 100 repeats of 10-fold-cross-validation. Based on a consensus for clinical significance, a benchmark was established that a prediction with an R^2 of 6.3 or greater is likely to make a meaningful difference in clinical setting (Uher et al., 2012d). For categorical outcomes, we indexed the accuracy of prediction as area under the receiver operating curve (AUC), which provides a summary measure for sensibility and specificity, averaged across the 100 repeats of 10-fold-cross-validation. To quantify categorical outcome prediction on a scale comparable with continuous outcomes, we also derived a log-likelihood pseudo R^2, which is robust to differences in base rate.

2.5. Cross-validation of the models

We used permutation testing to establish the statistical significance of the predictions (Perez-Guaita et al., 2015). A distribution of 10-fold-cross-validation R^2 values was computed in 1000 samples with the outcomes randomly permuted. Then the distribution was considered as a null distribution for the R^2, and used to derive the p values for the R^2 estimators obtained using real outcome. As the analysis was computationally intensive, we focused the calculation on the models that best predicted the primary outcomes in our sample. The use of compute engine from Google cloud platform, a high performance cloud computing resource, allowed us to complete such a high computing demand process.

3. Results

3.1. Predicting reduction in depressive symptoms

A model including 29 of the 60 predictors from combination I explained a 3.85% of the variance in MADRS scores change across treatment arms. The most relevant variables in the prediction were the baseline MADRS items apparent sadness and inability to feel, HRSD item psychic anxiety, and BDI items pessimism and indecisiveness (Table 1).

The accuracy of prediction was higher in the drug-specific analyses. In the escitalopram-treated group, 24 out of 121 variables in combination III explained 6.32% of variance in MADRS outcome, with BDI item indecisiveness, SCAN hopelessness and preoccupation with death, HRSD items work and interests and depressed

Table 1
Predicting reduction in depressive symptoms.

All	Both antidepressants		Escitalopram		Nortriptyline	
Sample size	n = 793		n = 465		n = 328	
	R^2		R^2		R^2	
Combination I	**0.038**		0.055		0.045	
Combination II	0.034		0.059		0.041	
Combination III	0.035		**0.064**		0.046	
Combination IV	0.026		0.061		**0.053**	
Best model:	I		III		IV	
Cross-drug prediction			R^2		R^2	
Escitalopram			–		0.004	
Nortriptyline			0.001		–	
No. predictors retained:	29		24		122	
Strongest effect size:	Predictor	beta	Predictor	beta	Predictor	beta
1.	Apparent Sadness (MADRS)	2.14	Indecisiveness (BDI)	−2.67	Apparent Sadness (MADRS)	5.49
2.	Inability to feel (MADRS)	1.79	Hopelessness (SCAN)	−2.19	Fatiguability (SCAN)	5.01
3.	Anxiety, psychic (HRSD)	1.62	Work and interests (HRSD)	−2.03	Inefficient thinking (SCAN)	−4.50
4.	Pessimism (BDI)	−1.57	Preoccupation with death (SCAN)	1.91	Depressed mood (HRSD)	−4.18
5.	Indecisiveness (BDI)	−1.57	Depressed mood (HRSD)	1.79	Anhedonia (SCAN)	−4.12
6.	Work and interests (HRSD)	−1.49	Problems with close people (LTE-Q)	1.68	Loss of interest (SCAN)	3.92
7.	Insomnia: initial (HRSD)	−1.32	Fatiguability (SCAN)	−1.57	Dual-action antidepressants (MHF)	3.73
8.	Psychomotor retardation (HRSD)	−1.28	Phobia (SCAN)	−1.39	SSRI antidepressants (MHF)	3.68
9.	Worthlessness (BDI)	−1.03	Early waking (SCAN)	−1.30	Neurovegetative (Factor)	−3.54
10.	Hypochondriasis (HRSD)	−1.00	Anxiety, psychic (HRSD)	1.08	Melancholic depression (SCAN)	3.27

Highest accuracy (R^2) across models trained in every sample is marked in bold.
R^2 = coefficient of determination (proportion of variance explained); beta = standardized regression coefficient (= a measure of effect size).

mood, and LTE-Q problems with close people contributing most strongly to the prediction (Table 1). This prediction was largely escitalopram-specific; cross-drug prediction by models derived from the escitalopram-treated training dataset explained only 0.14% of outcome variance in the nortriptyline-treated group. In the nortriptyline-treated group, a model including all variables in combination IV explained 5.32% of variance in outcome, with strongest contributions from MADRS item apparent sadness, SCAN fatigability, inefficient thinking, and anhedonia and HRSD depressed mood (Table 1). This model explained only 0.42% of outcome variance in the escitalopram-treated group.

When we restricted the analysis to randomly allocated patients, 48 predictors explained up to 5.13% (p value 0.03) of variance in MADRS change overall, 14 predictors explained 10.25% (p value 0.016) among escitalopram-treated participants and 29 predictors explained 6.49% (p value 0.235) among nortriptyline-treated participants. (Table 2, Figs. 1 and 2). Predictions remained largely drug specific, with cross-drug predictions explaining only 1.95% and 1.04% of outcome variance (Fig. 1).

3.2. Predicting remission

Variables selected from combinations II, III and IV predicted HRSD remission with an AUC of 0.72 and a pseudo R^2 around 0.16 across treatment arms. Among the 41 selected variables from combination IV, symptom dimensions appetite and interest-activity, SCAN item phobia, BMI and age contributed most strongly to predicting remission with an AUC 0.72, sensitivity 0.66, specificity 0.66 and a.

Pseudo R^2 of 0.15 in the entire sample (Table 3).

Among escitalopram-treated participants, 46 variables from Combination IV predicted remission with an AUC 0.72, sensitivity 0.65, specificity 0.69 and pseudo R^2 0.18, with symptom dimension of appetite and interest-activity, use of benzodiazepines, SCAN items fatigability, anhedonia, phobia, guilt and loss of interest, age and smoking status contributing the most. In cross-drug analyses, this model predicted remission in nortriptyline-treated participants with an AUC of only 0.53 and pseudo R^2 0.013. Among nortriptyline-treated participants, 52 variables from combination II predicted remission with an AUC 0.70, sensitivity 0.64, specificity 0.61 and a pseudo R^2 0.15, with strongest contributions from baseline BMI, symptom dimensions pessimism, loss of interest-activity and appetite, depression subtypes anxious-somatizing depression and melancholic, SCAN items phobia and loss of libido and age (Table 3). This prediction was almost entirely drug-specific, with cross-drug analyses showing prediction at chance level (AUC 0.50; pseudo R^2 0.022).

When we restricted the analysis to randomly allocated participants, the prediction further improved with highest AUC values 0.74, 0.75 and 0.72 and pseudo R^2 0.22 (p value < 0.001), 0.46 (p value < 0.001) and 0.20 (p value 0.296) for predicting remission in the entire sample, escitalopram-treated and nortriptyline-treated participants respectively (Table 4, Fig. 2). In all cases, the prediction from a combination of variables was several-fold more accurate than prediction from baseline severity alone (Fig. 1). Cross-drug analyses showed high drug-specificity (nortriptyline-to-escitalopram AUC of 0.5; pseudo R^2 = 0.029; escitalopram-to-nortriptyline AUC of 0.5 and a pseudo R^2 = 0.006) (Fig. 1).

3.3. Predicting completion of the trial

The highest AUC for predicting adequate completion of an antidepressant treatment trial in the whole sample was 0.63. The SCAN items preoccupation of death and retardation, BDI item feeling punished, and MADRS item inability to feel contributed the most. AUC were similar in drug-specific analyses and slightly higher in analyses of randomly allocated participants (Supplementary Tables S4 and S5).

3.4. Predicting treatment resistant depression

Features selected from combination II predicted TRD with an AUC 0.67 and strongest contributions from BDI item indecisiveness, SCAN guilty ideas of reference and depressed mood, symptom dimension appetite and HRSD-17 late insomnia. The accuracy of prediction was similar in drug-specific analyses. Prediction

Table 2
Predicting reduction in depressive symptoms among randomly allocated participants.

Randomly allocated	Both antidepressants		Escitalopram		Nortriptyline	
Sample size	n = 450		n = 232		n = 218	
	R^2		R^2		R^2	
Combination I	0.042		0.097		**0.065**	
Combination II	**0.051**		0.082		0.060	
Combination III	0.046		0.078		0.057	
Combination IV	0.046		**0.102**		0.058	
Best model:	II		IV		I	
Cross-drug prediction			R^2		R^2	
Escitalopram			–		0.019	
Nortriptyline			0.010		–	
No. predictors retained:	48		14		29	
Strongest effect size:	Predictor	beta	Predictor	beta	Predictor	beta
1.	Hopelessness (SCAN)	−3.13	Indecisiveness (BDI)	−4.56	Psychomotor retardation (HRSD)	−4.39
2.	Anxiety, psychic (HRSD)	3.10	Tricyclic antidepressants (MHF)	−3.42	BMI	−3.82
3.	Psychomotor retardation (HRSD)	−2.82	Phobia (SCAN)	−2.62	Low self-esteem, guilt (HRSD)	3.45
4.	Suicidal thoughts (MADRS)	2.65	Somatic symptoms, general (HRSD)	−2.35	Feeling punished (BDI)	−3.35
5.	Anxious-somatizing depr. (SCAN)	−2.65	Work and interests (HRSD)	−2.01	Depressed mood (HRSD)	−3.14
6.	Indecisiveness (BDI)	−2.63	Insomnia, initial (HRSD)	−1.28	Loss of Pleasure (BDI)	−2.76
7.	Phobia (SCAN)	−2.39	Dual-action antidepressants (MHF)	−0.94	Apparent Sadness (MADRS)	2.65
8.	BMI	−2.35	Depressed mood (HRSD)	0.87	Pessimism (BDI)	−2.17
9.	Melancholic symptoms (SCAN)	2.20	Suicidal thoughts (MADRS)	0.62	Concentration difficulties (MADRS)	2.13
10.	Concentration difficulties (MADRS)	1.83	Loss of energy (SCAN)	−0.36	Reduced appetite (MADRS)	1.77

Highest accuracy (R^2) across models trained in every sample is marked in bold.
R^2 = coefficient of determination (proportion of variance explained); beta = standardized regression coefficient (= a measure of effect size).

improved in drug-specific analyses of randomly allocated participants with AUC of 0.72 and a pseudo-R^2 over 15% (Supplementary Table S6 and S7).

4. Discussion

In a comparative study of treatment with two antidepressants, a combination of demographic and clinical variables predicted outcome to escitalopram with clinically significant accuracy. In conjunction with earlier studies (DeRubeis et al., 2014; Perlis, 2013), these results suggest that a combination of information from questionnaires and rating scales can meaningfully contribute to predict treatment outcome for individual patients.

Prediction models derived from training datasets explained 5% of variation in symptom improvement overall and10% of variation in symptom improvement with escitalopram in testing datasets that were not used in deriving the predictive algorithms. Permutation of outcomes confirmed that the overall and escitalopram-specific predictions were significantly more accurate than what could be expected by chance. Escitalopram-specific outcome prediction was more accurate than generic outcome prediction, and reached effect sizes that were above a previously established benchmark for clinical significance (Uher et al., 2012d). At least one-third of the prediction was escitalopram-specific suggesting that easy-to-obtain clinical variables can meaningfully contribute to predict outcome for those treated with this antidepressant drug. The prediction of outcomes of treatment with nortriptyline was also numerically more accurate than the overall prediction of outcome and nearly as strong as for escitalopram, but it could not be confirmed as statistically significant in the permutation test. This may be due to the fact that individuals allocated to nortriptyline were more likely to experience adverse effects and to end treatment early. Consequently, nortriptyline-treated participants contributed on average fewer data points than those treated with escitalopram. As a result, the measurement of nortriptyline treatment outcome may have been less accurate and the statistical power to determine robust prediction may have been reduced into he nortriptyline-treatment arm.

The prediction of remission was strong in both generic and drug-specific analyses, with at least comparable predictive power to previously reported results from an ethnically mixed sample (Perlis, 2013). Permutation analyses confirmed that the prediction overall and among escitalopram-treated participants was significantly distinct from chance. Combinations of variables predicted remission with sensitivity and specificity 0.66, and area under the curve of 0.72, explaining approximately 15% of variance in who achieves remission. Such prediction is clinically significant (Uher et al., 2012d). The estimation of variance explained from pseudo-R^2 would suggest that prediction of remission is more accurate than prediction of symptom reduction. This comparison needs to be interpreted with caution because the categorical outcome remission contains less information and because the comparability of pseudo R^2 from logistic analysis with R^2 from linear analysis depends on several assumptions. Nonetheless, it is clear that inclusion of additional symptom predictors improved the prediction of remission several-fold compared to prediction based on total score on the depression rating scale, which is known to have a strong relationship with remission.

Our statistical learning approach allowed to identify and rank the variables that contributed the most and understand the mechanism underlying multivariate outcome prediction. The prediction was largely driven by specific symptom profiles. Symptoms of reduced interest, decreased activity, anxiety, and indecisiveness contributed to the prediction of all types of outcomes. High scores on symptom dimension of interest and activity and its component symptoms contributed to all predictions. More severe ratings on this dimension were among the strongest predictors of non-remission to both drugs, in agreement with previously reported univariate analyses (Uher et al., 2012b). Anxious-somatizing subtype of depression strongly predicted poor citalopram treatment outcome in STAR*D (Fava et al., 2008), but the prediction did not replicate in GENDEP (Uher et al., 2011). Interestingly, the present analysis shows that anxious-somatizing depression plays an important role in the multivariate prediction of remission in GENDEP, suggesting a complex interplay between anxiety and other predictors in determining treatment outcomes. Other predictors were drug-dependent or outcome-dependent. For example, body mass index and loss of appetite specifically contributed to the

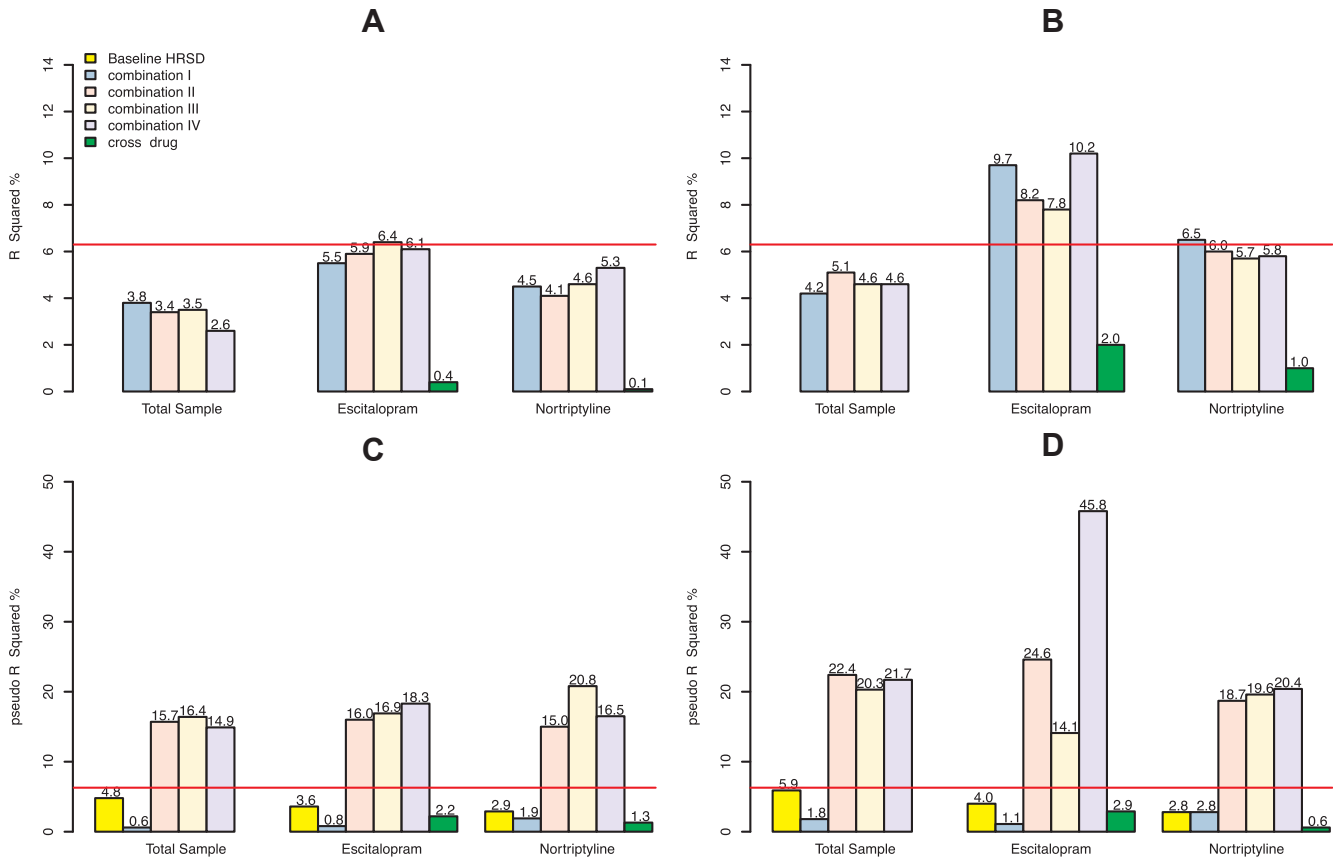

Fig. 1. Title: Predictive power of models predicting treatment outcomes from demographic, and clinical variables. Description: "Panel A: Prediction of change in MADRS in the no randomly allocated sample; Panel B: Prediction of change in MADRS in the Randomly allocated sample; Panel C: Prediction of remission (HRSD) in the no randomly allocated sample; Panel D: Prediction of remission (HRSD) in the Randomly allocated sample; The y-axes plot the proportion of variance in outcome explained in testing datasets estimated as R^2 for symptom reduction and pseudo R^2 for remission. For each outcome the left panel shows results in the whole sample and the right panel shows results in the randomly-allocated subsample. The red horizontal line marks the previously established benchmark or what is a clinically significant prediction. Baseline score on the HRSD scale, which has a known strong relationship with remission, was excluded from the variable combinations in the lower panels. For comparison, the first (bright yellow) bar marks the prediction of remission from baseline total score on the HRSD scale. The last (green) bar marks the cross-drug prediction of remission, i.e. the proportion of variance in outcome explained in the escitalopram-treated sample by the best model in predicting outcome in the nortriptyline-treated sample and the proportion of variance in outcome explained in the nortriptyline-treated sample by the best model in predicting outcome in the escitalopram-treated sample.". (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

prediction of treatment outcome with nortriptyline (Uher et al., 2009b). The prediction profiles confirm findings from previous univariate analyses and extend these to demonstrate their joint predictive power to optimize prediction at the level of individual.

Prediction of trial completion was substantially less accurate than prediction of remission. Prediction of treatment resistance was intermediate. This may reflect reduced power in predicting less common outcomes, heterogeneity in determinants of trial discontinuation or uncertainty in the definition of TRD, which partly depends on recollection of past treatment trials. Prediction of these outcomes may benefit from inclusion of additional variables.

We included most previously-reported predictors of treatment outcome in the predictive models, including demographic and physical variables (age, body mass index, social status), concurrent anxiety, symptom dimension of interest-activity, suicidal thoughts and behaviours, treatment history and life events (Keers et al., 2010; Uher et al., 2011; 2009a; 2009b; 2012b; 2012c). Other potentially predictive variables were not included because they were invariant in this relatively homogeneous sample (e.g. ethnicity, some comorbid disorders, family history of bipolar disorder) or the data were not available (e.g. personality disorders were not assessed in GENDEP). While this degree of inclusiveness provided a scope of predictors for multivariate analyses, the selection can by no means be considered exhaustive. At present, it remains unclear why some predictors' contribution depends on outcome (e.g. loss of appetite contributes to the prediction of remission but not to the prediction of symptom improvement). Other predictors, such as interest-activity symptom dimension, contribute consistently to predicting all types of outcomes and are likely to be included in further refinements of models predictive of depression treatment outcomes.

Our results should be interpreted with regard to the limitations of carrying a complex analysis in a finite dataset with imperfect measurement. We used a 10-fold cross-validation to identify predictors of outcome. Although this is a rigorous approach, in some cases cross-validation can lead to misestimation of the effect sizes (Kim, 2009). We have confirmed the specificity and robustness of our results in various ways. We averaged cross-validation estimators across 100 repetitions, performed cross-drug analyses and used permutation testing to provide a significance value for the prediction. Although testing the models in an independent set of unseen cases would be the most reliable way of estimating the generalizability of the predictive capacity of our predictors, the number of clinical variables that overlapped between GENDEP and other studies was much smaller than what we used (Investigators et al., 2013; Perlis, 2013). These studies also explored different drugs than GENDEP or a mixture of drugs, precluding the validation of drug-specific predictors. Alternatively, the statistical significance

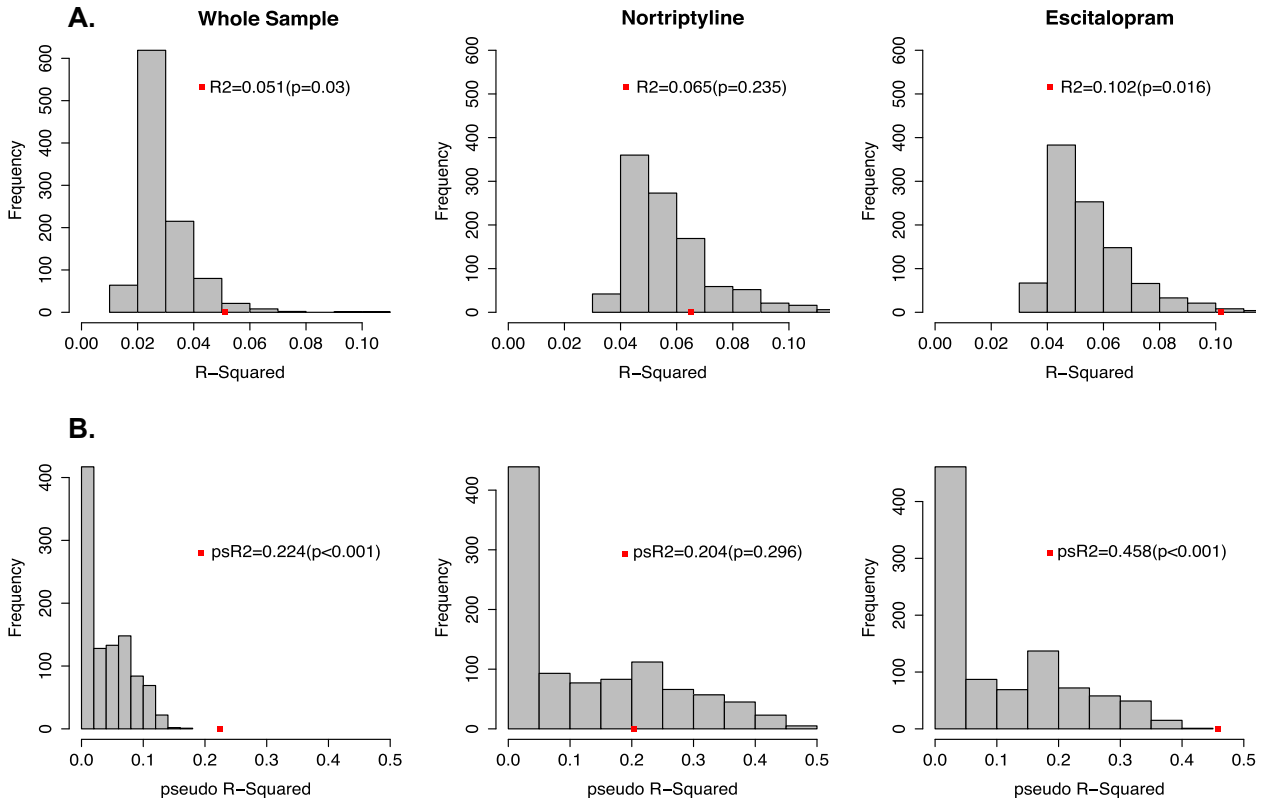

Fig. 2. Title: Null distribution for R^2 and pseudo-R^2 values for models predicting treatment outcome. Description: "Distribution of 10-fold-cross-validation R^2 and pseudo-R^2 values computed in 1000 samples with the outcomes randomly permuted. X-axes plot the R^2 and pseudo-R^2 range of values over 1000 samples with permuted outcome. Y-axes plot the histogram for the observed frequencies for R^2 and pseudo-R^2 over the 1000 samples. The distribution is considered as the null distribution for R^2 and pseudo-R^2, and used to derive the p values for the R^2 and pseudo-R^2 estimations obtained using real outcome, marked with a red square in the plot. Panel A: Null distribution for R^2 for models predicting percentage change in MADRS in the whole sample, and in the subgroups of patients randomly-allocated to nortriptyline or escitalopram arms. Panel B: Null distribution for pseudo-R^2 for models predicting remission (HRSD) in the whole sample, and in the subgroups of patients randomly-allocated to nortriptyline or escitalopram arms. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Predicting remission.

All	Both antidepressants		Escitalopram		Nortriptyline	
Sample size	n = 793		n = 465		n = 328	
	pseudo R^2	AUC	pseudo R^2	AUC	pseudo R^2	AUC
Baseline HRSD[a]	0.048	0.69	0.036	0.65	0.029	0.70
Combination I	0.006	0.61	0.008	0.60	0.019	0.63
Combination II	0.157	0.72	0.160	0.72	**0.150**	**0.70**
Combination III	0.164	0.71	0.169	0.72	0.208	0.70
Combination IV	**0.149**	**0.72**	**0.183**	**0.72**	0.165	0.70
Best model:	IV		IV		II	
Cross-drug prediction			pseudo R^2	AUC	pseudo R^2	AUC
Escitalopram			–	–	0.022	0.50
Nortriptyline			0.013	0.53	–	–
No. predictors retained:	41		46		52	
Strongest effect size:	Predictor	OR	Predictor	OR	Predictor	OR
1.	Appetite (Dimension)	0.82	Appetite (Dimension)	0.78	BMI	0.87
2.	Phobia (SCAN)	0.85	Benzodiazepine (yes)	0.83	Pessimism (Dimension)	0.90
3.	Interest-activity (Dimension)	0.85	Fatiguability (SCAN)	0.83	Anxious-somatizing depr. (SCAN)	0.90
4.	BMI	0.88	Anhedonia (SCAN)	1.15	Phobia (SCAN)	0.91
5.	Age	0.88	Phobia (SCAN)	0.86	Interest-activity (Dimension)	0.91
6.	Anxious-somatizing depr. (SCAN)	0.89	Age	0.87	Melancholic depression (SCAN)	1.09
7.	Benzodiazepine (MHF)	0.91	Interest-activity (Dimension)	0.88	Appetite (Dimension)	0.91
8.	Pathological guilt (SCAN)	1.09	Preoccupation with death (SCAN)	1.11	Loss of libido (SCAN)	0.92
9.	Cognitive (Factor)	0.91	Smoker (yes)	0.89	Age	0.92
10.	Hopelessness (SCAN)	0.92	Pathological guilt (SCAN)	1.10	Hopelessness (SCAN)	0.93

Highest accuracy (AUC) across models trained in every sample is marked in bold.
Pseudo R^2 = log-likelihood pseudo R^2 (estimated proportion of variance explained in logistic regression); AUC = area under the curve; OR = odds ratio.
[a] Baseline HRSD variables were excluded from all combinations.

Table 4
Predicting remission among randomly allocated participants.

Randomly allocated	Both antidepressants		Escitalopra		Nortriptyline	
Sample size	n = 450		n = 232		n = 218	
	pseudo R^2	AUC	pseudo R^2	AUC	pseudo R^2	AUC
Baseline HRSD[a]	0·059	0·69	0·04	0·70	0·028	0·65
Combination I	0·018	0·63	0·011	0·63	0·028	0·65
Combination II	**0.224**	**0.74**	0·246	0·72	**0.187**	**0.72**
Combination III	0·203	0·73	0·141	0·72	0·196	0·70
Combination IV	0·217	0·74	**0.458**	**0.75**	0·204	0·70
Best model:	II		IV		II	
Cross-drug prediction			pseudo R^2	AUC	pseudo R^2	AUC
Escitalopram			–	–	0·029	0·57
Nortriptyline			0·006	0·50	–	–
No. predictors retained:	41		46		52	
Strongest effect size:	Predictor	OR	Predictor	OR	Predictor	OR
1.	Interest-activity (Dimension)	0·77	Interest-activity (Dimension)	0·75	BMI	0·84
2.	Anxious-somatizing depr. (SCAN)	0·78	Phobia (SCAN)	0·76	Anxious-somatizing depr. (SCAN)	0·88
3.	Phobia (SCAN)	0·79	Appetite (Dimension)	0·77	Appetite (Dimension)	0·90
4.	Appetite (Dimension)	0·80	Suicidality (SCAN)	1·22	Hopelessness (SCAN)	0·91
5.	Age	0·82	Smoker (yes)	0·81	Interest-activity (Dimension)	0·91
6.	BMI	0·82	Anxious-somatizing depr. (SCAN)	0·81	Irritability (SCAN)	1·09
7.	Suicidality (SCAN)	1·18	Age	0·83	Restlessness (SCAN)	0·91
8.	Melancholic depression (SCAN)	1·16	Neurovegetative (Factor)	0·86	Age	0·91
9.	Loss of appetite (SCAN)	1·15	Anxiety, somatic (SCAN)	0·86	Pessimism (Dimension)	0·91
10.	Smoker (yes)	0·87	Cognitive (Factor)	0·88	Loss of libido (SCAN)	0·91

Highest accuracy (AUC) across models trained in every sample is marked in bold.
Pseudo R^2 = log-likelihood pseudo R^2 (estimated proportion of variance explained in logistic regression); AUC = area under the curve; OR = odds ratio.
[a] Baseline HRSD variables were excluded from all combinations.

values obtained from permutations tests definitely strengthened our results, suggesting a significant prediction for our clinical predictors.

In summary, the present study suggests that easily obtained demographic and clinical variables can predict response to escitalopram with clinically meaningful accuracy. The obtained results might have the potential to individualized prescription of this antidepressant. Combination with biochemical, genetic, electrophysiological and neuroimaging biomarkers may further increase the prediction accuracy. The present study provides a basis that can be used to test the added benefits of more intensive measurements.

The R code of the model to predict escitalopram outcome is available upon request from raquel.iniesta@kcl.ac.uk.

Acknowledgment

This work has been funded by the European Commission Framework 6 grant, EC Contract LSHB-CT-2003-503428 and an Innovative Medicine Initiative Joint Undertaking (IMI-JU) grant n°115008 of which resources are composed of European Union and the European Federation of Pharmaceutical Industries and Associations (EFPIA) in-kind contribution and financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013).

This article represents independent research part funded by the National Institute for Health Research (NIHR) 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.

Funding was also received from the European Community's FP7 Marie Curie Industry-Academia Partnership and Pathways, grant agreement n° 286213 (PsychDPC).

Open access for this article was funded by King's College London.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jpsychires.2016.03.016.

Conflict of interest

Iniesta, McGuffin, Farmer, Maier, Stahl, Rietschel, Dernovsek, Souery, Hauser, Mors, Dobson and Malki have no conflicts of interest. Lewis and Uher reports grants from European Commission, during the conduct of the study. Henigsberg reports grant from European Commission (through Institute of Psychiatry, King's College, London), and participation in clinical trials sponsored by pharmaceutical companies, including Lundbeck outside the submitted work. Aitchison reports grants and personal fees from Lundbeck/GlaxoSmithKline, outside the submitted work.

Authors' contributions

K Malki, D Stahl, R Dobson and C Lewis contributed to data analysis, interpretation of results, final approval of the version to be published, revised the manuscript for important content and warranted that any part of the work was appropriately investigated and resolved.

W Maier, M Rietschel, O Mors, J Hauser, N Henigsberg, MZ Dernovsek, D Souery, KJ Aitchison, A Farmer and P McGuffin contributed to study design, data collection, interpretation of results, revised the manuscript for important content, contributed to the final approval of the version to be published and warranted that any part of the work was appropriately investigated and resolved.

R Uher and R Iniesta contributed to study design of the work, data collection, data analysis, interpretation of results, writing of the manuscript, contributed to final approval of the version to be published and warranted that any part of the work was appropriately investigated and resolved.

Role of funding

The funding source had no role in study design, collection, analysis or interpretation of data, in the writing of the report or in the decision to submit the article for publication.

References

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., Erbaugh, J., 1961. An inventory for measuring depression. Arch. General Psychiatry 4, 561–571.
Brugha, T., Bebbington, P., Tennant, C., Hurry, J., 1985. The list of threatening experiences: a subset of 12 life event categories with considerable long-term contextual threat. Psychol. Med. 15, 189–194.
DeRubeis, R.J., Cohen, Z.D., Forand, N.R., Fournier, J.C., Gelfand, L.A., Lorenzo-Luaces, L., 2014. The personalized advantage index: translating research on prediction into individualized treatment recommendations. A demonstration. PLoS One 9, e83875.
Fava, M., Rush, A.J., Alpert, J.E., Balasubramani, G.K., Wisniewski, S.R., Carmin, C.N., et al., 2008. Difference in treatment outcome in outpatients with anxious versus nonanxious depression: a STAR*D report. Am. J. Psychiatry 165, 342–351.
Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33, 1–22.
Hamilton, M., 1967. Development of a rating scale for primary depressive illness. Br. J. Soc. Clin. Psychol. 6, 278–296.
Hastie, T., Tibshirani, R., Friedman, J.H., 2009. The Elements of Statistical Learning : Data Mining, Inference, and Prediction, second ed. Springer, New York.
Investigators, G., Investigators, M., Investigators, S.D., 2013. Common genetic variation and antidepressant efficacy in major depressive disorder: a meta-analysis of three genome-wide pharmacogenetic studies. Am. J. Psychiatry 170, 207–217.
Keers, R., Uher, R., Gupta, B., Rietschel, M., Schulze, T.G., Hauser, J., et al., 2010. Stressful life events, cognitive symptoms of depression and response to antidepressants in GENDEP. J. Affect. Disord. 127, 337–342.
Kim, J.H., 2009. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput. Stat. Data Anal. 53, 3735–3745.
Kohavi, R.A., 1995. Study of Cross-validation and Bootstrap for Accuracy Estimation and Model Selection. Morgan Kaufmann.
Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28, 1–26.
Kupfer, D.J., Frank, E., Phillips, M.L., 2012. Major depressive disorder: new clinical, neurobiological, and treatment perspectives. Lancet 379, 1045–1055.
Montgomery, S.A., Asberg, M., 1979. A new depression scale designed to be sensitive to change. Br. J. Psychiatry 134, 382–389.
Perez-Guaita, D., Kuligowski, J., Garrigues, S., Quintas, G., Wood, B.R., 2015. Assessment of the statistical significance of classifications in infrared spectroscopy based diagnostic models. Analyst 140, 2422–2427.
Perlis, R.H., 2013. A clinical risk stratification tool for predicting treatment resistance in major depressive disorder. Biol. Psychiatry 74, 7–14.
Simon, G.E., Perlis, R.H., 2010. Personalized medicine for depression: can we match patients with treatments? Am. J. Psychiatry 167, 1445–1455.
Uher, R., Farmer, A., Maier, W., Rietschel, M., Hauser, J., Marusic, A., et al., 2008. Measuring depression: comparison and integration of three scales in the GENDEP study. Psychol. Med. 38, 289–300.
Uher, R., Maier, W., Hauser, J., Marusic, A., Schmael, C., Mors, O., et al., 2009a. Differential efficacy of escitalopram and nortriptyline on dimensional measures of depression. Br. J. Psychiatry 194, 252–259.
Uher, R., Mors, O., Hauser, J., Rietschel, M., Maier, W., Kozel, D., et al., 2009b. Body weight as a predictor of antidepressant efficacy in the GENDEP project. J. Affect. Disord. 118, 147–154.
Uher, R., Perroud, N., Ng, M.Y., Hauser, J., Henigsberg, N., Maier, W., et al., 2010. Genome-wide pharmacogenetics of antidepressant response in the GENDEP project. Am. J. Psychiatry 167, 555–564.
Uher, R., Dernovsek, M.Z., Mors, O., Hauser, J., Souery, D., Zobel, A., et al., 2011. Melancholic, atypical and anxious depression subtypes and outcome of treatment with escitalopram and nortriptyline. J. Affect. Disord. 132, 112–120.
Uher, R., Carver, S., Power, R.A., Mors, O., Maier, W., Rietschel, M., et al., 2012a. Non-steroidal anti-inflammatory drugs and efficacy of antidepressants in major depressive disorder. Psychol. Med. 42, 2027–2035.
Uher, R., Perlis, R.H., Henigsberg, N., Zobel, A., Rietschel, M., Mors, O., et al., 2012b. Depression symptom dimensions as predictors of antidepressant treatment outcome: replicable evidence for interest-activity symptoms. Psychol. Med. 42, 967–980.
Uher, R., Perlis, R.H., Placentino, A., Dernovsek, M.Z., Henigsberg, N., Mors, O., et al., 2012c. Self-report and clinician-rated measures of depression severity: can one replace the other? Depress. Anxiety 29, 1043–1049.
Uher, R., Tansey, K.E., Malki, K., Perlis, R.H., 2012d. Biomarkers predicting treatment outcome in depression: what is clinically significant? Pharmacogenomics 13, 233–240.
Wall, R., Cunningham, P., Walsh, P., Byrne, S., 2003. Explaining the output of ensembles in medical decision support on a case by case basis. Artif. Intell. Med. 28, 191–206.
Whiteford, H.A., Degenhardt, L., Rehm, J., Baxter, A.J., Ferrari, A.J., Erskine, H.E., et al., 2013. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 382, 1575–1586.
Wing, J.K., Babor, T., Brugha, T., Burke, J., Cooper, J.E., Giel, R., et al., 1990. SCAN. Schedules for clinical assessment in neuropsychiatry. Arch. General Psychiatry 47, 589–593.
Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B 67, 301–320.