



Operationalization of Cochrane's Risk of Bias 2 Tool (RoB 2) in the Context of Psychotherapy Trials

Introduction & Guidance Document

Clara Miguel¹, Mathias Harrer^{2,3}, Eirini Karyotaki^{1,4}, Ethan Sakher^{4,5}, Masatsugu Sakata^{4,6}, Toshi A. Furukawa⁴, and Pim Cuijpers¹

¹Department of Clinical, Neuro and Developmental Psychology, World Health Organization Collaborating Center for Research and Dissemination of Psychological Interventions, Amsterdam Public Health Research Institute, Vrije Universiteit, Amsterdam, The Netherlands, ²Department of Psychiatry, TUM School of Medicine and Health, Technical University of Munich, Germany, ³Department of Clinical Psychology, University of Amsterdam, Amsterdam, Netherlands, ⁴Department of Health Promotion and Human Behavior, Kyoto University Graduate School of Medicine and School of Public Health, Kyoto, Japan, ⁵Population Health and Policy Research Unit, Medical Education Center, Kyoto University Graduate School of Medicine, Kyoto, Japan, ⁶Department of Neurodevelopmental Disorders, Nagoya City University Graduate School of Medial Sciences, Nagoya, Japan

Abstract

Background. Evaluating risk of bias (RoB) is a crucial step in systematic reviews and meta-analyses to ensure trustworthy evidence. Cochrane's RoB 2 tool is the most widely used approach for assessing risk of bias in randomized controlled trials. However, its application can be challenging, particularly in fields like psychotherapy, where standard trial procedures such as double blinding are often unfeasible.

Objective. This article presents a context-specific operationalization and implementation guideline for applying RoB 2 to psychotherapy trials, addressing common challenges and ambiguities in this field.

Method. This guideline was developed based on empirical evidence, an expert consensus process, and iterative piloting. It provides a structured interpretation of all RoB 2 domains and signalling questions, adapted to the common methodological features of psychotherapy trials.

Results. The operationalization resulted in 24 standardized RoB items with a harmonized scoring logic, accompanied by a practical handbook offering domain-specific advice and examples relevant to psychotherapy research. To support its use, two open-access digital assistant tools (web-based and spreadsheet formats) were developed to guide users through the rating process and facilitate integration with systematic review workflows (metapsy.org/rob).

Conclusion. This guidance offers a practical approach for applying RoB 2 in psychotherapy outcome research. It is intended to be used alongside the Cochrane manual and may help improve the transparency and reproducibility of RoB 2 assessments for psychotherapy trials.

Publication Status

Preprint – changes resulting from peer-review are possible

Correspondence to

Clara Miguel
clara.miguelsanz@vu.nl

Open Material

Online handbook, rating assistant, and other guidance available at metapsy.org/rob.

Supplementary Materials

A: Challenges Identified in the Preliminary Assessment of RoB 2. **B:** Rating Manual. **C:** Rating Sheet (MS Excel). **D:** Rating of 35 Trials (MS Excel).

INTRODUCTION

Systematic reviews and meta-analyses of randomized controlled trials (RCTs) are considered the gold standard in evidence-based decision making (Group, 1992), significantly influencing clinical guidelines and policy (Steinberg, Greenfield, Wolman, Mancher, & Graham, 2011). When systematic reviews and meta-analyses are conducted properly, the trustworthiness of their conclusions will ultimately depend on the quality of their included primary studies. RCTs are considered the best design to examine the effects of an intervention due to their ability to provide causal evidence. However, flaws in the design or execution, such as improper randomization or selective reporting, may lead to systematic errors or deviations from the true effect of an intervention, also known as biases. If not properly prevented, these problems can result in distorted estimates of treatment efficacy, by overestimation (exaggerated treatment efficacy) or underestimation (underrated treatment efficacy). Evaluating these biases is an essential step in any systematic review.

Cochrane's Risk of Bias (RoB) tool is the most well-known and widely used method to assess RoB in RCTs across disciplines (Jørgensen *et al.*, 2016). This tool, in its latest revised version (RoB 2) (Sterne *et al.*, 2019), evaluates risk of bias across five different domains of a clinical trial. Through a series of signalling questions, reviewers can assess bias arising from 1) the randomization process, 2) deviations from the intended interventions, 3) missing outcome data, 4) the measurement of the outcome, and 5) the selection of the reported result. Risk of bias judgements (low risk, some concerns, or high risk of bias) for each of these five domains can be reached through algorithms based on the signalling questions. Wherever appropriate, reviewers can override these algorithms with their own judgement. An overall risk of bias score for each study is obtained, which is usually the worst risk of bias score in any of the domains (e.g., overall high risk if one domain has a high risk score; Sterne *et al.*, 2019). Very detailed manuals for the application of the tool have been developed (Higgins, Savović, Page, Sterne, & Group, 2019).

RoB 2 was designed to address the most recent theoretical and empirical advances in the literature about bias. Nevertheless, some challenges have been described when applying the tool. Independent studies have reported low agreements between reviewers, time-intensive assessments (about 30 minutes per study), and an overall high complexity of application (Minozzi, Cinquini, Gianola, Gonzalez-Lorenzo, & Banzi, 2020). A recent study of complex interventions reached similar conclusions, finding that some signalling questions were too theoretical and needed more practical guidance (Crocker *et al.*, 2023). Domains that were particularly difficult to rate were missing outcome data, bias in the measurement of the outcome in unblinded trials, and deviations from intended interventions (Crocker *et al.*, 2023; Minozzi *et al.*, 2020).

Evaluating RoB is particularly challenging in disciplines like psychotherapy, where standard double-blind RCT procedures are not easily applicable. Unlike drug trials, psychotherapy trials cannot achieve full blinding due to the nature of the intervention. Therapists cannot be masked to the treatment being delivered, and control groups do not have a placebo equivalent, making masking of participants unattainable (Cuijpers, in press). This results in concerns for risk of bias, especially related to deviations from the intended interventions and the use of self-reports to measure the outcome. These conceptual difficulties result in reviewers using very different criteria for RoB or even skipping entire domains (Munder & Barth, 2018). Inconsistencies in applying the RoB 2 tool across reviews can lead to different ratings for the same trial, which can create conflicting conclusions in clinical guidelines.

Developing an implementation document tailored to the review context can facilitate the application of RoB 2, reduce time, and increase interrater reliability (Crocker et al., 2023; Minozzi, Dwan, Borrelli, & Filippini, 2022) (Tomlinson et al., 2024). This is also a recommendation given throughout the RoB 2 tool manual, which recommends reviewers to specify in advance how signalling questions will be evaluated based on the research context. To the best of our knowledge, no prior work has provided specific recommendations for implementing the Cochrane RoB 2 tool in the context of psychotherapy trials.

In the current article, we present an extensive operationalization of its signalling questions and domains, addressing the particularities of psychotherapy trials. The development of this operationalization was motivated by the need to conduct reliable and reproducible assessments in a large living meta-analytic database of depression psychotherapy trials (Cuijpers, Harrer, Miguel, Ciharova, & Karyotaki, 2023; metapsy.org). Additionally, we developed digital assistant tools to facilitate the application of this operationalization (metapsy.org/rob), which may also support other researchers working on similar projects.

Method

This article outlines an operationalization of Cochrane's RoB 2 tool for use in trials of psychological interventions, developed through expert meetings, piloting exercises, and meta-epidemiological evidence from psychotherapy trials. Additionally, we aimed to optimize its implementation in large meta-analytic projects by integrating a suite of digital assistant tools (metapsy.org/rob).

Preliminary Assessment of the RoB 2 tool

First, we conducted a preliminary assessment of the RoB 2 to identify aspects that required more specific guidance in the context of psychotherapy trials. For this, we rated a sample of 35 trials comparing psychotherapies for depression with control conditions. These trials are part of a living meta-analytic database comprising over 1,000 RCTs (Cuijpers et al., 2023; metapsy.org), to which we plan to apply RoB 2 ratings. The need for reliable and reproducible assessments within this database motivated the development of the current implementation manual.

Pairs of independent reviewers (all the authors from the current manuscript) were provided with the original Cochrane RoB 2 documentation and guidance to perform the ratings (Higgins *et al.*). Reviewers documented general challenges encountered during the assessment process, such as items where discrepancies and uncertainties were most common, as well as difficulties specific to the nature of psychological interventions. This pilot exercise provided the foundation for the operationalization of RoB 2 presented in this article.

Process of Operationalization of RoB 2

Following the preliminary assessment of implementation needs, we conducted a series of meetings in which experts in mental health and psychotherapy research discussed all signaling questions and domains of the RoB 2 tool. Special attention was given to the RoB indicators that generated more difficulties during the preliminary assessment. These meetings were held through monthly videoconferences for one year and involved the entire author team. We also sought additional input from experts within the team's network on specific indicators.

The discussions focused on developing recommendations for applying the RoB 2 signaling questions to the psychotherapy field, while also considering how to integrate the tool into the infrastructure of our living meta-analytic database. Throughout the process, several rounds of piloting exercises were conducted between the expert meetings. In these iterations, pairs of experts independently evaluated a sample of trials using the proposed criteria. These rounds continued until consensus was reached, resulting in the development of the final implementation manual.

Importantly, the recommendations presented here should not be considered as a substitution of the Cochrane's RoB 2 tool manual. Rather, our aim is to provide further recommendations and an operationalization framework to facilitate the ratings in the context of psychological trials. This aligns with the RoB 2 feature that allows researchers to override the standard algorithmic scores when additional judgment or evidence is available.

We focused our recommendations on *individually randomized parallel-group trials*. Some of the core specifications from this extension can be applied for cluster randomized trials or crossover trials, although these types of designs pose specific challenges that should be thoroughly evaluated. We refer the reader to the Cochrane's RoB 2 guidance tailored to these types of designs.

RESULTS

Preliminary Assessment of the RoB 2

The preliminary assessment of 35 trials highlighted several challenges in applying the RoB 2 tool to psychotherapy trials. Challenges were identified across all domains, and included inconsistencies in assessing baseline imbalances (Domain 1) and general difficulties with blinding and defining deviations from intended interventions (Domain 2), selecting appropriate analysis methods for missing data (Do-

main 3), interpretation of self-report instruments (Domain 4), and procedures in the evaluation of pre-specified outcomes (Domain 5). More details on these findings are reported in the **Supplement A**.

The Operationalization of RoB 2

A detailed document describing the operationalization of the tool is provided in the Supplement B and it is summarized in this section of the article. We operationalized the guidance for each of the five domains (and 22 signalling questions) of the RoB 2 tool with 24 items that assist the application of the tool. We divided some signalling questions into multiple questions to ease their standardization across trials and to be able to keep better track of decisions. We also unified some signalling questions that could be rated simultaneously into one item. Importantly, all items were reformulated so that all are scored in the same direction, in which "Yes/PY" (Yes/Probably Yes) answers are always indicative of lower risk of bias. Based on the answers to the 24 items, an RoB score in each domain can be calculated through the algorithms in the Supplement B or by means of our digital assistant tools. These algorithms follow the recommendations from the Cochrane RoB 2 manual, and are supported by empirical evidence and content expertise in the field of psychotherapy.

The questions conforming our operationalization of the RoB 2 tool are presented in **Table 1**. Detailed descriptions for the assessment of each of these questions can be found in the online handbook (metapsy.org/rob/handbook) or in its pdf version (Supplement B). In the following sections of this article, we will provide a general overview of the different RoB domains while contextualizing the tool to psychotherapy outcome research. Supplement D presents the results for the 35 trials rated during the pilot exercises, with assessments conducted using the operationalization presented in this manuscript.

Preliminary Considerations

One important point that reviewers should specify when using the RoB 2 tool is the nature of the effect of interest. The *effect of assignment to the interventions*, or intention-to-treat (ITT) effect, evaluates effectiveness regardless of whether participants receive or adhere to the interventions. This effect could be of interest when the aim is to evaluate the effect of offering an intervention in a health system. In contrast, the *effect of adhering to the intervention*, or per-protocol effect, evaluates the effectiveness of adhering to the intervention as specified in the trial protocol, which is more directly applicable to clinical decisions for individual patients (Higgins et al., 2019). Reviewers may also be interested in examining both effects. The choice of effect of interest has an impact on the assessment of Domain 2. In our operationalization, we focused on the *effect of assignment to intervention*, as this is typically prioritized by clinical guideline developers and policymakers. However, if reviewers wish to assess the effect of adhering to intervention, Domain 2 of the RoB 2 tool contains a section specific to that aim.

As recommended by the RoB 2 tool (Sterne *et al.*, 2019), risk of bias should be assessed for a specific result, outcome, and time point, although some biases may apply to the entire study.

Domain 1. *Bias Arising from the Randomization Process*

Randomization aims to balance prognostic factors between the groups, preventing confounding. To ensure successful randomization, participants should be allocated to groups through a process involving chance, such as a random numbers table (Higgins *et al.*, 2019). This process is known as allocation sequence generation. The next step that must be taken is to prevent selective allocation of participants based on prognostic factors. This can happen when trial staff and participants have knowledge about what the next assignment is and can therefore have influence on the allocation to the groups. A safeguard to prevent this is allocation sequence concealment, such as organizing the allocation process through an independent organization.

The RoB 2 tool also includes an assessment of baseline imbalances to detect problems with the randomization process. Severe imbalances may suggest deliberate alterations to the randomization process. Important imbalances to consider include large differences in group sizes (relative to the intended allocation ratio), excess in statistically significant differences in baseline characteristics, imbalances in key prognostic factors, and excessive similarity between groups at baseline. Importantly, these imbalances should exceed what could be expected by chance.

The specific considerations in psychotherapy trials pertain to defining the key prognostic factors that are specific to each psychotherapeutic intervention and mental health problem. We focused on two aspects: imbalance in the number of randomized participants and baseline clinical severity. Imbalance in randomized participants is examined through the proportion assigned to one group and its 99% confidence intervals. If the intervals do not cover the expected proportion (e.g., 50% in a 1:1 allocation), this may indicate imbalance. Baseline severity is examined using the standardized mean difference (SMD) between groups at baseline and its 99% confidence intervals. An SMD > 0.2 (with 99% CI lower limit > 0) could be suggestive of baseline imbalance.

We selected baseline severity as the key prognostic factor because it is consistently reported across trials, can be examined objectively, and it has been shown to be a consistent predictor of treatment outcome in various types of psychological interventions for common mental health conditions, such as depression (Driessen, Cuijpers, Hollon, & Dekker, 2010; Bower *et al.*, 2013; Karyotaki *et al.*, 2018; Reins *et al.*, 2021; Kuyken *et al.*, 2016; Furukawa *et al.*, 2018; Furukawa *et al.*, 2021; Cuijpers *et al.*, 2022), anxiety (e.g., Rosenfield *et al.*, 2019; Scholten *et al.*, 2023) or PTSD (de Haan *et al.*, 2024; Wright *et al.*, 2024). Other prognostic factors may emerge depending on the mental disorder and clinical context. In such cases, our operationalization should be adjusted to include additional variables.

Domain 2. *Bias Due to Deviations from Intended Interventions*

As described earlier, we focused on the effect of assignment to interventions. This domain assesses bias arising when the intended interventions are not delivered as planned. Reviewers should assess three key aspects (Higgins et al., 2019): (1) whether participants and personnel were blinded, (2) whether any deviations occurred due to the trial context, and (3) whether appropriate analyses were used to estimate the effect of assignment i.e., ITT analysis.

This domain focuses on deviations from the intended interventions as specified in the trial protocol, particularly those arising because the intervention was delivered in the context of a randomized trial rather than routine care (Higgins et al., 2019). The intended interventions must be pre-specified in detail in the trial protocol, including any changes allowed in response to participant status (e.g., switching if the patient worsens). Changes that align with the protocol should not be considered as bias. This includes a precise definition of the control condition. Unfortunately, these descriptions are often lacking or vague in trial protocols and publications. One notable example is care as usual, a highly heterogeneous control condition that is typically under-described in psychotherapy trials (Watts et al., 2015; Cuijpers et al., 2021).

One important type of deviation is non-protocol interventions (also known as co-interventions). These are interventions received during the trial that were not part of the specified plan and that may introduce bias if they are related to the outcome. Whether such interventions introduce bias depends on the review's research question (Higgins et al., 2019). Even when co-interventions are not fully specified in the trial protocol, they may still be acceptable if they align with the broader review protocol. For example: In a trial of psychotherapy for depression, a participant allocated to "treatment as usual" (e.g., psychoeducation by a general practitioner) was dissatisfied with the randomization outcome. This participant had expected to receive the promising intervention developed for this trial, as announced in a flyer. As a result, the participant seeks additional care from another clinic and receives the same type of psychoeducation as in the "treatment as usual" group.

This does not constitute a deviation. However, if the participant is referred to a psychiatrist who prescribes antidepressant medication, this is a deviation from the intended control condition, as antidepressant medication was not pre-specified in the protocol. Since antidepressant medication will likely affect the participant's symptoms, this deviation is expected to influence the outcome. If, however, the referral aligns with the broader definition of "treatment as usual" in the review, it may not compromise the validity or applicability of the trial's results in the context of the review. This highlights the importance of clearly defining what constitutes "treatment as usual" in the protocol and trial reports. A vague or flexible approach in defining this can obscure what the RCT is actually comparing.

This example illustrates a key factor that may give rise to deviations from the intended interventions: blinding. In pharmacological trials, participants and person-

nel are unaware of whether the participant is receiving the active treatment or a placebo. Successful blinding protects against deviations caused by knowledge of the assignment. However, attempts to blind participants and personnel do not always result in successful blinding (e.g., side effects in drug trials might break the blinding). Thus, the success of blinding should be carefully assessed (Lin *et al.*, 2022) (Tajika *et al.*, 2023). At the same time, failure of blinding does not automatically mean bias due to deviation from the intended interventions. In psychotherapy, blinding is rarely feasible and arguably counterproductive, as therapeutic engagement requires participant awareness and expectations (Baskin, Tierney, Minami, & Wampold, 2003; Munder & Barth, 2018). In fact, this awareness has been described as a necessary mechanism of therapeutic change (Doering, Glombiewski, & Rief, 2018; Wampold & Imel, 2015). However, this lack of masking may introduce bias through subtle psychological effects, such as disappointment. For example, participants in a waitlist control group may be less motivated to engage in solving their problems because they know they will eventually receive the active intervention. This has been evidenced in the literature, with waitlist-controlled trials being associated with an overestimation of psychotherapy effects (Furukawa *et al.*, 2014; Cuijpers, Miguel, Harrer, Ciharova, & Karyotaki, 2024; Michopoulos *et al.*, 2021). This risk can be mitigated by selecting more robust control conditions, such as high-quality protocolized usual care.

The influence of these subtle factors (expectations, disappointment) can often be inferred indirectly by examining deviations from protocolized interventions. Reviewers assessing this domain must consider whether deviations arose due to the trial context itself or whether they would likely have occurred in routine care. (Higgins *et al.*, 2019). For instance, dropouts due to side effects of antidepressants would be expected even outside the trial context and, therefore, would not introduce bias in an ITT analysis. If these influential deviations are unbalanced between groups, it is more likely that the intervention effect estimate is biased. Reviewers should make agreements on how an imbalance between groups is defined.

The last part of this domain evaluates whether there was an appropriate analysis to estimate the effect of assignment to intervention. In this case, an adequate analysis should follow the ITT principles (Hollis & Campbell, 1999; Detry & Lewis, 2014), analyzing participants according to the group to which they were originally assigned (regardless of the intervention finally received) and including all randomized participants in the analysis (regardless of the actual adherence of the participants; Higgins *et al.*, 2019). Approaches such as “per-protocol” analyses (exclusion of those who did not receive the allocated intervention) or “as treated” analysis (analysis of participants according to the intervention received, rather than the originally assigned) should be considered inappropriate. For example, it is not appropriate if participants are excluded due to not receiving a minimum number of treatment sessions. Reviewers should examine if such approach may have resulted in a substantial impact in the effect estimates. If the number of excluded participants or participants analyzed in the wrong group is very small, the impact will be less likely. A threshold of 5% could be taken as a rule of thumb; however, even exclusions of less than 5% can have a substantial impact if the outcome is rare or if exclusions

are linked to prognostic factors (Higgins et al., 2019). Therefore, a precise assessment of the available information in each trial is warranted.

Domain 3. *Bias Due to Missing Outcome Data*

Missing outcome data (i.e., missing data from the measured outcome at a given time point) poses a threat to the validity of trial results. It can compromise the core principle of randomization, which aims to balance prognostic factors between groups. For example, if more severely depressed participants drop out from the psychotherapy group than from the control group, and the analysis is based only on those who remain in the study, the results are likely to be biased. Therefore, the reasons for missing data are a very important source of information for understanding whether missing could be associated to bias. If the reason for missing data is unrelated to the outcome (e.g., funding termination or technical problems during assessment), the results will less likely be biased. However, if missing data is related to the intervention or the true value of the outcome (e.g., side effects, worsening symptoms), bias is more likely.

To rate this domain, reviewers should first evaluate whether outcome data is available for nearly all randomized participants. As generally outlined in the Cochrane RoB 2 manual (Higgins et al., 2019), we consider that bias is likely ruled out when data is available for 95% or more of the randomized participants. If missing data exceeds this threshold, further evaluation of how the missing data was handled is required. A careful assessment is recommended for rare binary outcomes, as even smaller proportions of missing data can lead to RoB.

Appropriate methods for handling the impact of missing data may include mixed models for repeated measures, also known as growth curves, using several time points (e.g., mid-treatment, post-test, follow-ups), multiple imputation following the state of the art procedures (application of Rubin's combination rules or similar methods that incorporate imputation uncertainty, use of important auxiliary variables in the imputation model, imputations generated separately by arms, or controlled/reference-based imputation; Barnard & Rubin, 1999; von Hippel & Bartlett, 2021; Sullivan, White, Salter, Ryan, & Lee, 2018; Carpenter, Roger, & Kenward, 2013; Cro, Morris, Kenward, & Carpenter, 2020), or sensitivity analyses testing a range of plausible assumptions for the missing mechanism, resulting in no substantial differences between them. Other methods for handling missing data may also be adequate, such as generalized estimating equations (Little et al., 2012). Methods for imputing data such as "last observation carried forward" should not be considered as appropriate.

If no appropriate methods for correcting the impact of missing data were used (or if this is unclear), RoB 2 users can further look at indirect evidence that may be indicative of risk of bias (Higgins et al., 2019). In our implementation guideline, we focus on examining a) whether the reported reasons for missing data indicate that missingness could depend on its true value (e.g., funding termination would not be

indicative of risk of bias), *b*) large imbalances in the proportion of missing data between the groups (considering as large imbalance a difference between the groups of above 20%, as a rule of thumb), and *c*) differing reasons for missing data between the groups (e.g., in the intervention group missing data is related to clinical severity, while in the control group participants with higher severity remain in the trial to get access to the intervention after the post-assessment).

Domain 4. Bias in Measurement of the Outcome

This domain evaluates biases that arise from systematic differences between experimental and control groups in the measurement of the outcome. To assess this domain, RoB 2 recommends evaluating *a*) whether the method of measuring the outcome is appropriate, *b*) whether it does not differ between groups, *c*) who is the assessor, *d*) whether the assessor is blinded, and *e*) whether the assessment of the outcome is likely influenced by knowledge of treatment allocation.

The evaluation of the first two points is straightforward. Trials must assess the outcome using an appropriate instrument (e.g., properly validated), and the assessment process must be consistent between groups. The third question, regarding who the assessor is, can be more complicated in psychotherapy trials. In essence, all mental health outcomes are patient-reported, either in a self-rated questionnaire or as answers during an interview with a clinician. In the first case, participants rate their symptoms by themselves through the completion of a questionnaire (e.g., Beck Depression Inventory II, BDI-II). In the second case, a clinician interviews the participant in a structured or semi-structured format. The advantage of involving a clinician in the assessment is that, in principle, as compared to self-reports, a further element of blinding can be included. For instance, using an independent assessor who is unaware of treatment allocation. Still, even when blinded clinicians are the evaluators, these outcomes are essentially patient-reported, and thus we cannot rule out an effect of unmasking. However, it has been generally assumed that even in those circumstances, masked clinicians can offer more objective assessments. For example, clinicians can interpret the changes in behavior, facial expressions, or verbal outputs of the patients. Masked clinician assessments are thus usually assessed as low risk of bias by reviewers in our field, while the use of self-reports has often been assessed very differently between reviews (ranging from low risk to high risk).

Given these divergences in the field, we empirically examined whether there are differences between these types of assessments across psychotherapy trials for depression (Miguel et al., 2025; Cuijpers, Li, Hofmann, & Andersson, 2010). We found that self-reports produced somewhat smaller effect sizes compared to clinician-rated instruments, probably not resulting in clinically meaningful differences when clinicians are masked. Depression scales administered by unmasked clinicians produced substantially larger effect estimates than self-reports, with a differential SMD of 0.20. These results indicate that the lack of blinding of participants does not seem to influence self-reported assessments compared to the gold stan-

dard measurement of masked clinicians. We cannot truly evaluate the difference between self-reports and a completely objective measure of depression, because such a measure does not exist in the context of psychotherapy.

Therefore, based on the comparable outcomes between self-reports and the gold-standard of masked clinician-rated scales, and based on the precondition that participants need to be aware of receiving a psychotherapeutic intervention, we consider that there is no evidence for bias arising from the use of self-reports. It is not likely that the assessment of the outcome is influenced by the knowledge of the intervention in this case.

Still, some considerations should be checked for self-reports. For example, there should be no evidence of participants misreporting their symptoms - or there should be a safeguard to prevent this. If there is concern that patients underreported their severity (e.g., to please the therapists or researchers when self-reports are completed in front of them therapist at the last session), reviewers should consider whether this could lead to a higher risk of bias.

Domain 5. Bias in Selection of the Reported Result

This domain assesses bias due to selecting the reported result. This means that the results that are reported from a trial are based on their direction, magnitude, or statistical significance. For instance, by selecting among multiple effect estimates from different outcome domains (e.g., anxiety severity), scale and time point (e.g., GAD-7 at 8-weeks post-randomization), or analyses (e.g., differences between groups in mean change from baseline to week 8).

The implementation of safeguards against this type of bias still has much room for improvement in the field of psychological interventions. In an in-depth examination of this bias across trials on psychotherapies for depression (Miguel et al., 2021), we found that only 21% of trials (75 out of 353) were prospectively registered. Of these 75 prospectively registered trials in which selective reporting could be reliably examined, 17% were judged to involve selection of the reported result. The method of aggregation or analysis metric could not be examined due to the small number of trials that pre-specified this information. Effect sizes diverged by an SMD of 0.27 between trials with and without selective reporting, suggesting that this bias may be associated with overestimated treatment effects. This lack of prospective registration is a widespread issue in psychological treatment research, also observed in trials on anxiety disorders (Papola et al., 2022), obsessive compulsive disorder (Wang et al., 2024), borderline personality disorder (Storebø et al., 2020), and post-traumatic stress disorder (Mavranouzouli et al., 2020), among others.

As recommended in the RoB 2 manual (Higgins et al., 2019), we assess selection based on a particular outcome measurement and on a particular analysis. The evaluation of this domain depends on the meta-analysts' protocol regarding the choice of outcome measures. Some meta-analysts may include all available mea-

surements for a given outcome domain (e.g., anxiety symptoms), even if this entails including multiple measurements from within a study. Others might decide on more restrictive approaches (e.g., including only assessor-rated anxiety severity). To guide the evaluation, reviewers can check if the trial has a registration, protocol, or statistical analysis plan. If available, it should be ascertained whether the outcomes were prospectively specified, (i.e., before unblinded outcome data were available). At least scale and time point should be pre-specified. An analysis plan should also be prospectively specified, although these are still often unavailable. Once a reliable source of pre-specified scale(s)/time point/analysis is found, reviewers can evaluate whether all outcomes of interest were appropriately reported and if the reported analysis was not selected.

Examples of selective reporting include selecting one time point from a range of measurements, reporting only one instrument from a set of pre-specified measures, or analyzing data as a binary outcome (e.g., remission) when an analysis based on continuous data (e.g., change from baseline in symptom severity) was pre-specified. Reviewers should judge whether these selections were influenced by the direction, magnitude, or statistical significance of the results. Also, they should consider whether data from the outcome of interest can be made available through other means, such as contacting study authors. Importantly, RoB 2 is assessed at the outcome level. Therefore, even if a trial shows signs of selective reporting through e.g., omitting a pre-specified outcome for anxiety in systematic review for depression, this does not lead automatically to high RoB if the omitted outcome is not considered of interest for the review.

Digital Assistant Tools

To facilitate the implementation of our operationalization, we developed two types of digital assistant tools tailored to different use cases. The most user-friendly option is the online *RoB Assistant* (metapsy.org/rob/assistant, **Figure 1**), an online questionnaire that allows users to answer each item directly. Brief guidance is provided with each item, with more detailed explanations available in the online handbook or its pdf version (**Supplement B**). Each question also includes links to related materials from Cochrane's handbook (Higgins *et al.*, 2019). The assistant allows reviewers to address relevant signalling questions step by step. Based on the given answers, the tool automatically skips questions that are not required to complete the rating algorithm. Once the questions are completed, the assistant calculates the domain and overall scores using the algorithms that run behind the web-app. The results for the rated study can be downloaded in various formats. We believe this web-app is particularly helpful for training novel reviewers in RoB rating.

For experienced reviewers and those working with large meta-analytic datasets, we developed a spreadsheet-based version of the tool (**Supplement C; D**), which offers more flexibility than the online RoB Assistant. This spreadsheet integrates with the *metapsyTools* package in RStudio (Harrer, Kuper, Sprenger, & Cuijpers, 2022), as well as its user-friendly RoB web-app (metapsy.org/rob/database-app). To use these functions, users need to provide the datasets used in their analyses and the predefined spreadsheet rating template (**Supplement C; D**).

Table 1: Questionnaire outlining RoB 2 specifications for use in psychotherapy trials.

RoB 2 specifications for the use in psychotherapy trials			
<p>The questionnaire below aims to facilitate the application of Cochrane's risk of bias tool version 2 in the context of psychotherapy trials. The 24 questions evaluate risk of bias arising from the five domains contained in the original tool, with additional specifications for psychological interventions. For a correct application, please consult the accompanying manual with all the specifications for each item, as well as the RoB 2 guidance manual (Higgins et al., 2019). For an interactive version of this questionnaire, please visit https://www.metapsy.org/rob/assistant/. Response options: every question can be responded with Yes/PY (Probably Yes), No/PN (Probably No), or NI (No Information).</p>			
Domain	RoB 2	Question	
1	1.1	1	Was the allocation sequence random?
	1.2	2	Was the allocation sequence concealed until participants were enrolled and assigned to interventions?
	1.3	3	Is the number of randomized participants balanced between groups?
		4	Is baseline clinical severity balanced between groups?
2	2.1	5	Were participants masked to the intervention?
	2.2	6	Were carers and people delivering the interventions masked?
	2.3	7	If "No/PN" or "NI" to #5 or #6, Did the interventions closely align with the intended plan, without any significant deviations that could affect the results?
	2.6	8	Was an appropriate analysis used to estimate the effect of assignment to intervention?
	2.7	9	If "No/PN" or "NI" to #8, is the inappropriate analysis unlikely to have an impact on the result?
3	3.1	10	Was data available for nearly all participants?
	3.2	11	If "No/PN" or "NI" to #10, was an appropriate analysis used for handling the impact of missing data?
	3.3, 3.4	12	If "No/PN" or "NI" to #11, Is the trial of moderate to large size (minimum 40 participants per arm reported data)?
		13	If "No/PN" or "NI" to #11, Did at least 70% of the randomized participants complete the assessments at the post-treatment?
		14	If "No/PN" or "NI" to #11, Is there indirect evidence indicating that missingness is unrelated to the outcome?
4	4.1, 4.2	15	Is the method of measurement appropriate and applied using similar procedures for both groups?
	4.3, 4.4, 4.5	16	Were self-report measures used?
		17	Were assessor-rated instruments used?
		18	If "Yes" to #17, were the assessors masked to treatment allocation?
5	5.1, 5.2, 5.3	19	Is there a trial registration, protocol, or statistical analysis plan (SAP) available?
		20	Write down the registration number, protocol, or link to SAP
		21	If "Yes" to #19, were all outcomes of potential interest for the meta-analysis (scale and time point) pre-specified before unmasked outcome data were available?
		22	If "Yes" to #21, are all outcomes of potential interest for the meta-analysis fully reported in the paper or available through other means? (e.g., contact with authors)
		23	If "Yes" to #19, Was there an analysis plan pre-specified?
		24	If "Yes" to #23, was data analyzed in accordance with the pre-specified plan?

Reference. Higgins, J. P., Savović, J., Page, M. J., Elbers, R. G., & Sterne, J. A. (2019). Assessing risk of bias in a randomized trial. *Cochrane handbook for systematic reviews of interventions*, 205-228. *Note.* Detailed descriptions for the assessment of each of these questions can be found in the online handbook ([metapsy.org/rob/handbook](https://www.metapsy.org/rob/handbook)) or in its pdf version (**Supplement B**). The algorithms for calculating overall and domain RoB scores are also available in Supplement B and in the online handbook. These algorithms follow the same logic as Cochrane's RoB 2, with the overall score being the worst RoB score in any of the domains (e.g., overall high risk if one domain has a high risk score).

Figure 1. Digital assistant tool to facilitate of risk of bias ratings in the context of psychotherapy trials.

Risk of Bias Assistant
Version 1.0-0 - Cite this tool

Start

Randomization Process

- Was the allocation sequence random?
- Was the allocation sequence concealed?
- Is the number of randomized participants balanced between groups?
- Is baseline clinical severity balanced between groups?

Notes

Intervention Deviations

Missing Data

Outcome Measurement

Selection of Results

Assessment

4 Is baseline clinical severity balanced between groups?

No/PN No Information Yes/PY

Explanation

Imbalance in baseline symptom severity is examined by the SMD at baseline and its 99% confidence intervals. An SMD > 0.2 (with 99% CI lower limit > 0) will be suggestive of baseline imbalance. In the Metapsy RoB addendum, we focus on baseline severity only because it is consistently reported across trials, can be examined objectively, and it is a known predictor of the outcome (for depression interventions). P-values and confidence intervals resulting from multiple significance testing (e.g., due to multiple clinical severity measures) can be adjusted using Bonferroni corrections.

Depending on the clinical field and mental health disorder, there might be key prognostic factors that should be added to this examination. In that case, the judgment of the reviewers can overwrite the algorithm and the proposed items and further variables can be examined for baseline imbalances.

Note. The digital assistant tool is openly available at metapsy.org/rob/assistant.

This setup allows to perform assessments for each specific trial result that will be meta-analyzed. Importantly, these functions can automatically pre-calculate some standardized items, such as unbalanced clinical severity between the groups, based on the collected review data (e.g., baseline means and standard deviations for the outcome of interest). Domain and overall scores are automatically generated and appended to the analysis dataset. Additionally, other useful functions from the *metapsyTools* package can be applied, such as “*checkRoBDiscrepancies*”, which flags discrepancies between raters and provides Cohen’s κ as an index of agreement.

Discussion

Evaluating risk of bias is an essential step in any well-conducted systematic review. By using risk of bias assessment tools such as RoB 2 (Sterne et al., 2019), reviewers are able to identify methodological flaws in RCTs that may distort the true effect of an intervention. An accurate assessment of bias requires careful consideration of the specific context of each review. It has been proposed that implementation manuals tailored to these contexts may improve the reliability of the assessments (Minozzi et al., 2020). In research domains where the conventional gold-standard procedures for RCTs are not easily applicable, such as in psychotherapy, these implementation manuals may provide particularly valuable guidance.

Given this context, and also based on our need to assess RoB in a large, living database of over 1,000 trials on psychotherapy for depression (www.metapsy.org) (Cuijpers et al., 2023), we developed an operationalization and implementation manual for Cochrane’s RoB 2 tool in the context of psychotherapy outcome research. This

operationalization was developed through consensus meetings and piloting exercises involving experienced mental health researchers, and was guided by meta-epidemiological evidence in the field. An additional important aim of this operationalization was to facilitate the use of RoB 2 in large meta-analytic projects by reducing application time and improving agreement among multiple reviewers. To facilitate its application, we accompanied the operationalization with open digital tools (metapsy.org/rob/), which may be particularly helpful to researchers involved in similar projects and can serve as a resource for training new reviewers.

In the near future, we aim to apply RoB 2 to all the trials included in this meta-analytic database. Given the large number of studies that need to be rated, we will approach this task in phases, starting with the largest, most homogeneous set of trials (psychotherapy vs. control conditions, >500 trials). This initial phase will allow us to test the tool's performance and examine some domains in depth. We hope that the application of the RoB 2 assessments to this complete cohort of trials will provide future insights to the meta-epidemiological literature in the psychotherapy field. For instance, we will be able to estimate the prevalence of different types of biases, as well as their association with under- or overestimation of treatment effects.

It is important to note that this operationalization is not intended to substitute Cochrane's RoB 2 tool but to supplement it, providing context-specific recommendations. Several limitations of our implementation manual for RoB 2 should also be taken into account. First, the use of cut-offs in evaluating some items (e.g., 5% of missing data, SMD > 0.2 for baseline severity imbalance, etc.) may not align with RoB 2 guidance, which recommends avoiding fixed cut-offs in favor of a more nuanced and individualized evaluation of each trial. However, in meta-analytic projects involving a large number of trials and multiple reviewers (as in our future implementation case in the Metapsy database), the use of predefined cut-offs might facilitate inter-reviewer agreement, as well as the development of digital assistant tools. The application of such cut-offs, along with the overall high level of item standardization, was a necessary and justifiable compromise because of the specific focus of our dataset and homogeneity of the trials therein.

Another limitation is that this article and the domain explanations primarily focus on individually randomized parallel-group trials, as they are currently the most common design in psychotherapy trials. Nevertheless, most of the items in this operationalization can also be applied to cluster trials, adding the signaling questions specific to these designs. Similarly, due to the nature of our project, we also focused our implementation guideline for reviews that aim at evaluating the "intention-to-treat effect" rather than the "per-protocol" effect. Reviewers focusing on the latter should consult the guidance in the RoB 2 manual for assessing Domain 2 (Higgins *et al.*, 2019).

One further limitation is the lack of inter-rater agreements on the initial ratings used to develop the final version of the tool. Given that this was not a pre-planned project, and the discrepancies were lengthy discussed in group until consensus was reached, we did not keep track of all the disagreements. We plan to present inter-

rater agreement data in future work, where we will rate the >500 trials comparing psychotherapies to control conditions. A further limitation is the lack of empirical research across psychotherapy trials to inform decision-making in some RoB 2 domains (e.g., deviations from intended interventions).

This gap could be a consequence of poor reporting in the primary trials. Much of the meta-epidemiological evidence guiding our operationalization comes from depression trials, although we did not intend to limit it to this field. We plan to update our online handbook as new evidence becomes available to inform and refine our recommendations. Finally, it is worth noting that RoB ratings can be adapted when reviewers have access to individual patient data (IPD) from the included trials. For example, missing outcome data may be adequately handled by the reviewer's statistical models, and ratings for selective outcome reporting might change if unpublished data becomes available via IPDs.

Due to the scope of this article, we have not focused on other general problems in psychotherapy research but that are also relevant for the design of future RCTs. For example, problems with the use of weak control groups (e.g., waitlist), the imprecise definition of "treatment as usual", researcher allegiance, small sample sizes, lack of assessment of adverse events, and poor reporting in general (Cuijpers & Cristea, 2016) (Cuijpers, *in press*) (Cuijpers, 2019) (Watts, Turnell, Kladnitski, Newby, & Andrews, 2015) (Cuijpers, Quero, Papola, Cristea, & Karyotaki, 2021). Moreover, there are other aspects related to the quality of implementation of treatments that are also important to consider in the psychotherapy field. The training and supervision of therapists, use of a manual, or the evaluation of treatment fidelity are relevant for evaluating the effect of adhering to intervention.

Despite these limitations, we hope that researchers in our field will benefit from the general conceptualization of risk of bias in psychotherapy trials discussed in this article, and that the presented materials serve as helpful guidance for future meta-analytical projects. The use of implementation manuals can contribute to a more efficient use of time and personnel resources, and most importantly, to reduce errors and increase consistency and transparency across the ratings (Minozzi *et al.*, 2022). Not accounting for the unique characteristics of psychotherapy trials may lead to researchers omitting or inadequately evaluating certain RoB domains, which has important implications. The certainty of the evidence, evaluated through means such as GRADE assessment (Guyatt *et al.*, 2008), partly depends on the risk of bias assessed in systematic reviews. If systematic reviews use very different standards for assessing risk of bias, the conclusions reached in clinical guidelines could be misleading. Thus, reaching some common ground for the implementation of RoB 2 in psychotherapy trials is essential, particularly for large systematic reviews that can significantly influence policy-making decisions.

□

References

- Barnard, J., & Rubin, D. (1999).** Miscellanea. Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86(4), 948-955. doi:10.1093/biomet/86.4.948
- Baskin, T. W., Tierney, S. C., Minami, T., & Wampold, B. E. (2003).** Establishing Specificity in Psychotherapy: A Meta-Analysis of Structural Equivalence of Placebo Controls. *Journal of Consulting and Clinical Psychology*, 71(6), 973-979. doi:10.1037/0022-006X.71.6.973
- Bower, P., Kontopantelis, E., Sutton, A., Kendrick, T., Richards, D. A., Gilbody, S., . . . Liu, E. T. (2013).** Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. *Bmj*, 346, f540. doi:10.1136/bmj.f540
- Carpenter, J. R., Roger, J. H., & Kenward, M. G. (2013).** Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *J Biopharm Stat*, 23(6), 1352-1371. doi:10.1080/10543406.2013.834911
- Cro, S., Morris, T. P., Kenward, M. G., & Carpenter, J. R. (2020).** Sensitivity analysis for clinical trials with missing continuous outcome data using controlled multiple imputation: A practical guide. *Statistics in Medicine*, 39(21), 2815-2842. doi:
- Crocker, T. F., Lam, N., Jordão, M., Brundle, C., Prescott, M., Forster, A., . . . Clegg, A. (2023).** Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. *Journal of clinical epidemiology*, 161, 39-45. doi:
- Cuijpers, P. (2019).** Targets and outcomes of psychotherapies for mental disorders: an overview. *World Psychiatry*, 18(3), 276-285. doi:
- Cuijpers, P. (in press).** Has the time come to stop using control groups in trials of psychosocial interventions? . *World Psychiatry*.
- Cuijpers, P., Ciharova, M., Quero, S., Miguel, C., Driessen, E., Harrer, M., . . . Karyotaki, E. (2022).** The Contribution of "Individual Participant Data" Meta-Analyses of Psychotherapies for Depression to the Development of Personalized Treatments: A Systematic Review. *J Pers Med*, 12(1). doi:10.3390/jpm12010093
- Cuijpers, P., Harrer, M., Miguel, C., Ciharova, M., & Karyotaki, E. (2023).** Five decades of research on psychological treatments of depression: A historical and meta-analytic overview. *American Psychologist*, No Pagination Specified-No Pagination Specified. doi:10.1037/amp0001250
- Cuijpers, P., Li, J., Hofmann, S. G., & Andersson, G. (2010).** Self-reported versus clinician-rated symptoms of depression as outcome measures in psychotherapy research on depression: a meta-analysis. *Clin Psychol Rev*, 30(6), 768-778. doi:10.1016/j.cpr.2010.06.001
- Cuijpers, P., Miguel, C., Harrer, M., Ciharova, M., & Karyotaki, E. (2024).** The overestimation of the effect sizes of psychotherapies for depression in waitlist controlled trials: a meta-analytic comparison with usual care controlled trials. *Epidemiology and Psychiatric Sciences*, 33, e56.
- Cuijpers, P., Quero, S., Papola, D., Cristea, I. A., & Karyotaki, E. (2021).** Care-as-usual control groups across different settings in randomized trials on psychotherapy for adult depression: a meta-analysis. *Psychol Med*, 51(4), 634-644. doi:10.1017/s0033291719003581

- de Haan**, A., Meiser-Stedman, R., Landolt, M. A., Kuhn, I., Black, M. J., Klaus, K., . . . Dalgleish, T. (2024). Efficacy and moderators of efficacy of cognitive behavioural therapies with a trauma focus in children and adolescents: an individual participant data meta-analysis of randomised trials. *Lancet Child Adolesc Health*, 8(1), 28-39. doi:10.1016/s2352-4642(23)00253-5
- Detry**, M. A., & Lewis, R. J. (2014). The intention-to-treat principle: how to assess the true effect of choosing a medical treatment. *Jama*, 312(1), 85-86.
- Doering**, B. K., Glombiewski, J. A., & Rief, W. (2018). Chapter Fourteen - Expectation-Focused Psychotherapy to Improve Clinical Outcomes. In L. Colloca (Ed.), *International Review of Neurobiology* (Vol. 138, pp. 257-270): Academic Press.
- Driessen**, E., Cuijpers, P., Hollon, S. D., & Dekker, J. J. (2010). Does pretreatment severity moderate the efficacy of psychological treatment of adult outpatient depression? A meta-analysis. *J Consult Clin Psychol*, 78(5), 668-680. doi:10.1037/a0020570
- Furukawa**, T. A., Efthimiou, O., Weitz, E. S., Cipriani, A., Keller, M. B., Kocsis, J. H., . . . Schramm, E. (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(3), 140-153. doi:10.1159/000489227
- Furukawa**, T. A., Noma, H., Caldwell, D. M., Honyashiki, M., Shinohara, K., Imai, H., . . . Churchill, R. (2014). Waiting list may be a nocebo condition in psychotherapy trials: a contribution from network meta-analysis. *Acta Psychiatr Scand*, 130(3), 181-192. doi:10.1111/acps.12275
- Furukawa**, T. A., Sukanuma, A., Ostinelli, E. G., Andersson, G., Beevers, C. G., Shumake, J., . . . Cuijpers, P. (2021). Dismantling, optimising, and personalising internet cognitive behavioural therapy for depression: a systematic review and component network meta-analysis using individual participant data. *Lancet Psychiatry*, 8(6), 500-511. doi:10.1016/s2215-0366(21)00077-8
- Group**, E.-B. M. W. (1992). Evidence-based medicine. A new approach to teaching the practice of medicine. *Jama*, 268(17), 2420-2425. doi:10.1001/jama.1992.03490170092032
- Guyatt**, G. H., Oxman, A. D., Vist, G. E., Kunz, R., Falck-Ytter, Y., Alonso-Coello, P., & Schünemann, H. J. (2008). GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. *Bmj*, 336(7650), 924-926. doi:10.1136/bmj.39489.470347.AD
- Harrer**, M., Kuper, P., Sprenger, A., & Cuijpers, P. (2022). metapsyTools: Several R Helper Functions For the "Metapsy" Database. doi:10.5281/zenodo.6566632
- Higgins**, Savović, Page, Sterne, & Group, T. R. D. (2019). Revised Cochrane risk-of-bias tool for randomized trials (RoB 2). Retrieved from
- Higgins**, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane Handbook for Systematic Reviews of Interventions* (Second ed.). Chichester (UK): John Wiley & Sons.
- Hollis**, S., & Campbell, F. (1999). What is meant by intention to treat analysis? Survey of published randomised controlled trials. *Bmj*, 319(7211), 670-674. doi:10.1136/bmj.319.7211.670
- Jørgensen**, L., Paludan-Müller, A. S., Laursen, D. R., Savović, J., Boutron, I., Sterne, J. A., . . . Hróbjartsson, A. (2016). Evaluation of the Cochrane tool for assessing risk of bias in randomized

clinical trials: overview of published comments and analysis of user practice in Cochrane and non-Cochrane reviews. *Systematic reviews*, 5, 1-13.

Karyotaki, E., Ebert, D. D., Donkin, L., Riper, H., Twisk, J., Burger, S., . . . Cuijpers, P. (2018). Do guided internet-based interventions result in clinically relevant changes for patients with depression? An individual participant data meta-analysis. *Clin Psychol Rev*, 63, 80-92. doi:10.1016/j.cpr.2018.06.007

Kuehn, R., Wang, Y., & Guyatt, G. (2024). Overly complex methods may impair pragmatic use of core evidence-based medicine principles. *BMJ Evidence-Based Medicine*, 29(3), 139-141. doi:10.1136/bmjebm-2024-112868

Kuyken, W., Warren, F. C., Taylor, R. S., Whalley, B., Crane, C., Bondolfi, G., . . . Dalgleish, T. (2016). Efficacy of Mindfulness-Based Cognitive Therapy in Prevention of Depressive Relapse: An Individual Patient Data Meta-analysis From Randomized Trials. *JAMA Psychiatry*, 73(6), 565-574. doi:10.1001/jamapsychiatry.2016.0076

Lin, Y.-H., Sahker, E., Shinohara, K., Horinouchi, N., Ito, M., Lelliott, M., . . . Furukawa, T. A. (2022). Assessment of blinding in randomized controlled trials of antidepressants for depressive disorders 2000-2013;2020: A systematic review and meta-analysis. *eClinicalMedicine*, 50. doi:10.1016/j.eclinm.2022.101505

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., . . . Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *N Engl J Med*, 367(14), 1355-1360. doi:10.1056/NEJMSr1203730

Mavranезouli, I., Megnin-Viggars, O., Daly, C., Dias, S., Welton, N. J., Stockton, S., . . . Pilling, S. (2020). Psychological treatments for post-traumatic stress disorder in adults: a network meta-analysis. *Psychological Medicine*, 50(4), 542-555. doi:10.1017/S0033291720000070

Michopoulos, I., Furukawa, T. A., Noma, H., Kishimoto, S., Onishi, A., Ostinelli, E. G., . . . Cuijpers, P. (2021). Different control conditions can produce different effect estimates in psychotherapy trials for depression. *J Clin Epidemiol*, 132, 59-70. doi:10.1016/j.jclinepi.2020.12.012

Miguel, C., Harrer, M., Karyotaki, E., Plessen, C. Y., Ciharova, M., Furukawa, T. A., . . . Cuijpers, P. (2025). Self-reports vs clinician ratings of efficacies of psychotherapies for depression: a meta-analysis of randomized trials. *Epidemiology and Psychiatric Sciences*, 34, e15. doi:10.1017/S2045796025000095

Minozzi, S., Cinquini, M., Gianola, S., Gonzalez-Lorenzo, M., & Banzi, R. (2020). The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *Journal of clinical epidemiology*, 126, 37-44. doi:

Minozzi, S., Dwan, K., Borrelli, F., & Filippini, G. (2022). Reliability of the revised Cochrane risk-of-bias tool for randomised trials (RoB2) improved with the use of implementation instruction. *Journal of clinical epidemiology*, 141, 99-105. Retrieved from

Munder, T., & Barth, J. (2018). Cochrane's risk of bias tool in the context of psychotherapy outcome research. *Psychother Res*, 28(3), 347-355. doi:10.1080/10503307.2017.1411628

- Papola, D., Ostuzzi, G., Tedeschi, F., Gastaldon, C., Purgato, M., Del Giovane, C., . . . Barbui, C.** (2022). Comparative efficacy and acceptability of psychotherapies for panic disorder with or without agoraphobia: systematic review and network meta-analysis of randomised controlled trials. *The British Journal of Psychiatry*, 221(3), 507-519. doi:10.1192/bjp.2021.148
- Reins, J. A., Buntrock, C., Zimmermann, J., Grund, S., Harrer, M., Lehr, D., . . . Ebert, D. D.** (2021). Efficacy and Moderators of Internet-Based Interventions in Adults with Subthreshold Depression: An Individual Participant Data Meta-Analysis of Randomized Controlled Trials. *Psychother Psychosom*, 90(2), 94-106. doi:10.1159/000507819
- Rosenfield, D., Smits, J. A. J., Hofmann, S. G., Mataix-Cols, D., de la Cruz, L. F., Andersson, E., . . . Otto, M. W.** (2019). Changes in Dosing and Dose Timing of D-Cycloserine Explain Its Apparent Declining Efficacy for Augmenting Exposure Therapy for Anxiety-related Disorders: An Individual Participant-data Meta-analysis. *Journal of Anxiety Disorders*, 68, 102149. doi:
- Scholten, W., Seldenrijk, A., Hoogendoorn, A., Bosman, R., Muntingh, A., Karyotaki, E., . . . Batelaan, N.** (2023). Baseline Severity as a Moderator of the Waiting List-Controlled Association of Cognitive Behavioral Therapy With Symptom Change in Social Anxiety Disorder: A Systematic Review and Individual Patient Data Meta-analysis. *JAMA Psychiatry*, 80(8), 822-831. doi:10.1001/jamapsychiatry.2023.1291
- Steinberg, E., Greenfield, S., Wolman, D. M., Mancher, M., & Graham, R.** (2011). Clinical practice guidelines we can trust: national academies press.
- Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., . . . Higgins, J. P. T.** (2019). RoB 2: a revised tool for assessing risk of bias in randomised trials. *Bmj*, 366, 14898. doi:10.1136/bmj.14898
- Storebø, O. J., Stoffers-Winterling, J. M., Völm, B. A., Kongerslev, M. T., Mattivi, J. T., Jørgensen, M. S., . . . Simonsen, E.** (2020). Psychological therapies for people with borderline personality disorder. *Cochrane Database Syst Rev*, 5(5), Cd012955. doi:10.1002/14651858.CD012955.pub2
- Sullivan, T. R., White, I. R., Salter, A. B., Ryan, P., & Lee, K. J.** (2018). Should multiple imputation be the method of choice for handling missing data in randomized trials? *Stat Methods Med Res*, 27(9), 2610-2626. doi:10.1177/0962280216683570
- Tajika, A., Furukawa, T. A., Shinohara, K., Kikuchi, S., Toyomoto, R., Furukawa, Y., . . . Leucht, S.** (2023). Blinding successfulness in antipsychotic trials of acute treatment for schizophrenia: a systematic review. *BMJ Ment Health*, 26(1). doi:10.1136/bmjment-2023-300654
- Tomlinson, E., Cooper, C., Davenport, C., Rutjes, A. W. S., Leeflang, M., Mallett, S., & Whiting, P.** (2024). Common challenges and suggestions for risk of bias tool development: a systematic review of methodological studies. *Journal of clinical epidemiology*, 171. doi:10.1016/j.jclinepi.2024.111370
- von Hippel, P. T., & Bartlett, J. W.** (2021). Maximum Likelihood Multiple Imputation: Faster Imputations and Consistent Standard Errors Without Posterior Draws. *Statistical Science*, 36(3), 400-420, 421. Retrieved from
- Wampold, B. E., & Imel, Z. E.** (2015). *The great psychotherapy debate : the evidence for what makes psychotherapy work* (Second edition. ed.). New York: Routledge.

Wang, Y., Miguel, C., Ciharova, M., Amarnath, A., Lin, J., Zhao, R., . . . Cuijpers, P. (2024). The effectiveness of psychological treatments for obsessive-compulsive disorders: a meta-analysis of randomized controlled trials published over last 30 years. *Psychological Medicine*, 54(11), 2838-2851. doi:10.1017/S0033291724001375

Watts, S. E., Turnell, A., Kladnitski, N., Newby, J. M., & Andrews, G. (2015). Treatment-as-usual (TAU) is anything but usual: A meta-analysis of CBT versus TAU for anxiety and depression. *Journal of Affective Disorders*, 175, 152-167. doi:10.1016/j.jad.2014.12.025

Wright, S. L., Karyotaki, E., Cuijpers, P., Bisson, J., Papola, D., Witteveen, A., . . . Sijbrandij, M. (2024). EMDR v. other psychological therapies for PTSD: a systematic review and individual participant data meta-analysis. *Psychological Medicine*, 54(8), 1580-1588. doi:10.1017/S0033291723003446