Accepted author's manuscript. Published in final edited form as: Pharmacology & Therapeutics 2020; 212(Aug): 107557. Publisher DOI: 10.1016/j.pharmthera.2020.107557

TITLE

Predicting treatment effects in unipolar depression: a meta-review

AUTHORS

Dr George Gillett[1, 2]

Dr Anneka Tomlinson[2, 4]

Dr Orestis Efthimiou[3]

Professor Andrea Cipriani[2, 4]

1. Oxford University Clinical Academic Graduate School, John Radcliffe Hospital, Oxford, OX3 9DU, UK
2. Department of Psychiatry, University of Oxford, Oxford OX3 7JX, UK
3. Institute of Social and Preventive Medicine, University of Bern, Switzerland
4. Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford OX3 7JX, UK

CORRESPONDING AUTHOR

Dr Anneka Tomlinson,
Department of Psychiatry,
Warneford Hospital,
University of Oxford,
Oxford OX3 7JX, UK
Email: anneka.tomlinson@psych.ox.ac.uk

ABSTRACT

There is increasing interest in clinical prediction models in psychiatry, which focus on developing multivariate algorithms to guide personalized diagnostic or management decisions. The main target of these models is usually the prediction of treatment response to different antidepressant therapies. This is because the ability to predict response based on patients' personal data may allow clinicians to make individualised treatment decisions, and to provide more efficacious or more tolerable medication to a specific patient. Here, we systematically search the literature for systematic reviews about treatment prediction in the context of existing treatment modalities for adult unipolar depression, until July 2019. Treatment effect is defined broadly to include efficacy, safety, tolerability and acceptability outcomes. We first focus on the identification of individual predictor variables that may predict treatment response, and second, we consider multivariate clinical prediction models. Our meta-review included a total of 10 systematic reviews; seven (from 2014-2018) focusing on individual predictor variables and three focusing on clinical prediction models. These identified a number of sociodemographic, phenomenological, clinical, neuroimaging, remote monitoring, genetic and serum marker variables as possible predictor variables for treatment response, alongside statistical and machine-learning approaches to clinical prediction model development. Effect sizes for individual predictor variables were generally small and clinical prediction models had generally not been validated in external populations. We identify the need for rigorous model validation in large external data-sets to prove the clinical utility of models. We also discuss potential future avenues in the field of personalized psychiatry, particularly the combination of multiple sources of data and the potential of the emerging field of artificial intelligence and digital mental health to identify new individual predictor variables.

KEYWORDS

Antidepressant drugs, Prediction; Unipolar depression; Treatment response; Clinical prediction model; Precision psychiatry; Personalized medicine

ABBREVIATIONS

BDI; Beck's Depression Inventory

BDNF; Brain-derived neurotrophic factor

CBASP; Cognitive behavioral analysis system of psychotherapy

CBT; Cognitive behavioral therapy

CRP; C-reactive protein

DOR; Diagnostic odds ratio

DSM; Diagnostic and Statistical Manual of Mental Disorders

EEG; Electroencephalography

HAM-D, HDRS; Hamilton Depression Rating Scale

ICD; International Classification of Diseases

IL-6; Interleukin 6

MADRS; Montgomery–Åsberg Depression Rating Scale

MDD; Major Depressive Disorder

NDST; Non-directive supportive therapy

RCT; Randomised controlled trial

rTMS; Repetitive transcranial magnetic stimulation

SMD; Standardised mean difference

SSRI; Selective serotonin reuptake inhibitor

TCA; Tricyclic antidepressant

tDCS; Transcranial direct current stimulation

TNF-alpha; Tumour necrosis factor alpha

CONTENTS

1. Introduction

1. INTRODUCTION

There is an increasing interest into the use of so-called 'precision' (or 'personalized') medicine in psychiatry, particularly to predict treatment effects (Cohen et al, 2018). This has led to the recent development of an increasing number of clinical prediction models, a term used to describe a multivariate algorithm that utilizes patient-level data in order to make individualized clinical predictions (Wessler et al, 2015). It is hoped that clinical prediction models may inform improved clinical decisions and offer patients more efficacious, safer or better tolerated treatments based on their personal data, especially in the context of digital mental health (Shinohara et al., 2019a).

Depression is a psychiatric disorder typically characterised by low mood, reduced energy and anhedonia in addition to a number of associated symptoms. Estimates suggest over 300 million people globally experience depression, making it the single largest factor contributing to disability worldwide (Liu et al., 2019). However, depression also exhibits heterogeneity. Based on DSM-5 criteria alone, there are 227 unique symptom profiles which meet criteria for a diagnosis of Major Depressive Disorder (MDD) (Fried et al., 2015). Likewise, depressive episodes may represent manifestations of different conditions, such as unipolar depression or bipolar affective disorder. Therefore, depression is commonly classified into a number of different categories, based on the severity, nature and type of symptoms present as well as their response to treatment (Table 1).

Precision medicine can be particularly relevant to unipolar depression, where a plethora of treatment modalities exist with potential effectiveness for any given individual (Table 2). In particular, the efficacy and adverse effect profiles of commonly-prescribed antidepressant treatments may differ between individuals (Cipriani et al., 2019). In the context of pharmacotherapy, a recent analysis of 87 eligible randomized placebo-controlled trials identified significantly more variability in response to antidepressant medications than to placebo, and this variability differed between different classes of medications (Maslej et al., 2020). Consequently, there is increasing interest in using clinical prediction models to better tailor treatment to each individual patient, based on his/her characteristics, to enhance treatment effectiveness, tolerability or acceptability (Tomlinson et al., 2019). This also mirrors interest in predicting long-term response from an individual's initial response in order to avoid protracted courses of potentially ineffective or harmful treatments (Hallgren et al, 2017).

An array of variables have been hypothesised to be useful in informing clinical prediction models in depression. These include sociodemographic, phenomenological, psychological, neuroimaging, genetic, immune, endocrine and remote monitoring data (Perlman et al, 2019). However, the predictive ability of these variables and their reliability across different clinical populations is still unclear (Bzdok & Meyer-Lindenberg, 2018). In this paper, we will employ a three-fold categorization of predictor variables (Simon & Perlis 2010). We will assume that a variable may act as a prognostic factor, a specific predictor of treatment response, or as an effect modifier. These terms are explained below:

- Prognostic factor - A variable is a prognostic factor when it moderates response but does not interact with treatment. It affects the outcome in the same way for all patients, irrespectively of the received treatment (including placebo).
- Specific predictor - A variable is a specific predictor when it affects the outcome in the same way for all patients receiving an active intervention, different however to

the way it affects outcome in patients receiving placebo. For example, a specific predictor moderates the efficacy of a intervention vs. placebo but does not moderate the effect of two different interventions.

- Effect modifier – A variable is an effect modifier when it moderates response and interacts with treatment. This suggests that the relative effects between any two treatments (active or placebo) depend on the value of the effect modifier.

We will also include studies that present data on variables that predict within-group treatment response. It is not possible to conclude that these variables are specific predictors or effect modifiers because they have not been assessed for a differential effect in multiple intervention or placebo arms. As it is not possible to assess the true nature of these within-group predictive variables within a population, we simply provide a description of what can be concluded from the current evidence.

All of the categories described above are formative to model development and in this review we include them under the broad category of "predictive factors" (Simon & Perlis, 2018). We include any patient-specific variable which may predict future treatment response, including both baseline variables and markers of early response following treatment initiation, as either could plausibly inform clinical decisions when deciding to initiate, change or stop treatments. Likewise, clinical prediction models have been developed using a variety of statistical and machine-learning approaches. The data ('training sets') that these models are developed from vary in scope, quality and clinical significance. Stern and colleagues argued that despite promising results, clinical prediction models have often not been extensively cross-validated in novel populations or tested in clinical settings

(Stern et al, 2018). Additionally, clinical prediction models are derived from a variety of different methods and are often evaluated against a variety of different metrics, making quantitative analysis difficult. Previous systematic reviews have been limited in scope to specific technical approaches used to develop clinical prediction models and therefore may not provide a comprehensive overview of the field (Lee et al, 2018).

This meta-review focuses on existing treatments including pharmacological, psychological, neuromodulatory and electroconvulsive therapies and aims to summarize the literature on the prognostic value of individual variables and clinical prediction models that forecast treatment effects in people with unipolar depression. We conceptualise treatment effects broadly, including efficacy, safety, tolerability and acceptability. Although there is overlap between these terms, safety typically refers to the occurrence of specific adverse events, tolerability to the number of people who stop treatment because of adverse events, whereas acceptability refers to dropouts from any-cause (Shinohara et al., 2019b). In our discussion, we also briefly outline the use of prediction in other disorders in psychiatry, as well as in other medical specialties such as oncology, neurology and cardiovascular medicine, to illustrate how predictive models may enhance future clinical practice relevant to psychiatry in general and unipolar depression in specific.

2. METHODS

We conducted a meta-review of the English-language literature on the topic of treatment response prediction in the context of unipolar depression in adults, considering existing systematic reviews and meta-analyses. We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) approach (Moher et al, 2009). Our review focuses on: i) individual predictor variables and ii) clinical prediction models of treatment effects for any treatment intervention in unipolar depression in adults. Our protocol is registered with PROSPERO (CRD42019141425).

2.1. Search strategy:

We searched Ovid MEDLINE, EMBASE and PsycINFO from inception to 15th July 2019, using the following keywords/terms: "prediction", "antidepressants", "psychological therapy", "psychotherapy", "electroconvulsive therapy", "transcranial magnetic stimulation", "vagal nerve stimulation", "unipolar depression", "major depressive disorder". Our complete search strategy is detailed in a freely available data repository (http://dx.doi.org/10.17632/3v49p2dtnx.1). Our search of electronic databases was complemented by a manual search of the reference lists of relevant publications.

2.2 Selection criteria:

The titles and abstracts of all references were screened for eligibility by three authors (GG, AT, AC). Full-texts of potentially eligible references were then retrieved and assessed for inclusion. Inclusion criteria was limited to systematic reviews of male and female adults (≥18 years) with a primary diagnoses of unipolar depression according to standard operationalised criteria such as DSM-III, DSM-III-R, DSM-IV, DSM-5, ICD-10 or Research Diagnostic Criteria who received any treatment modality for depression. Reviews of individuals with schizophrenia, bipolar disorder and dementia were excluded (studies with psychiatric co-morbidities were included only if participants had a primary diagnosis of unipolar depression or results were presented separately for unipolar depression). Reviews considering generic study design or non-patient factors (such as length of treatment, year of study) as predictor variables were excluded as they were not felt to be relevant to our review's focus. See protocol for further detail.

For the meta-review of individual predictor variables, studies were classified into seven groups, guided by their search strategy, presentation of results, critical appraisal and quantitative synthesis (Table 3). Our review focuses on the most rigorous systematic reviews, classified as level 4 (Table 4), although we also summarise reviews classified as level 3 (Table 5).

2.3 Data synthesis:

Relevant information was extracted from included reviews, including aim(s), intervention(s), population, variable(s) of interest, outcome predicted, methodology, types of clinical prediction model(s), their evaluation and validation. Data extraction for reviews of clinical prediction models mirrored guidelines set out by standardized checklist such as the CHARMS checklist, a data extraction tool specficially designed for systematic reviews of prediction modelling studies (Moons et al, 2014). We anticipated that, due to the heterogeneity of included studies, a quantitative synthesis would not be possible. We therefore presented a qualitative synthesis of results from the two areas of focus separately; i) individual predictor variables and ii) clinical prediction models.

2.4 Critical appraisal:

The AMSTAR-2 tool (Shea et al, 2017) is a popular instrument used to critically appraise systematic reviews with particular focus on a number of 'critical domains'; pre-registration of the review protocol, adequacy of the literature search, justifications for study exclusion, the appropriateness of meta-analytical methods, the risk of bias from individual studies, the consideration of this risk and the assessment of publication bias. The instrument was used to critically appraise all the included systematic reviews in this meta-review.

3. RESULTS

Our search returned 1,869 unique references and we retrieved the full-text of 205 articles. 118 references were initially deemed to be relevant to individual predictor variables, of which 21 were classified as level 3 or level 4 (Figure 1). Seven of these were classified as level 4 (including a single, well-defined population or treatment intervention) and are discussed in our results (Table 4). The remaining 14 reviews are summarized in Table 5. Three reviews concerning clinical prediction models were included (Table 6).

3.1 Individual predictor variables:

Of the seven included reviews, two focused on studies comparing cognitive behavioral therapy (CBT) with pharmacotherapy (Cuijpers et al, 2014; Cuijpers et al, 2017a), two focused on studies including all antidepressant medications (Polyakova et al, 2015; Wagner et al, 2017), one focused on two antidepressants, venlafaxine and sertraline (Gibiino et al, 2014), one focused on transcranial direct current stimulation in isolation or in addition to pharmacotherapy (Shiozawa et al, 2014) and one focused on cognitive-behavioral analysis system of psychotherapy (CBASP), pharmacotherapy or a combination of both (Furukawa et al, 2018).

3.1.1 Variables of interest

Variables of interest included demographic variables such as age (3 reviews), gender (3), childhood maltreatment (2), marital status (2), social adjustment (1), job (1), education level (1), clinical variables such as baseline depression severity (3), age of onset (2), duration of episode (2), subtype of depression (2), number of previous episodes (1), prior treatments received (1), treatment resistance (1), baseline anxiety severity (1), family history (1), early clinical improvement following treatment initiation (1) and biochemical variables such as serum and plasma levels of brain-derived neurotrophic factor (BDNF) levels following treatment initiation (1).

3.1.2 Methods

A variety of methods were used to assess possible relationships between predictor variables and outcomes. Two studies (Cuijpers et al, 2014; Cuijpers et al, 2017a) used a one-step individual patient data meta-regression to identify differential response to treatments, differential response to treatment and placebo, or response to individual treatments for a single predictor variable of interest (gender and subtype of depression, respectively). Both reviews reported estimated coefficients to present the relationship between predictor variables and treatment outcomes, adjusted for other covariates which might otherwise act as confounding factors. One study reported coefficients from a meta-regression analysis (Gibiino et al, 2014; Shiozawa et al, 2014). In contrast, Polyakova et al, 2015 assessed the effect of a single predictor variable (BDNF change) and considered treatment responders, remitters and non-responders as categorical groups, comparing BDNF change in each group. This review included studies which attempted to predict treatment response at end-point (week 6) from changes in BDNF levels at day seven; therefore BDNF levels were included as a predictor of future response, although not a baseline variable as considered by the reviews above (Dreimüller et al, 2012). Similarly, Wagner and colleagues (2017) assessed the role of a single predictor variable (early improvement) and reported outcomes including sensitivity, specificity and odds ratios of responding to treatment for categorical groups of early improvers and non-improvers.

3.1.3 Outcomes

All seven reviews reported outcome data related to efficacy, generally reporting on treatment response or remission as defined using a standardized depression scale including the Hamilton Depression Rating Scale (HAM-D), Beck's Depression Inventory (BDI), and the Montgomery–Åsberg Depression Rating Scale (MADRS). One review also reported on deterioration of depression symptoms (Furukawa et al, 2018). Only one review (Furukawa et al, 2018) considered acceptability (dropout rate) as an outcome, while none were designed to identify predictors of tolerability or the development of specific adverse events.

With regards to efficacy, Cuijpers et al, 2014, 2017a found no evidence that either subtype of depression (melancholia, atypical) or gender were associated with treatment response to CBT or pharmacotherapy as an effect modifier, specific predictor or within-group predictive variable. Gibiino et al, 2014 found that female gender (standardized mean difference (SMD) between groups: 1.43, p=0.007) was a within-group predictive vairable of response to venlafaxine, as were shorter duration of illness (SMD 0.98, p=0.001) and Caucasian ethnicity (SMD 2.57, p=0.0212), but there was weaker evidence of an association at week 6 (SMD 2.21, p=0.125). There was no evidence that baseline depression severity was associated with venlafaxine response. There was very weak evidence of an association with recurrent depression (SMD 1.58, p=0.352). None of the variables were strongly associated with sertraline response. Furukawa et al, 2018 found evidence that baseline depression and anxiety severity and use of prior medications were effect modifiers for CBASP, pharmacotherapy or combination therapy (predicted relative treatment effects ranged between -3.9 and 9.4 on HAM-D scores in sub-groups defined by these three variables). No variables were found to be predictors of deterioration in this analysis. There was little evidence that baseline depression severity can be used to predict antidepressant response or remission in the Wagner et al, 2017 review (explained variance in odds ratios: 0.6%, p=0.744 and 8.1%, p=0.285 respectively).

Polyakova and colleagues (2015) identified that serum, but not plasma, BDNF increased more in responders (Cohen's d=1.33, 95% CI 0.69–1.97) and remitters (d=0.85, 95% CI 0.39–1.29) following antidepressant medication including Selective Serotonin Reuptake Inhibitors (SSRIs), Selective Noradrenergic Reuptake Inhibitors, Tricylcic Antidepressants (TCAs) and atypical antidepressants compared to non-responders, while Wagner et al, 2017 found that patients with early improvement were more likely to achieve response (pooled OR 8.37, 95% CI: 6.97; 10.05) or remission (pooled OR 6.38, CI: 5.07; 8.02) compared to those without early improvement. Cohen's d is an estimate of effect size, and is often interpreted as small where effect size is >0.2, medium where effect size is >0.5 or large where effect size >0.8 (Cohen, 1988). Confidence intervals quantify the uncertainty estimated effects, by providing a range of values within which the true effect size is expected to lie 95% of the times. In the context of transcranial direct current stimulation, one study found no evidence of association between age, gender, baseline depression severity or treatment-resistance and treatment response (Shiozawa et al, 2014).

With regards to acceptability, Furukawa et al, 2018 found evidence that age and subtype of depression modify the effect of CBASP compared to combination of CBASP and pharmacotherapy, although it is difficult to disentangle the effects of individual covariates as this study modelled interactions between different combinations of predictor variables rather than considering each variable as a predictive factor in isolation.

3.1.4 Critical appraisal

AMSTAR-2 assessements are summarised in Table 4, with full details for each study available in a data repository (http://dx.doi.org/10.17632/3v49p2dtnx.1). All seven reviews contained two or more flaws in critical domains and therefore were considered to have "critically low" quality. Common areas of weakness included a lack of explicit statement detailing that methods were established prior to review commencement, failure to provide a list of excluded studies and a lack of consideration of risk of bias when interpreting results.

3.2 Clinical Prediction Models:

Our search returned three reviews meeting our inclusion criteria for assessment of clinical prediction models (Table 6). Bos et al, 2015 presented a broad review on the role of experience sampling and ecological momentary assessment in prescribing of psychotropic medications in MDD. The authors identified one study involving a sample of 49 patients receiving the tricyclic antidepressant imipramine, which found a clinical prediction model combining measures of early change in HDRS with early measures of positive affect (measured by experience-sampling) improved prediction of response and remission compared to single variables alone (Geschwind et al, 2011). The model accounted for 28 and 40% of explained variance in response and remission respectively. However, given that the clinical prediction model was developed using information of relatively few patients and also given that it was not validated in an external sample, its clinical usefulness remains questionable.

Lee et al, 2018 presented a review of clinical prediction models in the context of depression. Although criteria included both bipolar and unipolar depression, components of the qualitative and quantitative syntheses presented results exclusively from participants with unipolar depression. Twenty-six studies were included, two of which featured both bipolar and unipolar depression, with the remainder featuring solely unipolar depression. Clinical prediction models predicted a range of proxy-markers of treatment response, including patient- or observer-rated symptom scales, frequency of hospital admission or suicidal ideation. The majority (92%) of models used supervised-learning algorithms, including logistic regression, support vector machines, decision trees, linear discriminant analysis, gradient boosting machines, random forest algorithm and mixture of factor analysis (Iniesta et al., 2016, Redlich et al., 2016, Korgaonkar et al., 2015, Al-Kaysi et al., 2017, Chekroud et al., 2017, Kautzky et al., 2017, Khodayari-Rostamabad et al., 2013). Unsupervised approaches included neural networks (Serretti et al., 2007).

Candidate predictors in clinical prediction models were most commonly neuroimaging (defined as Magnetic Resonance Imaging (*MRI), functional* Magnetic Resonance Imaging *(fMRI) or Electroencephalography (EEG)*, phenomenological (defined as baseline symptom scores, functioning, number of previous depressive episodes and sociodemographic variables including employment, education, household income). Two studies focused exclusively on candidate genetic predictors (Belzeaux et al., 2016, Serretti et al., 2004) and three studies used phenomenological predictors in combination with genetic or neuroimaging predictors (Guilloux et al., 2015, Kautzky et al., 2015, Dysdale et al., 2017). Studies generally evaluated their models using classification accuracy. All studies reported a percentage rate of correct classification, apart from one study (Iniesta et al., 2016) that reported the area under the receiver-operator characteristic curve, a commonly used measure in medical decision-making to determine how well a model or tool distinguishes between groups (Hoo et al., 2017). Sensitivity and specificity were also commonly reported. Importantly, only four of 20 studies evaluated models in an external dataset and one study evaluated performance with hold-out validation. Quantitative pooling of phenomenological and combined prediction models reported classification accuracy of 0.76 (CI: 0.63; 0.87) and 0.93 (CI: 0.86;0.97) respectively, but pooled classification accuracy of neuroimaging and genetic prediction models were presented separately for participants with bipolar and unipolar illness. The authors note that application of commonly-used methods (Egger's test and funnel plot asymmetry) provided some indication of small study effects and publication bias in their included studies, suggesting that smaller studies gave larger estimates. This might be the case for example when studies with negative results were less likely to be published than those with positive findings.

Five studies included in the review by Lee et al, 2018 compared results from machine-learning approaches with conventional statistical analyses of the same dataset (Bailey et al., 2018, Liu et al., 2012, Serretti et al., 2004, Serretti et al., 2007, van Waarde et al., 2014,). *Three* neuroimaging studies failed to identify baseline predictors of treatment response with univariate analysis while the machine-learning algorithms predicted response with a classification accuracy of between 78–91% (Bailey et al., 2018, Liu et al., 2012, van Waarde et al., 2014,). However, multiple regression analysis in another study did identify clinical and demographic predictor variables such as the number of previous depressive episodes, age of onset and the duration of the current episode, consistent with machine-learning methodology (Serretti et al., 2007).

Finally, Rosenblat et al, 2017 presented another approach to clinical prediction models in a review focusing on whether pharmacogenomic clinical prediction models improved treatment response. The review included five studies from three separate commercial pharmacogenomic models featuring a variety of candidate predictor genetic variants. Due to these tools' commercial nature and study of design, the exact outcome predicted, or advice outputted, by each tool is poorly reported but guided treatment selection and dosing. However, the reviewers focused on the clinical validation of these tools, assessing changes in depression severity, response or remission rates among participants whose clinician used the tool. While no tolerability or acceptability data were reported in the review, the candidate predictor variables included genes supposedly associated with adverse reactions, alongside those associated with treatment efficacy and drug metabolism. The review included four controlled trials (two randomized and two non-randomized) and one naturalistic study lacking a control group (Hall-Flavin et al., 2012, Hall-Flavin et al., 2013, Winner et al., 2013, Singh, 2015, Brennan et al., 2015). Two open-label non-randomized trials showed an improvement in response and remission rates when using a pharmacogenomic tool, however this result was not replicated for the same tool in a randomized-control blinded study (Hall-Flavin et al., 2012, Hall-Flavin et al., 2013, Winner et al., 2013). A randomized-controlled double-blinded trial of a different tool did show improvements in remission compared to an unguided group, however reviewers caution interpreting results due to the study funding arising from the developers of the tool (Singh, 2015). Results were not independently replicated. The naturalistic study suggested improvement compared to baseline but lacked a control group for comparison.

In terms of critical appraisal using AMSTAR-2, all three reviews contained two or more flaws in critical domains and therefore were considered to have “critically low” quality (Table 6). Common areas of weakness were a lack of explicit statement detailing that methods were established prior to review commencement, failure to provide a list of excluded studies and a lack of consideration of risk of bias when discussing results.

4. DISCUSSION

4.1 Individual predictor variables:

This meta-review identified seven reviews meeting our inclusion criteria for individual predictor variables. No single variable was found to consistently predict treatment response across multiple reviews. It is possible that this finding represents specificity of individual predictive variables to specific treatments (Simon & Perlis 2010). Alternatively, due to the high number of retrospective analyses performed for a variety of candidate predictor variables, it is also possible that some of the studies’ findings arose from chance (Head et al, 2015). However, the lack of consistent statistically significant findings across different studies does not in itself demonstrate disagreement, as it may simply result from meta-analyses being under-powered to detect the effects of the explored variables to the outcome of interest. Also of particular note are cases where the effects estimated for the individual predictor factors had small effect sizes. In such cases, the clinical significance of predictor variables might be questionable. For example, Gibiino et al, 2014 found that older age predicted worse outcome when considered as a continuous variable, but the difference in outcome associated with older age was so small that it would unlikely be recognized by patients or clinicians without the use of numerical rating scale.

All seven included reviews focused on efficacy, with only one review presenting acceptability data (Furukawa, et al., 2018). Broadening our search to include level 3 reviews (Table 5), identified one further review of tolerability in the context of amitriptyline (Chen et al, 2018) and two further reviews of acceptability in the context of psychotherapy (Karyotaki et al, 2015, Cooper et al, 2015). This may represent a relatively unexplored area in this field, especially given the variety of potentially under-utilized treatments licensed for unipolar depression (Table 2) which exhibit potentially clinically relevant differerences in terms of efficacy and acceptability (Cipriani et al, 2018). Other medical specialties, such as cardiovascular medicine, have struggled to predict adverse effects due to their relatively low frequency in clinical trial data (van der Leeuw et al, 2014), although there is limited precedent for doing so, with the example of natalizumab (Tysabri) in multiple sclerosis. In this case, data from postmarketing sources, clinical studies, and a national registry identified antibodies which predicted occurrence of a rare but potentially life-threatening progressive multifocal leukoencephalopathy (PML) following treatment with the drug (Bloomgren et al, 2012).

Only three of the seven reviews of individual predictor variables listed in Table 4 were designed to identify variables acting as effect modifiers, with gender, subtype of depression and clinical factors (such as age of onset, number of precious depressive episodes) investigated in this manner (Cuijpers et al, 2014, 2017a; Furukawa et al, 2018). Three reviews reported on specific predictors, by assessing variables which may predict differential response in the intervention arm compared to a placebo arm. All seven reviews assessed for within-group predictive variables, whereby the predictive efficacy of variables were not compared with a second active intervention or placebo arm. Variables investigated in such a way, including age, gender, ethnicity and duration of depressive episode, may therefore represent prognostic factors, specific predictors or effect modifiers (Gibiino et al, 2014). Furthermore, we found that these terms were used inconsistently, with some studies reporting on "moderator" variables without an appropriate study design to distinguish effect modifiers from prognostic factors (Gibiino et al, 2014). For example, while female gender was shown to be associated with response to venlafaxine and there was no evidence of an association with response to sertraline, this cannot be considered evidence of effect modification. For this, a formal statistical analysis would be required (Cuijpers et al, 2014, 2017a and Furukawa et al, 2018). Indeed, in this example it is possible that gender may be a generic prognostic factor and that certain samples were underpowered to detect an association, rather than there being a differential effect.

Likewise, in our broader search (Table 5) we identified a number of reviews assessing whether variables predicted response in a pooled grouping of multiple treatment interventions or placebo. Such analyses are not adequate to elucidate prognostic factor, specific predictor or within-group effects since they do not clarify whether variables affect response differentially in each treatment arm or placebo as they only present the pooled effect size for all component treatment arms or placebo. Analyses containing hetergenous treatment interventions not assessed independently are therefore of limited value and are not the focus of our reported results. The importance of study design to identify and classify effect modifiers, specific predictors and prognostic factors in order to isolate the individual relationships between predictor variables and specific treatments rather than more generic predictors of outcome has been discussed in the literature previously (Simon and Perlis, 2010).

There was also heterogeneity in how treatment response was conceived and reviewed across different studies. Approaches included categorical variables (response defined as an >50% improvement symptoms) or continuous variables (symptom severity change pre- and post-intervention using a variety of different rating scales), measured at different time-points and using different cut-offs. Heterogeneity also existed in the individual predictor variables themselves. For serum and plasma markers such as BDNF, samples were drawn at different time-points between studies (Polyakova et al, 2015). In particular, the clinical utility of the results presented in that review may be questionable due to the heterogeneity of time-points where treatment response and BDNF levels were measured, since BDNF can only be used as a predictor of future response if it is measured significantly in advance of clinical response measurements. Heterogeneity in study population and treatment setting also exist, raising questions about the specific contexts in which individual predictor variables and models are valid.

4.2 Clinical prediction models:

This meta-review identified three reviews of clinical prediction models (Bos et al., 2015, Lee et al., 2018, Rosenblat et al., 2017). The clinical prediction models we identified were generally poorly validated, and commonly not evaluated on a separate data-set to that which the models were derived from. Even in the rare cases that models were externally validated, the test data was often derived from clinical trial settings with similar methodology and populations, raising questions about its external validity. The importance of internal-external and external validation has been discussed in the literature and is important to avoid over-fitting and to improve external validity (Steyerberg and Harrell, 2016). Therefore, caution should be advised when interpreting results for clinical prediction models which haven't been properly validated.

The lack of external validation is compounded by the need of large datasets to develop clinical prediction models. There is evidence that training-sample size is the most robust predictor of model performance (Popovici et al, 2010). Therefore, while it is possible to split data-sets into training, cross-validation and test sets in order to better evaluate and validate models, doing so reduces the sample size which the model is built from, potentially compromising its performance. A compromise must often be reached between optimising model performance and rigorous validation, especially given that data-set size is already a major limitation in precision psychiatry. Indeed, Cuijpers et al, 2012 estimated that in order to perform sufficiently powered analyses of individual predictor variables predicting pharmacotherapy or psychotherapy response, another 254 studies would have to be conducted.

Another limitation of clinical prediction model development is that they do not neccessarily inform our understanding of which candidate predictor variables are significant. For instance, clinical prediction model development using machine-learning methods such as neural networks do not provide individual coefficients for each variable inputted into the model. These models also often include high order interactions between covariates, such that one cannot identify the effect of a single covariate to the outcome. Meanwhile commercially developed models often do not share the weight assigned to each predictor variable in model development. These factors limit our understanding of which candidate predictor variables deserve further research interest. This is especially important, given that in oncology the selection of candidate predictor variables and data inputted has been shown to be a more significant determiner of model performance than the type of algorithm used (Jang et al, 2014).

The vast majority of prediction models identified in our meta-review focused on individual predictors of just one domain, such as genetic, neuroimaging, demographic or clinical data. This raises the difficulty of how to combine prediction models from different domains to make more accurate predictions. While an ensemble method using stacking may prove useful in combining the predictions of a number of separate prediction models into one, our meta-review did not identify any examples of this within the context of unipolar depression (Wan et al, 2014). Therefore, in depression, a significant challenge exists in standardizing and combining predictor variables of different domains to predict treatment response, compared to other specialties such as oncology where pathology is thought to arise primarily from genetics (Kelloff and Sigman, 2012).

Finally, similarly to reviews of individual predictor variables, we found that significant heterogeneity exists in reviews of clinical prediction models. Alongside heterogeneity in the definitions of treatment response and candidate predictor variables, there is heterogeneity in the specific context in which a clinical prediction model is valid in. For instance, it is possible that a model which accurately predicts treatment response to a drug after two failed interventions would fail to accurately predict response to the same drug in the context of a new presentation of unipolar depression, or vice versa. This consideration emphasises the importance of rigorous model validation in different populations and clinical contexts.

All included reviews of individual predictor variables and clinical prediction models were deemed to be *"critically low"* according to the AMSTAR-2 instrument, which is the label given to reviews with flaws in at least two critical domains and indicative that reviews "should not be relied on to provide an accurate and comprehensive summary of the available studies" (Shea et al., 2017). However, it is worth noting that this was often the result of not providing a list of excluded studies, which no review did, or not explicitly referencing a pre-registered protocol in the article, which only one review did. While these criteria represent good practice they are not necessarily evidence that the reviews contained bias or were of poor quality. Furthermore, the suitability of the AMSTAR-2 tool for some of our included studies could be questioned. Although we felt it is important to systematically use a single tool to standardize quality assessment across included reviews, it can be argued that AMSTAR-2 does not comprehensively capture all the quality issues relevant to each review, and may assess aspects not relevant to some reviews. For instance, reviews developing a model (such as Furukawa et al, 2018) might be better assessed by the CHARMS checklist, a tool specifically designed for critical appraisal of prediction modelling studies (Moons et al, 2014). Furthermore, although a review might be classified as high quality, the included studies within that review may be of low quality and therefore review quality alone may present a deceiving impression of the overall quality of evidence. In particular, many of our included reviews failed to consider their risk of bias assessment results when interpreting their results, raising the possibility that meta-analytic findings may conceal biases present in individual component studies.

4.3 Limitations:

Although our meta-review is wide-ranging in its scope, it exhibits limitations. Our focus on treatment response and remission meant that it was not possible to include reviews focusing on prediction of relapse following treatment cessation. Most of our included reviews measure treatment response within eight weeks of commencing an intervention, a relatively short time in the context of unipolar depression (Penninx, 2011). However, an equally relevant question is which variables predict relapse or durability of response following the cessation of an intervention and how we might design clinical prediction models to guide clinical decisions to discontinue treatments (Berwian et al, 2017, Kedzior et al, 2015). These reviews are not discussed in detail here but are worthy of consideration due to the known chronicity of unipolar depression.

Another limitation of our meta-review is our focus on the most rigorous reviews featuring a single, well-defined population or treatment intervention. This may compromise external validity, given that the well-defined populations of our included reviews may be dissimilar to the wider clinical population. However, this approach was necessary due to the heterogeneity of the studies, the methodological rigour of our meta-review and the need to clarify the effects of candidate predictor variables. Nonetheless, this concern emphasises the importance of proper validation of clinical prediction models in external samples prior to their use in clinical practice.

Finally, this meta-review is limited in the clinical prediction models it identified. It is likely that recent clinical prediction models have been published that have not yet been identified by systematic reviews and were therefore missed by our literature search. Likewise, clinical prediction models may be derived from meta-analysis, such as the Furukawa et al, 2018 review, which was included in our review of individual predictor variables since it focused on relative effectiveness research rather than performance of multiple clinical prediction models. While an updated search of primary studies would likely be fruitful in identifying further clinical prediction models, it was deemed to be beyond the scope of this review. Rather, our meta-review presents a broader overview of the synthesized literature on individual predictor variables and clinical prediction models in unipolar depression.

4.4. Recommendations for future research

Our findings highlight encouraging efforts towards the prediction of treatment response in unipolar depression, in-keeping with the wider interest of applying precision medicine methods in psychiatry more generally (Cipriani and Tomlinson, 2019). Our meta-review leads us to the following recommendations for future research.

First, we believe the field would benefit from more consistent terminology when characterising predictor variables. Here, we use the terms “prognostic factor,” “specific predictor” and “effect modifier” to distinguish types of predictor variables, although heterogeneity exists in the literature (Simon & Perlis 2010). Similarly, we believe the field may benefit from the use of more consistent and clinically meaningful definitions of treatment outcomes. Prediction of safety and tolerability appear to be relatively underexplored areas worthy of more structured and standardised investigation that can take into account preferences and values of end users (Kernot et al., 2019).

Second, it is noteworthy that the majority of data in our included reviews come from randomized-controlled trials. Combining data from other sources, such as observational studies and “real-world” clinical data, may aid the identification and development of new and possibly stronger candidate variables associated with treatment outcomes and prediction models (Tomlinson et al., 2019). If these variables exhibit true effects, we would expect findings from randomized study populations to be replicated in large observational clinical dataset. Obtaining information from multiple sources would not only provide opportunity to optimise model performance and identify further predictor variables, but to also thoroughly assess the clinical utility of models and elucidate the specific clinical contexts and populations in which they are valid (Vaci et al., 2020). A recent study in cardiology highlighted that differences between the population in which prognostic models are developed and the populations in which models are tested represents a key determinant of model validity (Iwakami et al., 2020). We therefore emphasise the pressing need to undertake external validation of clinical prediction models to thoroughly assess their performance, clinical utility and to guide their appropriate clinical application.

Other specialties may inform methods to increase access to large data-sets. Oncology has benefitted from public repositories of genomic data in its development of targeted therapies, while neurology has benefitted from international patient databases when developing personalized therapies for multiple sclerosis (Azuaje, 2017, Kalincik et al, 2017). Oncology has also used novel trial designs such as adaptive clinical trials to more efficiently identify predictive response biomarkers (Kelloff and Sigman, 2012). Meanwhile networks have been designed based on similarities in genomics and drug structures to predict drug response, methods which may prove useful in overcoming limited data (Zhang et al, 2015). However, it is worth noting that while cancer may be understood as a “genetic disease” (Kelloff and Sigman, 2012), the same may not be true of psychiatric disorders where a more diverse array of predictor variables are likely to contribute to the treatment response (Gómez-Carrillo, 2018). Consequently, while personalized therapies in oncology benefitted from methods such as co-clinical trials with mouse models, such approaches may not prove as successful in psychiatry (Huang et al, 2014).

One particularly exciting area is digital mental health. The emerging field of digital phenotyping offers a wealth of behavioral data which is both cost-effective and practical to collect (Torous et al, 2018). Although our meta-review identified just one example of a prediction model using remote monitoring featuring active data collection (Bos et al, 2015), the role of passive data collection not requiring continuous active patient engagement may provide abundant clinically relevant data to inform future prediction models (Gillett & Saunders, 2019). Such data may prove useful in implementing measurement-based-care, in which a patient’s response to a treatment could inform further predictions on a continuous basis (Lewis et al, 2019). Our meta-review identified one example of using early improvement (Wagner et al, 2017) as a predictor of future response, although this was based on established standard symptom rating scales (HAM-D or MADRS) rather than remotely collected data. If properly validated, the emergence of clinically-relevant digital data may therefore identify new candidate predictor variables and allow clinical prediction models to be validated on large digital datasets. We therefore welcome interest into the identification of standardised digital metrics as candidate variables in the prediction of treatment effects.

4.5 Conclusions:

To conclude, this review summarized the evidence on a number of predictive factors for treatment response in adult unipolar depression and identified interest in an array of predictor variables, especially in the context of efficacy. We found that clinical prediction models had generally not been validated in external populations and discuss potential future avenues in the field, particularly the need for rigorous external validation, the combination of multiple sources of data and the emerging field of digital mental health.

5. CONFLICT OF INTEREST STATEMENT

Two authors (AC & OE) are authors on one of the systematic reviews included in our meta-review (Furukawa et al, 2018). The authors have no other conflicts of interest to disclose.

6. FUNDINIG

The study was funded by the National Institute for Health Research (NIHR) Oxford Health Biomedical Research Centre (grant BRC-1215-20005). The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health.

7. ACKNOWLEDGEMENTS

OE was supported by project grant No. 180083 from the Swiss National Science Foundation (SNSF). AC is supported by the National Institute for Health Research (NIHR) Oxford

Cognitive Health Clinical Research Facility, by an NIHR Research Professorship (grant RP-2017-08-ST2-006) and by the NIHR Oxford Health Biomedical Research Centre (grant BRC-1215-20005). The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the UK Department of Health.

8. AUTHORSHIP AND CONTRIBUTORSHIP

Protocol was designed and registered by GG, AT, AC and OE. Search was performed by GG and reference screening undertaken by GG, AT and AC. Data extraction performed by GG. The manuscript was written by GG and reviewed by AT, AC and OE. All authors gave final approval for the work to be published.

9. REFERENCES

Al-Kaysi, A.M., Al-Ani, A., Loo, C.K., Powell, T.Y., Martin, D.M., Breakspear, M. et al. (2017). Predicting tDCS treatment outcomes of patients with major depressive disorder using automated EEG classification. J. Affect. Disord. 208, 597–603.

Allan CL, Herrmann LL, Ebmeier KP. (2011). Transcranial Magnetic Stimulation in the Management of Mood Disorders. Neuropsychobiology, 64(3):163-9.

Azuaje, F. (2017). Computational models for predicting drug responses in cancer research. *Brief Bioinform, 18*, 820-829.

Bailey, N.W., Hoy, K.E., Rogasch, N.C., Thomson, R.H., McQueen, S., Elliot, D. et al. (2018). Responders to rTMS for depression show increased fronto-midline theta and theta connectivity compared to non-responders. Brain Stimul. 11 (1), 190–203.

Belzeaux, R., Lin, C.W., Ding, Y., Bergon, A., Ibrahim, E.C., Turecki, G., et al. (2016). Predisposition to treatment response in major depressive episode: a per-ipheral blood gene coexpression network analysis. J. Psychiatr. Res. 81, 119–126

Berwian, I. M., Walter, H., Seifritz, E., & Huys, Q. J. (2017). Predicting relapse after antidepressant withdrawal - a systematic review. *Psychol Med, 47*, 426-437.

Bloomgren, G., Richman, S., Hotermans, C., Subramanyam, M., Goelz, S., Natarajan, A et al. (2012). Risk of natalizumab-associated progressive multifocal leukoencephalopathy. *N Engl J Med, 366*, 1870-1880.

Borrione L, Moffa AH, Martin D, Loo CK, Brunoni AR. (2018). Transcranial Direct Current Stimulation in the Acute Depressive Episode: A Systematic Review of Current Knowledge, 34(3):153-63.

Bos, F. M., Schoevers, R. A., & aan het Rot, M. (2015). Experience sampling and ecological momentary assessment studies in psychopharmacology: A systematic review. *European Neuropsychopharmacology, 25*, 1853-1864.

Brennan FX., Gardner KR., Lombard J., Perlis RH., Fava M., Harris HW., et al. A Naturalistic Study of the Effectiveness of Pharmacogenomic Testing to Guide Treatment in Psychiatric Patients With Mood and Anxiety Disorders. (2015). The primary care companion for CNS disorders, 17(2)

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine Learning for Precision Psychiatry: Opportunities and Challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging, 3*, 223-230.

Chekroud, A.M., Gueorguieva, R., Krumholz, H.M., Trivedi, M.H., Krystal, J.H., McCarthy, G. (2017). Reevaluating the efficacy and predictability of antidepressant treatments: a symptom clustering approach. JAMA Psychiatry 74 (4), 370–378.

Chen, L. W., Chen, M. Y., Lian, Z. P., Lin, H. S., Chien, C. C., Yin, H. L. et al. (2018). Amitriptyline and Sexual Function: A Systematic Review Updated for Sexual Health Practice. *American Journal of Mens Health, 12*, 370-379.

Cipriani, A., Furukawa, T. A., Salanti, G., Chaimani, A., Atkinson, L. Z., Ogawa, Y. et al (2018). Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet, 391*, 1357-1366.

Cipriani A, Tomlinson A (2019). Providing the most appropriate care to our individual patients. Evidence-Based Mental Health, 22(1):1-2.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Cohen, Z. D., & DeRubeis, R. J. (2018). Treatment Selection in Depression. *Annu Rev Clin Psychol, 14*, 209-236.

Cooper, A. A., & Conklin, L. R. (2015). Dropout from individual psychotherapy for major depression: A meta-analysis of randomized clinical trials. *Clinical Psychology Review, 40*, 57-65.

Cowen, P., & Anderson, I. (2015). New approaches to treating resistant depression. *BJPsych Advances*, 21(5), 315-323.

Cuijpers, P., Reynolds, C. F., 3rd, Donker, T., Li, J., Andersson, G., & Beekman, A. (2012). Personalized treatment of adult depression: medication, psychotherapy, or both? A systematic review. *Depress Anxiety, 29*, 855-864.

Cuijpers, P., Weitz, E., Twisk, J., Kuehner, C., Cristea, I., David, D. et al (2014). Gender as predictor and moderator of outcome in cognitive behavior therapy and pharmacotherapy for adult depression: an "individual patient data" meta-analysis. *Depress Anxiety, 31*, 941-951.

Cuijpers, P., de Wit, L., Weitz, E., Andersson, G., & Huibers, M. J. H. (2015). The combination of psychotherapy and pharmacotherapy in the treatment of adult depression: A comprehensive meta-analysis. [References]. *Journal of Evidence Based Psychotherapies, 15*, 147-168.

Cuijpers, P., Ebert, D. D., Acarturk, C., Andersson, G., & Cristea, I. A. (2016). Personalized Psychotherapy for Adult Depression: A Meta-Analytic Review. *Behavior Therapy, 47*, 966-980.

Cuijpers, P., Weitz, E., Lamers, F., Penninx, B. W., Twisk, J., DeRubeis, R. J. et al (2017a). Melancholic and atypical depression as predictor and moderator of outcome in cognitive behavior therapy and pharmacotherapy for adult depression. *Depress Anxiety, 34*, 246-256.

Cuijpers P, Noma H, Karyotaki E, Cipriani A, Furukawa TA. Effectiveness and Acceptability of Cognitive Behavior Therapy Delivery Formats in Adults With Depression: A Network Meta-analysis [published correction appears in JAMA Psychiatry. 2019 Jul 17;:]. JAMA Psychiatry. 2019;76(7):700–707. doi:10.1001/jamapsychiatry.2019.0268

Cuijpers P, Noma H, Karyotaki E, Vinkers CH, Cipriani A, Furukawa TA. A network meta-analysis of the effects of psychotherapies, pharmacotherapies and their combination in the treatment of adult depression. World Psychiatry. 2020;19(1):92–107. doi:10.1002/wps.20701

Dreimüller, N., Schlicht KF, Wagner S, Peetz D, Borysenko L, Hiemke C et al. (2012). Early reactions of brain-derived neurotrophic factor in plasma (pBDNF) and outcome to acute antidepressant treatment in patients with Major Depression. *Neuropharmacology*, *62(1)*: 264-9

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y. et al. (2017). Resting-state connectivity biomarkersdefine neurophysiological subtypes of depression. Nat. Med. 23 (1), 28–38

Ebert, D. D., Donkin, L., Andersson, G., Andrews, G., Berger, T., Carlbring, P. et al. (2016). Does Internet-based guided-self-help for depression cause harm? An individual participant data meta-analysis on deterioration rates and its moderators in randomized controlled trials. *Psychol Med, 46*, 2679-2693.

Fischer, S., Strawbridge, R., Vives, A. H., & Cleare, A. J. (2017). Cortisol as a predictor of psychological therapy response in depressive disorders: systematic review and meta-analysis. *British Journal of Psychiatry, 210*, 105-109.

Fornaro M, Giosuè P. (2010). Current nosology of treatment resistant depression: a controversy resistant to revision. *Clin Pract Epidemiol Ment Health*, 6:20–24

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;172:96–102.

Furukawa, T. A., Efthimiou, O., Weitz, E. S., Cipriani, A., Keller, M. B., Kocsis, J. H. et al (2018). Cognitive-Behavioral Analysis System of Psychotherapy, Drug, or Their Combination for Persistent Depressive Disorder: Personalizing the Treatment Choice Using Individual Participant Data Network Metaregression. *Psychother Psychosom, 87*, 140-153.

Geschwind, N., Nicolson, N. A., Peeters, F., van Os, J., Barge-Schaapveld, D., & Wichers, M. (2011).

Early improvement in positive rather than negative emotion predicts remission from depression after pharmacotherapy. European Neuropsychopharmacology, 21, 241-247.

Gibiino, S., Marsano, A., & Serretti, A. (2014). Specificity profile of venlafaxine and sertraline in major depression: metaregression of double-blind, randomized clinical trials. *International Journal of Neuropsychopharmacology, 17*, 1-8.

Gillett, G. & Saunders, KEA. (2019). Remote Monitoring for Understanding Mechanisms and Prediction in Psychiatry. Current Behavioral Neuroscience Reports. 6(2), 51-56.

Glue, P., Donovan, MR., Kolluri, S. & Emir, B.(2010). Meta-analysis of relapse prevention antidepressant trials in depressive disorders, *Australian and New Zealand Journal of Psychiatry*, 44:8, 697-705

Gómez-Carrillo, A., Langlois-Therien, T., & Kirmayer, L. J. (2018). Precision Psychiatry-Yes, but Precisely What? *JAMA Psychiatry, 75*, 1302-1303.

Guilloux, J.-P., Bassi, S., Ding, Y., Walsh, C., Turecki, G., Tseng, G., Cyranowski, J.M.,Sibille, E. (2015). Testing the predictive value of peripheral gene expression fornonremission following citalopram treatment for major depression.Neuropsychopharmacology 40 (3), 701–710.

Hall-Flavin DK, Winner JG, Allen JD, Jordan JJ, Nesheim RS, Snyder KA, et al. (2012). Using a pharmacogenomic algorithm to guide the treatment of depression. Translational psychiatry, 2:e172.

Hall-Flavin DK, Winner JG, Allen JD, Carhart JM, Proctor B, Snyder KA, et al. (2013). Utility of integrated pharmacogenomic testing to support the treatment of major depressive disorder in a psychiatric outpatient setting. Pharmacogenetics and genomics, 23(10):535-48.

Hallgren, K. A., Bauer, A. M., & Atkins, D. C. (2017). Digital technology and clinical decision making in depression treatment: Current findings and future opportunities. *Depress Anxiety, 34*, 494-501.

Head, ML., Holman, L., Lanfear, R., Kahn, AT., Jennions, MD. (2015). The Extent and Consequences of P-Hacking in Science. PLoS Biology 13(3): e1002106.

Huang, M., Shen, A., Ding, J., & Geng, M. (2014). Molecularly targeted cancer therapy: some lessons from the past decade. *Trends Pharmacol Sci, 35*, 41-50.

Hoo ZH, Candlish J, Teare D. (2017). What is an ROC curve? Emergency medicine journal, 34(6):357-9.

Iniesta, R., Malki, K., Maier, W., Rietschel, M., Mors, O., Hauser, J. et al. (2016). Combining clinical variables to optimize prediction of antidepressant treatment outcomes. J. Psychiatr. Res. 78, 94–102

Iwakami N, Nagai T, Furukawa TA, Tajika A, Onishi A, Nishimura K, et al. (2020) Optimal sampling in derivation studies was associated with improved discrimination in external validation for heart failure prognostic models. Journal of Clinical Epidemiology, 121:71-80.

Jang, I. S., Neto, E. C., Guinney, J., Friend, S. H., & Margolin, A. A. (2014). Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. *Pac Symp Biocomput*, 63-74.

Jauhar, S. & Morrison, P. (2019). Esketamine for treatment resistant depression. BMJ, 366:l5572

Johnsen, T. J., & Friborg, O. (2015). The effects of cognitive behavioral therapy as an anti-depressive treatment is falling: A meta-analysis. *Psychological Bulletin, 141*, 747-768.

Kalincik, T., Manouchehrinia, A., Sobisek, L., Jokubaitis, V., Spelman, T., Horakova, D. et al. (2017). Towards personalized therapy for multiple sclerosis: prediction of individual treatment response. *Brain, 140*, 2426-2443.

Karyotaki, E., Kleiboer, A., Smit, F., Turner, D. T., Pastor, A. M., Andersson, G. et al. (2015). Predictors of treatment dropout in self-guided web-based interventions for depression: an 'individual patient data' meta-analysis. *Psychol Med, 45*, 2717-2726.

Kautzky, A., Baldinger, P., Souery, D., Montgomery, S., Mendlewicz, J., Zohar, J., Serretti,A., Lanzenberger, R., Kasper, S. (2015). The combined effect of genetic polymorphismsand clinical parameters on treatment outcome in treatment-resistant depression. Eur.Neuropsychopharmacol. 25 (4), 441–453

Kautzky, A., Dold, M., Bartova, L., Spies, M., Vanicek, T., Souery, D. et al. (2017). Refining prediction in treatment-resistant depression: results of machine learning analyses in the TRD III sample. J. Clin. Psychiatry 79 (1).

Kedzior, K. K., Reitz, S. K., Azorina, V., & Loo, C. (2015). Durability of the antidepressant effect of the high-frequency repetitive transcranial magnetic stimulation (rTMS) In the absence of maintenance treatment in major depression: a systematic review and meta-analysis of 16 double-blind, randomized, sham-controlled trials. *Depress Anxiety, 32*, 193-203.

Kelloff, G. J., & Sigman, C. C. (2012). Cancer biomarkers: selecting the right drug for the right patient. *Nat Rev Drug Discov, 11*, 201-214.

Kernot C, Tomlinson A, Chevance A, Cipriani A (2019). One step closer to personalised prescribing of antidepressants: using real-world data together with patients and clinicians' preferences. Evidence-Based Mental Health. 22(3):91-92.

Khodayari-Rostamabad, A., Reilly, J.P., Hasey, G.M., de Bruin, H., Maccrimmon, D.J. (2013). A machine learning approach using EEG data to predict response to SSRI treatment for major depressive disorder. Clin. Neurophysiol. 124 (10), 1975–1985

Korgaonkar, M.S., Rekshan, W., Gordon, E., Rush, A.J., Williams, L.M., Blasey, C. et al (2015). Magnetic resonance imaging measures of brain structure to predict antidepressant treatment outcome in major depressive disorder. EBioMedicine 2 (1), 37–45

Lee, Y., Ragguett, R. M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A. et al. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression:

A meta-analysis and systematic review. *J Affect Disord, 241*, 519-532.

Lewis, C. C., Boyd, M., Puspitasari, A., Navarro, E., Howard, J., Kassab, H. et al.(2019). Implementing Measurement-Based Care in Behavioral Health: A Review. *JAMA Psychiatry, 76*, 324-335.

Liu, F., Guo, W., Yu, D., Gao, Q., Gao, K., Xue, Z. et al. (2012). Classification of different therapeutic responses of major depressive disorder with multivariate pattern analysis method based on structural MR scans. PLoS ONE 7 (7), e40968.

Liu Q, He H, Yang J, Feng X, Zhao F, Lyu J. (2019). Changes in the global burden of depression from 1990 to 2017: Findings from the Global Burden of Disease study. *Journal of Psychiatric Research*. pii: S0022-3956(19)30738-1. doi: 10.1016/j.jpsychires.2019.08.002. [Epub ahead of print]

Łojko D, Rybakowski JK. (2017). Atypical depression: current perspectives. *Neuropsychiatr Dis Treat*, 13:2447-2456

Lv H, Zhao YH, Chen JG, Wang DY, Chen H. (2019). Vagus Nerve Stimulation for Depression: A Systematic Review. Front Psychol, 10:64

Maslej, M., Furukawa, T., Cipriani, A., Andrews, P., Mulsant, B. (2020). Individual Differences in Response to Antidepressants: A Meta-analysis of Placebo-Controlled Randomized Clinical Trials. *JAMA Psychiatry*, doi:10.1001/jamapsychiatry.2019.4815

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med, 6*, e1000097.

Moons, K. G., de Groot, J. A., Bouwmeester, W., Vergouwe, Y., Mallett, S., Altman, D. G. et al.(2014). Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med, 11*, e1001744.

Nanni, V., Uher, R., & Danese, A. (2012). Childhood maltreatment predicts unfavorable course of illness and treatment outcome in depression: a meta-analysis. *American Journal of*

Psychiatry, 169, 141-151.

Palpacuer, C., Gallet, L., Drapier, D., Reymann, J.-M., Falissard, B., & Naudet, F. (2017). Specific and non-specific effects of psychotherapeutic interventions for depression: Results from a meta-analysis of 84 studies. [References]. *Journal of Psychiatric Research, 87*, 95-104.

Penninx, B. W., Nolen, W. A., Lamers, F., Zitman, F. G., Smit, J. H., Spinhoven, P. et al. (2011). Two-year course of depressive and anxiety disorders: results from the Netherlands Study of Depression and Anxiety (NESDA). *J Affect Disord, 133*, 76-85.

Perlman, K., Benrimoh, D., Israel, S., Rollins, C., Brown, E., Tunteng, J. F. et al. (2019). A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder. *J Affect Disord, 243*, 503-515.

Polyakova, M., Stuke, K., Schuemberg, K., Mueller, K., Schoenknecht, P., & Schroeter, M. L. (2015). BDNF as a biomarker for successful treatment of mood disorders: A systematic & quantitative meta-analysis. *J Affect Disord, 174*, 432-440.

Popovici, V., Chen, W., Gallas, B. G., Hatzis, C., Shi, W., Samuelson, F. W. et al. (2010). Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res, 12*, R5.

Redlich, R., Opel, N., Grotegerd, D., Dohm, K., Zaremba, D., Bürger, C. et al. (2016). Prediction of individual response to electroconvulsive therapy via machine learning on structural magnetic resonance imaging data. JAMA Psychiatry 73 (6), 557–564

Rosenblat, J. D., Lee, Y., & McIntyre, R. S. (2017). Does Pharmacogenomic Testing Improve Clinical Outcomes for Major Depressive Disorder? A Systematic Review of Clinical Trials and Cost-Effectiveness Studies. *J Clin Psychiatry, 78*, 720-729.

Serretti, A. & Smeraldi, E. (2004). Neural network analysis in pharmacogenetics of mooddisorders. BMC Med. Genet. 5, 27.

Serretti, A., Zanardi, R., Mandelli, L., Smeraldi, E., Colombo, C., 2007. A neural network model for combining clinical predictors of antidepressant response in mood disorders. J. Affect. Disord. 98 (3), 239–245.

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J. et al. (2017). AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *Bmj, 358*, j4008.

Shinohara, K., Tanaka S., Imai H., Noma H., Maruo K., Cipriani A. et al. (2019a). Development and validation of a prediction model for the probability of responding to placebo in antidepressant trials: a pooled analysis of individual patient data. *Evidence-Based Mental Health*, 22(1); 10-16

Shinohara, K., Efthimiou, O., Ostinelli, EG., Tomlinson, A., Geddes, JR., Nierenberg, AA. et al. (2019b). Comparative efficacy and acceptability of antidepressants in the long-term treatment of major depression: protocol for a systematic review and networkmeta-analysis. BMJ Open, 9(5):e027574. Published 2019 May 19. doi:10.1136/bmjopen-2018-027574

Shiozawa, P., Fregni, F., Bensenor, I. M., Lotufo, P. A., Berlim, M. T., Daskalakis, J. Z. et al. (2014). Transcranial direct current stimulation for major depression: An updated systematic review and meta-analysis. [References]. *International Journal of Neuropsychopharmacology, 17*, 1443-1452.

Simon, G. E., & Perlis, R. H. (2010). Personalized medicine for depression: can we match patients with treatments? *Am J Psychiatry, 167*, 1445-1455.

Singh AB. (2015) Improved Antidepressant Remission in Major Depression via a Pharmacokinetic Pathway Polygene Pharmacogenetic Report. Clin Psychopharmacol Neurosci, 13(2):150–156.

Stern, S., Linker, S., Vadodaria, K. C., Marchetto, M. C., & Gage, F. H. (2018). Prediction of response to drug therapy in psychiatric disorders. *Open Biol, 8*.

Steyerberg, E. W., & Harrell, F. E., Jr. (2016). Prediction models need appropriate internal, internal-external, and external validation. *J Clin Epidemiol, 69*, 245-247.

Strawbridge, R., Arnone, D., Danese, A., Papadopoulos, A., Herane Vives, A., & Cleare, A. J. (2015). Inflammation and clinical response to treatment in depression: A meta-analysis. *European Neuropsychopharmacology, 25*, 1532-1543.

The UK ECT Review Group (2003). Efficacy and safety of electroconvulsive therapy in depressive disorders: a systematic review and meta-analysis. The Lancet. 361(9360):799-808.

Tomlinson A, Boaden K, Cipriani A (2019). Withdrawal, dependence and adverse events of antidepressants: lessons from patients and data. Evidence-Based Mental Health. 22(4):137-138.

Tomlinson A, Furukawa TA, Efthimiou O, Salanti G, De Crescenzo F, Singh I, Cipriani A (2019. Personalise antidepressant treatment for unipolar depression combining individual choices, risks and big data (PETRUSHKA): rationale and protocol. Evidence-Based Mental Health. Published Online First: 23 October 2019. doi: 10.1136/ebmental-2019-300118

Torous, J., Staples, P., Barnett, I. et al. Characterizing the clinical relevance of digital phenotyping data quality with applications to a cohort with schizophrenia. npj Digital Med 1, 15 (2018). https://doi.org/10.1038/s41746-018-0022-8

Tunvirachaisakul, C., Gould, R. L., Coulson, M. C., Ward, E. V., Reynolds, G., Gathercole, R. L., et al. (2018). Predictors of treatment outcome in depression in later life: A systematic review and meta-analysis. *J Affect Disord, 227*, 164-182.

Vaci N, Liu Q, Kormilitzin A, et al. Natural language processing for structuring clinical text data on depression using UK-CRIS. Evid Based Ment Health. 2020;23(1):21–26. doi:10.1136/ebmental-2019-300134

Vázquez, G., Tondo, L., Undurraga, J., Zaratiegui, R., Selle, V., & Baldessarini, R. (2014). Pharmacological treatment of bipolar depression. *Advances in Psychiatric Treatment*, 20(3), 193-201.

van der Leeuw, J., Ridker, P. M., van der Graaf, Y., & Visseren, F. L. (2014). Personalized cardiovascular disease prevention by applying individualized prediction of treatment effects. *Eur Heart J, 35*, 837-843.

van Waarde, J.A., Scholte, H.S., van Oudheusden, L.J.B., Verwey, B., Denys, D., van Wingen, G.A. (2014). A functional MRI marker may predict the outcome of electroconvulsive therapy in severe and treatment-resistant depression. Mol. Psychiatry 20, 609

Wagner, S., Engel, A., Engelmann, J., Herzog, D., Dreimuller, N., Muller, M. B. et al. (2017). Early improvement as a resilience signal predicting later remission to antidepressant treatment in patients with Major Depressive Disorder: Systematic review and meta-analysis. [References]. *Journal of Psychiatric Research, 94*, 96-106.

Wan, Q., & Pal, R. (2014). An Ensemble Based Top Performing Approach for NCI-DREAM Drug Sensitivity Prediction Challenge. *PLOS One*, 9(6): e101183.

Wessler, B. S., Lai Yh, L., Kramer, W., Cangelosi, M., Raman, G., Lutz, J. S. et al. (2015). Clinical Prediction Models for Cardiovascular Disease: Tufts Predictive Analytics and Comparative Effectiveness Clinical Prediction Model Database. *Circ Cardiovasc Qual Outcomes, 8*, 368-375.

Widge, A. S., Bilge, M. T., Montana, R., Chang, W., Rodriguez, C. I., Deckersbach, T. et al. (2019). Electroencephalographic biomarkers for treatment response prediction in major depressive illness: A meta-analysis. *American Journal of Psychiatry, 176*, 44-56.

Winner JG, Carhart JM, Altar CA, Allen JD, Dechairo BM. (2013). A prospective, randomized, double-blind study assessing the clinical impact of integrated pharmacogenomic testing for major depressive disorder. Discovery medicine, 16(89):219-27.

Zhang, N., Wang, H., Fang, Y., Wang, J., Zheng, X., & Liu, X. S. (2015). Predicting Anticancer Drug Responses Using a Dual-Layer Integrated Cell Line-Drug Network Model. *PLoS Comput Biol, 11*, e1004498.

Table 1
Summary of classifications of depression.

Type	Description	Diagnostic criteria[a]	Typical treatment recommendation
Depressive episode	A period lasting at least two weeks, typically characterised by low mood, reduced energy and loss of interest or pleasure in normally enjoyable activities. Associated with impairment in social or occupational function. Typically, sub-categorised by severity (determined by type, severity and number of individual symptoms).	*Mild:* At least two core symptoms & two additional symptoms. **Core symptoms:** Low mood, anhedonia, fatiguability. **Additional symptoms:** Reduced concentration, reduced self-esteem, ideas of guilt and unworthiness, hopelessness, ideas of self-harm or suicide, disturbed sleep, changes to appetite.	Psychosocial interventions generally preferred to pharmacological interventions due to risk-benefit ratio (Cuijpers et al., 2020).
		Moderate: At least two core symptoms & three additional symptoms.	Pharmacological treatment, typically a selective-serotonin reuptake inhibitor (SSRI) due to tolerability in addition to psychosocial intervention (Cipriani et al., 2018). Patients who fail to respond should be switched to a 2nd antidepressant medication (Cowen & Anderson, 2015).
		Severe: All three core symptoms & four additional symptoms.	As (above) for moderate depression.
Recurrent unipolar depression	A diagnosis given to individuals who have experienced two distinct depressive episodes.	Patient should have experienced two distinct depressive episodes separated by several months without significant mood disturbances.	In addition to treatment for a depressive episode, antidepressant medications are often continued for relapse prevention, although the optimal treatment duration is unclear (Glue, Donovan, Kolluri, & Emir, 2010).
Bipolar depression	A depressive episode seen in the context of a bipolar disorder illness, where patients experience mania, hypomania or mixed affective episodes in addition to depressive episodes.	As well as meeting the criteria for a depressive episode, patients should have at least one hypomanic, manic or mixed affective episode in the past.	If using lithium or a mood-stabiliser, dosing should be optimised. Antipsychotics (with or without) SSRI may be offered, or alternatively an anticonvulsant, however good evidence is scarce (Vázquez et al., 2014).
Dysthymia	Chronic low mood experienced by patients where symptoms do not meet the criteria due to symptom severity or duration. Generally, symptoms are chronic but associated with less functional impairment than a typical depressive episode. Cyclothymia describes a similar condition which involves mild elation episodes as well as low mood episodes.	Described as very long-standing depression of mood which is never (or very rarely) enough to fill the criteria for recurrent depressive disorder.	Evidence-base similar to mild depression; psychosocial interventions are generally preferred.
Atypical depression	A debated subgroup of depression, with features that may include; mood reactivity to positive events, increased appetite and sleep, pronounced anxiety and heaviness of limbs.	Classified under “other depressive episodes” In ICD-10. “With atypical features” is a specifier in DSM-5.	Historically it was believed atypical depression responded preferably to Monoamine Oxidase Inhibitors (MAOIs), however there is no strong evidence to support this (Łojko & Rybakowski, 2017, Singh et al., 2006).
Melancholic depression (or depression with somatic symptoms)	A debated subgroup of depression, where anhedonia and lack of mood reactivity predominate alongside biological symptoms; diurnal variation in mood (worse in morning), early-morning waking, psychomotor agitation or retardation, weight loss and guilt.	Known as “somatic syndrome” in ICD-10, usually requiring four somatic symptoms to be present. “With melancholic features” is a specifier in DSM-5	No strong evidence to treat melancholic depression differently to standard treatment for depressive episodes (Łojko & Rybakowski, 2017).
Psychotic depression	Severe depressive episodes with psychotic symptoms. May feature delusions, hallucinations, or depressive stupor.	As well as meeting criteria for a severe depressive episode, with delusions, hallucinations or depressive stupor also present.	Antidepressant in combination with an antipsychotic thought to be most effective (Cowen & Anderson, 2015).
Treatment-resistant depression	A term used to describe depression that fails to respond to standard treatment, typically defined as two sequential antidepressant medication trials.	Not defined in diagnostic manuals (DSM-5 or ICD-10). Research criteria usually defines treatment-resistance as a failure to respond to two or more antidepressant medications sequentially at adequate dose and duration. Debates persist around defining response and adequacy of dosing and duration (Fornaro & Giosuè, 2010).	Combination (2 or more antidepressant medications) or augmentation (antidepressant with an atypical antipsychotic or lithium) is recommended. Augmentation with atypical antipsychotic has strongest evidence base (Cowen & Anderson, 2015).

[a] ICD-10 classifications are used preferentially, although DSM-5 classifications are also described where necessary.

Table 2
Summary of treatment modalities for unipolar depression.

Category	Type	Description/examples	Summary of evidence
Pharmacological	Selective serotonin reuptake inhibitors (SSRIs)	Commonly used medications include citalopram, escitalopram, fluoxetine, sertraline, paroxetine.	Medications from all classes shown to be more effective than placebo. Small relative differences between individual medications exist for both efficacy and acceptability although overall there are few differences between individual antidepressants (Cipriani et al., 2018).
	Selective noradrenergic reuptake inhibitors (SNRIs)	Commonly used medications include venlafaxine, duloxetine, atomoxetine.	
	Tricyclic antidepressants (TCA)	Commonly used medications include amitriptyline, nortriptyline, imipramine and clomipramine.	
	Monoamine oxidase inhibitors (MOAIs)	Commonly used medications include phenelzine, selegiline and rasagiline.	
	Atypical antidepressants	Medications not well characterised by other groupings. Commonly used medications include vortioxetine, mirtazapine, bupropion, trazodone and agomelatine	
	Rapid-acting antidepressants	Recent interest into using esketamine for rapid-acting treatment, in combination with established antidepressant medication. Esketamine recently approved by U.S. Food and Drug Administration (FDA)	A small number of randomised controlled trials suggest the possibility of increased response compared to placebo, but no difference in remission demonstrated (Jauhar & Morrison, 2019).
Psychological[a]	Cognitive-behavioral therapy (CBT)	Focus on dysfunctional beliefs and their impact on behaviour. Aim is to restructure these beliefs.	All the listed psychological therapies have been shown to be effective in treating depression, with comparable effectiveness across individual modalities (Cuijpers, Noma, Karyotaki, Cipriani, & Furukawa, 2019).
	Behavioral activation therapy	Focus is on encouraging regular and routine activities, which may be pleasant or functional in nature.	
	Interpersonal psychotherapy	Focus is on interpersonal issues in depression.	
	Problem-solving psychotherapy	Focus is on identifying problems, formulating solutions, acting upon them and reviewing outcomes.	
	Non-directive counselling	Unstructured therapy focused on offering empathy to patients sharing their experiences and emotions.	
	Psychodynamic psychotherapy	Focused on a psycho-analytic framework of a patient's experiences and behaviour.	
Convulsive	Electroconvulsive therapy (ECT)	Used in addition to pharmacotherapy in some cases of treatment-resistant depression.	Meta-analysis evidence that ECT is more effective than simulated ECT or pharmacotherapy alone (The UK ECT Review Group, 2003).
Neuromodulatory	Vagal nerve stimulation	Electrode implanted and attached to vagus nerve to deliver electrical impulses.	No good evidence that vagal nerve superior to sham in sham-controlled trials (Lv, Zhao, Chen, Wang, & Chen, 2019)
	Transcranial Magnetic Stimulation (TMS)	Uses a magnetic field to generate electrical current in specific cortical regions.	Evidence for moderate effect in favour of TMS compared to sham, although there is significant heterogeneity between studies (Allan, Herrmann, & Ebmeier, 2011).
	Transcranial Direct Current Stimulation (tDCS)	Uses scalp electrodes to deliver electrical current to specific cortical regions.	Robust efficacy has not been consistently demonstrated and heterogeneity exists between studies (Borrione, Moffa, Martin, Loo, & Brunoni, 2018)

[a] This selection has been guided by previous reviews in the literature (Cuijpers et al., 2020) and summarises the main forms of psychological interventions.

Table 3
Outline of level categorization process for systematic reviews of individual predictor variables.

Level	Database of search	Date range of search	PRISMA diagram or equivalent	Critical Appraisal	Quantitative synthesis	Single population or intervention
0	✓					
1	✓	✓				
2a	✓	✓	✓			
2b	✓	✓		✓		
2(a + b)	✓	✓	✓	✓		
3	✓	✓	✓	✓	✓	
4	✓	✓	✓	✓	✓	✓

Table 4
Included reviews of individual predictor variables of sufficient quality, transparency and rigour.

Author & year	Intervention	Variable(s) of interest	Outcome[a]	Methods	Population	Interpretation	Quality[b]
Cuijpers et al., 2014	CBT, pharmacotherapy or placebo	Gender	Efficacy as measured by HAM-D-17. Where dichotomous outcomes reported, standardised mean difference was calculated. [E]	Searched for RCTs comparing CBT vs pharmacotherapy in individuals with depression where HAM-D-17 was reported. One-step individual-patient-data meta-analysis mixed effects model used to assess association between gender & treatment response (HAM-D-17). Model adjusted for age, minority status, marital status, education, trial characteristics & quality.	Adults with depression receiving CBT (individual or group), pharmacotherapy (SSRI, TCA or other) or placebo. 1 study included individuals with dysthymia. 1 study focused on individuals with multiple sclerosis, one focused exclusively on women. $n = 14$ studies. 1766 participants.	**Effect modifier:** No strong evidence of gender being an effect modifier of CBT vs pharmacotherapy (adjusted coefficient 0.72, p 0.47). **Specific predictor:** No strong evidence of gender being a specific predictor for CBT or pharmacotherapy outcome (CBT: 0.01, p 1.00, pharmacotherapy: −0.54, p 0.68). **Within-group predictive variable:** No strong evidence of gender being a predictor of within-group response for CBT or pharmacotherapy outcome (CBT: −0.48, p 0.54, pharmacotherapy: −0.85, p 0.20). Analysis repeated to exclude each study once & results replicated.	Critically low
Cuijpers et al., 2017	CBT, pharmacotherapy or placebo	Melancholic & atypical subtypes of depression (DSM-IV)	Efficacy as measured by HAM-D-17. Where dichotomous outcomes reported, standardised mean difference was calculated. BDI-II also reported in 3 of 4 studies. [E]	Searched for RCTs comparing CBT vs pharmacotherapy in individuals where depression subtype was measured. One-step individual-patient-data meta-analyses mixed effects model used to assess association between melancholia/atypical depression & treatment response (HAM-D-17 or BDI-II). Performed on intention-to-treat and completers samples separately. Model adjusted for gender, minority status, marital status, education, trial quality.	Adults with depression, receiving CBT, pharmacotherapy (SSRI) or placebo. USA & Canada. $n = 4$ studies. 805 participants.	**Effect modifier:** No strong evidence of melancholia being an effect modifier of CBT vs SSRI outcome (adjusted coefficient − 0.38, p 0.74). No strong evidence of atypical being an effect modifier of CBT vs SSRI outcome (adjusted coefficient − 1.83, p 0.15). In unadjusted completers sample, results suggest atypical depression responds better to SSRI (coefficient − 2.71, p 0.048), although caution advised. **Specific predictor:** No strong evidence of melancholia being a specific predictor for CBT or SSRI outcome (CBT: 1.14, p 0.56, SSRI: 1.30, p 0.46). No strong evidence of atypical being a specific predictor for CBT or SSRI outcome (CBT: 0.45, p 0.86, SSRI: −2.48, p 0.24). **Within-group predictive variable:** No strong evidence of melancholia being a predictor of within-group response to CBT or SSRI outcome (CBT: −0.32, p 0.71, SSRI: −0.01, p 0.99). No strong evidence of atypical being a predictor of within-group response to CBT or SSRI outcome (CBT: −0.49, p 0.61, SSRI: 1.33, p 0.13). **Other:** Outcome differences between melancholia or atypical depression patients versus those without were consistently small (effect sizes $g < 0.10$).	Critically low
Furukawa et al., 2018	CBASP, pharmacotherapy or combination	Age, childhood maltreatment, marital status, social adjustment/function. Age at onset, length of current episode, number of previous episodes, prior treatments with antidepressants or psychotherapies. Subtype of depression, Baseline severity, baseline anxiety, comorbid personality disorder.	Depression severity measured on an observed-rated scale for depression, converted to 24-item HAM-D. Deterioration as measured by observer-rated scale for depression. Dropouts for any reason. [E], [A]	Searched for RCTs of CBASP, pharmacotherapy or combination in context of recurrent depressive disorder, depression, dysthymia in Cochrane CENTRAL, Pubmed, Scopus & PsycINFO. Synthesized data using individual-patient-data meta-analysis and used penalized regression model to identify covariates (as effect modifiers or prognostic factors) for development of predictive model. First and second-order combinations of the variables of interest used in this model.	Adults with persistent depressive disorder (DSM-5), major depression or dysthymia (DSM-4) or corresponding condition, receiving CBASP, pharmacotherapy or combination. $n = 3$ studies. 1036 participants.	**Change in depression severity: Effect modifier:** Subgrouping by baseline HAM-D, baseline anxiety and having received prior medications associated with a small modifier effect for all three treatments. **Within-group predictive variable:** Baseline anxiety, baseline HAM-D, prior medication and neglect included in model as prognostic factors alone or in combination. Generally weak regression coefficients, baseline HAM-D was the most influential covariate. **Deterioration: Effect modifier:** No strong evidence of effect modifiers identified. **Within-group predictive variable:** Baseline HAM-D, baseline anxiety, social function, marital status included in model as prognostic factors alone or in combination. Generally weak regression coefficients & baseline HAM-D was most influential. **Dropouts: Effect modifier:** Age & depression subtype (chronic MDD) associated with small modifier effect for CBASP vs combination. **Within-group predictive variable:** Baseline HAM-D, age, prior medication, depression subtype (chronic MDD, dysthymia), marital status included in model as prognostic factors alone or in combination. Generally weak regression coefficients, baseline HAM-D and depression subtype played a prominent role.	Critically low

Gibiino et al., 2014	Venlafaxine or sertraline	Demographics: Age, sex, job, educational level, marital status. Clinical data: age of onset, duration, type, family history.	Efficacy as measured by standardised mean difference of HDRS or MADRS scores from baseline to week 6 (and week 8 as a secondary outcome). [E]	Searched for RCTs of venlafaxine or sertraline on Medline, EMBASE and Cochrane Library. Meta-regression analysis used to evaluate potential predictor variables using random-effects model. Categorised variables expressed as a percentage of patients enrolled in each study. Patient data considered on intention-to-treat basis.	Adults with diagnosis of Major Depressive Disorder (excluding dysthymia) enrolled in RCT receiving venlafaxine or sertraline. $n = 59$ unique trials for sertraline 6029 participants $n = 57$ unique trials for venlafaxine 6375 participants	**Within-group predictive variable:** Female gender was associated with within-group response to venlafaxine (SMD = 1.43, p = .007 and $p = .004$ at weeks 6 and 8). Age was extremely weakly associated with within-group response to venlafaxine (SMD = -0.01, $p = .040$), unlikely to be clinically significant. Caucasian ethnicity predicted better within-group response to venlafaxine at week 8 (SMD = 2.57, $p = .0212$), but no strong evidence of predicting response at week 6 (SMD = 2.21, $p = .125$). Duration of episode less than 6 months predicted better within-group response to venlafaxine (SMD = 0.98, $p = .001$ and p = .004) and duration over 1 year predicted worse response to venlafaxine at week 6 (SMD = -1.09, $p = .0004$). Baseline severity and recurrent depression were not strongly associated with within-group response to venlafaxine. No individual patient clinical or demographic factor was strongly associated with within-group response to sertraline. Study design not set-up to assess effect modifiers.	Critically low
Polyakova et al., 2015	Antidepressant medication	Serum or plasma BDNF change (pre and post-treatment initiation)	Treatment response, defined as 50% reduction of scores on HDRS or MADRS. Remission, defined as scores <7 HDRS, <8 MADRS. [E]	Searched for case-control or longitudinal studies of MDD or BD patients receiving standardised treatment in which serum BDNF levels were measured in PubMed, ISI Web of Science, PsycINFO. Bipolar disorder was included in review, but results presented separately and not discussed here. Standardised mean difference was calculated from BDNF measurements. Subdivided MDD sample into non-responders, responders, remitters. Random-effects meta-analysis used to assess association between BDNF & treatment response.	Adults with a diagnosis of Major Depressive Disorder receiving pharmacotherapy in a case-control or longitudinal therapy study, excluding patients with somatic illnesses, pregnant women. Studies using transcranial magnetic stimulation, electroconvulsive therapy or psychotherapy were excluded. **Total:** $n = 48$ studies (MDD) 3365 participants (MDD) **Relevant to prediction of treatment response:** $n = 21$ studies 553 patients	**Within-group predictive variable:** Serum BDNF levels increased following treatment initiation for treatment responders (d = 1.33, 95% CI 0.69–1.97) and treatment remitters (d = 0.85, 95% CI 0.39–1.29), but not non-responders. Serum BDNF levels increased more in remitters and responders compared to non-responders ($p = .036$, $p = .012$).	Critically low
Shiozawa et al., 2014	tDCS	Age, gender, baseline depression severity, treatment-resistant depression.	Depression scores following treatment, as reported on standardised scale. If more than one timepoint reported, last blinded score used. Treatment response defined as >50% depression improvement from baseline to outcomes. Remission as defined in individual included studies (HAMD<8, MADRS<10, MADRS≤10 or unreported). [E]	Searched for randomised, sham-controlled trials of tDCS for Major Depressive Disorder in Medline database. Main focus of review on general efficacy of tDCS. Meta-regression using random effects model also performed to assess association between predictor variables and outcomes.	Adults with MDD enrolled in a randomised trial receiving tDCS alone or in combination with pharmacotherapy. $n = 7$ studies. 259 participants	**Within-group predictive variable:** Meta-regression suggested no strong evidence of age, gender, baseline depression severity or treatment-resistant depression on outcomes.	Critically low

(continued on next page)

Table 4 (*continued*)

Author & year	Intervention	Variable(s) of interest	Outcome[a]	Methods	Population	Interpretation	Quality[b]
Wagner et al., 2017	Antidepressant medication or placebo	Early improvement, defined as HAMD/MADRS reduction of 20, 25% from baseline to day 14. Baseline depression severity measured on standardised scale.	Treatment response defined as >50% reduction in depression severity from initiation to discharge. Remission defined as cut of ≤7 in HAMD or ≤ 12 in MADRS at end of treatment. [E]	Search of randomised, double-blind, placebo-controlled studies of antidepressant medications vs placebo or another antidepressant in MDD in Medline, EMBASE, Web of Science, CINAHL, PsycINFO, CENTRAL and numerous clinical trials databases. Sensitivity, specificity, positive and negative predictive values, false positives and negatives and diagnostic odds ratio were calculated for all antidepressants, as well as separated by class and individual drug. Meta-analysis and meta-regression calculated odds ratios to test the association between early improvement and baseline severity and outcomes.	Adults with acute MDD according to DSM-IV, DSM-III-R or DSM-III receiving antidepressant medication in RCT setting. $n = 17$ studies included in sensitivity/specificity calculations 14,779 participants	**Specific predictor:** Early improvement of all antidepressant medications combined predicted response (sensitivity 83%, specificity 54%, DOR 19.15 [95% CI 18.4–19.9]) and remission (sensitivity 86%, specificity 42%, DOR 15.70 [95% CI 14.9–16.5]) at endpoint, with high sensitivity and low-to-moderate specificity. Sensitivities and specificities felt to be similar between antidepressants except SSRIs, where early improvement may have lower predictive value. Early improvement of placebo group predicted response (sensitivity 79%, specificity 67%, DOR 27.29 [95% CI 25.0–29.7]) and remission (sensitivity 76%, specificity 61%, DOR 15.82 [95% CI 14.3–17.4]) at endpoint. This suggests differential response between placebo and pharmacotherapy arms for response but not remission. **Within-group predictive variable:** Patients with early improvement were more likely to achieve response (pooled OR = 8.37, $p < .000$) or remission (OR = 6.38, p < .000) compared to those without early improvement. The effect of baseline severity on response ($p = .744$) and remission ($p = .285$) was reported as explaining 8.1% of variance of odds ratios for remission.	Critically low

[Prognostic factor: a variable which moderates response but does not interact with treatment; Specific predictor: a variable which affects outcome differentially for active intervention compared to placebo; Effect modifier: a variable which moderates response and interacts with treatment; Within-group predictive variable: a variable which predicts within-group effects but is not compared across treatment arms or placebo].

[a] [E] = Efficacy, [T] = Tolerability, [A] = Acceptability.

[b] Rating determined using AMSTAR-2 instrument.

Table 5
Reviews of individual predictor variables lacking appropriate quality, transparency or rigour (classified as “level 3” evidence).

Author & year	Intervention	Variable(s) of interest	Outcome[a]	Population	Sample	Interpretation	Reason classified as level 3
Chen et al., 2018	Amitriptyline or placebo	Gender, presence of other adverse effects	Standardised scales of sexual function [T]	Adults enrolled in a randomised-trial setting. 80% had depression, other diagnoses included irritable bowel syndrome and interstitial cystitis.	$n = 8$ randomised controlled trials 685 participants receiving amitriptyline, 418 controls	**Specific predictor** Odds ratio for sexual dysfunction for male patients was 2.6 ($\chi2 = 6.03$, $p < .025$) and 0.37 ($\chi2 = 4.27$, $p < .05$) for female patients. **Within-group predictive variable** Positive linear correlation between sexual dysfunction and insomnia ($r^2 = 0.996$, F = 231.5, $p < .05$) and a biphasic correlation with somnolence ($r^2 = 0.9342$, F = 56.8, $p < .01$) and nausea ($r^2 = 0.9107$, F = 30.6, $p < .05$). No correlation with all other adverse effects including dry mouth, constipation, tremor, and agitation.	Heterogeneous population
Cooper & Conklin, 2015	Psychotherapy with or without pharmacotherapy or placebo	Sex, race, cohabitation status, age and concurrent personality disorder	Percentage dropout. [A]	Adults with a diagnosis of MDD or post-partum depression enrolled in a randomised trial in an outpatient setting.	$n = 54$ articles 3394 patients, across 80 psychotherapy treatment conditions	**Pooled analysis predictive variable** Percentage of patients of minority racial status predicted higher dropout (Freeman-Turkey estimate: 0.432, 95% CI: 0.04, 0.82). Percentage of patients with comorbid personality disorder was strongly predictive of higher dropout, but only provided in approximately one quarter of treatment conditions (Freeman-Turkey estimate: 0.976, 95% CI: 0.56, 1.39). Sex, cohabitation status and age not strongly associated with dropout.	Heterogeneous interventions
Cuijpers et al., 2012	Pharmacotherapy, psychotherapy or combined	13 characteristics, including type of depression, sociodemographic variables (older women, poor minority women, cohabitation status), co-morbidities.	Efficacy, measured by change on standardised instrument (e.g. HAM-D, BDI). [E]	Participants with an established depressive disorder (including dysthymia) enrolled in a randomised trial.	$n = 52$ studies 4734 participants	**Effect modifier** Pharmacotherapy more effective than psychotherapy in patients with dysthymia ($g = -0.28$; 95% CI: -0.53-0.04), postnatal depression and infertile women, however. No strong association with other predictors, although analyses generally lacked statistical power. Combination more effective than pharmacotherapy in older patients, chronic depression, treatment-resistant depression, impaired cognitive function and comorbid borderline personality disorder. No strong association with other predictors, but analyses generally lacked power. Combination more effective than psychotherapy in chronic depression. No strong association with other predictors, but analyses generally lacked power. Many findings based on one study, so caution advised in interpretation.	Heterogeneous population
Cuijpers, de Wit, Weitz, Andersson, & Huibers, 2015	Pharmacotherapy, psychotherapy, placebo or combination.	Baseline severity (HAM-D)	Efficacy, measured by change on standardised instrument (e.g. HAM-D, BDI). [E]	Adults enrolled in a randomised controlled trial diagnosed with major depressive disorder (including dysthymia) or with depressive symptoms above cut-off on a rating scale.	$n = 53$ studies 4740 participants	**Pooled analysis predictive variable** No evidence that baseline HAM-D severity was a predictor of within-group outcome (coefficient 0.03, CI:-0.02–0.08).	Heterogeneous interventions & population.

(continued on next page)

Table 5 (*continued*)

Author & year	Intervention	Variable(s) of interest	Outcome[a]	Population	Sample	Interpretation	Reason classified as level 3
Cuijpers, Ebert, Acarturk, Andersson, & Cristea, 2016	Psychological therapy	27 characteristics, including age, occupation, co-morbidities, income, caregivers, depression severity, gender.	Efficacy, measured by change on standardised instrument (e.g. HAM-D, BDI). [E]	Adults enrolled in a randomised controlled trial diagnosed with major depressive disorder (including dysthymia) or with depressive symptoms above cut-off on a rating scale.	$n = 41$ studies 2741 participants	**Effect modifier** Characteristics were limited to those with sufficient power for comparisons to find a clinically significant effect size. Three characteristics were found to be associated with response; CBT more effective than other psychotherapies in older adults ($g = 0.29$), in patients with comorbid alcohol problems ($g = 0.31$), and in university students ($g = 0.46$). However, authors advise caution due to inclusion of studies with high risk of bias.	Heterogeneous population
Ebert et al., 2016	Internet-based guided self-help or control (waiting list, treatment-as-usual)	Sex, age, education, co-morbid anxiety, depression severity	Deterioration, defined as a significant increase on standardised instrument (e.g. BDI) [E]	Adults with a diagnosis of depression in an acute depressive episode.	$n = 18$ RCTs 2079 participants	**Within-group predictive variable** Lower vs moderate/high education level was a predictor of within-group deterioration (low education RR: 1.72, 95% CI: 0.53–5.61, moderate/high education RR: 0.39, 95% CI: 0.22–0.68). No strong evidence that other variables predicted within-group deterioration.	Heterogeneous intervention
Fischer, Strawbridge, Vives, & Cleare, 2017	Psychological therapy	Cortisol (hair, urine, saliva or blood)	Response to therapy, measured by symptom rating scales (e.g. HAM-D, BDI). [E]	Patients with a diagnosis of a depressive disorder as per DSM or ICD criteria.	n = 8 studies 212 participants	**Pooled analysis predictive variable** Higher basal cortisol predicted poor within-group response to treatment (mean ES = 0.264, 95% CI 0.047–0.481, Z = 2.382, $P = .017$).	Heterogeneous interventions, population includes adolescents
Johnsen & Friborg, 2015	CBT or control (waiting list, treatment-as-usual)	Gender, age, proportion of participants on medication, comorbidity, baseline depression severity.	Response to therapy, measured by symptom rating scales (e.g. HAM-D, BDI). [E]	Adults with unipolar depressive disorder enrolled in trial setting.	$n = 70$ studies 2426 participants	**Within-group predictive variable** Percentage of men enrolled in study predicted worse treatment effect (change coefficient: −0.0104, 95% CI: 0.019, 0.001). Other variables were not strongly associated with outcome. Caution advised in interpreting results due to small number of studies.	Heterogeneous population
Karyotaki et al., 2015	Psychological web-based intervention	Gender, age, education, employment status, relationship status, comorbid anxiety, depression severity.	Dropout from intervention arm. [A]	Adults with a diagnosis of depressive disorder in an RCT setting.	$n = 10$ studies 2705 participants	**Pooled analysis predictive variable** Male gender (RR 1.08, CI 1.03–1.13), lower educational level (primary education: RR 1.26, CI 1.14–1.39), older age (RR 0.94, CI 0.87–1.02) and comorbid anxiety (RR 1.18, 95% CI 1.01–1.38) predicted treatment dropout, but effect sizes were small. Baseline depression severity, employment status and relationship status were not strongly associated with dropout.	Heterogeneous interventions
Nanni, Uher, & Danese, 2012	Pharmacotherapy, psychotherapy, combination or placebo	Childhood maltreatment (physical abuse, sexual abuse, neglect, family conflict or violence)	Response to therapy, using standardised symptom rating scales or categorical outcome of response or remission. [E]	Participants with a diagnosis of depressive disorder in population or community samples.	n = 10 trials 3098 participants	**Effect modifier** In pharmacological and combination treatments, maltreatment history predicted poorer outcome (odds ratio = 1.26, 95% CI = 1.01–1.56) and (odds ratio = 1.90, 95% CI = 1.40–2.58) respectively. In psychotherapy treatment, maltreatment history not strongly associated with poorer outcome (odds ratio = 1.12, 95% CI = 0.68–1.85). However, no direct statistical analysis between different treatments. **Pooled analysis predictive variable** Individuals with a history of maltreatment more likely have poor within-group treatment outcome (odds ratio = 1.43, 95% CI = 1.11–1.83).	Heterogeneous interventions, population includes adolescents

Palpacuer et al., 2017	Psychotherapy, waiting list, treatment-as-usual, placebo	Age, gender	Response to therapy, using clinician or self-report standardised scale (HDRS, MADRS, BDI) [E]	Adults diagnosed with depression as defined or with symptoms scored on standardised scale. Excluded studies with a large number of older adults (>65), postnatal or atypical depression.	$n = 84$ studies 4213 (intervention arm) & 2617 (control) participants	**Pooled analysis predictive variable** Both age (coefficient − 0.02, CI: −0.04;0.00) and gender (coefficient 0.00, CI:-0.01;0.01) not strongly associated with treatment outcome.	Heterogeneous interventions
Strawbridge et al., 2015	Pharmacotherapy or psychotherapy	IL6, TNF-alpha, CRP	Treatment response, defined as ≥50% reduction in symptoms on standardised instrument. [E]	Adults in a depressive episode as determined by clinician-rated standardised measure.	$n = 35$ articles 1908 participants	**Pooled analysis predictive variable** No strong differences identified between responders and non-responders for baseline TNF-alpha (effect size: 0.08, 95% CI: 0.34, 0.17), CRP (effect size: 0.03, 95% CI: 0.22, 0.16), and IL-6 (effect size: 0.83, 95% CI: 0.41, 2.07).	Heterogeneous interventions
Tunvirachaisakul et al., 2018	Any treatment for depression, treatment-as-usual, placebo	Age, depression severity, co-morbid anxiety, physical comorbidity and executive functioning.	Treatment outcome categorised as response, remission or change in score or severity on a depression questionnaire. [E]	Adults over the age of 60 with a diagnosis of major depression according to DSM/ICD enrolled in a randomised trial.	$n = 67$ studies	**Effect modifier** Age, baseline depression and anxiety and physical comorbidity appeared to have a small moderator effect in a subgroup analysis when interventions grouped by biological, psychosocial or both. **Pooled analysis predictive variable** 65 predictor variables reported in individual studies, of which seven were reported in over three studies. These were age, baseline severity, early improvement, current episode duration, baseline anxiety, physical comorbidity and set shifting in the trail making test. In meta-analysis, only baseline anxiety, baseline depression and the trail-making test were strong predictors of within-group treatment response. Significant publication bias noted by authors.	Heterogeneous interventions
Widge et al., 2019	Any treatment for depression	Quantitative EEG analysis	Clinical response to antidepressant treatment. [E]	Human subjects receiving a treatment modality for depressive illness.	$n = 76$ articles	**Pooled analysis predictive variable** Overall meta-analytic sensitivity was 0.72 (CI = 0.67–0.76), specificity was 0.68 (CI = 0.63–0.73), and log(diagnostic odds ratio) was 1.89 (CI = 1.56–2.21), corresponding to area under the curve of 0.76 (CI = 0.71–0.80). **Effect modifier** Meta-analytic subgroup analysis found no strong difference for predictive utility between treatment groups. Diagnostic odds ratio for rTMS (2.19 CI 1.22;3.15), pharmacotherapy (1.89 CI 1.53;2.26) and other treatment modalities (1.37 CI 0.26;2.49).	Heterogeneous interventions

[Prognostic factor: a variable which moderates response but does not interact with treatment; Specific predictor: a variable which affects outcome differentially for active intervention compared to placebo; Effect modifier: a variable which moderates response and interacts with treatment; Within-group predictive variable: a variable which predicts within-group effects but is not compared across treatment arms or placebo; Pooled analysis predictive variable: a variables which predicts response in a pooled grouping of multiple treatment interventions and/or placebo].

[a] Efficacy = [E], Tolerability = [T], Acceptability = [A].

Table 6
Included reviews of clinical prediction models.

Author & year	Review aim(s)	Methods	Population	Intervention	Candidate predictors in model(s)	Outcome predicted[*], including timespan	Prediction model & evaluation	Validation	Interpretation	Quality[**]
Bos, et al., 2015	To review all literature relating to experience sampling and ecological momentary assessment and psychotropic medications for DSM-III-R & DSM-IV disorders	Searched PsycINFO & MEDLINE databases for relevant references. Inclusion criteria included the examination of a pharmacological intervention, involvement of experience sampling method/ecological momentary assessment (defined as repeated measures outside the laboratory more than once a day for over one day) in presence of DSM-III-R or DSM-IV diagnosis.	Identified 7 studies in MDD adult population, derived from three distinct samples. 133 participants in total with MDD or acute unipolar depression. One study featured a multivariable clinical prediction model. **For relevant study (1):** 49 patients with DSM-IV diagnosis of MDD.	Pharmacotherapy alone or in combination with supportive pharmacotherapy. **For relevant study (1):** imipramine vs placebo)	**For relevant study (1):** Early changes in HDRS, negative affect and positive affect, measured at week one following treatment initiation.	**For relevant study (1):** Response, defined as 50% reduction in HDRS from baseline to week 6 and remission, defined as HDRS score ≤7 at week 6. Continuous HDRS score also recorded. [E]	**Type of prediction model:** **For relevant study (1):** Linear regression with covariates and logistic regression used for dichotomous outcomes. Predictor variables combined in model and compared to predictions based on single variables alone. **Model evaluation:** **For relevant study (1):** Model combining early change in positive affect with early change in HDRS was compared to predictions based on single variables alone. Not validated on external sample.	**For relevant study (1):** Not externally validated, nor internally validated on a separate test-set.	Methodology of review considered both individual predictor variables and prediction models. One study (Geschwind t al., 2011) focused on a clinical prediction model. In a sample of 49 depressed patients, the association between change in positive and negative emotions and severity of depression at week 6 was examined. Early change in HDRS combined with early change in positive affect improved prediction of response and remission compared to early HDRS change alone (response: chi-square = 6.24, $p < 0.05$; remission chi-square = 14.72, $p < 0.001$). Proportion of explained variance for "combined early prediction model" 28% for response and 40% for remission.	Critically low
Lee et al., 2018	To review literature on the use of machine-learning algorithms to predict therapeutic outcomes in mood disorder populations.	Searched MEDLINE, Cochrane Library, ClinicalTrials.gov & Google Scholar for prospective or retrospective, open-label or controlled clinical study (+/- randomisation or blinding) which applied a machine learning algorithm to assess patient or group level data as predictors of a therapeutic outcome	Adults in depressive episode receiving a treatment intervention for depression. Included both bipolar and unipolar depression in quantitative synthesis, but results presented separately in qualitative synthesis and	Pharmacotherapy (SSRI, TCA, selective-noradrenaline reuptake inhibitor) (n=16), psychotherapy (n=1), neuromodulation (rTMS, tDCS, electroconvulsive therapy) (n=6), combined pharmacotherapy & antidepressants) (n=2).	Neuroimaging (EEG n=5, MRI or fMRI n=8), phenomenological (n=8, including overall mood symptom severity, anxiety, anhedonia, functioning, number of previous mood episodes and demographic variables including employment, education, household income), genetic (n=3), combined (n	Change in depression related outcome, defined as any standardised measure as a proxy of therapeutic improvement (including patient and clinician-rated scales, hospital admission frequency or suicidal ideation. [E]	**Type of prediction model:** Supervised-learning algorithms in 92% of included studies. Logistic regression, support vector machinet, neural network, decision trees, mixture of factor analysis, linear discriminant analysis, random forests, gradient boosting machine. **Model evaluation:**	Only 4 of 20 studies evaluated model performance in a dataset not used in training or cross-validation. One study evaluated model performance with hold-out validation.	**Quantitative:** Pooled classification accuracy for phenomenological predictors: 0.76 [0.63; 0.87]. Significant heterogeneity. Pooled classification accuracy for combined prediction models (phenomenological in combination with genetic or neuroimaging): 0.93 [0.86; 0.97]	Critically low

(continued on next page)

Table 6 (*continued*)

Author & year	Review aim(s)	Methods	Population	Intervention	Candidate predictors in model(s)	Outcome predicted*, including timespan	Prediction model & evaluation	Validation	Interpretation	Quality**
		or cluster subjects based on a therapeutic outcome to devise a predictive model Measures of classification accuracy were reported & pooled, including percentage rate, receiver operating characteristic, area-under-curve.	some of quantitative synthesis. n = 26 studies (qualitative synthesis), 17,449 participants n = 20 studies (quantitative synthesis), 6325 participants All studies included MDD patients, two studies including BD patients.		= 2, e)		Classification accuracy (reported as percentage rate in all but one study reporting receiver operating characteristic & area-under-curve), sensitivity and specificity. Classification accuracy assessed by nested, leave-n-out or k-fold cross-validation in all but two studies.		Not possible to comment on pooled classification accuracy for neuroimaging & genetic predictors, or overall, due to inclusion of bipolar disorder patients.	
Rosenblat et al., 2017	To review the impact of pharmacogenomic clinical prediction models on clinical outcomes and cost-effectiveness in MDD.	Searched MEDLINE and Google Scholar for primary studies or reviews evaluating the impact of pharmacogenomic testing on MDD treatment outcomes and cost-effectiveness or utility. No restrictions placed on study quality, randomisation or control group.	Adults with MDD receiving treatment guided by pharmacogenomic testing. n=5 clinical trials evaluating efficacy, n=5 studies evaluating cost-effectiveness.	Pharmacotherapy guided by commercial pharmacogenomic testing (GeneSight, Geneecpt, CNSDosing), based on pharmacodynamics and pharmacokinetic profile of participants.	**GeneSight:** Genotyping of five copies of both genes based on pharmacodynamic considerations of treatment response or antidepressant metabolism. Genes: CYP2D6, CYP2C19, CYP1A2, SLC6A4, HTR2A T102C SNP. **Genecept:** Tests for 22 genes and copy number, associated with psychiatric presentation, treatment efficacy or adverse reactions. **CNSDosing:** SNPs not reported in paper	Review assessed changes in depression severity, response or remission rates as a result of prescribing guided by pharmacogenomic testing [E]	**Type of prediction model:** Commercial pharmacogenomics tools, making predictions about treatment response or tolerability, or making treatment recommendations guided by these factors. **GeneSight:** Tool generates report advising clinicians to 'use as directed', 'use with caution' and 'use with caution or more frequent monitoring **Genecept:** Report generated by pharmacogenomic tool not discussed in detail in paper. **CNSDosing:** Tool generated report advising on dosing of antidepressant.	Focus of review was on clinical validation of pharmacogenomic models. Main outcome (depression severity, response and remissions) were a proxy for this. Classification accuracy not directly reported. Findings often not independently replicated.	4 controlled studies, 1 uncontrolled study. **GeneSight:** Two open-label, nonrandomised studies identified GeneSight improving response and remission, but this was not replicated in a randomised, controlled blinded study. **CNSDosing:** Randomised-controlled trial of CNSDosing showed association with remission. **Genecept:** Uncontrolled prospective cohort study of Genecept Assay lacked a control group.	Critically low

Model evaluation: Pharmacogenomic models tested in MDD populations to assess if guided treatment superior to unguided treatment. Remission or response rates or improvement in depression severity compared between populations with and without pharmacogenomic tool.

[*] [E] = Efficacy, [T] = Tolerability, [A] = Acceptability

[**] As determined using AMSTAR-2 instrument.

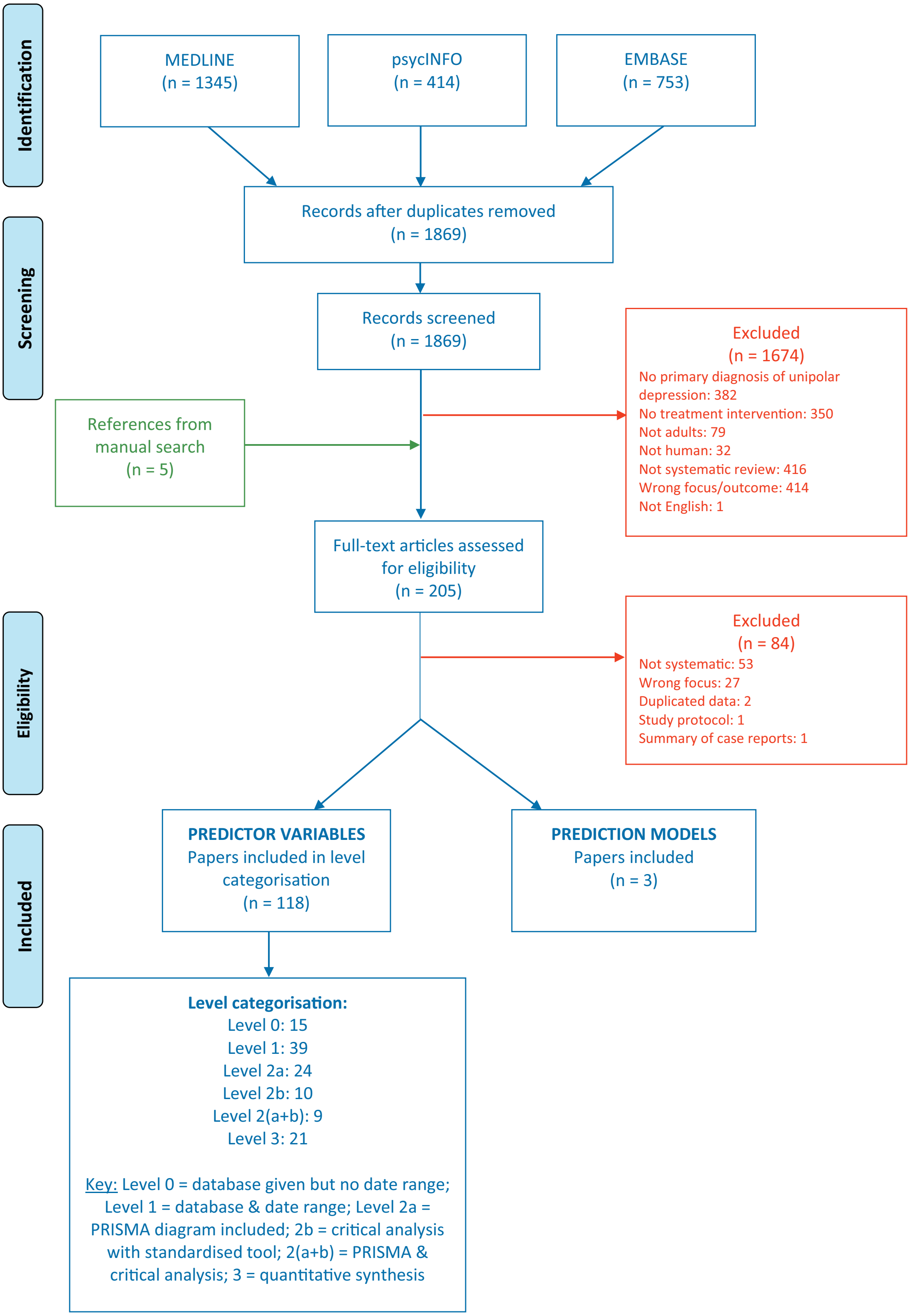
Identification
Screening
Eligibility
Included
MEDLINE
(n = 1345)
psycINFO
(n = 414)
EMBASE
(n = 753)
Records after duplicates removed
(n = 1869)
Records screened
(n = 1869)
Excluded
(n = 1674)
No primary diagnosis of unipolar depression: 382
No treatment intervention: 350
Not adults: 79
Not human: 32
Not systematic review: 416
Wrong focus/outcome: 414
Not English: 1
References from manual search
(n = 5)
Full-text articles assessed for eligibility
(n = 205)
Excluded
(n = 84)
Not systematic: 53
Wrong focus: 27
Duplicated data: 2
Study protocol: 1
Summary of case reports: 1
PREDICTOR VARIABLES
Papers included in level categorisation
(n = 118)
PREDICTION MODELS
Papers included
(n = 3)
Level categorisation:
Level 0: 15
Level 1: 39
Level 2a: 24
Level 2b: 10
Level 2(a+b): 9
Level 3: 21
Key: Level 0 = database given but no date range; Level 1 = database & date range; Level 2a = PRISMA diagram included; 2b = critical analysis with standardised tool; 2(a+b) = PRISMA & critical analysis; 3 = quantitative synthesis

Fig. 1. PRISMA flow diagram of meta-review process