BMJ Open

Researcher allegiance in research on psychosocial interventions: meta-research study protocol and pilot study

Whitney Rose Yoder,[1] Eirini Karyotaki,[1] Ioana-Alina Cristea,[1,2] Daniëlle van Duin,[3] Pim Cuijpers[1]

[1]Department of Clinical, Neuro and Developmental Psychology Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[2]Department of Clinical Psychology and Psychotherapy, Babes-Bolyai University, Cluj-Napoca, Romania
[3]Trimbos-instituut, Utrecht, The Netherlands

Correspondence to
Mrs Whitney Rose Yoder;
w.r.yoder@vu.nl

ABSTRACT

Introduction One potential source of bias in randomised clinical trials of psychological interventions is researcher allegiance (RA). The operationalisation of RA differs strongly across studies, and there is not a generally accepted method of operationalising or measuring it. Furthermore, it remains unclear as to how RA affects the outcomes of trials and if it results in better outcomes for a preferred intervention. The aim of this project is to develop and validate a scale that accurately identifies RA, contribute to the understanding of the impact that RA has in a research setting and to make recommendations for addressing RA in practice.

Methods and analysis A scale will first be developed and validated to measure RA in psychotherapy trials. The scale will be validated by surveying authors of psychotherapy trials to assess their opinions, beliefs and preferences of psychotherapy interventions. Furthermore, the scale will be validated for use outside the field of psychotherapy. The validated checklist will then be used to examine two potential mechanisms of how RA may affect outcomes of interventions: publication bias (by assessing grants) and risk of bias (RoB). Finally, recommendations will be developed, and a feasibility study will be conducted at a national mental health agency in The Netherlands. Main analyses comprise inter-rater reliability of checklist items, correlations to examine the relationship between checklist items and author survey (convergent validity) as well as checklist items and trial outcomes and multivariate meta-regression techniques to assess potential mechanisms of how allegiance affects trial outcomes (publication bias and RoB).

Ethics and dissemination This study has been reviewed and approved by the Scientific and Ethical Review Board (VCWE) at the Vrije Universiteit Amsterdam. Study result and advancements will also be published on the Open Science Framework. Furthermore, main findings will be disseminated through articles in international peer-reviewed open access journals. Results and recommendations will be communicated to the Cochrane Collaboration, the Campbell Collaboration and other funding agencies.

INTRODUCTION

One potential source of bias in randomised clinical trials (RCTs) of psychological interventions is researcher allegiance (RA). RA

Strengths and limitations of this study

► This is the first attempt to validate a tool for assessing researcher allegiance.
► Findings will contribute to the understanding of how researcher allegiance affects psychotherapy trial results.
► Comprehensive findings will lead to recommendations for clinical guidelines about how to address researcher allegiance in practice.
► A limitation that may impact response rate is the negative connotation that many researchers have about researcher allegiance.
► Using a self-report survey to assess authors' allegiances is a limitation as it is not able to capture latent or unconscious features that may distort results.

is defined as 'the belief in superiority of an intervention and of the superior validity of the theory of change that is associated with the treatment'[1] (p. 55). In psychotherapy research, RA can be acknowledged as a specific intellectual conflict of interest that is consistent with one's professional or personal commitment for one type of therapy. This allegiance may unintentionally reduce objectivity, lead to questionable research practices and may consequently distort the outcomes (or the interpretation of outcomes) of RCTs examining psychological interventions.[2–6] However, it may be that RA is beneficial in psychology as it simply reflects a higher level of skills in those who are well-trained in delivering an intervention.[3 7]

RA has been a common topic of interest in psychological literature throughout the past 50 years. However, research on this subject has failed to adhere to a shared definition of RA, as well as how it should be identified and measured (eg, refs 2 3 8–10). In the past, many researchers have assessed RA by means of the reprint method that allows one to rate RA at the study level or intervention level based on information presented in a publication's introduction and methods section. For

example, the reprint method involves assessing a publication for indicators of RA such as if the author developed the intervention, if the author provides an extensive amount of information about one intervention compared with the other(s), if the author refers to previous research showing superiority of one intervention and if the author advocates the intervention through their writing (see online supplementary appendix A for a complete overview of reprint method indicators). Assessing RA at the intervention level by rating each intervention for RA separately is beneficial as it allows for the detection of directional or balanced RA in each study; balanced meaning that authors have the same amount of allegiance towards each of the interventions.

Although it is commonly agreed in the literature that an important indicator of RA is if an author developed or first introduced an intervention,[5 8 11 12] other indicators of RA differ across studies. Furthermore, Leykin and DeRubeis[1] warrant caution regarding the reprint method as it remains unclear as to how positive trial results influence the presence of common RA indicators in a published paper. If introductions are written in light of the study results rather than an author's pre-existing allegiances, the reprint method leads to an identification of RA for whichever of the interventions had been found superior in the study.[8 13] Furthermore, these commonly used RA indicators of the reprint method have not been investigated for their validity and have been found to have a low correlation (r=0.10) with researcher's self-reported ratings of their own allegiances.[2]

In addition to operationalisation problems, it remains unclear as to how RA affects the outcomes of trials and if it results in better outcomes for a preferred intervention (namely the RA effect).[1 3 12] Researchers have been able to show an RA effect where an author's RA is correlated with their study results,[2 14–16] thus concluding that a statistical correction is necessary when analysing trial results. Furthermore, in a meta-analysis on trauma-focused therapies for post-traumatic stress disorder, Munder et al[4] found RA to be a significant predictor of the treatment effect and showed that an increase of 1 point on the RA scale was associated with an increase in the effect size (d=0.109) in favour of the preferred treatment. In contrast, other meta-analyses have concluded that RA is not an important source of bias as it was found not to influence the relative treatment effect.[8 9 17] In a case such as this, a statistical correction may actually introduce another source of bias.[18]

There is a need in the field for an operationalisation of RA and a reliable and valid method for assessing it. Furthermore, research must be conducted to understand how RA works and the impact that it has on research practices (ie, publication bias, risk of bias (RoB) and conflict of interest), study outcomes and patient results. This protocol reports on a comprehensive ongoing project (May 2017–May 2019), which aims to:

1. Develop and validate a checklist that measures RA in psychotherapy trials (in progress).
2. Externally validate the RA checklist for use in the biomedical field.
3. Examine how RA is related to publication bias in psychotherapy research.
4. Examine the relationship between RA and RoB in psychotherapy research.
5. Formulate and test recommendations on how to address RA in practice.

Figure 1 provides an overview of the project's goals, as well as the current status of the project.

METHODS AND ANALYSIS

To conduct this study, we are using an up-to-date existing database that was built through systematic searches in bibliographical databases (PubMed, Embase, PsycINFO and Cochrane Library). It includes trials comparing psychological interventions (for depression, anxiety disorders, and borderline personality disorder) with either an untreated control group (absolute efficacy comparisons) or an alternative intervention (relative comparisons). The database contains effect sizes, RoB item ratings and other study characteristics. Details about the database development and characteristics have been published elsewhere.[19–21]
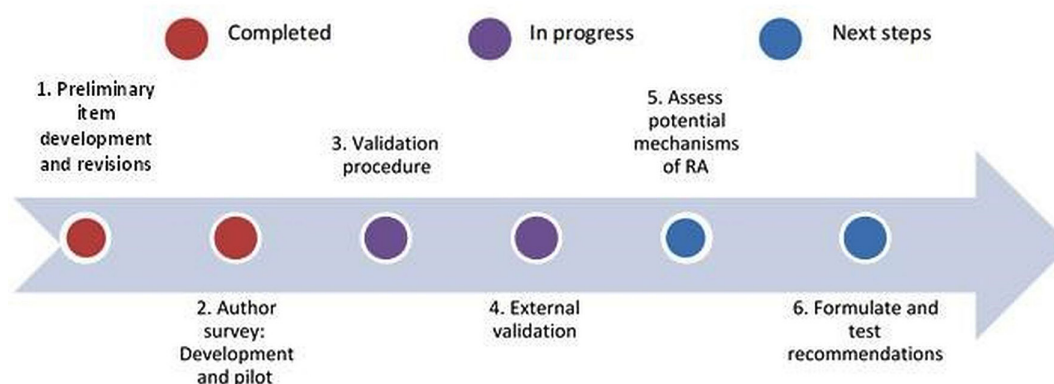

Figure 1 Goals and status of current project. RA, researcher allegiance.

Table 1 Researcher allegiance domains and definitions

RA domain	Definition
Effectiveness	The author believes that the intervention is effective in the treatment of the diagnosis of interest.
Superiority	The author believes that the intervention is effective beyond other interventions in the treatment of the diagnosis of interest.
Advocacy	The author advocates the intervention for the treatment of the diagnosis of interest.
Development/ contribution	The author has developed or contributed to the development or the enhancement of the intervention for the treatment of the diagnosis of interest.
Methodology	The author was involved in the trial in a way that could have influenced the implementation of the intervention.

RA, researcher allegiance.

Checklist development and validation pilot
Preliminary item development and revisions
Since we will examine both absolute efficacy (control comparison) and relative comparison (head-to-head) trials, two comprehensive checklists were developed based on previous operationalisation, measurements and indicators of RA found in the literature.[3 8–10 12 22] Selected items were categorised to five RA domains in line with previously defined direct and indirect indicators of RA[3] (table 1): effectiveness, superiority, advocacy, development/contribution and methodology. During this development phase, feedback was requested from experts in the field and used to inform decision making. Items were given answer categories of 'yes', 'no' or 'not clear'. Methodology items were given additional categories of 'not reported' or 'not applicable'.

The checklists require that each of the interventions in each trial be rated for RA. For example, if there are two interventions (intervention A and intervention B), then each intervention will receive an RA score by summing the items scores (yes=1, no=0, NR/NA=0) per intervention. Item weights (no=0, yes = ½, 1 or 2) were predetermined during the development phase based on the items' content and theoretical contribution to RA. For example, the item 'Is the title of the intervention mentioned in the title of the paper' only receives ½ point if the rating is 'yes'. In contrast, the item 'Did one of the authors develop the intervention' receives 2 points if rated as 'yes'. These item weights will be tested during the validation phase.

To calculate a total RA score at the study level, the difference between each of the interventions' RA score is taken (A minus B). This allows for the assessment of directional (positive difference is in favour of intervention A and negative difference is in favour of intervention B) or balanced RA (a difference of zero). If more than two interventions are examined in the trial, interventions are grouped together if possible (ie, if they were grouped together for the calculation of the effect size in the study) and if their individual RA scores do not differ. If it is not possible to group interventions, a total RA score is calculated for each comparison.

After the checklists were developed, the items were piloted and tested for their ratability and inter-rater reliability. To do this, a small subset of trials (n=11) comparing different interventions with supportive therapy for the treatment of depression were selected and rated by two independent raters (WRY, EK, or MS) using the preliminary RA checklists. Inter-rater reliability (percent agreement) was calculated per item, and all disagreements were discussed and resolved. Items were then revised or deleted based on the percentage agreement scores and rater discussions (see online supplementary appendix B for overview of decisions).

Author survey: development and pilot
Convergent validity refers to the extent to which a measurement corresponds to that of an already established measurement of the same construct. As there is not yet an established measure or 'gold standard' to assess RA, the convergent validity of the RA checklists will be examined by asking authors of psychotherapy trials to complete a survey about their career history and beliefs related to psychotherapy. This author survey will then be used to assess the convergent validity of the checklists by comparing author responses and RA ratings from associated trials.

In preparation for the validity assessment, an author survey was drafted with items that are in line with the identified RA domains (table 1) and piloted in a small sample of authors (all authors of all papers rated during the development of the RA checklist) in order to assess response rates and author reactions. All authors were contacted via email when a valid email address was available. Authors that did not respond after 2 weeks were sent a reminder email. Furthermore, all authors received and signed an electronic informed consent. The survey was created using Qualtrics and was approved by an internal review board (see online supplementary appendix C for IRB approval). Any useful feedback that was received from authors during this pilot phase was taken into account when revising the survey for the next round of data collection.

Validation procedure
In order to assess the convergent validity of the RA checklists, the updated author survey was sent to authors of 100 trials examining treatments for depression (50 head-to-head comparisons, 50 with control comparisons) that were randomly selected from our database. The random selection of trials was conducted by randomly generating a number (n) and by taking every nth publication from the alphabetised list of all trials in our database. If a selected study was not eligible (ie, due to publication date), the next study in the list was taken. Once the end of the list was reached, this process was continued, each time drawing

a new randomly generated number, until 100 trials were selected. Selected trials were published in the year 2001 or later to account for the implementation of Consolidated Standards for Reporting Trials (CONSORT) trial guidelines. CONSORT guidelines offer a standard way for authors to conduct, report and interpret their trials in a transparent and systematic method.[23] Therefore, it was important to select published trials that follow these guidelines as to ensure that equal comparisons were made. Furthermore, by selecting more recent studies, we were more likely to find current email addresses for the authors, and thus ensure a better response rate.

Author survey

The author survey was sent via email to first, second and last authors of the 100 randomly selected trials. This method was selected since first, second and last authors are most often more involved in designing, supervising and publishing a trial. Furthermore, it allowed for the assessment of the relationship between an author's involvement in the study (as denoted by authorship order) and the effect that individual RA has on trial outcomes. When study authors do not respond, despite two reminders, that trial will not be included in the final analysis. Furthermore, in this case, data collection will be continued using the same method for additional randomly selected trials from the database until 100 responses are received.

Authors who respond to the author survey are also asked to send their CV so we can better assess their connections (ie, development, implementation, training and advocacy of a treatment) to the respective psychotherapy of interest (table 2). Activities listed on the CV are only evaluated if they were dated before or up to 1 year after the date of the published trial being assessed for that author. This allows for the assessment of RA indicators at the time the study was conducted. Authors who are

Table 2 Data extracted from authors' CVs

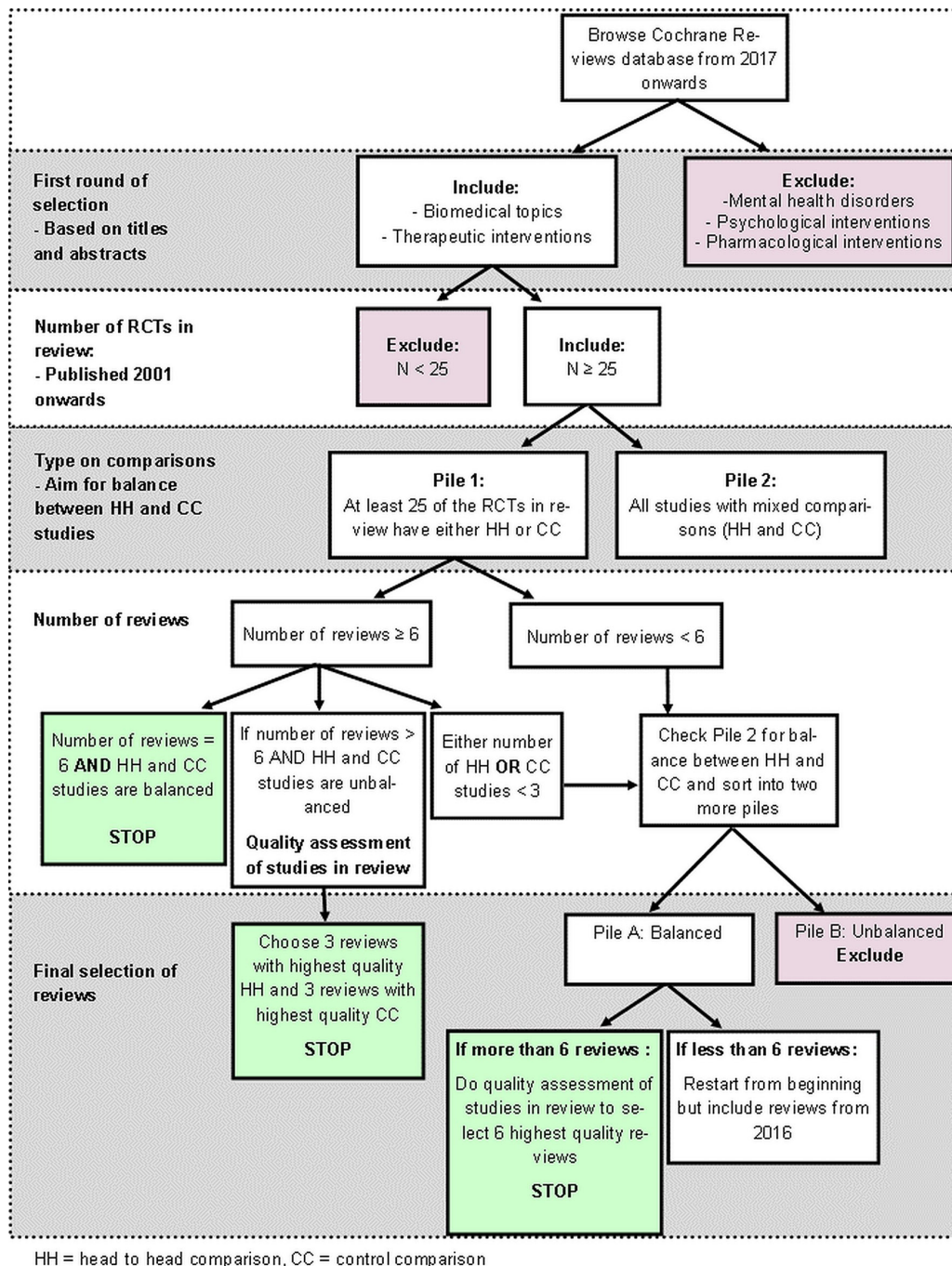| Activities assessed	RA domain
Teaching/training of an intervention (ie, workshops or university classes)	Proponent
Supervision of therapists for an intervention	Proponent
Patents or copyrights related to an intervention	Development/contribution
Form of training/supervision of an intervention (did they receive training/supervision from the developer, did they receive training/supervision at an institute, are they certified)	Proponent
Involvement on a board of directors or as a coordinator or committee chair of a professional society related an intervention	Proponent

RA, researcher allegiance.

not willing to send their CV are prompted to answer a few additional survey items in line with the data that is extracted from CVs.

After data collection is complete, responses from the survey and the data extracted from CVs will be scored (each item receiving one point per type of therapy) and combined, therefore providing a score of *true* RA for each author for the type of therapy that is examined in their corresponding trial.

Trial ratings

After all author responses are received, the associated trial publication will be rated with the RA checklist (head-to-head or control comparison) by two independent raters (WRY and EK). Any disagreements among the raters will be discussed. If an agreement cannot be made, a senior researcher will be consulted (PC).

If not already in the current database, the effect sizes (Hedges' g; standardised mean difference adjusted for small sample size[24]) for each trial will also be calculated to represent the difference between the intervention and control group or alternative intervention.

Final selection of items

A final list of RA checklist items will be selected based on:
1. The relationship of the RA checklist (at item and domain level) and scores from the author survey (convergent validity).
2. Inter-rater reliability (Cohen's kappa ≥0.6).
3. The relationship of the RA checklist (at item and domain level) and the effect sizes of the associated trials.

It is the aim that this process will result in a final checklist that is usable (with the fewest number of items), reliable and valid.

External validation

Although RA has been examined mostly in the field of psychological interventions of mental disorders, it is likely that the phenomenon is not limited to this field. To assess this, we will validate the RA checklist outside the field of mental disorders. We have selected trials contained in six recent Cochrane Reviews in a systematic procedure (figure 2). Selected reviews focus on non-pharmacological, surgical or behavioural (ie, exercise for cancer and virtual reality for stroke rehabilitation) interventions in the biomedical field. Exclusion criteria included interventions for mental health disorders, psychotherapy interventions and pharmacological interventions. Reviews of RCTs have been selected with the goal of assessing an equal number of head-to-head and control comparison trials. Studies of the selected reviews were rated using the newly developed RA checklists. Results will be compared with those of the validation results of the psychotherapy trials. Furthermore, conclusions will be drawn regarding the checklists' use as a valid instrument for assessing RA in the biomedical field.

Figure 2 Selection process of biomedical reviews for external validation. RCTs, randomised controlled trials.

Assess potential mechanisms of RA

To gain an understanding of how RA works in psychotherapy research, two potential mechanisms of RA will be assessed.

Publication bias

One mechanism through which RA may work is through publication bias in that an author does not publish the results of negative trials. Negative trials are defined as RCTs in which an intervention is compared with a non-intervention control group and in which no significant difference between the intervention and the control group is found, or as a head-to-head comparison where

no significant difference between the two interventions is found. It is hypothesised that negative trials are more often not published when authors are allegiant compared with when authors are not.

To assess this, we will search for and collect grants on all trials on interventions related to treating depression, social anxiety disorder, generalised anxiety disorder, panic and borderline personality disorder. A comprehensive search will be conducted in the National Institutes of Health (NIH), the National Institute of Health Research (NIHR), the Wellcome Trust (UK) and the Deutsche Forschungsgemeinschaft (Germany) databases

to find applicable grant-funded studies that were funded up until 2010. Next, grant-funded studies will be assessed for whether they resulted in published articles. For those grants that did not result in publication, two independent raters will rate each grant application (which can be requested from the appropriate institution) using our validated RA checklists. This will allow for the assessment of whether negative studies with RA are published less often than negative studies without RA.

If possible, it will also be an aim of this project to collect primary data from the grants that did not result in publications (but did collect data) from the researchers that conducted the trials. If this is feasible, effect sizes for published and unpublished trials can be calculated, thus providing the means to differentiate between those studies with and without RA.

Risk of bias

RoB domains refer to the internal validity of clinical trials. These biases can be categorised as selection bias, performance bias, detection bias, attrition bias, reporting bias and other biases that do not fit into these categories.[25] As a high RoB may provide leeway for researchers to influence results, we hypothesise that the effects of RoB items on outcomes are stronger in trials with RA than in trials without RA. All trials in our database will be assessed and scored for RoB using the Cochrane Risk of Bias Assessment.[25]

The remaining trials in our database that were not used in the development and validation processes will be assessed for RA with the validated RA checklist. Next, the interaction of RA and RoB in the prediction of effect sizes will be assessed to evaluate if RA works through RoB. This will allow for the assessment of whether RA results in larger intervention effects in clinical trials.

Formulate and test recommendations on how to handle RA in practice

Guidelines will be formulated as to how RA should be accounted for in randomised trials with the aim of reducing its impact on research outcomes and associated clinical intervention guidelines.

Recommendations

Based on the results of this project, we will write recommendations on how RA should be reported by authors of trials and meta-analyses. A report of the main findings of our project will be created and sent with our recommendations to the Cochrane Collaboration, the Campbell Collaboration and other local funding agencies. These organisations will be advised to require applicants and authors to indicate possible sources of RA when writing grants or meta-analyses.

Feasibility study

A feasibility study will be conducted by the Trimbos Institute in The Netherlands (DvD). The Trimbos Institute has a leading role in the development of clinical guidelines and standards of care in the field of mental healthcare in the Netherlands. In the past years, they have developed clinical evidence-based guidelines for all major psychological disorders. In this pilot, the recommendations for assessing and reporting RA will be tested for feasibility in the process of the development of an intervention guideline. Because this feasibility study will be conducted when the consensus document is available, it is not possible yet to decide which intervention guideline will be used. However, the Trimbos Institute is involved in several guideline development projects and by that time one of those will be chosen for the current project.

The technical review team of the guideline panel (who are responsible for the assessment of the evidence that is used for the guideline) will use the consensus document and the RA checklist in the assessment of the evidence they are reviewing. Qualitative interviews with the panel members will also be held to ask for their experiences and to assess the feasibility and acceptability of the consensus document. Based on this feasibility study, the consensus document will be revised as needed.

Patient and public involvement

Consent from authors is obtained before they complete the author survey. However, this is not a clinical study where patients or the public are involved in any aspect.

DATA MANAGEMENT AND ANALYSIS
Data management

All information obtained from the survey will be stored in a password-protected electronic format that only authorised personnel can access. Names and any other personal information will not be included in any publications or presentations based on these data, and author responses to the survey will remain confidential. Results of this study will only be published and presented at study level.

Data analysis

All data analysis will be conducted in R (2009–2015) and Comprehensive Meta-Analysis (CMA; V.3.3.070). In all analyses, an alpha level of <0.05 will be considered statistically significant. Furthermore, a Cohen's kappa ≥0.60 will be considered as adequate reliability.[26]

To calculate the effect sizes, the outcome instruments that measure the symptoms of the disorder that is being targeted will be used. For example, when we are assessing RA in depression studies, we will calculate the effect size for the primary outcome measure for depression that is used. If a secondary outcome is considered relevant or if more than one relevant outcome measure is used, the effect sizes within the study will first be pooled. Effect sizes will then be pooled across studies. When a measure is used in 10 or more studies, sensitivity analyses will also be conducted where effect sizes are included based on one outcome measure. Post-test scores of the intervention and comparison groups will be used to calculate effect sizes, or if available, the change scores between pretest and post-test. If these data are not available, effect sizes

will be calculated based on dichotomous outcomes using the methods implemented in CMA.

Checklist development and validation

Interitems correlations will be calculated to assess if checklist items overlap. Next, convergent validity will be assessed by calculating correlations between the author survey/CV scores (*true* RA) and the rated RA checklist items (both at item and scale level) for the associated publication. We will separately assess the responses of first, second and last authors. If more than one author from a study responds, the average of the responses will be calculated in order to examine the relationship between the RA checklist and the author groups' average RA. Furthermore, we will conduct sensitivity analyses where we compare the total RA checklist score with and without the predefined item weights. When assessing the convergent validity of the checklist items, we will select the best RA indicators for the final checklist by considering any significant correlation, even of a small magnitude, as an indicator of a relationship between the examined variables. Construct reliability will also be assessed by calculating the correlation between each of the checklist items and the effect sizes found in the studies. Finally, inter-rater reliability of each item will be calculated as the per cent agreement of raters and Cohen's kappa.

External validation

Correlations between RA (from the checklist) and effect sizes of the selected biomedical trials will be calculated and assessed for significance and strength. Findings will be compared with those of psychological intervention trials.

Assessing mechanisms of RA
Publication bias

Crosstabs and χ^2 tests will be conducted to test if RA is more common among grants that did and did not result in publications. This will be done separately for each of the RA items, domains and total RA score. Next, multivariate meta-regression analyses will be conducted (with the effect size as dependent variable) in which the two variables (RA score; published yes/no) and their interaction will be entered as predictors.

Risk of bias

Multivariate meta-regression analyses will be conducted with the effect sizes as the dependent variable, and RA score, RoB and the interaction between RA and RoB as predictors.

DISCUSSION

RA may not be representative of a purposeful attempt to skew results as it is simply human nature to hold beliefs in ways that can compromise objectivity.[27] However, overlooking RA could be considered as a methodological issue in psychotherapy research.[22] Even if researchers begin declaring their allegiances to particular psychotherapies as they do their financial (and other) conflict of interests, there is still no way to know whether their RA influenced the research methods, data analysis or interpretation of study results.[27] This study attempts to advance RA research through the development of valid methods to measure and account for RA in psychotherapy trials, thus aiding in the understanding of the impact that RA has in psychotherapy research. Finally, this study will contribute to the development of clinical trial guidelines and enhance the field of psychotherapy research and practice. Once a valid method exists to measure RA, future research should be devoted to further studying different mechanisms of bias (ie, quality of delivered therapy and control conditions in clinical trials) and the associated relationship with RA.

ETHICS AND DISSEMINATION

Study results and advancements will also be published on the Open Science Framework. Furthermore, main findings will be disseminated through articles in international peer-reviewed open access journals. Results and recommendations will be communicated to the Cochrane Collaboration, the Campbell Collaboration and other local funding agencies (ie, ZonMw).

REFERENCES

1. Leykin Y, DeRubeis RJ. Allegiance in psychotherapy outcome research: separating association from bias. *Clin Psychol Sci Prac* 2009;16:54–65.
2. Luborsky L. The researcher's own therapy allegiances: a ''wild card'' in comparisons of treatment efficacy. *Clin Psychol Sci Prac* 1999;6:95–106.
3. Munder T, Gerger H, Trelle S, *et al*. Testing the allegiance bias hypothesis: a meta-analysis. *Psychother Res* 2011;21:670–84.
4. Munder T, Flückiger C, Gerger H, *et al*. Is the allegiance effect an epiphenomenon of true efficacy differences between treatments? a meta-analysis. *J Couns Psychol* 2012;59:631–7.

5. Munder T, Brütsch O, Leonhart R, *et al*. Researcher allegiance in psychotherapy outcome research: an overview of reviews. *Clin Psychol Rev* 2013;33:501–11.
6. Cuijpers P, Cristea IA. How to prove that your therapy is effective, even when it is not: a guideline. *Epidemiol Psychiatr Sci* 2016;25:428–35.
7. Ioannidis JP. Most psychotherapies do not really work, but those that might work should be assessed in biased studies. *Epidemiol Psychiatr Sci* 2016;25:436–8.
8. Gaffan EA, Tsaousis I, Kemp-Wheeler SM. Researcher allegiance and meta-analysis: the case of cognitive therapy for depression. *J Consult Clin Psychol* 1995;63:966–80.
9. Tolin DF. Is cognitive-behavioral therapy more effective than other therapies? A meta-analytic review. *Clin Psychol Rev* 2010;30:710–20.
10. Cuijpers P, Driessen E, Hollon SD, *et al*. The efficacy of non-directive supportive therapy for adult depression: a meta-analysis. *Clin Psychol Rev* 2012;32:280–91.
11. Miller S, Wampold B, Varhely K. Direct comparisons of treatment modalities for youth disorders: a meta-analysis. *Psychother Res* 2008;18:5–14.
12. Wampold BE, Budge SL, Laska KM, *et al*. Evidence-based treatments for depression and anxiety versus treatment-as-usual: a meta-analysis of direct comparisons. *Clin Psychol Rev* 2011;31:1304–12.
13. Berman JS, Reich CM. Investigator allegiance and the evaluation of psychotherapy outcome research. *Eur J Psychother Couns* 2010;12:11–21.
14. Berman JS, Miller RC, Massman PJ. Cognitive therapy versus systematic desensitization: is one treatment superior? *Psychol Bull* 1985;97:451–61.
15. Luborsky L, Singer B, Luborsky L. Comparative studies of psychotherapies. Is it true that "everywon has one and all must have prizes"? *Arch Gen Psychiatry* 1975;32:995–1008.
16. Botella L, Beriain D. Allegiance effects in psychotherapy research: a constructivist approach. *Eur J Psychother Couns* 2010;12:55–64.
17. Spielmans GI, Berman MI, Usitalo AN. Psychotherapy versus second-generation antidepressants in the treatment of depression: a meta-analysis. *J Nerv Ment Dis* 2011;199:142–9.
18. Weisz JR, Weiss B, Han SS, *et al*. Effects of psychotherapy with children and adolescents revisited: a meta-analysis of treatment outcome studies. *Psychol Bull* 1995;117:450–68.
19. Cuijpers P, van Straten A, Warmerdam L, *et al*. Psychological treatment of depression: a meta-analytic database of randomized studies. *BMC Psychiatry* 2008;8:36.
20. Cuijpers P, Cristea IA, Karyotaki E, *et al*. How effective are cognitive behavior therapies for major depression and anxiety disorders? A meta-analytic update of the evidence. *World Psychiatry* 2016;15:245–58.
21. Cristea IA, Gentili C, Cotet CD, *et al*. Efficacy of psychotherapies for borderline personality disorder: a systematic review and meta-analysis. *JAMA Psychiatry* 2017;74:319–28.
22. Dragioti E, Dimoliatis I, Evangelou E. Disclosure of researcher allegiance in meta-analyses and randomised controlled trials of psychotherapy: a systematic appraisal. *BMJ Open* 2015;5:e007206.
23. Moher D, Schulz KF, Altman DG, *et al*. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *Ann Intern Med* 2001;134:657–62.
24. Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics* 1981;6:107–28.
25. Higgins JPT, Altman DG, Sterne JAC, *et al*. Chapter 8: Assessing risk of bias in included studies. In: Higgins JPT, Green S, eds. *Cochrane handbook for systematic reviews of interventions Version 5.1.0*: The Cochrane Collaboration, 2011. Available from. www.handbook.cochrane.org
26. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012;22:276–82.
27. Lomangino KM. Countering cognitive bias: tips for recognizing the impact of potential bias on research. *J Acad Nutr Diet* 2016;116:204–7.