PLOS One

STUDY PROTOCOL

How trustworthy and applicable is the evidence from systematic reviews of depression treatments: Protocol for systematic examination

Iwo Fober[1], Lidia Baran[1,2], Myrto Samara[3], Spyridon Siafis[4], David Robert Grimes[5], Bartosz Helfer[1,2]*

1 Meta-Research Centre, University of Wroclaw, Wroclaw, Poland, **2** Institute of Psychology, University of Wroclaw, Wroclaw, Poland, **3** Department of Psychiatry, Faculty of Medicine, University of Thessaly, Larissa, Greece, **4** Technical University of Munich, TUM School of Medicine and Health, TUM University Hospital, Department of Psychiatry and Psychotherapy, Munich, Germany, **5** TCD Biostatistics Unit, School of Medicine, Trinity College Dublin, Dublin, Ireland

* bartosz.helfer@gmail.com

Abstract

Background

Depression is a common mental disorder significantly impacting daily functioning. Standard treatments include drugs, psychotherapies, or a combination of both. Treatment selection relies on scientific evidence, though the trustworthiness and applicability of this evidence can vary.

Objectives

This protocol presents a method to evaluate evidence from systematic reviews for pharmacological and psychological treatments for depression, focusing on trustworthiness and applicability structured into five components: quality of conduct and reporting, risk of bias, spin in abstract conclusions, robustness of meta-analytical results, heterogeneity and clinical diversity.

Methods

We will conduct a systematic search of systematic reviews in MEDLINE, Embase, PsycInfo, and Cochrane Database of Systematic Reviews. Our focus will be on systematic reviews of first-line treatments for depression in adults, including antidepressants, psychotherapy, or combined treatments, compared to either active or inactive comparators. We will extract information needed for a comprehensive methodological evaluation using qualitative tools, including AMSTAR 2, ROBIS, Conflict-of-Interest assessment, Referencing Framework for SRs, Spin Measure, and heterogeneity exploration assessment. For quantitative analyses, such as Fragility Index, Ellipse of Insignificance, Region of Attainable Redaction, GRIM test, Leave-N-Out analysis,

OPEN ACCESS

Citation: Fober I, Baran L, Samara M, Siafis S, Grimes DR, Helfer B (2025) How trustworthy and applicable is the evidence from systematic reviews of depression treatments: Protocol for systematic examination. PLoS One 20(6): e0325384. https://doi.org/10.1371/journal.pone.0325384

Editor: Johannes Hönekopp, Northumbria University, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Received: February 13, 2025

Accepted: May 12, 2025

Published: June 6, 2025

Data availability statement: All data are in the manuscript and/or Supporting information files.

Funding: This research was funded in whole by the National Science Centre, Poland (grant number 2021/41/B/HS6/02844).

Competing interests: The authors have declared that no competing interests exist.

and prediction intervals, we will select and recalculate two meta-analyses per review. We define a set of outcomes to enable practical and intuitive interpretation of these analyses' results. Descriptive statistics, non-parametric statistical tests, and narrative summaries will be used to synthesize and compare outcomes across several pre-specified subgroups.

Expected outcomes

We expect these analyses to provide an enhanced perspective on the practice of evidence synthesis in the field of mental health, offer methodological guidance for future systematic reviews and meta-analyses, and contribute to improved informed decision-making by clinicians and patients.

OSF registration

osf.io/7f9cj and osf.io/ynejs

Introduction

Background

With approximately 280 million people worldwide affected by depressive disorders [1], there is a constant demand for effective interventions. According to evidence-based practice (EBP) and evidence-based medicine (EBM), the utilization of the best evidence in clinical decision-making is not only ethical but also necessary to address worldwide treatment demands [2,3]. The ethicality of applying these approaches in mental health is subject to debate [4], nevertheless systematic reviews (SRs) and randomized controlled trials (RCTs) still occupy the highest tiers in many evidence hierarchies and are instrumental in forming recommendations and treatment guidelines, including psychotherapy and pharmacotherapy for depression [5–9]. However, despite decades of research on depression treatments [10,11], concerns persist regarding the evidence base.

For instance, RCTs' reports often lack detailed descriptions of control conditions such as 'treatment as usual,' [12] and industry-funded studies tend to report higher effectiveness, especially in pharmacological research [13]. Many studies do not reflect real-world conditions [14,15], and underpowered trials limit the detection of true treatment effects [16]. Additionally, the risk of bias due to issues such as unblinded designs or inadequate randomization is common [17,18]. A high risk of bias and low-quality of evidence from trials can, in turn, lead to incorrect estimation of treatment effects [19].

In evidence synthesis, it is essential to address and account for the limitations of primary studies when drawing conclusions; otherwise, the findings may be misleading. However, in addition to inheriting some methodological flaws from the primary studies, SRs also face review-specific challenges that reduce their trustworthiness and applicability. Specifically, SRs are susceptible to multiple factors impacting their quality of conduct and reporting. These include issues at the design stage (e.g., failure to account for previous reviews, lack of prespecified methods, conflicts of interest), with prospectively registered protocol being particularly important [20], and execution stage (e.g., insufficiently comprehensive search strategies, single study selection and data extraction) [21]. The risk of bias might be introduced to SR by inappropriate or unclear inclusion criteria, or flawed data synthesis and analysis, among others [22]. Of particular concern are inaccuracies in interpretation, which can lead to spin – the presentation or interpretation of findings in a way that emphasizes favourable results or downplays unfavourable ones, potentially leading to biased conclusions [23]. When a quantitative synthesis is performed, the robustness of overall effect size estimation should receive as much attention as its magnitude and statistical significance, as many meta-analyses are fragile to even slight changes in trials results, or their statistical significance relay on a single study [24,25]. Robustness may also be compromised when data in primary studies are redacted or when studies with reporting anomalies are excluded – issues that have increasingly drawn the attention of the public and researchers in recent years, particularly in psychology and biomedical sciences [26–28]. Finally, evidence-based approaches have been criticized for focusing on average treatment effects, which may not adequately reflect individual patient outcomes due to variability in responses [29,30]. In addition, the diagnosis of mental health issues, which forms the basis for inclusion in clinical trials, may rely on self-reported scales, clinical judgment, or structured interviews based on various sets of criteria that are periodically updated. The lack of biomarkers further complicates the standardization of patient samples. Moreover, psychological treatments are complex interventions, which are difficult to standardize and may be compared against a range of heterogeneous comparators (e.g., other psychotherapies, waiting lists, or treatment as usual). To address these limitations, it is crucial to sufficiently account for between-study heterogeneity in SRs while considering the clinical diversity of patients seeking help for mental health issues. This approach is essential for gaining a better understanding of treatment effects, ensuring the generalizability of evidence, and translating it into practice effectively [31–33].

Hypotheses

Based on the overview of common issues in systematic reviews [34,35], and drawing on prior analyses conducted in other areas of medicine [36–42] and mental health [43–46], showing critically low quality of 53%–99% and 68%–88% SRs respectively (see S1 Appendix), we hypothesize that a significant proportion of SRs on depression treatments will show low overall quality of conduct and reporting, with issues like lack of pre-registered protocols, incomplete risk of bias evaluations, and insufficient justification for excluded studies. Most SRs will demonstrate low or unclear overall risk of bias, with some shortcomings in reporting eligibility criteria and search strategies, reflecting previous findings [47–49]. A substantial portion of SRs will exhibit spin, potentially distorting the interpretation of findings, aligning with previous evidence from psychotherapy [50] and adolescent depression trials [51]. Many meta-analyses will display low robustness, where statistically significant results could be easily reversed with minimal changes in trial events [24], or the removal of a single study significantly alters the overall effect, undermining the stability of conclusions [52,53]. Supplementing meta-analytic results with prediction intervals (PIs) will alter or add to conclusions about the safety and efficacy of depression treatments. Specifically, in some reviews PIs will be much wider than reported confidence intervals, offering a broader perspective on result uncertainty, potentially including null effects or effects in the opposite direction to those reported, capturing both positive and negative effects of similar size within the interval. We also anticipate that clinical diversity and statistical heterogeneity will be inadequately addressed, limiting the generalizability of findings [31–33,54–56]. Conclusions of these analyses will vary between subgroups of reviews defined by factors like interventions, comparators, or methodological quality.

Anticipated new evidence

This study aims to evaluate trustworthiness and applicability structured into five components: quality of conduct and reporting, risk of bias, spin in abstract conclusions, robustness of meta-analytical results, heterogeneity and clinical diversity.

The qualitative assessment was designed to replicate and expand on findings from analyses partially overlapping with our project [43–46]. To date, SRs in mental health have primarily been evaluated using the AMSTAR tool. Our analysis incorporates ROBIS, as well as additional tools that enable a more nuanced examination of key concepts we believe warrant closer attention – namely referencing, conflicts of interest, spin, and heterogeneity exploration practices.

To the best of our knowledge, the quantitative assessments we propose have not been applied to such a large body of evidence on depression or any other mental disorder. Examining the fragility of meta-analyses will enrich the assessment of evidence certainty by adding a new dimension of robustness – an aspect whose importance is increasingly recognized in other areas of medicine, both in the evaluation of primary studies [57] and meta-analyses [24,58–61], as well as in its influence on clinical guidelines [62–67]. In addition to calculating the fragility for the meta-analysis, we will also contextualize it using dropout rates from the included clinical trials. Calculating prediction intervals will enhance understanding of the impact of heterogeneity on conclusions drawn from meta-analyses of depression treatments [68]. To interpret them meaningfully, we adopted the most practical approaches based on analysing the relationship between prediction intervals, confidence intervals, and the line of null effect. This is particularly valuable given the clear need among mental healthcare professionals for a framework that supports the implementation of EBM and EBP in clinical practice [69].

Through additional analyses, we will be able to offer further insight into certain issues specific to clinical research and evidence synthesis in psychiatry and clinical psychology, such as differences in evaluating pharmaceutical and psychological interventions for the same indication, and the impact of how the research question was framed or which comparator was selected.

Our approach, grounded in systematic and transparent evaluation using state-of-the-art tools that are rarely or never applied in this field, will provide a new, in-depth perspective on the implementation of EBP and EBM methods in mental health. Ultimately, this will reduce research waste by offering methodological guidance for future systematic reviews and meta-analyses, strengthen the evidence base for clinical guidelines, and support more informed clinical decision-making and personalized patient care.

Materials and methods

This observational meta-research study employs systematic approach for data collection, analysis, and reporting, and follows PRISMA guidelines [70]. A visual summary of the materials and methods is presented in Figs 1 and 2.

Given the exploratory nature of this analysis and the need to balance the significance of our study's conclusions with the feasibility of the project, we defined our eligibility criteria to obtain a sample of systematic reviews applicable to the broadest possible patient population, focusing on well-established first-line treatments and addressing the fundamental questions of efficacy and safety. Reviews must clearly indicate in the title or abstract that their primary focus is on treatments for depression or depressive symptoms, as defined below. To ensure a systematic and transparent study selection process while maintaining feasibility, we decided that eligibility for this study would be determined based on the inclusion and exclusion criteria prespecified in the reviews' methods sections, rather than the characteristics of the trials actually included. Table 1. contains a summary of eligibility criteria.

Inclusion criteria

Population. The inclusion criteria for this study will encompass systematic reviews focused on depressive disorders, regardless of how they are described or defined by the authors (e.g., 'depression,' 'major depressive disorder,' 'unipolar depression,' 'elevated depression symptoms'). We will include reviews based on self-reports, structured diagnostic criteria or clinical judgment, as well as those that do not specify the diagnostic criteria for depression in their inclusion criteria. Additionally, we will include reviews investigating specific features of depression in populations where the primary diagnosis is a depressive disorder. The target population is adults (over 18 years of age), and if age is not explicitly mentioned and the target group is unclear, we will assume the reviews focus on adults.

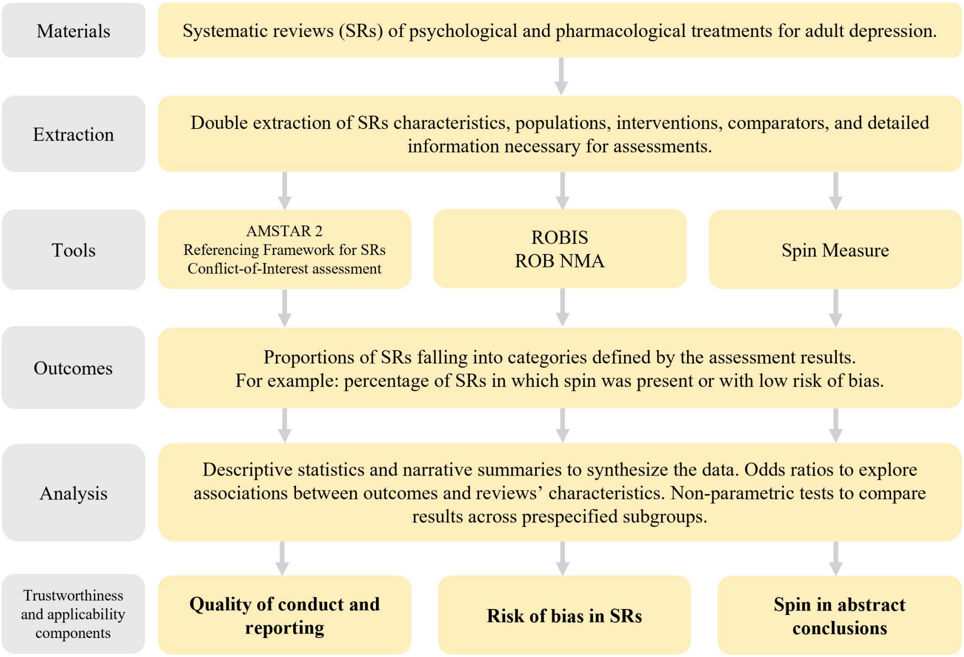
Materials
Systematic reviews (SRs) of psychological and pharmacological treatments for adult depression.
Extraction
Double extraction of SRs characteristics, populations, interventions, comparators, and detailed information necessary for assessments.
Tools
AMSTAR 2
Referencing Framework for SRs
Conflict-of-Interest assessment
ROBIS
ROB NMA
Spin Measure
Outcomes
Proportions of SRs falling into categories defined by the assessment results.
For example: percentage of SRs in which spin was present or with low risk of bias.
Analysis
Descriptive statistics and narrative summaries to synthesize the data. Odds ratios to explore associations between outcomes and reviews' characteristics. Non-parametric tests to compare results across prespecified subgroups.
Trustworthiness and applicability components
Quality of conduct and reporting
Risk of bias in SRs
Spin in abstract conclusions

Fig 1. Study workflow leading to the assessment of quality of conduct and reporting, risk of bias and spin in abstract conclusions.

https://doi.org/10.1371/journal.pone.0325384.g001

Interventions. We will include reviews on the acute treatment of depression. If the phase of treatment is not explicitly mentioned, we will assume the review is on acute treatment. Eligible pharmacotherapies include drugs commonly referred to as 'antidepressants,' whether in mono- or polytherapy. This includes individual drugs (e.g., 'sertraline', 'mirtazapine'), pharmacological groups (e.g., 'tricyclic antidepressants'), or antidepressants as a drug class. Additionally, we will include psychological treatments in the form of psychotherapies ('talking therapies' that consist mainly of verbal communication with specialist) of all formats (e.g., 'individual', 'group', or 'family therapies') and modes of delivery (e.g., 'face-to-face', 'videoconference', 'telephone call'). Eligible theoretical approaches will include but will not be limited to cognitive behavioural therapy, third-wave cognitive behavioural therapy (e.g., 'dialectical behaviour therapy,' 'acceptance and commitment therapy'), problem-solving therapy, interpersonal therapy, psychodynamic therapy, behavioural activation therapy, and life review therapy. Reviews of combinations of all the above treatments (used simultaneously or sequentially) will be included.

Comparators. We will include reviews that consider any of the interventions described above as an active comparator. Additionally, control conditions such as 'care as usual,' 'minimal treatment,' 'no treatment,' 'placebo pill,' 'psychological placebo,' and 'waiting list' will be included.

Outcomes. The focus of eligible reviews must be on safety and/ or efficacy, measured by any related outcomes (e.g., 'response,' 'remission,' 'symptom reduction,' 'dropouts' rate,' 'adverse events'). We will include reviews aiming to explore moderators or predictors only if the overall effect size is reported.

Trials' design. Only reviews of RCTs will be included

Review type and data availability. In the qualitative assessment (i.e., quality of conduct and reporting, risk of bias, spin in abstract conclusions), we will include any study meeting the eligibility criteria and identified by the authors as a systematic review, regardless of whether it incorporates a meta-analysis. Additionally, studies that broadly follow a systematic approach will also be eligible. Specifically, they should address a clearly defined research question. It should be evident that the included studies were identified through searches conducted in medical databases or clinical trial

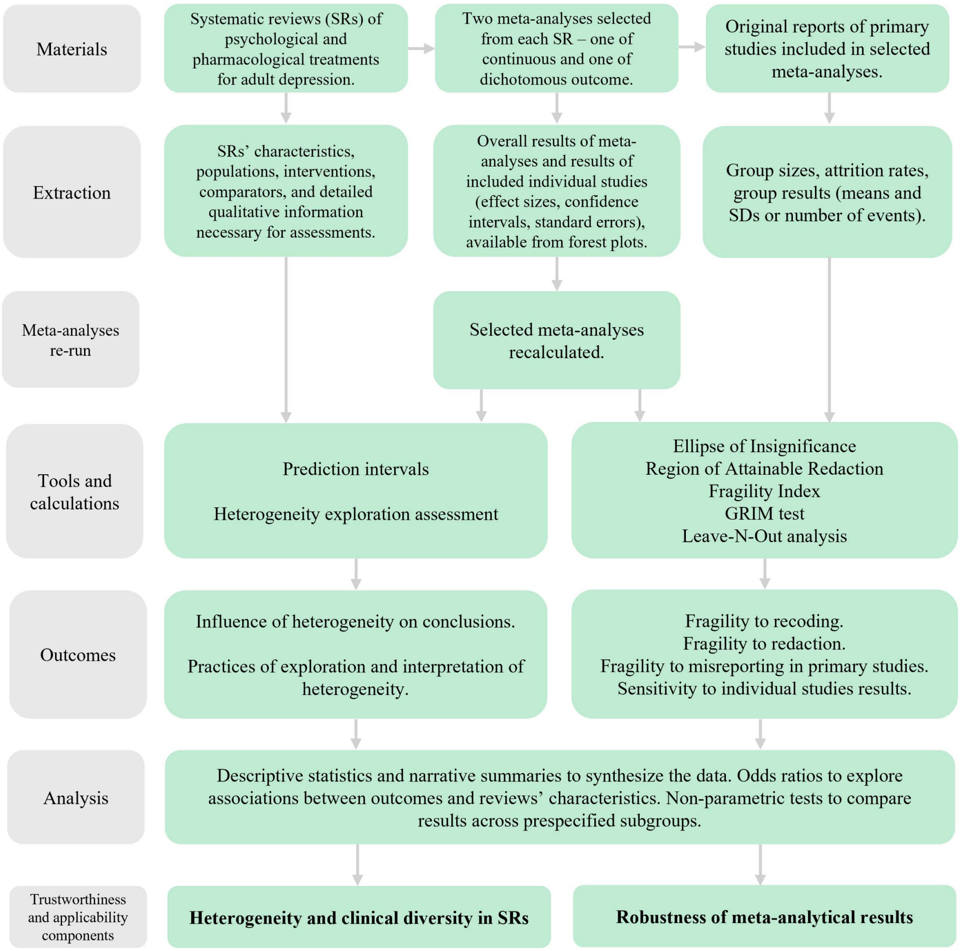
Materials
Systematic reviews (SRs) of psychological and pharmacological treatments for adult depression.
Two meta-analyses selected from each SR – one of continuous and one of dichotomous outcome.
Original reports of primary studies included in selected meta-analyses.
Extraction
SRs' characteristics, populations, interventions, comparators, and detailed qualitative information necessary for assessments.
Overall results of meta-analyses and results of included individual studies (effect sizes, confidence intervals, standard errors), available from forest plots.
Group sizes, attrition rates, group results (means and SDs or number of events).
Meta-analyses re-run
Selected meta-analyses recalculated.
Tools and calculations
Prediction intervals
Heterogeneity exploration assessment
Ellipse of Insignificance
Region of Attainable Redaction
Fragility Index
GRIM test
Leave-N-Out analysis
Outcomes
Influence of heterogeneity on conclusions.
Practices of exploration and interpretation of heterogeneity.
Fragility to recoding.
Fragility to redaction.
Fragility to misreporting in primary studies.
Sensitivity to individual studies results.
Analysis
Descriptive statistics and narrative summaries to synthesize the data. Odds ratios to explore associations between outcomes and reviews' characteristics. Non-parametric tests to compare results across prespecified subgroups.
Trustworthiness and applicability components
Heterogeneity and clinical diversity in SRs
Robustness of meta-analytical results

Fig 2. Study workflow leading to the assessment of heterogeneity and clinical diversity, and robustness of meta-analytical results.

https://doi.org/10.1371/journal.pone.0325384.g002

Table 1. Summary of eligibility criteria.

	Inclusion criteria	Exclusion criteria
Population	General adult population with depression.	Other, e.g., specific populations (older adults, postpartum depression, Latinos, students), bipolar or psychotic depression.
Intervention	Antidepressants, psychotherapy or combined treatment.	Other, e.g., antipsychotics, supplements, exercises, self-guided programs.
Comparators	Any treatment eligible as intervention, plus: placebo (pill and psychological), waiting list, treatment as usual, no or minimal treatment.	Other, e.g., complementary and alternative treatments, like acupuncture, physical exercises, light therapy.
Outcomes	Related to safety and/or efficacy.	Other, e.g., bias assessments.
Review type	Systematic reviews of randomized controlled trials (RCTs).	Other, e.g., systematic reviews including non-RCTs, methodological reviews, pooled analyses.

https://doi.org/10.1371/journal.pone.0325384.t001

registries, as opposed to, for example, analysing studies conducted by a single drug manufacturer. Eligibility criteria should be defined at least in terms of population, intervention, and comparator. The results should report the number and characteristics of the included studies. If no meta-analysis was performed, the authors should provide an unbiased narrative summary of the review's findings, covering aspects such as efficacy, safety, quality, or quantity of available evidence, and offering recommendations for clinical practice or future research. Reviews employing Network Meta-Analyses (NMAs) will also be included in the qualitative assessment.

For the quantitative analysis (i.e., robustness of meta-analytical results, heterogeneity evaluation with prediction intervals), we will include reviews that report pairwise meta-analyses, provided that the extraction of necessary data is possible either from forest plots or from the original reports of included primary studies. We aim to maintain a focus on direct pairwise comparisons with clear methodological assumptions. Therefore, Network Meta-Analyses (NMAs) and Individual Participants Data Meta-Analyses (IPDMAs) will be excluded from this part of the project. However, a recommended and widely adopted practice when conducting review with NMAs is to first perform a standard direct pairwise meta-analysis. Some IPDs also report such results. In these cases, NMAs and IPDMAs will be treated as sources of eligible meta-analyses, which will be extracted and included in the quantitative assessment as if derived from a standard SRs.

Exclusion criteria

Population. We will exclude reviews focused on treatment-resistant depression, psychotic depression, schizoaffective disorder, and bipolar depression. Additionally, reviews targeting specific adult groups (e.g., 'older adults,' 'late life depression', 'peripartum depression,' 'depression in patients with physical illness,' 'students,' 'specific races or ethnic groups') will be excluded. Reviews that include both eligible and ineligible populations will be excluded. Reviews with a broad (e.g., 'depression and anxiety disorders') or unspecified scope (e.g., 'common mental disorders') will be excluded. The exclusion of other affective disorders with a depressive component and specific populations is due to differences in presumed pathophysiology, clinical presentation, management, and prognosis [71–75]. Their inclusion would also conflict with our core objective of assessing pieces of evidence applicable to the broadest possible population and would negatively impact the feasibility of the study.

Interventions. We will exclude reviews focused on continuation and maintenance treatment, relapse, and recurrence prevention. Additionally, pharmacotherapies involving drugs other than antidepressants (e.g., 'antipsychotics', 'mood stabilizers'), phytopharmaceuticals, and dietary supplements (e.g., 'St. John's wort,' 'fatty acids'), as well as psychedelics (e.g., 'psilocybin', 'LSD', 'ketamine') and psychedelic-assisted psychotherapy, will be excluded, as they are not first-line treatments. Self-guided and Internet-delivered programs (without therapist engagement) will be excluded. Reviews focused on both the included and excluded therapies mentioned above will be excluded.

Comparators. Reviews considering comparators other than specified in the inclusion criteria for this study (e.g., 'acupuncture,' 'physical exercises', 'self-guided programs,' 'light therapy') or both eligible and ineligible comparators, will be excluded.

Outcomes. Reviews reporting solely outcomes irrelevant to safety and/ or efficacy will be excluded. We will exclude reviews exploring moderators and predictors (unless the overall effect size is reported) and methodological reviews (addressing aspects of depression research like 'differences in baseline severity' or 'types of outcomes measures').

Trials' design. Reviews including studies other than RCTs or both RCTs and non-RCTs, will be excluded.

Review type and data availability. We will exclude publications synthesising trials selected in a non-systematic manner (e.g., 'pooled analysis' of all trials performed by a manufacturer). IPDMAs will be excluded unless they report a pairwise meta-analysis, as described in the inclusion criteria.

Information sources and search strategy

We performed a comprehensive search in electronic databases, including MEDLINE (via PubMed), Embase, and the Cochrane Library, as the most extensive and widely used general medical databases for evidence synthesis, and PsycINFO, as a subject specific database – as recommended by Cochrane Handbook [76]. The search strategy combined general and specific terms associated with depression, psychotherapy, and pharmacotherapy. For PubMed and Embase we used a validated methodological filter for systematic reviews [77,78]. We used build-in filters for systematic reviews in

Cochrane Library and PsycNet. We combined free-text search terms with structured vocabularies (MeSH and Emtree). For the antidepressant search component, we used the strategy published in Cipriani et al [79]. No time or language restrictions were applied. Search strategies are reported in the S2 Appendix.

Reviews selection

Search results were pooled and deduplicated using EndNote software, following the Bramer Method [80]. The screening was facilitated by Covidence software and occurred in two stages: 1) title and abstract screening with two independent reviewers evaluating titles and abstracts for eligibility based on the inclusion and exclusion criteria; 2) full-text screening, where reviews were assessed independently by two reviewers, and discrepancies resolved through consensus (discussion or consultation with a third reviewer). The PRISMA Flow Diagram is reported in S3 Appendix. In total, we included 153 reviews (see S4 Appendix for list of included reviews and S5 Appendix for excluded reviews with reasons for exclusion).

Data extraction

Data extraction will be conducted independently by two reviewers, using standardized forms in Google Sheets. Disagreements will be resolved through consensus or, if necessary, by consulting a third reviewer. Extracted data will include SRs' characteristics, populations, interventions, comparators, and information required for qualitative assessments. List of extracted variables is reported in the S6 Appendix. We will also select two meta-analyses per review – one of a binary outcome and one of a continuous outcome, regardless of the measure of effect size (e.g., odds ratio, risk ratio, standardized mean difference). The following predefined selection approach will be applied. We will use a meta-analysis of the primary outcome relevant to efficacy or safety. If multiple comparisons are eligible, we will prioritize the comparison to inactive control conditions. If the primary outcome is not defined or not relevant, we will use the first relevant meta-analysis reported. If the first selected outcome is binary, we will select the first relevant continuous outcome as the second analysis, and vice versa. We will use one meta-analytic result if only one type of outcome is analysed. For the selected meta-analyses, we will extract overall results and the results of included individual studies (numbers of participants in experimental and control groups, means and standard deviations or numbers of events, number of dropouts in both groups). This will be done from forest plots (if available) or original study reports.

Assessment methods and tools

The tools we will use to assess each component of trustworthiness and applicability, and their descriptions are presented in Table 2. Two reviewers will perform assessments independently, and conflicts will be resolved by consensus with a third reviewer. A pilot assessment will be conducted on a sample of studies to ensure consistency. In cases where systematic reviews do not contain a protocol or a list of excluded studies, we will contact the corresponding authors to request this information, as the presence of both is assessed within the so-called critical domains of the AMSTAR 2 tool – domains that, if rated poorly, can significantly affect the overall quality rating. These actions are based on the assumption that the lack of access to a pre-registered protocol or a list of excluded studies, while a limitation, does not necessarily mean that they were not prepared. Allowing authors to share this information helps prevent unfairly negative bias against reviews, particularly older ones conducted before protocol registration and reporting guidelines became widespread. First, we will email the corresponding author using the published contact information. If no response is received within one week, we will follow up by emailing both the corresponding and one other author (preferably first or second). If there is no response after two weeks, we will finalize the procedure and consider the attempt concluded. The absence of other essential information for assessment (e.g., search strategy) in the review report or protocol (whether published or obtained through author contact) will not be supplemented in this manner and will result in a lower rating.

Calculations, outcomes and data analysis

Calculations will be performed for meta-analyses selected at the data extraction stage.

Table 2. Tools to assess trustworthiness and applicability components of evidence from systematic reviews of depression treatments.

Tool	Description
Quality of conduct and reporting	
Measurement Tool to Assess Systematic Reviews (AMSTAR 2)	Allows for the assessment of the quality of conduct of the systematic reviews (SRs), with 16 questions, of which seven constitute critical quality domains. Based on answers to those questions, overall confidence in the results of the review can be established as high, moderate, low, or critically low [21].
Conflict-of-Interest assessment (COI)	Allows investigating if included SRs have low, incompletely reported, present, or high conflict of interest based on three criteria (i.e., sources of funding, disclosure of interests, statistical analysis responsibility) [81].
Referencing Framework for SRs (RF)	Focused on the presence of references to relevant previous studies which are categorized as cited, described, or discussed by SRs authors [82].
Risk of bias in SRs	
Risk of Bias in Systematic Reviews (ROBIS)	Allows assessment of concerns regarding four domains (i.e., study eligibility criteria, identification and selection of studies, data collection, and study appraisal, synthesis of findings) and overall risk of bias of the included SRs. For each domain and for the overall assessment, SR can be appraised as having a low, high, or unclear risk of bias [22].
Risk of Bias in Network Meta-Analysis (RoB NMA)	Allows assessment of risk of bias in network meta-analyses (NMA), addressing both traditional systematic review domains (i.e., study eligibility criteria, identification and selection of studies, data collection and appraisal, synthesis of findings) and NMA-specific components such as the assumption of transitivity and biases related to indirect comparisons. Each domain and the overall NMA can be rated as having a low, high, or unclear risk of bias [83].
Spin in abstract conclusions	
Spin Measure (SM)	Evaluates the presence of spin in reporting based on consistency between results for the primary outcome described in the text and abstract of the SR [84].
Robustness of meta-analytical results	
Fragility Index (FI)	Measures how many event-status modifications (e.g., changing a non-event to an event or vice versa) are required to shift the pooled treatment effect from statistically significant to nonsignificant (or vice versa). The calculation involves iteratively modifying single events in individual trials included in the meta-analysis and recalculating the pooled result until the statistical significance changes [24].
Ellipse of Insignificance (EOI)	A geometric refinement of the FI. Allows the assessment of the robustness of results in dichotomous outcome trials by considering simultaneous recoding in both experimental and control arms to determine the minimal changes required to alter statistical significance [25].
Region of Attainable Redaction (ROAR)	An extension of the EOI analysis, designed to evaluate the impact of data redaction (removing or censoring events in the experimental or control groups), whether accidental or deliberate, on statistical significance in dichotomous outcome trials [26].
Granularity-related inconsistency of means (GRIM) test	Identifies inconsistencies in reported means calculated from integer data, such as Likert scales, given a specific sample size and number of items [28].
Leave-N-Out analysis (LNO)	Evaluates how the exclusion of one or more studies from the meta-analysis affects the statistical significance of the overall results.
Heterogeneity and clinical diversity in SRs	
Prediction intervals (PIs)	Estimate the range where the effect size of a future study is likely to fall, incorporating both sampling variability and between-study heterogeneity. Offer a more practical perspective on the consistency of effects across studies than other heterogeneity statistics like I^2 or τ^2 [68,85].
Heterogeneity exploration assessment (HE)	Allows investigation of methods used by reviewers to explain heterogeneity in meta-analyses. Based on methodological guidelines and previous approaches [31–33,54–56,68].

https://doi.org/10.1371/journal.pone.0325384.t002

Calculations for robustness analyses

For meta-analysis of 2 x 2 dichotomous outcome trials, the iterative method of Atal et al. [24] will be employed to estimate the fragility of systematic reviews. This will be cross-validated with a meta-analytic extension of EOI (Ellipse of Insignificance) analysis [25] to analytically determine fragility, and where applicable a ROAR (Region of Attainable Redaction) analysis [26] to estimate the effects of missing data.

EOI analysis will be used in our study in two ways. Firstly, we will apply it to data from individual RCTs included in the meta-analyses, extracted from the original trial reports to ascertain on an individual level how robust constituent trials are. Secondly, we will also apply it to aggregate data. For all meta-analyses, the crude risk ratio (RR-Crude) and the Cochran-Mantel-Haenszel risk ratio (RR-CMH) are calculated. When these differ by less than 10%, it is appropriate to treat data as pooled, and from this an EOI analysis can be performed on the pooled data to ascertain what fragility fraction of the aggregated studies. Alongside this, we deploy Atal's method for estimating meta-analytic fragility. This is in effect a greedy algorithm, which modifies the studies that have the biggest immediate effect on the meta-analytic result, in each step finding the study where flipping an event status would cause the largest movement toward changing the result and modifies and re-evaluates until the threshold is crossed. This greedy approach makes the algorithm much faster than brute-force, but it may not always find the absolute minimum number of changes if a different set of edits elsewhere would have been more optimal. It does however tend to find a unique and minimum set of specific modifications that would flip the results of a meta-analysis, whereas EOI finds the general degree of recoding required to flip conclusions. Thus, Atal's algorithm is deployed here as a lower absolute bound on fragility while EOI serves to estimate the pooled fragility of all studies.

To assess potential anomalies in the reported results of studies included in the SRs, we will apply the GRIM test [28] to chosen continuous outcomes. For each study, we will extract the relevant summary statistics: sample size, SD and mean and evaluate whether the reported results are consistent with the mathematical expectations derived from these parameters. Studies that fail the GRIM test will be excluded from sensitivity analyses and the meta-analytic results recalculated to assess whether the overall effect size and statistical significance remain robust to their removal.

We will also implement a leave-N-out sensitivity analysis to quantify the sensitivity of the SRs' results. For this, we will systematically exclude N studies from the analysis, iterating through all possible combinations (up to $N = 5$). For each recalculated meta-analytic result, we will record whether the significance or direction of the effect changes. The fragility of the result will be defined as the minimum number of excluded studies required to change the overall statistical significance.

Calculations for heterogeneity analyses

We will recalculate selected meta-analytic findings using a random-effects model with the Hartung-Knapp-Sidik-Jonkman (HKSJ) method. The HKSJ method is a random-effects meta-analytic approach that adjusts for small sample sizes and accounts for uncertainty in heterogeneity, making it superior to conventional random-effects models that may underestimate this variability [68,86,87]. In cases where meta-analyses calculated confidence intervals at significance levels other than 95%, we will use the original levels to examine the effect of the HKSJ method on statistical significance changes, and then calculate 95% confidence intervals. All PIs will be calculated at the 95% significance level using the R package meta with the *metagen* function [88]. In addition, we will calculate probabilities of true effect in a new study to be below null effect and of the opposite size using the metafor package with the *pt* function, employing a t-distribution with k-2 degrees of freedom [89].

These analyses will be automated using R scripts.

Outcomes

The outcomes for the qualitative parts of our study will be the percentages of systematic reviews in each category, based on the results of assessments using specific tools (e.g., presence or absence of spin, level of risk of bias, use of meta-regression to explore heterogeneity). The outcomes for the quantitative parts will include both categorical outcomes (based on prespecified criteria) and continuous outcomes. Definitions of the most important outcomes are presented in Table 3. Tools for assessments are described in Table 2.

We used the CINeMA approach to assess heterogeneity in pairwise meta-analyses, as it relies on forest plot interpretation and does not account for indirectness – making it suitable despite being developed for NMAs. We will also analyse the impact of using the HKSJ method on the imprecision of pooled effect estimates. Like the assessment of

Table 3. Outcomes for the trustworthiness and applicability evaluation.

Outcome name	Outcome measures
Quality of conduct and reporting	
Assessment results by AMSTAR 2, COI and RF tools.	1. The percentage of SRs having one, two, three, four, five, six or seven critical domains judged as negative. 2. The percentage of SRs of critically low, low, moderate, or high quality. 3. The percentage of SRs with low, incompletely reported, present or high conflict of interests. 4. The percentage of SRs that cited, described, and discussed relevant previous studies.
Risk of bias in SRs	
Assessment results by ROBIS and RoB NMA tools.	1. The percentage of SRs with low, high, or unclear risk in each of the four domains of ROBIS. 2. The percentage of SRs with low, high, or unclear overall risk of bias by ROBIS. 3. The percentage of SRs for which the assessment of risk of bias in the results of the systematic review with NMA is rated as 'low', 'high', or 'some concerns' by RoB NMA. 4. The percentage of SRs for which the assessment of risk of bias in the conclusions of the systematic review with NMA is rated as 'low', 'high', or 'some concerns' by RoB NMA. 5. The percentage of SRs for which the assessment of risk of bias in each of the three domains of RoB NMA is rated as ‘low’, ‘high’ or ‘some concerns’.
Spin in conclusions	
Assessments results by SM.	1. The percentage of SRs, with abstract conclusions being consistent or inconsistent with the primary outcome result.
Robustness of meta-analytical results	
Fragility to recoding (for meta-analyses of dichotomous endpoints).	1. The percentage of meta-analyses meeting the following criterion: in any of the studies included in the meta-analysis, the number of participants in the intervention group or the comparator group or both groups combined, who would need to be recoded for the significance to disappear, is at least as large as the number of those who dropped out from these groups (based on EOI and FI). 2. The average percentage of participants in trials included in meta-analyses, that would need to be recoded for the significance to disappear in the intervention group, comparator group and both groups combined (based on EOI and FI).
Fragility to redaction (for meta-analyses of dichotomous endpoints).	1. The average percentage of participant samples in the meta-analyses that would need to be redacted for the overall results to lose significance (based on ROAR).
Fragility to misreporting in primary studies (for meta-analyses of continuous endpoints).	1. The percentage of meta-analyses meeting the following criterion: statistical significance of the overall meta-analysis result changes after exclusion of the studies that showed anomalies in reporting detected by GRIM test.
Sensitivity to individual studies results (for meta-analyses of continuous endpoints).	1. The percentage of meta-analyses whose statistical significance changed after excluding up to five individual studies in LNO analysis. 2. The average number and percentage of studies that would have to be excluded from the meta-analysis to change significance (based on LNO analysis).
Heterogeneity and clinical diversity in SRs	
Influence of heterogeneity on conclusions.	1. The percentage of meta-analyses that fall into each of the following categories, based on the CINeMA approach for assessing heterogeneity [90]: no concerns, some concerns, major concerns; as the range of equivalence, we will use 0.8 to 1.25 for binary outcomes and -0,1 to 0,1 for continuous outcomes. 2. The percentage of meta-analyses that fall into each of the following categories describing change in the PIs’ position in relation to the CIs and the null: change (i.e. calculated PI includes null while calculated CI doesn’t), no change. 3. The average probability that the true effect equals null. 4. The percentage of meta-analyses whose calculated PI contains the effect opposite to the pooled summary effect. 5. The average probability that the true effect equals the opposite of the point estimation of the effect.
Practices of exploration and interpretation of heterogeneity.	Methods of statistical heterogeneity assessment, values of relevant statistics, heterogeneity thresholds, exploration of possible reasons for high heterogeneity (e.g., statistical models, effect modifiers), variables used for heterogeneity exploration, with a focus on clinical diversity variables that are crucial for the generalizability of evidence, based on the CDIM tool [32] and our expertise.

AMSTAR, The Measurement Tool to Assess systematic Reviews; COI, Conflict-of-Interest assessment; RF, Referencing Framework for systematic reviews; SRs, systematic reviews; ROBIS, Risk of Bias in Systematic Reviews; RoB NMA, Risk of Bias in Network Meta-Analysis; SM, Spin Measure; EOI, Ellipse of Insignificance; FI, Fragility Index; ROAR, Region of Attainable Redaction, GRIM, Granularity-related Inconsistency of Means; LNO, Leave-N-Out; CINeMA, Confidence in Network Meta-Analysis; PI, prediction interval, CI, confidence interval; CDIM, Clinical Diversity In Meta-analyses.

https://doi.org/10.1371/journal.pone.0325384.t003 heterogeneity's impact, we will first determine the proportion of meta-analyses of which pooled effect size changed – from significant to non-significant (and vice versa) or showed no change. To gain more practical insight, we will then assess the proportion of meta-analyses for which the assessment of imprecision would change in the Grading of Recommendations Assessment, Development, and Evaluation (GRADE) – a prominent approach for assessing certainty of evidence for healthcare recommendations [91]. Additionally, we will characterize calculated PIs in terms of their width, alone and in relation to corresponding confidence intervals. Finally, results for robustness analysis obtained with EOI will be compared to Atal's et al. method employing FI for cross-validation.

Data analysis

We will use descriptive statistics and narrative summaries to synthesize the data. To explore associations between outcomes and reviews' characteristics listed in Table 4 we will use odds ratios. In addition, we will use regression analysis to examine the relationship between the outcomes and the year of publication. We will conduct subgroup analyses to explore how the results vary between groups defined by factors outlined in Table 4. Where possible, we will conduct statistical analyses to explore differences between groups, using non-parametric tests, including Kruskal-Wallis (across 3 groups with post-hoc pairwise comparisons if significant) and Mann-Whitney U Test (for 2 groups). We will use the chi-squared test to analyse the proportion of reviews in each category as described above. Given the exploratory nature of these analyses, no adjustments for multiple testing will be employed.

Data management, open science, and dissemination plan

Upon completion of the analyses and after peer review and accompanying publication, we will make the data and analysis scripts openly available through a suitable data repository or supplementary materials, adhering to open science principles to facilitate transparency and reproducibility.

Ethical considerations

As this study involves the analysis of published data, ethical approval is not required. There are no safety considerations applicable to this study.

Table 4. Description of subgroups for analysis.

Subgroup	Comparisons
Intervention	1. Pharmacotherapy vs all comparators. 2. Psychotherapy vs all comparators. 3. Combined treatment (simultaneous or sequential use of psychotherapy and pharmacotherapy) vs all comparators.
Control conditions	1. Inactive (e.g., placebo) vs active (another treatment) vs mixed (active and inactive single treatments and combined treatments).
Research question	1. 'Narrow' (specific drug or strictly defined psychotherapeutic modality as the intervention studied) vs 'broad' (other than narrow, e.g., pharmacological group of drugs or psychotherapy in general).
Risk of bias	1. SRs (or meta-analyses selected from systematic reviews) with low vs unclear vs high risk of bias, based on our assessment with ROBIS tool.
Type of outcome	1. Continuous versus dichotomous.
Number of studies in meta-analysis	1. At least 10 vs less than 10.
Pre-registration	1. Pre-registered SRs vs. non–pre-registered SRs.
Year of publication	1. Before 2010 vs after 2010. The 2010 cut-off was established based on the publication timing of the first PRISMA statement, which was released in mid-2009 [92].

https://doi.org/10.1371/journal.pone.0325384.t004

Status and timeline

At the time of this protocol submission, the project successfully completed pilot testing of the data extraction processes and assessments, and OSF protocols have been developed and registered. The projected timeline for data collection is May 2025, and the second quarter of 2025 for project completion.

Discussion

This study aims to address challenges in evaluating treatments for depression by conducting a methodologically focused analysis of systematic reviews. We propose a framework to assess trustworthiness and applicability, structured into five components: quality of conduct and reporting, risk of bias, spin in abstract conclusions, robustness of meta-analytical results, and heterogeneity and clinical diversity.

The sample of evidence has both strengths and limitations. It will include a large selection of systematic reviews and meta-analyses, whose findings are expected to be generalizable to a broad population of adults with depression. However, narrowing the scope to psychotherapies and antidepressants may overlook other effective and relevant for clinical practice interventions, such as antipsychotics or transcranial stimulation. Additionally, the restrictive approach to selecting reviews based solely on their eligibility criteria, while beneficial for the feasibility of the study, might exclude certain articles considered seminal. These limitations could be addressed in future research of this kind by focusing on the most influential reviews, such as those informing clinical guidelines – an approach we plan to undertake in subsequent studies. Excluding NMAs from part of the analyses can be seen as another limitation, as these analyses provide indirect evidence and comparisons between multiple interventions. However, direct pairwise comparisons are easier to interpret and rely on simpler, clearer methodological assumptions. NMAs require complex assumptions about transitivity and homogeneity, which are harder to verify and can introduce additional bias. By focusing on direct comparisons, our study design ensures more robust, reliable, and clinically relevant results, free from the added complexities often found in NMAs.

The tools selected for this analysis also have their limitations. The qualitative tools provide a multifaceted perspective on the evidence but occasionally require subjective judgments. To mitigate the influence of such judgments and ensure transparency, we will document and publicly share the rationale behind them. Quantitative tools, on the other hand, rely on various assumptions which will be carefully considered when interpreting the results.

We define a set of outcomes that will enable the comparison and evaluation of the utility and relevance of state-of-the-art meta-research tools, as well as the intuitive and meaningful interpretation of the assessments' results. Our planned subgroup analyses may help identify directions for investigating reasons of potential methodological challenges. However, it should be noted that all analyses remain observational in nature and are intended as exploratory.

Our analysis, while comprehensive, does not address all methodological challenges encountered in evidence synthesis in mental health. For instance, publication bias is considered a potentially significant factor when drawing conclusions about the efficacy of both medications and psychotherapies [93,94]. However, we have opted not to analyse it due to the limitations of existing methods for its detection and correction, whose validity is particularly affected by between-study heterogeneity [95] – a factor we anticipate to be substantial in the meta-analyses included in our review.

Regarding dissemination plans, upon completion and after peer review and publication, we will make our data and analysis scripts openly available through a suitable data repository, adhering to open science principles. This transparency will facilitate reproducibility and allow other researchers to build upon our work, contributing to improved methodological standards in the field.

Any amendments to the study protocol will be documented transparently. Should adjustments be necessary – such as changes to eligibility criteria or analytical methods – we will update our registered protocol accordingly and report these modifications in our final publication, providing justifications and discussing potential impacts on our findings.

In conclusion, our intended study seeks to evaluate how trustworthy and applicable is the evidence from systematic reviews of depression treatments. Despite inherent limitations, we believe that results of this analysis will be well-positioned to form a basis for future recommendations on how to enhance the methodological rigor of SRs on treatments for depression by addressing key issues related to conduct, reporting and interpretation. Therefore, by highlighting methodological strengths and weaknesses in current SRs, we aim not to criticize the laudable efforts of past reviewers, but to eventually strengthen their toolkit and contribute to developing future methods for evidence evaluation. Such methodological advances are not trivial, often leading to improved treatment strategies for individuals with depression, as evidenced by the continuous developments in evidence-based medicine and research synthesis.

Supporting information

S1 Appendix. Results of AMSTAR 2 assessments in mental health and other medical fields.
(PDF)

S2 Appendix. Search strategies.
(PDF)

S3 Appendix. PRISMA Flow diagram.
(PDF)

S4 Appendix. List of included reviews.
(PDF)

S5 Appendix. List of excluded reviews.
(PDF)

S6 Appendix. Extracted variables.
(PDF)

Author contributions

Conceptualization: Iwo Fober, Lidia Baran, Myrto Samara, Spyridon Siafis, David Robert Grimes, Bartosz Helfer.

Funding acquisition: Bartosz Helfer.

Methodology: Iwo Fober, Lidia Baran, Spyridon Siafis, David Robert Grimes, Bartosz Helfer.

Supervision: Bartosz Helfer.

Writing – original draft: Iwo Fober, Lidia Baran, Bartosz Helfer.

Writing – review & editing: Iwo Fober, Lidia Baran, Myrto Samara, Spyridon Siafis, David Robert Grimes, Bartosz Helfer.

References

1. Depressive disorder (depression). Accessed 2025 January 29. https://www.who.int/news-room/fact-sheets/detail/depression
2. McKibbon KA. Evidence-based practice. Bull Med Libr Assoc. 1998;86(3):396–401. PMID: 9681176
3. Tenny S, Varacallo M. Evidence-based medicine. In: StatPearls. Treasure Island (FL): StatPearls Publishing; 2024.
4. Gupta M. Is evidence-based psychiatry ethical? Oxford: Oxford University Press; 2014.
5. Cleare A, Pariante CM, Young AH, Anderson IM, Christmas D, Cowen PJ, et al. Evidence-based guidelines for treating depressive disorders with antidepressants: a revision of the 2008 British Association for Psychopharmacology guidelines. J Psychopharmacol. 2015;29(5):459–525. https://doi.org/10.1177/0269881115581093 PMID: 25969470
6. Parikh SV, Segal ZV, Grigoriadis S, Ravindran AV, Kennedy SH, Lam RW. Canadian Network for Mood and Anxiety Treatments (CANMAT) clinical guidelines for the management of major depressive disorder in adults. II. Psychotherapy alone or in combination with antidepressant medication. J Affect Disord. 2009;117(Suppl 1):S15–25.

7. Depression in adults: treatment and management. London: National Institute for Health and Care Excellence (NICE); 2022. http://www.ncbi.nlm.nih.gov/books/NBK583074/

8. World Health Organization. Mental Health Gap Action Programme (mhGAP) guideline for mental, neurological and substance use disorders. Geneva: World Health Organization; 2023: 1.

9. Gelenberg AJ, Freeman MP, Markowitz JC, Rosenbaum JF, Thase ME, Trivedi MH. Practice guideline for the treatment of patients with major depressive disorder. 2010. https://psychiatryonline.org/pb/assets/raw/sitewide/practice_guidelines/guidelines/mdd-1410197717630.pdf

10. Cuijpers P. Four decades of outcome research on psychotherapies for adult depression: an overview of a series of meta-analyses. Canadian Psychology/ Psychologie Canadienne. 2017;58(1):7–19.

11. Luo Y, Chaimani A, Furukawa TA, Kataoka Y, Ogawa Y, Cipriani A, et al. Visualizing the evolution of evidence: cumulative network meta-analyses of new generation antidepressants in the last 40 years. Res Synth Methods. 2021;12(1):74–85. https://doi.org/10.1002/jrsm.1413 PMID: 32352639

12. Petersson E-L, Forsén E, Björkelund C, Hammarbäck L, Hessman E, Weineland S, et al. Examining the description of the concept "treatment as usual" for patients with depression, anxiety and stress-related mental disorders in primary health care research - A systematic review. J Affect Disord. 2023;326:1–10. https://doi.org/10.1016/j.jad.2023.01.076 PMID: 36708952

13. Cristea IA, Gentili C, Pietrini P, Cuijpers P. Sponsorship bias in the comparative efficacy of psychotherapy and pharmacotherapy for adult depression: meta-analysis. Br J Psychiatry. 2017;210(1):16–23. https://doi.org/10.1192/bjp.bp.115.179275 PMID: 27810891

14. Rutherford BR, Cooper TM, Persaud A, Brown PJ, Sneed JR, Roose SP. Less is more in antidepressant clinical trials: a meta-analysis of the effect of visit frequency on treatment response and dropout. J Clin Psychiatry. 2013;74(7):703–15. https://doi.org/10.4088/JCP.12r08267 PMID: 23945448

15. Weisz JR, Gray JS. Evidence-based psychotherapy for children and adolescents: data from the present and a model for the future. Child Adolesc Ment Health. 2008;13(2):54–65. https://doi.org/10.1111/j.1475-3588.2007.00475.x PMID: 32847169

16. Cuijpers P. Are all psychotherapies equally effective in the treatment of adult depression? The lack of statistical power of comparative outcome studies. Evid Based Ment Health. 2016;19(2):39–42. https://doi.org/10.1136/eb-2016-102341 PMID: 26984413

17. Arroll B, Chin W-Y, Martis W, Goodyear-Smith F, Mount V, Kingsford D, et al. Antidepressants for treatment of depression in primary care: a systematic review and meta-analysis. J Prim Health Care. 2016;8(4):325–34. https://doi.org/10.1071/HC16008 PMID: 29530157

18. Furukawa TA, Noma H, Caldwell DM, Honyashiki M, Shinohara K, Imai H, et al. Waiting list may be a nocebo condition in psychotherapy trials: a contribution from network meta-analysis. Acta Psychiatr Scand. 2014;130(3):181–92.

19. Cuijpers P, Miguel C, Harrer M, Plessen CY, Ciharova M, Papola D, et al. Psychological treatment of depression: a systematic overview of a 'Meta-Analytic Research Domain.' J Affective Disorders. 2023 Aug 15;335:141–51.

20. Ge L, Tian J-H, Li Y-N, Pan J-X, Li G, Wei D, et al. Association between prospective registration and overall reporting and methodological quality of systematic reviews: a meta-epidemiological study. J Clin Epidemiol. 2018;93:45–55. https://doi.org/10.1016/j.jclinepi.2017.10.012 PMID: 29111471

21. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J. AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ. 2017;358:j4008.

22. Whiting P, Savović J, Higgins JPT, Caldwell DM, Reeves BC, Shea B, et al. ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol. 2016;69:225–34. https://doi.org/10.1016/j.jclinepi.2015.06.005 PMID: 26092286

23. Yavchitz A, Ravaud P, Altman DG, Moher D, Hrobjartsson A, Lasserson T, et al. A new classification of spin in systematic reviews and meta-analyses was developed and ranked according to the severity. J Clin Epidemiol. 2016;75:56–65. https://doi.org/10.1016/j.jclinepi.2016.01.020 PMID: 26845744

24. Atal I, Porcher R, Boutron I, Ravaud P. The statistical significance of meta-analyses is frequently fragile: definition of a fragility index for meta-analyses. J Clin Epidemiol. 2019;111:32–40. https://doi.org/10.1016/j.jclinepi.2019.03.012 PMID: 30940600

25. Grimes DR. The ellipse of insignificance, a refined fragility index for ascertaining robustness of results in dichotomous outcome trials. Boonstra P, Zaidi M, Boonstra P, Jiang F, editors. eLife. 2022;11:e79573.

26. Grimes DR. Region of attainable redaction, an extension of ellipse of insignificance analysis for gauging impacts of data redaction in dichotomous outcome trials. eLife. 2024;13:e93050.

27. Grimes DR, Heathers J. The new normal? Redaction bias in biomedical science. Royal Society Open Science. 2021;8(12):211308.

28. Brown NJL, Heathers JAJ. The GRIM test: a simple technique detects numerous anomalies in the reporting of results in psychology. Social Psychol Personality Science. 2017;8(4):363–9.

29. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q. 2004;82(4):661–87. https://doi.org/10.1111/j.0887-378X.2004.00327.x PMID: 15595946

30. Siegel JS, Zhong J, Tomioka S, Ogirala A, Faraone SV, Szabo ST. Estimating heterogeneity of treatment effect in psychiatric clinical trials. medRxiv. 2024. https://doi.org/2024.04.23.24306211

31. Gagnier JJ, Moher D, Boon H, Beyene J, Bombardier C. Investigating clinical heterogeneity in systematic reviews: a methodologic review of guidance in the literature. BMC Med Res Methodol. 2012;12:111. https://doi.org/10.1186/1471-2288-12-111 PMID: 22846171

32. Barbateskovic M, Koster TM, Eck RJ, Maagaard M, Afshari A, Blokzijl F, et al. A new tool to assess Clinical Diversity In Meta-analyses (CDIM) of interventions. J Clin Epidemiol. 2021;135:29–41. https://doi.org/10.1016/j.jclinepi.2021.01.023 PMID: 33561529

33. Chess LE, Gagnier JJ. Applicable or non-applicable: investigations of clinical heterogeneity in systematic reviews. BMC Med Res Methodol. 2016;16:19. https://doi.org/10.1186/s12874-016-0121-7 PMID: 26883215

34. Uttley L, Quintana DS, Montgomery P, Carroll C, Page MJ, Falzon L, et al. The problems with systematic reviews: a living systematic review. J Clin Epidemiol. 2023;156:30–41. https://doi.org/10.1016/j.jclinepi.2023.01.011 PMID: 36796736

35. Uttley L, Weng Y, Falzon L. Yet another problem with systematic reviews: a living review update. J Clin Epidemiol. 2025;177.

36. Storman M, Storman D, Jasinska KW, Swierz MJ, Bala MM. The quality of systematic reviews/meta-analyses published in the field of bariatrics: a cross-sectional systematic survey using AMSTAR 2 and ROBIS. Obes Rev. 2020;21(5):e12994. https://doi.org/10.1111/obr.12994 PMID: 31997545

37. Ou SL, Luo J, Wei H, Qin XL, Du SY, Wang S. Safety and efficacy of programmed cell death 1 and programmed death ligand-1 inhibitors in the treatment of cancer: an overview of systematic reviews. Front Immunol. 2022;13:953761.

38. Pereira A, Martins C, Campos J, Faria S, Notaro S, Poklepović-Peričić T. Critical appraisal of systematic reviews of intervention studies in periodontology using AMSTAR 2 and ROBIS tools. J Clin Exp Dent. 2023:e678–94.

39. Ferri N, Ravizzotti E, Bracci A, Carreras G, Pillastrini P, Di Bari M. The confidence in the results of physiotherapy systematic reviews in the musculoskeletal field is not increasing over time: a meta-epidemiological study using AMSTAR 2 tool. J Clin Epidemiol. 2024;169:111303. https://doi.org/10.1016/j.jclinepi.2024.111303 PMID: 38402999

40. Rotta I, Diniz JA, Fernandez-Llimos F. Assessing methodological quality of systematic reviews with meta-analysis about clinical pharmacy services: a sensitivity analysis of AMSTAR-2. Res Social Adm Pharm. 2025;21(2):110–5. https://doi.org/10.1016/j.sapharm.2024.11.002 PMID: 39643474

41. Karakasis P, Bougioukas KI, Pamporis K, Fragakis N, Haidich A-B. Appraisal methods and outcomes of AMSTAR 2 assessments in overviews of systematic reviews of interventions in the cardiovascular field: a methodological study. Res Synth Methods. 2024;15(2):213–26. https://doi.org/10.1002/jrsm.1680 PMID: 37956538

42. Rainkie DC, Abedini ZS, Abdelkader NN. Reporting and methodological quality of systematic reviews and meta-analysis with protocols in Diabetes Mellitus Type II: a systematic review. PLoS One. 2020;15(12):e0243091. https://doi.org/10.1371/journal.pone.0243091 PMID: 33326429

43. De Santis KK, Lorenz RC, Lakeberg M, Matthias K. The application of AMSTAR2 in 32 overviews of systematic reviews of interventions for mental and behavioural disorders: a cross-sectional study. Res Synth Methods. 2022;13(4):424–33. https://doi.org/10.1002/jrsm.1532 PMID: 34664766

44. Chung VCH, Wu XY, Feng Y, Ho RST, Wong SYS, Threapleton D. Methodological quality of systematic reviews on treatments for depression: a cross-sectional study. Epidemiol Psychiatr Sci. 2018;27(6):619–27. https://doi.org/10.1017/S2045796017000208 PMID: 28462754

45. Desaunay P, Eude L-G, Dreyfus M, Alexandre C, Fedrizzi S, Alexandre J, et al. Benefits and risks of antidepressant drugs during pregnancy: a systematic review of meta-analyses. Paediatr Drugs. 2023;25(3):247–65. https://doi.org/10.1007/s40272-023-00561-2 PMID: 36853497

46. Matthias K, Rissling O, Pieper D, Morche J, Nocon M, Jacobs A. The methodological quality of systematic reviews on the treatment of adult major depression needs improvement according to AMSTAR 2: a cross-sectional study. Heliyon. 2020;6(9):e04776. https://doi.org/10.1016/j.heliyon.2020.e04776

47. Health Quality Ontario. Internet-delivered cognitive behavioural therapy for major depression and anxiety disorders: a health technology assessment. Ont Health Technol Assess Ser. 2019;19(6):1–199. PMID: 30873251

48. Ribeiro ELA, de Mendonça Lima T, Vieira MEB, Storpirtis S, Aguiar PM. Efficacy and safety of aripiprazole for the treatment of schizophrenia: an overview of systematic reviews. Eur J Clin Pharmacol. 2018;74(10):1215–33. https://doi.org/10.1007/s00228-018-2498-1 PMID: 29905899

49. Mangolini VI, Andrade LH, Lotufo-Neto F, Wang Y-P. Treatment of anxiety disorders in clinical practice: a critical overview of recent systematic evidence. Clinics (Sao Paulo). 2019;74:e1316. https://doi.org/10.6061/clinics/2019/e1316 PMID: 31721908

50. Stoll M, Mancini A, Hubenschmid L, Dreimüller N, König J, Cuijpers P, et al. Discrepancies from registered protocols and spin occurred frequently in randomized psychotherapy trials-A meta-epidemiologic study. J Clin Epidemiol. 2020;128:49–56. https://doi.org/10.1016/j.jclinepi.2020.08.013 PMID: 32828837

51. Narum S. Antidepressiva til ungdom - en kritisk analyse av bivirkningsbeskrivelser og nytte-risikovurderinger i en RCT En dokumentanalyse. 2018. https://www.duo.uio.no/handle/10852/67187

52. Gan DZQ, McGillivray L, Han J, Christensen H, Torok M. Effect of engagement with digital interventions on mental health outcomes: a systematic review and meta-analysis. Front Digit Health. 2021;3:764079. https://doi.org/10.3389/fdgth.2021.764079 PMID: 34806079

53. Gong X, Fenech B, Blackmore C, Chen Y, Rodgers G, Gulliver J. Association between noise annoyance and mental health outcomes: a systematic review and meta-analysis. Int J Environ Res Public Health. 2022;19(5):2696.

54. Brini S, Brudasca NI, Hodkinson A, Kaluzinska K, Wach A, Storman D. Efficacy and safety of transcranial magnetic stimulation for treating major depressive disorder: An umbrella review and re-analysis of published meta-analyses of randomised controlled trials. Clin Psychol Rev. 2023;100:102236.

55. Siemens W, Meerpohl JJ, Rohe MS, Buroh S, Schwarzer G, Becker G. Reevaluation of statistically significant meta-analyses in advanced cancer patients using the Hartung-Knapp method and prediction intervals-A methodological study. Res Synth Methods. 2022;13(3):330–41. https://doi.org/10.1002/jrsm.1543 PMID: 34932271

56. Faggion CM Jr, Atieh MA, Tsagris M, Seehra J, Pandis N. A case study evaluating the effect of clustering, publication bias, and heterogeneity on the meta-analysis estimates in implant dentistry. Eur J Oral Sci. 2024;132(1):e12962. https://doi.org/10.1111/eos.12962 PMID: 38030576

57. Holek M, Bdair F, Khan M, Walsh M, Devereaux PJ, Walter SD, et al. Fragility of clinical trials across research fields: a synthesis of methodological reviews. Contemp Clin Trials. 2020;97:106151. https://doi.org/10.1016/j.cct.2020.106151 PMID: 32942056

58. Mun KT, Bonomo JB, Liebeskind DS, Saver JL. Fragility index meta-analysis of randomized controlled trials shows highly robust evidential strength for benefit of <3 hour intravenous alteplase. Stroke. 2022;53(6):2069–74.

59. Anand S, Kainth D. Fragility index of recently published meta-analyses in pediatric urology: a striking observation. Cureus. 2021;13(7):e16225.

60. Schröder A, Muensterer OJ, Oetzmann von Sochaczewski C. Meta-analyses in paediatric surgery are often fragile: implications and consequences. Pediatr Surg Int. 2021;37(3):363–7. https://doi.org/10.1007/s00383-020-04827-5 PMID: 33454848

61. Won J. Robustness of meta-analysis results in Cochrane systematic reviews: a case for acupuncture trials. Integr Med Res. 2022;11(4):100890. https://doi.org/10.1016/j.imr.2022.100890 PMID: 36338607

62. Sorigue M, Kuittinen O. Robustness and pragmatism of the evidence supporting the European Society for Medical Oncology guidelines for the diagnosis, treatment, and follow-up of follicular lymphoma. Expert Rev Hematol. 2021;14(7):655–68.

63. Huang X, Chen B, Thabane L, Adachi JD, Li G. Fragility of results from randomized controlled trials supporting the guidelines for the treatment of osteoporosis: a retrospective analysis. Osteoporos Int. 2021;32(9):1713–23. https://doi.org/10.1007/s00198-021-05865-y PMID: 33595680

64. Tignanelli CJ, Napolitano LM. The fragility index in randomized clinical trials as a means of optimizing patient care. JAMA Surg. 2019;154(1):74–9. https://doi.org/10.1001/jamasurg.2018.4318 PMID: 30422256

65. Dey S, Saikia P, Choupoo NS, Das SK. How robust are the evidences that formulate surviving sepsis guidelines? An analysis of fragility and reverse fragility of randomized controlled trials that were referred in these guidelines. Indian J Crit Care Med. 2021;25(7):773–9.

66. Otalora-Esteban M, Delgado-Ramirez MB, Gil F, Thabane L. Assessing the fragility index of randomized controlled trials supporting perioperative care guidelines: a methodological survey protocol. PLoS One. 2024;19(9):e0310092. https://doi.org/10.1371/journal.pone.0310092 PMID: 39264894

67. Gaudino M, Hameed I, Biondi-Zoccai G, Tam DY, Gerry S, Rahouma M, et al. Systematic evaluation of the robustness of the evidence supporting current guidelines on myocardial revascularization using the fragility index. Circ Cardiovasc Qual Outcomes. 2019;12(12):e006017. https://doi.org/10.1161/CIRCOUTCOMES.119.006017 PMID: 31822120

68. IntHout J, Ioannidis JPA, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open. 2016;6(7):e010247. https://doi.org/10.1136/bmjopen-2015-010247 PMID: 27406637

69. Steele RG, McGuire AB, Kingston N. The meta-analysis application worksheet: a practical guide for the application of meta-analyses to clinical cases. Prof Psychol Res Pr. 2024;55(5):405–16. https://doi.org/10.1037/pro0000565 PMID: 39619795

70. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.

71. Husain-Krautter S, Ellison JM. Late life depression: the essentials and the essential distinctions. Focus (Am Psychiatr Publ). 2021;19(3):282–93. https://doi.org/10.1176/appi.focus.20210006 PMID: 34690594

72. Mitchell B, Martin N, Medland SE. Genetic and environmental predictors of treatment resistant depression. European Neuropsychopharmacol. 2024;87:25–6.

73. Mullen S. Major depressive disorder in children and adolescents. Mental Health Clin. 2018;8(6):275–83.

74. Field T. Prenatal depression risk factors, developmental effects and interventions: a review. J Preg Child Health. 2017;04(01).

75. Goodwin GM. Depression and associated physical diseases and symptoms. Dialogues Clin Neurosci. 2006;8(2):259–65. https://doi.org/10.31887/DCNS.2006.8.2/mgoodwin PMID: 16889110

76. Chapter 4: Searching for and selecting studies [Internet]. [cited 2025 April 24]. https://training.cochrane.org/handbook/current/chapter-04.

77. Avau B, Van Remoortel H, De Buck E. Translation and validation of PubMed and Embase search filters for identification of systematic reviews, intervention studies, and observational studies in the field of first aid. J Med Libr Assoc. 2021;109(4).

78. Salvador-Oliván JA, Marco-Cuenca G, Arquero-Avilés R. Development of an efficient search filter to retrieve systematic reviews from PubMed. J Med Libr Assoc. 2021;109(4).

79. Cipriani A, Furukawa TA, Salanti G, Chaimani A, Atkinson LZ, Ogawa Y, et al. 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. 2018;391(10128):1357–66. https://doi.org/10.1016/S0140-6736(17)32802-7 PMID: 29477251

80. Bramer WM, Giustini D, De Jonge GB, Holland L, Bekhuis T. De-duplication of database search results for systematic reviews in EndNote. J Med Libr Assoc. 2016;104(3).

81. Helfer B, Leonardi-Bee J, Mundell A, Parr C, Ierodiakonou D, Garcia-Larsen V, et al. Conduct and reporting of formula milk trials: systematic review. BMJ. 2021;375:n2202. https://doi.org/10.1136/bmj.n2202 PMID: 34645600

82. Helfer B, Prosser A, Samara MT, Geddes JR, Cipriani A, Davis JM. Recent meta-analyses neglect previous systematic reviews and meta-analyses about the same topic: a systematic examination. BMC Med. 2015;13:82.

83. Lunny C, Higgins JPT, White IR, Dias S, Hutton B, Wright JM, et al. Risk of Bias in Network Meta-Analysis (RoB NMA) tool. BMJ. 2025;388:e079839.

84. Bero L, Oostvogel F, Bacchetti P, Lee K. Factors associated with findings of published trials of drug-drug comparisons: why some statins appear more efficacious than others. PLoS Med. 2007;4(6):e184. https://doi.org/10.1371/journal.pmed.0040184 PMID: 17550302

85. Borenstein M. Avoiding common mistakes in meta-analysis: understanding the distinct roles of Q, I-squared, tau-squared, and the prediction interval in reporting heterogeneity. Res Synth Methods. 2024;15(2):354–68. https://doi.org/10.1002/jrsm.1678 PMID: 37940120

86. IntHout J, Ioannidis JPA, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14:25. https://doi.org/10.1186/1471-2288-14-25 PMID: 24548571

87. Wang Z, Alzuabi MA, Morgan RL, Mustafa RA, Falck-Ytter Y, Dahm P. Different meta-analysis methods can change judgements about imprecision of effect estimates: a meta-epidemiological study. BMJ Evid Based Med. 2023;28(2):126–32.

88. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evid Based Ment Health. 2019;22(4):153–60. https://doi.org/10.1136/ebmental-2019-300117 PMID: 31563865

89. Viechtbauer W. Conducting meta-analyses in R with the metafor package. J Stat Soft. 2010;36(3).

90. Nikolakopoulou A, Higgins JPT, Papakonstantinou T, Chaimani A, Del Giovane C, Egger M, et al. CINeMA: an approach for assessing confidence in the results of a network meta-analysis. PLoS Med. 2020;17(4):e1003082. https://doi.org/10.1371/journal.pmed.1003082 PMID: 32243458

91. Guyatt GH, Oxman AD, Vist GE, Kunz R, Falck-Ytter Y, Alonso-Coello P. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336(7650).

92. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. 2009;6(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097

93. Turner EH. Publication bias, with a focus on psychiatry: causes and solutions. CNS Drugs. 2013;27(6):457–68. https://doi.org/10.1007/s40263-013-0067-9 PMID: 23696308

94. Driessen E, Hollon SD, Bockting CLH, Cuijpers P, Turner EH. Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health-Funded Trials. PLoS One. 2015;10(9):e0137864. https://doi.org/10.1371/journal.pone.0137864 PMID: 26422604

95. van Aert RCM, Wicherts JM, van Assen MALM. Publication bias examined in meta-analyses from psychology and medicine: a meta-meta-analysis. PLoS One. 2019;14(4):e0215052. https://doi.org/10.1371/journal.pone.0215052 PMID: 30978228