



Combining endpoint and change data did not affect the summary standardised mean difference in pairwise and network meta-analyses: An empirical study in depression

Edoardo G. Ostinelli^{1,2,3}  | Orestis Efthimiou^{1,4,5}  | Yan Luo⁶  |
Clara Miguel⁷  | Eirini Karyotaki⁷  | Pim Cuijpers^{7,8}  |
Toshi A. Furukawa⁶  | Georgia Salanti⁵  | Andrea Cipriani^{1,2,3} 

¹Department of Psychiatry, University of Oxford, Oxford, UK

²Oxford Precision Psychiatry Lab, NIHR Oxford Health Biomedical Research Centre, Oxford, UK

³Oxford Health NHS Foundation Trust, Warneford Hospital, Oxford, UK

⁴Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland

⁵Institute of Social and Preventive Medicine (ISPM), University of Bern, Bern, Switzerland

⁶Department of Health Promotion and Human Behaviour, School of Public Health in the Graduate School of Medicine, Kyoto University, Kyoto, Japan

⁷Department of Clinical Neuro- and Developmental Psychology, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

⁸Babeş-Bolyai University, International Institute for Psychotherapy, Cluj-Napoca, Romania

Correspondence

Edoardo G. Ostinelli, Department of Psychiatry, University of Oxford, Warneford Hospital, Warneford Ln, Headington, Oxford OX3 7JX, UK.
Email: edoardo.ostinelli@psych.ox.ac.uk

Funding information

National Institute for Health and Care Research Applied Research Collaboration Oxford and Thames Valley; NIHR Oxford Health Biomedical Research Centre, Grant/Award Number: BRC-1215-20005; Brasenose College Senior Hulme Scholarship; National Institute for Health and Care Research Oxford Health Clinical Research Facility; Swiss National Science Foundation, Grant/Award Numbers: 179158, 180083; National Institute for Health and Care Research, Grant/Award Number: RP-2017-08-ST2-006

Abstract

When studies use different scales to measure continuous outcomes, standardised mean differences (SMD) are required to meta-analyse the data. However, outcomes are often reported as endpoint or change from baseline scores. Combining corresponding SMDs can be problematic and available guidance advises against this practice. We aimed to examine the impact of combining the two types of SMD in meta-analyses of depression severity. We used individual participant data on pharmacological interventions (89 studies, 27,409 participants) and internet-delivered cognitive behavioural therapy (iCBT; 61 studies, 13,687 participants) for depression to compare endpoint and change from baseline SMDs at the study level. Next, we performed pairwise (PWMA) and network meta-analyses (NMA) using endpoint SMDs, change from baseline SMDs, or a mixture of the two. Study-specific SMDs calculated from endpoint and change from baseline data were largely similar, although for iCBT interventions 25% of the studies at 3 months were associated with important differences between study-specific SMDs (median 0.01, IQR -0.10, 0.13) especially in smaller trials with baseline imbalances. However, when pooled, the differences between endpoint and change SMDs were negligible. Pooling only the more favourable of the two SMDs did not materially affect meta-analyses, resulting in differences of pooled SMDs up to 0.05 and 0.13 in the pharmacological and iCBT

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Authors. *Research Synthesis Methods* published by John Wiley & Sons Ltd.

datasets, respectively. Our findings have implications for meta-analyses in depression, where we showed that the choice between endpoint and change scores for estimating SMDs had immaterial impact on summary meta-analytic estimates. Future studies should replicate and extend our analyses to fields other than depression.

KEYWORDS

change, continuous outcome, depression, follow-up, meta-analysis, standardised mean difference

Highlights

What is already known

- Available aggregate continuous outcome data from randomised controlled trials (RCTs) often come as a mixture of endpoint and change from baseline scores across different scales.
- Combining these data in meta-analytical models using standardised mean differences (SMDs) might be problematic.
- Most available meta-analyses combine them, although available guidance advises against mixing endpoint and change from baseline when meta-analysing SMDs.
- Whether meta-analysts should lump these data or not is debated.

What is new

- We used individual participant data from RCTs in the field of depression to ensure that the same participants provide both endpoint and change from baseline score data across studies.
- At an individual study level, SMDs based on endpoint and change from baseline data can differ, especially in the presence of baseline imbalances.
- We showed, however, that the use of endpoint and change from baseline data produced negligible effects on the pooled standardised mean differences from pairwise and network meta-analyses.

Potential impact for *Research Synthesis Methods* readers

- Endpoint and change from baseline scores can be combined in meta-analyses of interventions for depression using standardised mean differences.
- Future studies should replicate and extend our analyses to fields other than depression.

1 | INTRODUCTION

In depression, symptom severity can be measured and reported using different rating scales.^{1,2} This multiplicity of instruments challenges meta-analysts when pooling estimates and interpreting outcomes, especially in fields relying on subjective or semi-objective outcomes (e.g., symptom or adverse outcome intensity, patient-reported outcomes) where scale-to-scale conversions are scarce and often non-linear.³

When a scale-to-scale conversion is not possible, the pooled treatment effect from multiple scales is commonly

estimated using the standardised mean difference (SMD).^{4,5} The calculation of SMD from each study requires the mean difference of outcomes across treatment arms and the pooled standard deviation (SD). In practice, however, this is complicated because studies included in meta-analyses often report a mixture of either mean endpoint scores (i.e., the mean outcome at the study's endpoint for each treatment arm) or change from baseline (i.e., mean change in outcome, endpoint minus baseline, across all patients in each treatment arm).⁶ Common approaches to perform meta-analyses in such cases are: (i) choose one of the two metrics (i.e., SMD on

endpoint or SMD on change) and disregard studies not providing data that can be used to calculate it, (ii) analyse both measures separately and appraise their findings jointly, or (iii) combine both types of SMD into a single aggregate data MA, that is, ignoring the difference in their definition. The first option will result in data loss. The second option may decrease statistical power because it splits the dataset into two parts and requires estimation and appraisal of two separate models. The third option has the advantage of not throwing away information, leading to a single set of estimates.

However, the third approach is currently discouraged by the Cochrane Handbook,⁴ where it writes: “... *post-intervention value and change scores should not in principle be combined using standard meta-analysis approaches when the effect measure is an SMD.*” This hesitation in combining SMDs calculated using endpoint and change scores stems from the fact that using data from the same study, the two SMDs may differ. These two different ways of calculating an SMD may give very different answers, for example, when the SD for change from baseline scores is very different to the SD of the outcome at endpoint. In that case, combining the two SMD types in meta-analysis might be seen as a case of mixing apples with oranges.

In more detail, if in a study there are no imbalances of baseline scores across treatment groups, the mean difference for endpoint will equal that of change scores. In practice, for randomised trials, even if there are some imbalances, they are not expected to be very large, particularly for larger sample sizes. Next, the SD for change scores is given by the following formula⁴:

$$SD_{\text{change}}^2 = SD_{\text{baseline}}^2 + SD_{\text{endpoint}}^2 - 2 \times \rho \times SD_{\text{baseline}} \times SD_{\text{endpoint}}$$

where ρ denotes the pre–post correlation, that is, the correlation between baseline and endpoint scores across patients. This formula suggests that the two SDs used for calculating the two types of SMDs (i.e., SD_{endpoint} and SD_{change}) are in general unequal. They will be approximately equal when $SD_{\text{baseline}}^2 - 2 \times \rho \times SD_{\text{baseline}} \times SD_{\text{endpoint}} \ll SD_{\text{endpoint}}^2$. This may happen, for instance, when SD_{baseline} is very small compared with SD_{endpoint} , or when $SD_{\text{endpoint}} \approx SD_{\text{baseline}}$ but also $\rho \approx 0.5$. Such conditions, however, may not hold in practice. In that case, SDs of change data may be systematically different than those of endpoint data. These differences could significantly impact the pooled SD required to estimate SMD.^{6–8}

There is little available literature on the impact of choosing between endpoint and change from baseline

scores on pooled treatment effects. Fu and Holmer reported some evidence of discrepancy between pooled mean differences from endpoint and change from baseline scores across several meta-analyses in diabetes, mental health and pain,⁹ but they based their conclusions mainly on comparing statistical significance, which is problematic for comparing agreement between different analyses.¹⁰ By contrast, da Costa and colleagues reported some evidence that it may be valid to pool treatment effects from endpoint and change from baseline scores of fixed range scales in meta-analyses of pain.¹¹ The authors could not identify, on average, relevant differences between the endpoint and change SMDs, although a selective choice (i.e., selecting the type of SMD showing the largest or smallest treatment effect in each study) introduced considerable differences in pooled estimates.¹¹ These studies, however, relied on aggregate data only, so that they were only able to compare endpoint SMDs and change SMDs in the same meta-analysis but based on slightly different datasets, as not all studies provided both endpoint and changes SMDs in their original reports. Thus, there is a need to explore whether combining endpoint with change of baseline SMD is justifiable in practice.

The aim of this article is to provide empirical evidence on the impact of choosing between endpoint-based and change from baseline SMDs in meta-analyses and network meta-analyses. We used IPD data from 150 studies that reported results using various symptoms scales of depression to evaluate differences between study-level SMDs estimated using endpoint and change from baseline values, as well as the impact of mixing the two types of SMD on the estimated treatment effect from pairwise and network meta-analyses. Further, we aimed to assess the impact of selective reporting at the meta-analysis.

2 | METHODS

2.1 | Description of datasets

We extracted patient-level baseline and endpoint scores of depression severity scales from two datasets. The first focused on the acute treatment with pharmacological interventions for adults with depression.¹ This dataset comprised 89 double-blind randomised controlled trials including 27,409 participants. Endpoint data was extracted at 4, 6 and 8 weeks. The second dataset was on short- and long-term efficacy of internet-based cognitive behavioural therapy (iCBT) for adults with depression. Out of 62 studies included in the original publication, 61 randomised controlled trials including 13,687 participants provided data at baseline and at least one follow-up

time point.² Endpoint data was extracted at 3, 6 and 12 months. For both datasets, at each time point of interest only participants providing data at baseline and at the time point of interest contributed to the analyses. Additional details on the datasets are available in the Supporting Information 0.1–2.

2.2 | Data harmonisation

2.2.1 | Pharmacological interventions

Outcome data were reported as the total score of either the Hamilton Depression Rating Scale (HAMD, range = 0–52; number of studies, $k = 81$ RCTs, $n = 23,856$ participants at baseline) or the Montgomery–Åsberg Depression Rating Scale (MADRS, range = 0–60; $k = 10$ RCTs, $n = 3553$ participants at baseline). For participants with both HAMD and MADRS scores at the same time point, we extracted data from the former as it was the most frequently reported scale.

2.2.2 | Internet-based cognitive behavioural therapy

Outcome data were reported as the total score of Beck's Depression Inventory (BDI-I or II, range = 0–63; $k = 27$ RCTs, $n = 5174$ participants at baseline), Patient Health Questionnaire 9 (PHQ9, range = 0–27; $k = 17$ RCTs, $n = 4675$ participants at baseline), Center for Epidemiological Studies Depression (CESD, range = 0–60; $k = 14$ RCTs, $n = 3086$ participants at baseline), MADRS ($k = 3$ RCTs, $n = 672$ participants at baseline) and Edinburgh Postnatal Depression Scale (EPDS, range = 0–30; $k = 2$ RCTs, $n = 80$ participants at baseline). For participants with scores on multiple scales simultaneously, we prioritised the scale in the following order: BDI, PHQ9, CESD, MADRS and EPDS.

Next, we calculated change scores using individual baseline and endpoint data. Finally, we aggregated the datasets at the arm and study level. This approach ensured that the aggregated data (i.e., mean, standard deviation and number of completers) for baseline, endpoint and change from baseline scores were consistently reported for the same participants. The format of these datasets emulates the data required to perform an aggregate meta-analysis, bypassing any selective reporting issues.

2.3 | SELECTION OF ANALYTICAL MODELS

We visually compared the distribution of study-specific standard deviations of baseline, endpoint and change

from baseline scores across all available time points. We reported median, interquartile range (IQR) and range of scale- and time-point-specific standard deviations. We estimated study-specific standardised mean differences (Hedges' g) and their standard error using endpoint ($SMD_{\text{endpoint}}^{\text{study}}$, $seSMD_{\text{endpoint}}^{\text{study}}$, respectively) and change from baseline ($SMD_{\text{change}}^{\text{study}}$, $seSMD_{\text{change}}^{\text{study}}$) scores from the pairwise meta-analysis datasets. We reported median, IQR and range of the absolute difference between $SMD_{\text{endpoint}}^{\text{study}}$ and $SMD_{\text{change}}^{\text{study}}$, and visually compared them in scatterplots.

2.3.1 | Pairwise and network meta-analyses

Aiming to assess the impact of the choice of SMD type at the meta-analysis level, we performed a series of random-effects pairwise and network meta-analyses using the following data:

- Only endpoint data—we performed a PWMA and an NMA at each time point to calculate $SMD_{\text{endpoint}}^{\text{pooled}}$.
- Only change from baseline data—we performed a PWMA and an NMA at each time point to calculate $SMD_{\text{change}}^{\text{pooled}}$.
- Assuming different proportions of studies contributing with $SMD_{\text{endpoint}}^{\text{study}}$ rather than $SMD_{\text{change}}^{\text{study}}$ (from 10% to 90% with 10% intervals for pairwise meta-analyses and network meta-analyses). For each proportion, the studies contributing with endpoint data were selected via random sampling without replacement. This process was repeated 100 times per choice of proportion, resulting in 900 pairwise meta-analyses and 900 network meta-analyses per time point after baseline. Overall, we performed 2700 pairwise meta-analyses and 2700 network meta-analyses for the dataset on pharmacological interventions and 2700 pairwise meta-analyses and 2700 network meta-analyses for the internet-based cognitive behavioural therapy dataset. We present the distribution of $SMD_{\text{mixed}}^{\text{pooled}}$, their standard errors ($seSMD_{\text{mixed}}^{\text{pooled}}$) and the between-studies heterogeneity standard deviations (τ) values.
- We investigated whether selective reporting could result in clinically important differences.¹¹ We explored two extreme scenarios. We selected from each study the type of SMD ($SMD_{\text{endpoint}}^{\text{study}}$ or $SMD_{\text{change}}^{\text{study}}$) that favoured the intervention (antidepressants or active iCBT in PWMA; paroxetine or guided iCBT in NMAs) or control conditions (placebo or control in PWMA; placebo or control in NMAs) in what we call optimistic ($SMD_{\text{optimistic}}^{\text{pooled}}$) and pessimistic ($SMD_{\text{pessimistic}}^{\text{pooled}}$) scenarios, respectively. For studies comparing interventions other than those above specified, we randomly sampled

$SMD_{\text{endpoint}}^{\text{study}}$ or $SMD_{\text{change}}^{\text{study}}$ prior to their inclusion in the NMAs.

- We investigated how the number of studies (k) included in the analysis affected the difference between $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$. To do this, a fixed number of studies ($k = 3, 10, 20, 30$ and 40) were randomly sampled and meta-analysed using both endpoint and change from baseline data; the process was repeated 100 times, resulting in a total of 3000 PWMAs (1000 PWMA per time point) on pharmacological interventions and 1800 PWMAs (1000 PWMAs at 3 months, 400 at 6 and 12 months; at 6 and 12 months, the total number of available studies limited this analysis to random sampling at $k = 3$ and 10). We restricted this analysis to PWMAs as the random sampling would introduce connectivity issues in the NMAs. We then plotted the difference between $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$ against k , to investigate the effect of the number of studies in the pooled result.

Pharmacological interventions

For PWMAs, we lumped individual drugs as an active intervention and compared it with pill placebo, thus excluding head-to-head trials. For NMAs, we considered each active intervention and pill placebo as separate nodes in the network. We present the results of paroxetine versus pill placebo. We chose paroxetine as it was the most studied active drug ($k = 45$ RCTs and $n = 5655$ participants).

Internet-based cognitive behavioural therapy

For PWMAs, we compared any active versus any control interventions. For NMAs, we considered guided and unguided iCBT, and control as separate nodes in the network. We reported the results of guided iCBT versus TAU, as guided iCBT was the active intervention most often found in the network ($k = 42$ RCTs and $n = 4431$ participants), and TAU was the only control intervention available at both 6 and 12 months.

Absolute differences between $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$ were considered important if ≥ 0.2 (threshold often used to indicate small effect sizes).^{11,12} We performed the analyses in R (version 4.2.2) using the *meta* and *netmeta* packages.^{13,14}

3 | RESULTS

3.1 | Description of available data

We analysed data on 41,096 participants from 150 RCTs on either pharmacological interventions or internet-based cognitive behavioural therapy for depression.

3.1.1 | Pharmacological interventions

The median SDs of endpoint were comparable to the SDs of change for HAMD (6.47 vs. 6.42 at 4 weeks, 6.66 vs. 6.69 at 6 weeks, 6.62 vs. 6.67 at 8 weeks) and MADRS (8.78 vs. 8.61 at 4 weeks, 8.86 vs. 8.80 at 6 weeks, 8.61 vs. 8.79 at 8 weeks). Additional details (IQR and range) are available in Supporting information Section 1.1.0. The fact that SDs of change was very close to the SD of endpoint was because $SD_{\text{baseline}}^2 - 2 \times \rho \times SD_{\text{baseline}} \times SD_{\text{endpoint}} \ll SD_{\text{endpoint}}^2$ (see formula in Section 1). $SMD_{\text{endpoint}}^{\text{study}}$ and $SMD_{\text{change}}^{\text{study}}$ and their corresponding standard errors values were comparable for most studies across the considered time points (Figure 1; additional details in Supporting Information 4.1.1–2). Important differences (difference in SMDs larger than 0.2) were observed in one study (2%) at 4 weeks (distribution of $SMD_{\text{endpoint}}^{\text{study}} - SMD_{\text{change}}^{\text{study}}$: median 0.02; IQR $-0.01, 0.06$; range $-0.24, 0.19$), one study (2%) at 6 weeks (distribution of $SMD_{\text{endpoint}}^{\text{study}} - SMD_{\text{change}}^{\text{study}}$: median 0.01; IQR $-0.03, 0.07$; range $-0.21, 0.19$), and three studies (8%) at 8 weeks (distribution of $SMD_{\text{endpoint}}^{\text{study}} - SMD_{\text{change}}^{\text{study}}$: median 0.01; IQR $-0.03, 0.05$; range $-0.19, 0.22$). Additional details are available in Supporting Information 4.1.3. Of studies with data at baseline and follow-up, 45 (90%), 40 (90.9%) and 34 (94.4%) studies had a $SMD_{\text{baseline}}^{\text{study}}$ lower than 0.2 at 4, 6 and 8 weeks, respectively. Additional information on baseline imbalance and its impact on $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$ are provided in Supporting information 8.1.0.

3.1.2 | Internet-based cognitive behavioural therapy

The median SDs of endpoint versus change from baseline scores were overall comparable over time (BDI: 9.92 vs. 9.08 at 3 months, 10.00 vs. 9.64 at 6 months, 10.57 vs. 9.19 at 12 months; PHQ-9: 4.81 vs. 5.33 at 3 months, 5.48 vs. 5.39 at 6 months, 5.93 vs. 6.04 at 12 months). Additional details and information on other scales are available in Supporting Information 8.1.0. SDs of change and endpoint were very close also in this dataset (see formula in Section 1). We did not observe substantial differences between $SMD_{\text{endpoint}}^{\text{study}}$ and $SMD_{\text{change}}^{\text{study}}$ and their corresponding standard errors for most studies (Figure 2; additional details in Supporting Information 4.2.1–2). Important differences (difference in SMDs larger than 0.2) were observed in 14 studies (25%) at 3 months (distribution of $SMD_{\text{endpoint}}^{\text{study}} - SMD_{\text{change}}^{\text{study}}$: median 0.01; IQR $-0.10, 0.13$; range $-0.58, 0.60$), one study (8%) at 6 months (distribution of $SMD_{\text{endpoint}}^{\text{study}} - SMD_{\text{change}}^{\text{study}}$: median -0.01 ; IQR $-0.07, 0.03$; range $-0.25, 0.15$) and

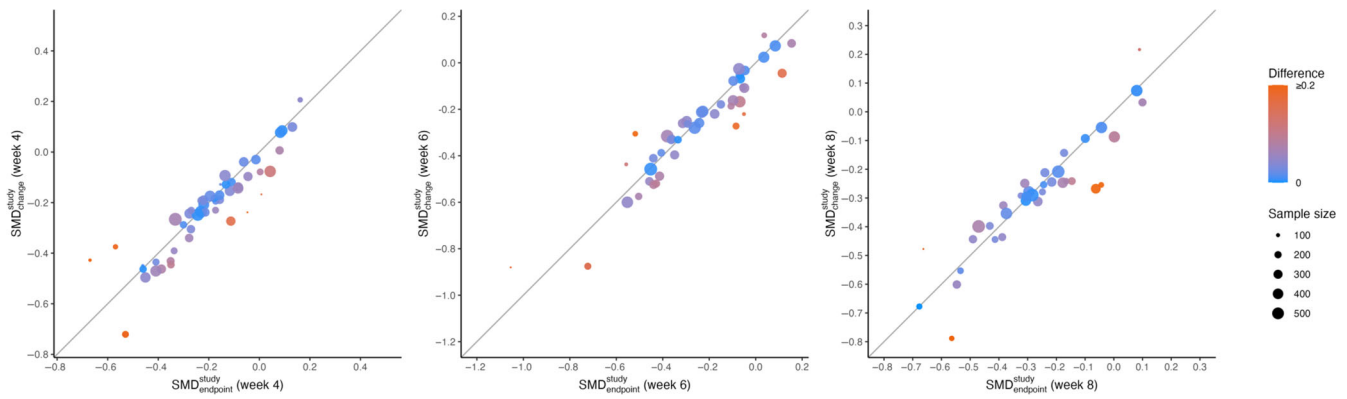


FIGURE 1 Comparison between study-specific $SMD_{\text{endpoint}}^{\text{study}}$ and $SMD_{\text{change}}^{\text{study}}$ in the pharmacological interventions dataset. Colours represent the absolute difference between $SMD_{\text{endpoint}}^{\text{study}}$ and $SMD_{\text{change}}^{\text{study}}$ (blue colour indicates values close to 0, orange indicates values close or beyond 0.2). Point sizes represent the sample size of the trial. SMD, standardised mean difference.

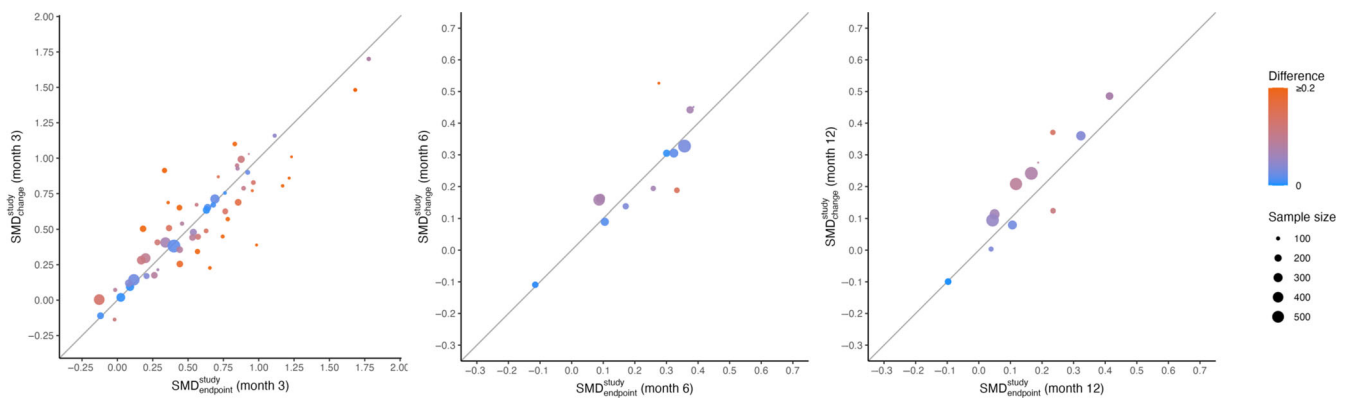


FIGURE 2 Comparison between study-specific $SMD_{\text{endpoint}}^{\text{study}}$ and $SMD_{\text{change}}^{\text{study}}$ in the internet-based cognitive behavioural therapy dataset. Colours represent the absolute difference between $SMD_{\text{endpoint}}^{\text{study}}$ and $SMD_{\text{change}}^{\text{study}}$ (blue colour indicates values close to 0, orange indicates values close or beyond 0.2). Point sizes represent the sample size of the trial. SMD, standardised mean difference.

zero studies at 12 months (distribution of $SMD_{\text{endpoint}}^{\text{study}} - SMD_{\text{change}}^{\text{study}}$: median -0.06 ; IQR $-0.08, 0.01$; range $-0.14, 0.11$). Additional details are available in Supporting Information 4.2.3. Of studies with data at baseline and follow-up, 44 (77.2%), 12 (92.3%) and 11 (91.7%) studies had a $SMD_{\text{baseline}}^{\text{study}}$ lower than 0.2 at 3, 6 and 12 months, respectively. Additional information on baseline imbalance and its impact on $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$ are provided in Supporting Information 8.2.0.

3.2 | Results from pairwise and network meta-analyses

3.2.1 | Pharmacological interventions

Pairwise meta-analyses

The $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$ for PWAs (antidepressants versus placebo) were -0.19 (95% CI $-0.23; -0.14$) and -0.21 (95% CI $-0.26; -0.16$) at 4 weeks ($k = 50$;

$n = 16,139$); -0.23 (95% CI $-0.30; -0.18$) and -0.25 (95% CI $-0.31; -0.20$) at 6 weeks ($k = 44$; $n = 16,301$); and -0.26 (95% CI $-0.32; -0.20$) and -0.27 (95% CI $-0.33; -0.22$) at 8 weeks ($k = 36$; $n = 10,840$), respectively.

The distribution of $SMD_{\text{mixed}}^{\text{pooled}}$, $seSMD_{\text{mixed}}^{\text{pooled}}$ and τ parameters when a mixture of endpoint and change from baseline scores was used showed small variation across different proportions of studies contributing with $SMD_{\text{endpoint}}^{\text{study}}$ and the considered time points. The difference between the lowest and highest median $SMD_{\text{mixed}}^{\text{pooled}}$ across different proportions at 4, 6 and 8 weeks was 0.02 ($k = 50$; $n = 16,139$), 0.01 ($k = 44$; $n = 16,301$) and 0.01 ($k = 36$; $n = 10,840$), respectively. Additional details on $SMD_{\text{mixed}}^{\text{pooled}}$, $seSMD_{\text{mixed}}^{\text{pooled}}$ and τ are available in the Supporting Information 5.1.1–5.

The difference between $SMD_{\text{optimistic}}^{\text{pooled}}$ and $SMD_{\text{pessimistic}}^{\text{pooled}}$ from the PWAs was not considered relevant at any time point (difference between $SMD_{\text{optimistic}}^{\text{pooled}}$ and $SMD_{\text{pessimistic}}^{\text{pooled}}$ of 0.05 at 4, 6 and 8 weeks) (additional details in Supporting Information 6.1.1–6).

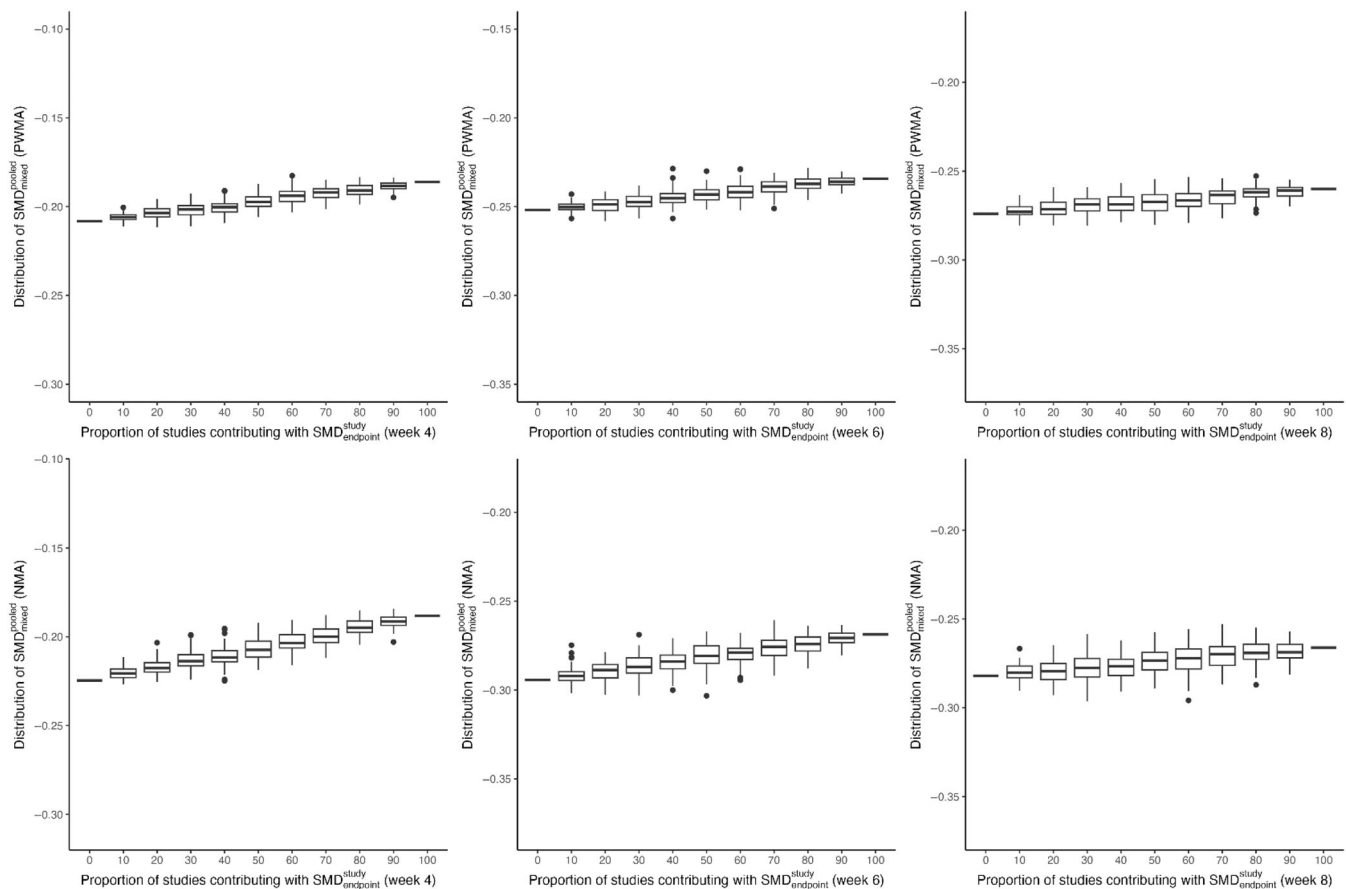


FIGURE 3 Distributions of pooled standardised mean differences ($SMD_{\text{mixed}}^{\text{pooled}}$) when randomly sampling 100 times studies contributing with endpoint ($SMD_{\text{endpoint}}^{\text{study}}$) over change from baseline ($SMD_{\text{change}}^{\text{study}}$) data in the pharmacological interventions dataset for pairwise and network meta-analyses (100 pairwise and 100 network meta-analyses per proportion and time point).

The number of included studies did not considerably affect the difference between $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$ across the considered time points. Additional details are available in Supporting Information 7.1.0.

Network meta-analyses

Overall, differences between the $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$ from NMAs (for the comparison of paroxetine versus pill placebo) were not considered clinically relevant: the two SMDs were estimated at -0.19 (95% CI $-0.25; -0.13$) and -0.23 (95% CI $-0.29; -0.17$) at 4 weeks ($k = 86; n = 23,413$); -0.27 (95% CI $-0.34; -0.20$) and -0.29 (95% CI $-0.36; -0.23$) at 6 weeks ($k = 82; n = 21,247$); and -0.27 (95% CI $-0.34; -0.20$) and -0.28 (95% CI $-0.35; -0.21$) at 8 weeks ($k = 56; n = 15,457$) respectively.

The distributions of $SMD_{\text{mixed}}^{\text{pooled}}$, $seSMD_{\text{mixed}}^{\text{pooled}}$ and τ parameters from the 2700 network meta-analyses showed small variations (Figure 3). The maximum difference between median $SMD_{\text{mixed}}^{\text{pooled}}$ across different proportions was 0.03 ($k = 86; n = 23,413$), 0.02 ($k = 82; n = 21,247$)

and 0.01 ($k = 56; n = 15,457$) across considered time points (4 weeks, 6 weeks and 8 weeks, respectively). Additional details on $SMD_{\text{mixed}}^{\text{pooled}}$, $seSMD_{\text{mixed}}^{\text{pooled}}$ and τ are available in the Supporting Information 5.1.6–13.

We found comparable results between the $SMD_{\text{optimistic}}^{\text{pooled}}$ and $SMD_{\text{pessimistic}}^{\text{pooled}}$ for the paroxetine versus placebo comparison (difference between $SMD_{\text{optimistic}}^{\text{pooled}}$ and $SMD_{\text{pessimistic}}^{\text{pooled}}$ of 0.04 at 4 weeks, and 0.05 at 6 and 8 weeks) (additional details in Supporting Information 6.1.7–12).

3.2.2 | Internet-based cognitive behavioural therapy

Pairwise meta-analyses

The estimated $SMD_{\text{endpoint}}^{\text{pooled}}$ and $SMD_{\text{change}}^{\text{pooled}}$ from the PWMA (active versus control interventions) were -0.53 (95% CI $-0.63; -0.43$) and -0.52 (95% CI $-0.61; -0.43$) at 3 months ($k = 57, n = 9194$); -0.21 (95% CI $-0.30; -0.12$) and -0.23 (95% CI $-0.30; -0.15$) at 6 months

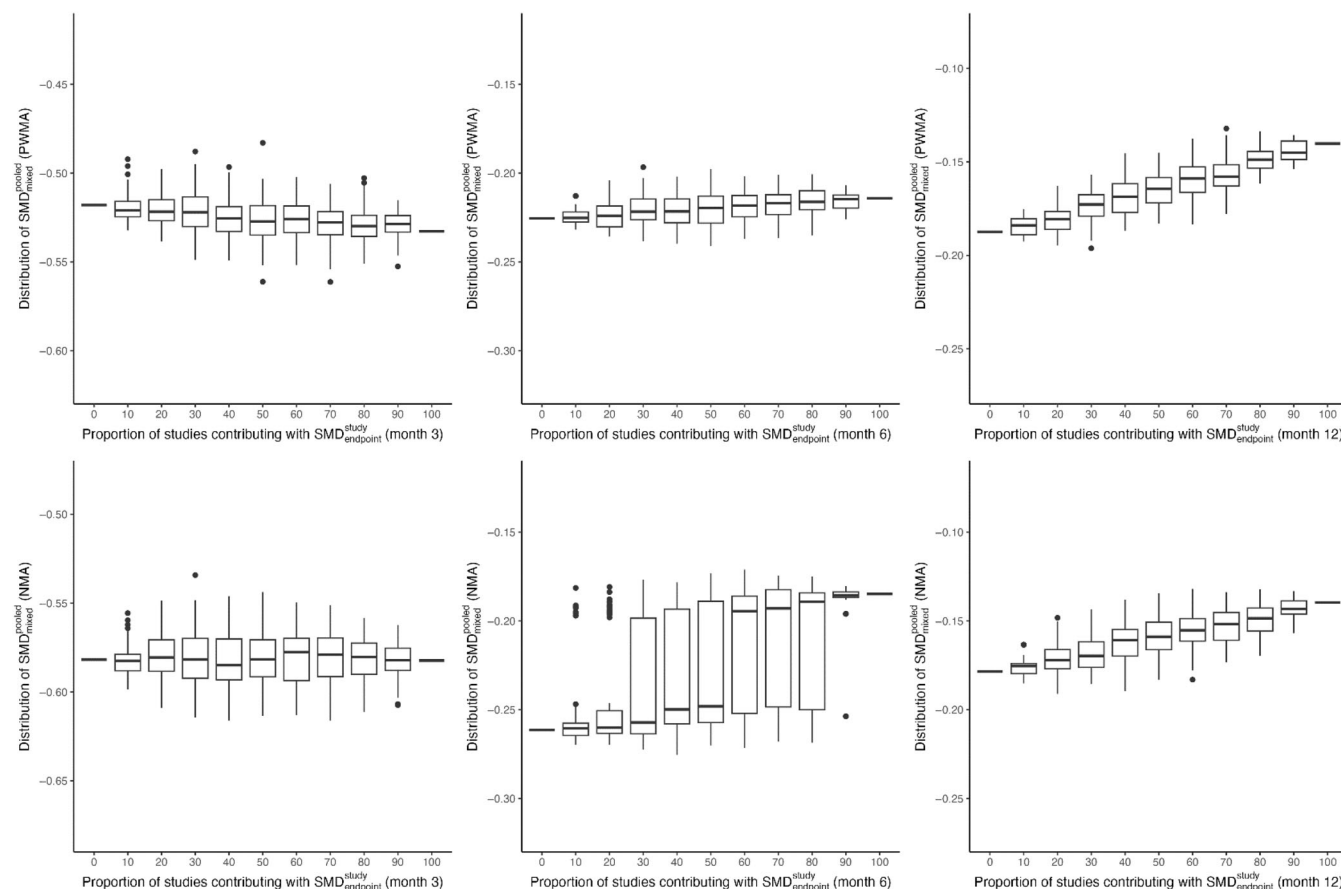


FIGURE 4 Distributions of pooled standardised mean differences (SMD_{mixed}^{pooled}) when randomly sampling 100 times studies contributing with endpoint ($SMD_{endpoint}^{study}$) over change from baseline (SMD_{change}^{study}) data in the internet-based cognitive behavioural therapy dataset for pairwise and network meta-analyses (100 pairwise and 100 network meta-analyses per proportion and time point).

($k = 13, n = 3341$); and -0.14 (95% CI $-0.21; -0.07$) and -0.19 (95% CI $-0.27; -0.11$) at 12 months ($k = 12, n = 3320$), respectively.

The distribution of SMD_{mixed}^{pooled} , $seSMD_{mixed}^{pooled}$ and τ parameters from the 2700 PWMA models showed limited variability across the investigated proportions of studies contributing with endpoint scores at different time points. The maximum difference between median SMD_{mixed}^{pooled} across different proportions was 0.01 ($k = 57, n = 9194$), 0.01 ($k = 13, n = 3341$) and 0.04 ($k = 12, n = 3320$) at 3, 6 and 12 months, respectively (additional details are available in Supporting Information 5.2.1–5).

The difference between $SMD_{optimistic}^{pooled}$ and $SMD_{pessimistic}^{pooled}$ from the PWMA models was 0.13 at 3 months, 0.04 at 6 months and 0.07 at 12 months. Additional details on the optimistic and pessimistic models are available in the Supporting Information 6.2.0.

The inclusion of fewer studies resulted in greater differences between $SMD_{endpoint}^{pooled}$ and SMD_{change}^{pooled} , but these differences were generally low (with 3 studies or more, $|SMD_{endpoint}^{pooled} - SMD_{change}^{pooled}|$ was 0.2 or larger 4% of the time at 3 months, and never at 6 and 12 months). Additional details are available in Supporting Information 7.2.0.

Network meta-analyses

The estimated $SMD_{endpoint}^{pooled}$ and SMD_{change}^{pooled} in NMAs (guided iCBT vs. control comparison) were -0.58 (95% CI $-0.69; -0.48$) and -0.58 (95% CI $-0.68; -0.48$) ($k = 60, n = 10,515$) at 3 months; -0.19 (95% CI $-0.31; -0.06$) and -0.26 (95% CI $-0.39; -0.13$) ($k = 14, n = 3958$) at 6 months; and -0.14 (95% CI $-0.22; -0.06$) and -0.18 (95% CI $-0.27; -0.09$) ($k = 14, n = 4124$) at 12 months, respectively.

We observed limited variability in SMD_{mixed}^{pooled} , $seSMD_{mixed}^{pooled}$ and estimates for τ from the 2700 NMAs across various proportion of studies contributing with endpoint scores and time points (Figure 4). When considering SMD_{mixed}^{pooled} across different proportions, the maximum difference between median values across different proportions was 0.01 ($k = 60, n = 10,515$), 0.07 ($k = 14, n = 3958$) and 0.03 ($k = 14, n = 4124$) at 3, 6 and 12 months, respectively (additional details in Supporting Information 5.2.6–13).

For NMAs, the difference between $SMD_{optimistic}^{pooled}$ and $SMD_{pessimistic}^{pooled}$ (guided iCBT vs. control comparison) was smaller than 0.2: 0.11 at 3 months, 0.09 at 6 months and 0.05 at 12 months. Additional details on the optimistic

and pessimistic models are available in the Supporting Information 6.2.0.

4 | DISCUSSION

Using individual patient data from 150 RCTs (41,096 participants), we showed that the choice between using endpoint scores ($SMD_{\text{endpoint}}^{\text{study}}$), change from baseline scores ($SMD_{\text{change}}^{\text{study}}$), or a mixture of the two, did not materially affect the estimation of the pooled SMD for the effect of pharmacological and non-pharmacological interventions on depression severity. This finding was robust in both PWMA and NMA, across different time points, and when different number of trials contributed to the analyses. Further, non-random selection of study-specific SMDs (i.e., driven by what is more advantageous for a specific treatment) had very small impact on the pooled SMD estimates in both examples and across all analysed time points.

In studies with imbalances in baseline scores, there may be important differences between endpoint and change from baseline SMDs.¹⁵ However, if randomisation is performed properly, such imbalances are expected to be smaller for larger sample sizes.¹⁶ One of the possible causes accounting for baseline imbalances in a randomised controlled trial is a small sample size. In our study, 25% of the studies investigating internet-based cognitive therapy at 3 months had important differences between endpoint and change from baseline study-specific SMDs. As expected, these studies tended to be of a smaller sample size. Nonetheless, also as expected, their impact on the summary SMD estimates was negligible due to the small weights allocated in the meta-analytical models, which in turn resulted from their small sample sizes. For meta-analysts, proposed strategies to address baseline imbalances in aggregated data are to (i) use study-specific ANCOVA estimates adjusting for baseline imbalances¹⁷; (ii) perform a meta-regression using the mean baseline scores as a covariate,¹⁵ in case study-specific ANCOVA-adjusted estimates are not available; or (iii) limit the meta-analytical model to studies with no baseline imbalances (e.g., excluding studies at high risk of bias in the sensitivity analyses of a meta-analysis).¹⁷ For trialists, this serves as a reminder to adjust their treatment effects for any baseline imbalances while adding to the debate on whether it is ethical to conduct underpowered studies given their future contribution in a meta-analytical model.^{18,19}

When individual study data on the SD of symptom severity at follow-up are not available, possible approaches to address this problem are to impute the missing SD using information from other studies, to use

instead the SD for the change from baseline SD from the same study, or to use the baseline SD from the same study.^{4,20} The latter has been discouraged because if the baseline SD values are smaller, this approach would overestimate treatment effects when using SMD.⁴ In our dataset, the median SD at baseline was indeed smaller than any follow-up, while the SD distributions of endpoint and change data were similar. Such differences are likely to reflect the application of a threshold cut-off (i.e. eligibility criteria) to the natural distribution of depressive symptom severity at baseline, heterogeneity of intervention effects across patients, and floor effects due to scores being bounded by upper lower limits. Therefore, using baseline SD instead of endpoint SD might overestimate treatment effect in many situations.

Another scenario that meta-analysts often face when dealing with aggregate data is the case when different studies report outcomes using a mixture of endpoints or change from baseline data, and in multiple scales. Such a scenario opens the question whether these effects should be lumped in a pooled SMD. In depression, this is of particular interest, given the considerable number of available trials and interventions, the lack of a consensus on how to report treatment effects, and the more than 280 rating scales developed to measure depressive severity.^{1,3,21} Our findings clearly support the combination of endpoint and change SMDs in a single meta-analytical model. The resulting pooled estimate will benefit from all the available data, potentially increasing precision and allowing the inclusion of more treatments. This approach strengthens the initial findings of da Costa and colleagues and confirm the robustness of key meta-analyses published in the field of depression.^{1,22}

Our results were consistent across multiple time points and for two IPD datasets, including 41,096 participants from 150 RCTs and six rating scales, overlapping with previous attempts not restricted to depression.¹¹ Moreover, we tested our findings when limited data was available (smaller number of studies) or maximising the differences (optimistic and pessimistic scenarios). A limitation of our findings is that they are specific to the considered datasets of pharmacological and psychological interventions for depression. The generalisability of these findings beyond depression should be investigated by replicating our analyses on other conditions where multiple types of rating scales are often used (e.g., schizophrenia, anxiety, pain).

In conclusion, we meta-analysed 150 studies on depression and found that the choice between SMDs estimated from endpoint and change from baseline data had immaterial impact on the summary meta-analytic estimate. We recommend future authors to pre-specify in their systematic review's protocol and method

section whether they plan to combine SMD^{study}_{endpoint} and SMD^{study}_{change} and potentially additional analyses to assess the robustness of their findings. Our findings should inform available guidance on how to synthesise outcome data from different scales in depression.

AUTHOR CONTRIBUTIONS

Edoardo G. Ostinelli: Conceptualization; investigation; methodology; writing – original draft; writing – review and editing; software; data curation; formal analysis; project administration. **Orestis Efthimiou:** Methodology; writing – review and editing. **Yan Luo:** Methodology; writing – review and editing. **Clara Miguel:** Project administration; data curation; writing – review and editing. **Eirini Karyotaki:** Data curation; writing – review and editing. **Pim Cuijpers:** Writing – review and editing. **Toshi A. Furukawa:** Methodology; supervision; writing – review and editing. **Georgia Salanti:** Methodology; writing – review and editing. **Andrea Cipriani:** Methodology; supervision; writing – review and editing.

ACKNOWLEDGMENTS

None.

FUNDING INFORMATION

Edoardo G. Ostinelli and Andrea Cipriani were supported by the National Institute for Health and Care Research (NIHR) (grant RP-2017-08-ST2-006), by the National Institute for Health Research (NIHR) Applied Research Collaboration Oxford and Thames Valley (ARC OxTV) at Oxford Health NHS Foundation Trust, by the National Institute for Health and Care Research Oxford Health Clinical Research Facility, and by the NIHR Oxford Health Biomedical Research Centre (grant BRC-1215-20005). Edoardo G. Ostinelli was supported by the Brasenose College Senior Hulme scholarship. Orestis Efthimiou was supported by project grant number 180083 from the Swiss National Science Foundation (SNSF). Georgia Salanti was supported by the Swiss National Science Foundation (grant/award number 179158). The views expressed are those of the authors and not necessarily those of the UK National Health Service, the NIHR, or the Department of Health and Social Care.

CONFLICT OF INTEREST STATEMENT

Edoardo G. Ostinelli received research and consultancy fees from Angelini Pharma. Toshi A. Furukawa reports personal fees from Boehringer-Ingelheim, DT Axis, Kyoto University Original, Shionogi and SONY, and a grant from Shionogi, outside the submitted work; in addition, Toshi A. Furukawa has patents 2020-548587 and

2022-082495 pending, and intellectual properties for Kokoro-app licenced to Mitsubishi-Tanabe. Andrea Cipriani received research, educational and consultancy fees from the Italian Network for Paediatric Trials, CARI-PLO Foundation, Lundbeck and Angelini Pharma. All the other authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from original data owner(s). Restrictions apply to the availability of these data, which were used under license for this study. Data are available from the author(s) with the permission of original data owner(s).

ORCID

Edoardo G. Ostinelli  <https://orcid.org/0000-0002-8717-0832>

Orestis Efthimiou  <https://orcid.org/0000-0002-0955-7572>

Yan Luo  <https://orcid.org/0000-0002-5271-5126>

Clara Miguel  <https://orcid.org/0000-0001-5563-5896>

Eirini Karyotaki  <https://orcid.org/0000-0002-0071-2599>

Pim Cuijpers  <https://orcid.org/0000-0001-5497-2743>

Toshi A. Furukawa  <https://orcid.org/0000-0003-2159-3776>

Georgia Salanti  <https://orcid.org/0000-0002-3830-8508>

Andrea Cipriani  <https://orcid.org/0000-0001-5179-8321>

REFERENCES

1. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 21 antidepressant drugs for the acute treatment of adults with major depressive disorder: a systematic review and network meta-analysis. *Lancet*. 2018; 391(10128):1357-1366.
2. Karyotaki E, Efthimiou O, Miguel C, et al. Internet-based cognitive behavioral therapy for depression: a systematic review and individual patient data network meta-analysis. *JAMA Psychiatry*. 2021;78(4):361-371.
3. Wahl I, Lowe B, Bjorner JB, et al. Standardization of depression measurement: a common metric was developed for 11 self-report depression measures. *J Clin Epidemiol*. 2014; 67(1):73-86.
4. Higgins JPT, Green S, Cochrane C. *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane Collaboration; 2022.
5. Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *BMJ*. 2019;364:k4817.
6. Vickers AJ, Altman DG. Statistics notes: Analysing controlled trials with baseline and follow up measurements. *BMJ*. 2001; 323(7321):1123-1124.
7. Luo Y, Funada S, Yoshida K, Noma H, Sahker E, Furukawa TA. Large variation existed in standardized mean difference estimates using different calculation methods in clinical trials. *J Clin Epidemiol*. 2022;149:89-97.
8. Norman GR. Issues in the use of change scores in randomized trials. *J Clin Epidemiol*. 1989;42(11):1097-1105.

9. Fu R, Holmer HK. Change score or follow-up score? Choice of mean difference estimates could impact meta-analysis conclusions. *J Clin Epidemiol*. 2016;76:108-117.
10. Gelman A, Stern H. The difference between “significant” and “not significant” is not itself statistically significant. *Am Stat*. 2006;60(4):328-331.
11. da Costa BR, Nuesch E, Rutjes AW, et al. Combining follow-up and change data is valid in meta-analyses of continuous outcomes: a meta-epidemiological study. *J Clin Epidemiol*. 2013;66(8):847-855.
12. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. Elsevier Science; 2013.
13. Balduzzi S, Rucker G, Nikolakopoulou A, et al. Netmeta: an R package for network meta-analysis using frequentist methods. *J Stat Softw*. 2023;106(2):1-40.
14. Balduzzi S, Rucker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health*. 2019;22(4):153-160.
15. Trowman R, Dumville JC, Torgerson DJ, Cranny G. The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study. *J Clin Epidemiol*. 2007;60(12):1229-1233.
16. Roberts C, Torgerson DJ. Understanding controlled trials: baseline imbalance in randomised controlled trials. *BMJ*. 1999;319(7203):185.
17. Riley RD, Kauser I, Bland M, et al. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. *Stat Med*. 2013;32(16):2747-2766.
18. Guyatt GH, Mills EJ, Elbourne D. In the era of systematic reviews, does the size of an individual trial still matter. *PLoS Med*. 2008;5(1):e4.
19. Halpern SD, Karlawish JH, Berlin JA. The continuing unethical conduct of underpowered clinical trials. *JAMA*. 2002;288(3):358-362.
20. Furukawa TA, Barbui C, Cipriani A, Brambilla P, Watanabe N. Imputing missing standard deviations in meta-analyses can provide accurate results. *J Clin Epidemiol*. 2006;59(1):7-10.
21. Santor DA, Gregus M, Welch A. Focus article: eight decades of measurement in depression. *Meas-Interdiscip Res Perspect*. 2006;4(3):135-155.
22. Cuijpers P, Noma H, Karyotaki E, Vinkers CH, Cipriani A, Furukawa TA. A network meta-analysis of the effects of psychotherapies, pharmacotherapies and their combination in the treatment of adult depression. *World Psychiatry*. 2020;19(1):92-107.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ostinelli EG, Efthimiou O, Luo Y, et al. Combining endpoint and change data did not affect the summary standardised mean difference in pairwise and network meta-analyses: An empirical study in depression. *Res Syn Meth*. 2024;1-11. doi:[10.1002/jrsm.1719](https://doi.org/10.1002/jrsm.1719)