







RESEARCH ARTICLE  

# Synthesis of depression outcomes reported on different scales: A comparison of methods for modelling mean differences

Beatrice C. Downing <sup>1</sup>, Nicky J. Welton <sup>1</sup>, Hugo Pedder <sup>1</sup>, Ifigenia Mavranzouli <sup>2</sup>,  
Odette Megnin-Viggars <sup>2</sup> and A.E. Ades <sup>1</sup>

<sup>1</sup>Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK

<sup>2</sup>Centre for Outcomes Research and Effectiveness, Research Department of Clinical, Educational & Health Psychology, University College London, London, UK

**Corresponding author:** Beatrice C. Downing; Email: [beatrice.downing@bristol.ac.uk](mailto:beatrice.downing@bristol.ac.uk)

**Received:** 24 May 2024; **Revised:** 14 November 2024; **Accepted:** 14 January 2025

**Keywords:** meta-regression against baseline severity; network meta-analysis; patient and clinician reported outcomes; ratio of means; standardised mean difference



## Abstract

Several methods have been proposed for the synthesis of continuous outcomes reported on different scales, including the Standardised Mean Difference (SMD) and the Ratio of Means (RoM). SMDs can be formed by dividing the study mean treatment effect either by a study-specific (Study-SMD) or a scale-specific (Scale-SMD) standard deviation (SD). We compared the performance of RoM to the different standardisation methods with and without meta-regression (MR) on baseline severity, in a Bayesian network meta-analysis (NMA) of 14 treatments for depression, reported on five different scales. There was substantial between-study variation in the SDs reported on the same scale. Based on the Deviance Information Criterion, RoM was preferred as having better model fit than the SMD models. Model fit for SMD models was not improved with meta-regression. Percentage shrinkage was used as a scale-independent measure with higher % shrinkage indicating lower heterogeneity. Heterogeneity was lowest for RoM (20.5% shrinkage), then Scale-SMD (18.2% shrinkage), and highest for Study-SMD (16.7% shrinkage). Model choice impacted which treatment was estimated to be most effective. However, all models picked out the same three highest-ranked treatments using the GRADE criteria. Alongside other indicators, higher shrinkage of RoM models suggests that treatments for depression act multiplicatively rather than additively. Further research is needed to determine whether these findings extend to Patient- and Clinician-Reported Outcomes used in other application areas. Where treatment effects are additive, we recommend using Scale-SMD for standardisation to avoid the additional heterogeneity introduced by Study-SMD.

## Highlights

### What is already known

1. To synthesise continuous outcomes measured on different scales, mean treatment effects can be standardised by dividing by the study-specific SD or by a scale-specific SD. Alternatively, a Ratio of Means (RoM) approach expresses all treatment effects as a ratio.

  This article was awarded Open Data and Open Materials badges for transparent practices. See the Data availability statement for details.

© The Author(s), 2025. Published by Cambridge University Press on behalf of The Society for Research Synthesis Methodology. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (<https://creativecommons.org/licenses/by/4.0>), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

2. Standardisation assumes treatments act additively; RoM assumes multiplicative effects.
3. Treatment effects measured on scales that consist of sums of correlated 2-, 3-, or 4-item subscales are expected to act multiplicatively.

#### **What is new**

1. In a network meta-analysis of treatments for depression, heterogeneity (measured by shrinkage) was lowest for RoM, suggesting treatments for depression may act multiplicatively on the commonly used scales.
2. Standardisation by scale-specific SD (Scale-SMD) was superior to standardisation by study-specific SD (Study-SMD), giving better model fit and lower heterogeneity.
3. We suggest five markers of multiplicative treatment effects, four of which are satisfied by depression scores.
4. The choice of scale model had a limited impact on treatment recommendations.

#### **Potential impact for RSM readers outside the authors' field**

1. Prior to selecting the approach for synthesising continuous outcomes reported on different outcome scales, it is necessary to establish whether interventions act in an additive or multiplicative fashion.
2. For outcome scales that are additive, we recommend standardising by a scale-specific SD, to avoid the additional heterogeneity introduced by study-specific standardisation.

## **1. Introduction**

Several methods have been proposed for the synthesis of continuous outcomes measured on different scales.<sup>1</sup> These can be sorted into two broad groups. In the first group, the mean treatment effect from each trial is “standardised” by dividing it by a constant, with the aim of expressing all treatment effects in the same units. This assumes that the scales are linearly related and that treatment acts in an additive fashion. Methods of this type can be further subdivided, according to whether the dividing constant is trial-specific,<sup>2</sup> or scale-specific<sup>3,4</sup>; this is discussed below. The second approach is the synthesis of the Ratio of Means (RoM).<sup>5</sup> While standardisation assumes that treatment acts on measurements in an additive way, RoM assumes that treatment acts to multiply or divide scores.

Most tutorial and methodological guidance papers have described multiplicative and additive approaches as alternative modelling options, without recommending one over the other.<sup>4,6–8</sup> A study of 232 systematic reviews<sup>9</sup> failed to show a clear advantage of either method over the other on measures of between-study heterogeneity. This may, however, have been due to the wide variety of continuous outcomes included in that review. It may be expected that the properties of the specific measurement scales determine whether additive or multiplicative approaches are the most appropriate. For example, many biological measurements, such as cell counts and concentrations in body fluids, are frequently transformed to the log-scale prior to statistical analysis, suggesting that multiplicative models are more appropriate. In studies of depression, a positive relation between the treatment effect and baseline severity is frequently reported,<sup>10–13</sup> which could be interpreted as indicating that treatment effects are proportional rather than additive. Moreover, depression scales are sums of correlated 2-, 3-, or 4-category sub-scales, with a zero origin: based on statistical theory they would be expected to be Poisson-distributed, heteroscedastic, and transformed to normality by log transformations. Patient- and Clinician-Reported outcomes (PCROs) of this type contrast with many of the measurement scales used in educational research, where meta-analytic methods originated,<sup>2,14</sup> that have been constructed to be additive on a natural scale.

By far the most common form of standardisation is to divide mean treatment differences by the study-specific standard deviation (SD), to produce the classic Standardised Mean Difference (SMD). Study-based standardisation assumes that all studies using the same scale have the same SD.<sup>15</sup> However, due to population differences and sampling error, SDs inevitably vary across studies reporting results on the same scale, introducing noise,<sup>6</sup> and potentially bias if they vary systematically. To avoid this, standardisation based on division by a scale-specific constant, rather than a study-specific constant, has been proposed. One option is to divide mean differences by the Minimal Clinically Important

Difference (MID) for each scale.<sup>1,4,16</sup> In a similar vein, Hunter and Schmidt (2004) recommended dividing by the SD in a reference population (the reference SD) for that scale, recognising that variation in SD, which they termed “range variation”, would create artefacts that should be removed to reveal the true treatment effects.<sup>17</sup> Unfortunately, neither MID’s nor reference SD’s have been published for the vast majority of outcome scales. A simple workaround,<sup>18</sup> which we adopt here, is to approximate a reference SD for each scale by taking the average baseline SD over all the trials reporting results for that scale in the evidence synthesis. Alternative SD and MID approaches are described in the Discussion.

To summarise the properties of the scales, we start from the standard model in which mean treatment effects are seen as the sum of a study component  $\mu$  and relative effect  $\delta$ . The hypothesis proposed by Study-SMD is that with mean outcomes  $\bar{X}_i, \bar{X}_j$ , measured on two scales  $i, j$ , there are standardizing constants  $s_i, s_j$  such that differences between the standardised mean outcomes are a constant, independent of treatment:

$$\frac{\bar{X}_i}{s_i} = \mu_i + \delta; \quad \frac{\bar{X}_j}{s_j} = \mu_j + \delta; \quad \therefore \frac{\bar{X}_i}{s_i} - \frac{\bar{X}_j}{s_j} = \mu_i - \mu_j = \text{constant}$$

This assumes that the standardising constants are constant between trials, contrary to the facts of range variation. The Scale-SMD hypothesis is exactly the same, except that range variation is eliminated by fixing the standardising constants for each scale.

The RoM hypothesis is that the ratio of the mean outcomes is constant, independent of treatment.

$$\log(\bar{X}_i) = \mu_i + \delta; \quad \log(\bar{X}_j) = \mu_j + \delta; \quad \therefore \frac{\bar{X}_i}{\bar{X}_j} = \exp(\mu_i - \mu_j) = \text{constant.}$$

In this paper, we compare the performance of different standardisation and ratio methods in a network meta-analysis of 14 treatments for depression, reported on five different scales. We compare Study-SMD and Scale-SMD, with and without meta-regression (MR) against baseline SD, and RoM. Our objective is to develop methods for deciding which of these models provides the best fit to the data, the most precise estimates, and the lowest between-study heterogeneity. We illustrate these methods by applying them to a frequently used form of PCRO data. We also investigate the impact of model choice on treatment recommendations.

## 2. Methods

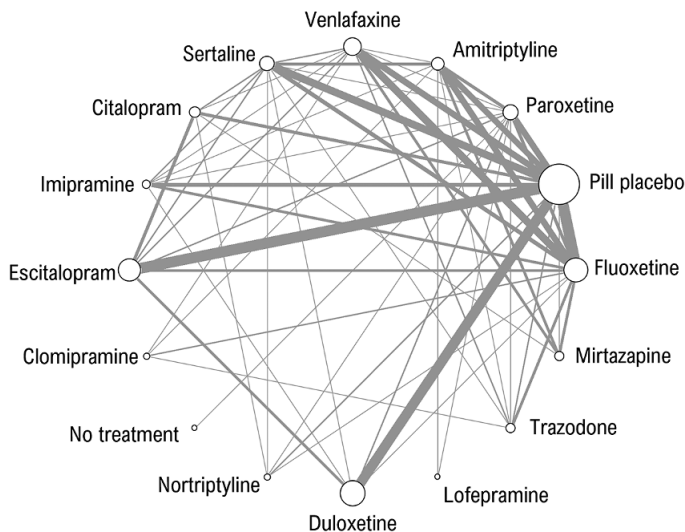
### 2.1. Dataset: Pharmacological treatments for depression

We use a dataset of 161 studies which were a subset of studies from NICE guideline NG222 that compared pharmacological treatments for more severe (moderate and severe) depression in adults from a larger review.<sup>19</sup> Studies were included only if they reported findings on the Hamilton Depression Rating Scale (HAMD-17, HAMD-21, HAMD-24), the Beck Depression Inventory (BDI), or the Montgomery-Asberg Depression Scale (MADRS). There were comparisons between 14 active treatments, pill placebo, and “no treatment”. Two studies that reported exceptionally low standard deviations (0.46 and 0.65 on the HAMD-21 scale) were excluded. The evidence network was highly connected with multiple trials of each active treatment (Figure 1). There was direct evidence on 59 of the possible 120 pair-wise comparisons, and they were informed by between 1 and 14 trials (median 2 trials). Because we wished to compare models with and without meta-regression, only studies reporting baseline severity for the whole sample were included.

Data were extracted for NG222 in one of three different formats in the following priority order:

- i) mean change-from-baseline (CFB) with standard error for each study arm
- ii) mean depression scores at baseline and follow-up with standard errors for each study arm
- iii) mean depression score at follow-up, with standard error for each study arm

following guidance in Guideline Methodology Document 2<sup>18</sup> for data extraction to fit SMD models.



**Figure 1.** Network of evidence on 14 pharmacological treatments of more severe depression, no treatment and pill placebo. The thickness of edges is proportional to the number of studies and the size of nodes is proportional to the number of participants receiving the treatment.

Network diagrams of trials with data in these three formats, along with the types of data (by format and scale) extracted for each treatment, are shown in the [Supplementary Materials \(Supplementary Figure S1 and Supplementary Table S1, respectively\)](#). However, when fitting a RoM model, the recommended<sup>18</sup> priority order of data extraction would instead be formats (ii), (iii), then (i), since an assumption of additivity is inherent in calculating CFB, and therefore it is least preferred when fitting a multiplicative model. We, therefore, ran a sensitivity analysis using this alternative prioritisation of data formats; however, this was constrained by available data in the NG222 dataset, as we did not re-extract data from these studies.

For studies reporting baseline and follow-up measures, we need to assume a value for the correlation,  $\rho_i$ , between baseline and follow-up outcomes to form the standard error of the CFB, and also to estimate RoM from studies that only report in CFB data format (see [Supplementary Materials, Appendix 1](#)). The correlation was assumed to be 0.3, which was the median correlation in the studies where it could be estimated in the NG222.<sup>19</sup> As a sensitivity analysis, we also examined results with the correlation set to 0.5.

## 2.2. Descriptive analyses

Two descriptive analyses were run, to test or to assess the assumptions of the different models. Bartlett's test of equality of variances was applied to assess the assumption of equality of SDs in different studies reported on the same scale.<sup>20</sup> The relation between mean scores at baseline (or follow-up) and the SD of scores at baseline (or follow-up) is an indicator of heteroscedasticity.

## 2.3. Standardised Mean Difference (SMD) measures

The standardised mean difference divides the mean difference between treatment arms by a standardising constant  $s$ .

For our depression model, the mean difference represents the difference between treatments in the mean CFB. Let  $mean_{B,k}$  be the mean outcome at baseline and  $mean_{F,k}$  the mean outcome at follow-up for treatment  $k$ , then the SMD for the mean CFB is:

$$SMD = \frac{(mean_{F,2} - mean_{B,2}) - (mean_{F,1} - mean_{B,1})}{s} \quad (1)$$

We prioritise modelling the CFB when possible, as it accounts for baseline differences, but we can use follow-up means if no other data are available.

We explore two alternative standardising constants  $s$ , one using the study-specific SD and the other using a scale-specific SD, and refer to the resulting SMD as Study-SMD and Scale-SMD respectively.

### 2.3.1. Study-specific SMD

The most common approach is to standardise by the pooled baseline SD specific to the study, termed Cohen's  $d$ .<sup>21</sup> In this case, the SDs at baseline from each arm,  $k$ , are combined to give a single pooled SD for the study  $i$ :

$$sd_i^{study} = \sqrt{\frac{\sum_k (n_{i,k} - 1) sd_{B,i,k}^2}{\sum_k n_{i,k} - k}} \tag{2}$$

where  $n_{i,k}$  is the number of patients randomised, and  $sd_{B,i,k}$  is the baseline SD for study  $i$  arm  $k$ . We use the baseline SD to compute the standardising constant because it is not influenced by treatment, so better reflects the SD of the outcome scale in the population recruited to the study and can be estimated from all trial arms. Note, that we do not consider the situation where interest is in the treatment effect on standard deviation of outcomes.

### 2.3.2. Scale-specific SMD

Ideally, a scale-specific standardisation constant would be obtained from a large representative population where all scales have been measured. However, in the absence of such a study, we can pool the study-specific baseline SDs,  $sd_i^{study}$ , across all those studies that report on a specific scale  $j$ :

$$sd_j^{scale} = \frac{\sum_{i \in j} sd_i^{study}}{n_j} \tag{3}$$

where  $i \in j$  indicates studies reporting scale  $j$ , and  $n_j$  the number of studies that report scale  $j$ .  $sd_j^{scale}$  is used as the standardising constant for those studies reporting scale  $j$ .

## 2.4. Ratio of Means (RoM) measure

If both baseline and follow-up measures are available then a multiplicative effect measure for the ratio change from baseline is the Ratio of Ratio of Means (RoRoM), calculated as:

$$RoRoM = \frac{mean_{F,2}/mean_{B,2}}{mean_{F,1}/mean_{B,1}} \tag{4}$$

If only follow-up measures are available, then the Ratio of Means (RoM) is:

$$RoM = \frac{mean_{F,2}}{mean_{F,1}} \tag{5}$$

Similarly to CFB, RoRoM adjusts for baseline imbalances, and so is preferred over RoM when it can be calculated. However, both RoRoM and RoM can be pooled to obtain an estimate of the treatment effect.

The RoM is not easily interpretable when outcome scores can have both positive and negative signs. However, if baseline values are available, they can be combined with the CFB to obtain baseline and follow-up means so that the RoRoM in equation (4) can be calculated (see Appendix 1).<sup>18</sup>

### 2.5. Network Meta-Analysis (NMA) model

Network meta-analysis (NMA) enables evidence to be pooled on multiple treatments where the evidence forms a connected network (Figure 1). We used a Bayesian NMA model,<sup>22,23</sup> where we adapted the likelihood and link functions to pool the various data formats to inform the SMD and RoM models (see Supplementary Materials, Appendix 1). The NMA model is put on the parameter  $\theta_{i,k}$  for arm  $k$  of study  $i$  which represents the standardised mean for the SMD models, and represents the log-mean outcome for the ROM model. The NMA model is the same for the SMD and RoM models, but the parameters have different interpretations:

$$\begin{aligned}\theta_{i,k} &= \mu_i + \delta_{i,k} \\ \delta_{i,k} &\sim N\left(d_{t_{i,k}} - d_{t_{i,1}}, \tau^2\right)\end{aligned}\quad (6)$$

where  $\mu_i$  is the standardised mean (or log-mean) for the treatment on arm 1 of study  $i$ , and  $\delta_{i,k}$  is the study-specific SMD (or log RoM) for arm  $k$  relative to arm 1 of study  $i$ . For a random effects NMA it is assumed that the study-specific SMDs (or log RoMs) come from a Normal distribution, where  $t_{i,k}$  indicates the treatment on arm  $k$  of study  $i$ ,  $d_k$  represents the SMD (or log RoM) for treatment  $k$  relative to treatment 1 (the reference treatment for the network), and  $\tau$  represents the between-study standard deviation on the linear predictor scale. Based on previous analyses of treatments for depression it is expected that there will be a degree of heterogeneity, and so we did not fit fixed effect models.

Estimates of the pooled RoM for treatment  $k$  relative to treatment 1 is obtained by exponentiating the log RoM:

$$RoM_k = e^{d_k} \quad (7)$$

By using the exact same dataset and likelihood for all the models, we are able to combine multiple data types (follow-up scores and CFB) and to provide valid comparisons of goodness of fit statistics for all five models.

### 2.6. Meta-regression of SMD on baseline severity

Versions of both Scale- and Study-SMD models were created in which a regression term for baseline severity was introduced in all active vs inactive comparisons, which we term Scale-SMD-MR and Study-SMD-MR respectively (see Appendix 1 for details).

### 2.7. Model comparison and selection

The models we compared were: SMD standardised with study-specific SD (Study-SMD); SMD standardised with scale-specific SD (Scale-SMD); Study-SMD with meta-regression for severity (Study-SMD-MR); Scale-SMD with meta-regression for severity (Scale-SMD-MR); and RoM. Models were compared using the posterior mean residual deviance,  $\bar{D}$ , as a measure of model fit, and the Deviance Information Criterion (DIC), which penalises the deviance by a measure of model complexity, the number of effective parameters,  $pD$ .<sup>24</sup>  $pD$  is calculated as the sum over study arms of the difference between the posterior mean deviance and the deviance evaluated at the posterior mean of the mean value of  $\theta_{i,k}$ . Models with lower  $\bar{D}$  and DIC are preferred.

We report the posterior median and 95% credible intervals for the between-study standard deviation, as a measure of heterogeneity. However, it is important to note that these are not comparable across different outcome scales. To overcome this, we introduce a scale-independent measure of heterogeneity, percentage shrinkage, that can be calculated using  $pD$ .

For a fixed effect model, all study estimates are equal for the same comparison, so there is no heterogeneity ( $\tau = 0$ ), 100% shrinkage and the effective number of model parameters is  $pD = ns + nt - 1$ ,

for the  $ns$  study baseline parameters, and  $nt - 1$  treatment effects relative to reference treatment 1. At the other extreme, for an “independent effects” model where each study effect is independent of all other study effects, there is a high value of  $\tau$ , which will depend on the scale of analysis, 0% shrinkage, and the effective number of model parameters is  $p_D = na$ , for the number of study arms  $na$ , reflecting a different parameter for each study arm, which is an upper bound for  $p_D$ . Random effects models with a value of  $p_D$  lying between these extremes exhibit a degree of heterogeneity, with values closer to the upper bound indicating higher levels of heterogeneity. We measure this using the % shrinkage defined as:

$$Shrinkage = 100 * \left( 1 - \frac{p_D - (ns + nt - 1)}{na - (ns + nt - 1)} \right) \tag{8}$$

with low heterogeneity having a % shrinkage closer to 100%, and those with higher heterogeneity having a % shrinkage closer to 0%.

Where baseline and CFB or baseline and follow-up values are used in a bivariate likelihood, there are two data points per study arm, so  $na$  is replaced with  $(na^{univ} + 2na^{biv})$  in (9), where  $na^{univ}$  is the number of univariate study arms and  $na^{biv}$  is the number of bivariate study arms.

In the depression example,  $ns = 161$ ,  $nt = 16$ , and  $na = 340$ , with 271 study arms reported as CFB or baseline and follow-up,  $na^{biv}$ , and 69 study arms reported as final values,  $na^{uni}$ , so  $p_D$  lies between 176 and 611.

**2.8. Transformation of SMDs and RoM to a common measurement scale (HAMD-17)**

To compare treatment effects from SMD and RoM models, the relative treatment effects  $d_k$  were transformed to a mean difference on the most frequently reported scale: HAMD-17. This required an assumption about the mean and SD on the reference treatment 1 on the HAMD-17 scale. Treatment effects estimated from SMD models,  $d_k^{SMD}$ , were back-transformed using the mean pooled baseline SD for the HAMD-17 scale,  $sd_B^{HAMD-17}$ :

$$d_k^{HAMD-17} = d_k^{SMD} * sd_B^{HAMD-17} \tag{9}$$

Similarly, between-study SD for the SMD models was back-transformed onto the HAMD-17 scale using the mean pooled baseline SD for the HAMD-17 scale,  $sd_B^{HAMD-17}$ .

Treatment effects estimated from the RoM model were back-transformed to a mean difference on the HAMD-17 using a representative mean HAMD-17 score on the reference treatment,  $\bar{\mu}_{HAMD-17}$ :

$$d_k^{HAMD-17} = \left( \bar{\mu}_{HAMD-17} * e^{d_k} \right) - \bar{\mu}_{HAMD-17} \tag{10}$$

where  $\bar{\mu}_{HAMD-17}$  was the CFB calculated from the mean at follow-up values (−6.4 and 15.7 respectively) taken from the largest study reporting depression scores for pill placebo (the reference treatment) on the HAMD-17 scale.<sup>25,26</sup>

**2.9. Treatment recommendations**

We also compare the impact of the different models on resulting treatment recommendations based on five different decision rules:

- i) the treatment with the highest posterior mean estimate of efficacy.
- ii) all treatments where the 95% credible interval (CrI) of the treatment effect relative to pill placebo did not include zero.



**Table 1.** Variation in baseline SD within each scale and regression of baseline SD against baseline severity.

Scale	N studies (participants)	Pooled SD at baseline		Difference max–min score	Ratio max:min pooled SD	Bartlett test	Slope: baseline SD vs baseline severity
		Mean	Range				
HAMD–17	78 (150,90)	4.01	1.55, 7.22	52	4.7	$x^2=1725$ , df = 77, $p < 0.0001$	0.08 (–0.01, 0.18)
HAMD–21	25 (3,807)	4.63	1.78, 6.40	52	3.6	$x^2=437$ , df = 24, $p < 0.0001$	0.15 (0.01, 0.29)
HAMD–24	5 (865)	4.75	3.36, 5.65	76	1.7	$x^2=26.0$ , df = 4, $p < 0.0001$	0.16 (–0.39, 0.72)
MADRS	52 (11,627)	5.15	3.35, 10.56	60	3.2	$x^2=1413$ , df = 51, $p < 0.0001$	–0.17 (–0.30, –0.04)
BDI-I	1 (97)	6.37	–	63	–	Not estimable	Not estimable

- iii) all treatments where the 95% CrI of the treatment effect relative to pill placebo was greater than a clinically important difference of 1 unit on the HAMD-17 scale, which is approximately 0.25 SD units (Table 1).
- iv) As (iii) above, but limited to treatments that were not inferior to the best treatment (ie the 95%CrI for the difference did not include zero).
- v) The set of treatments that satisfy decision rule (iii) and which were not inferior to any other treatment (ie the 95% credible interval on the difference does not include 0). This rule is a version of a proposal from the GRADE Working Party, with the threshold value set to zero.<sup>27</sup>

## 2.10. Software, model estimation, and priors

Models were estimated by Bayesian Markov chain Monte Carlo in WinBUGS 1.4.3.<sup>28</sup> Models were run with three chains with the first 40,000 iterations discarded as burn-in. Chain convergence was assessed with Brooks–Gelman–Rubin plots and chain mixing with trace and history plots. Following burn-in and convergence, 80,000 iterations from each of the three chains were used for inference.

Vague priors were provided for estimates of treatment effect,  $d_k$ , study-level baselines,  $\mu_i$ , and between-study SD,  $\tau$ .

$$d_k \sim N(0, 100^2)$$

$$\mu_i \sim N(0, 100^2)$$

$$\tau \sim U(0, 5) \quad (11)$$

Data visualisation was performed in R version 4.3.1<sup>27</sup> using packages ggplot2, ggpubr, and viridis.



### 3. Results

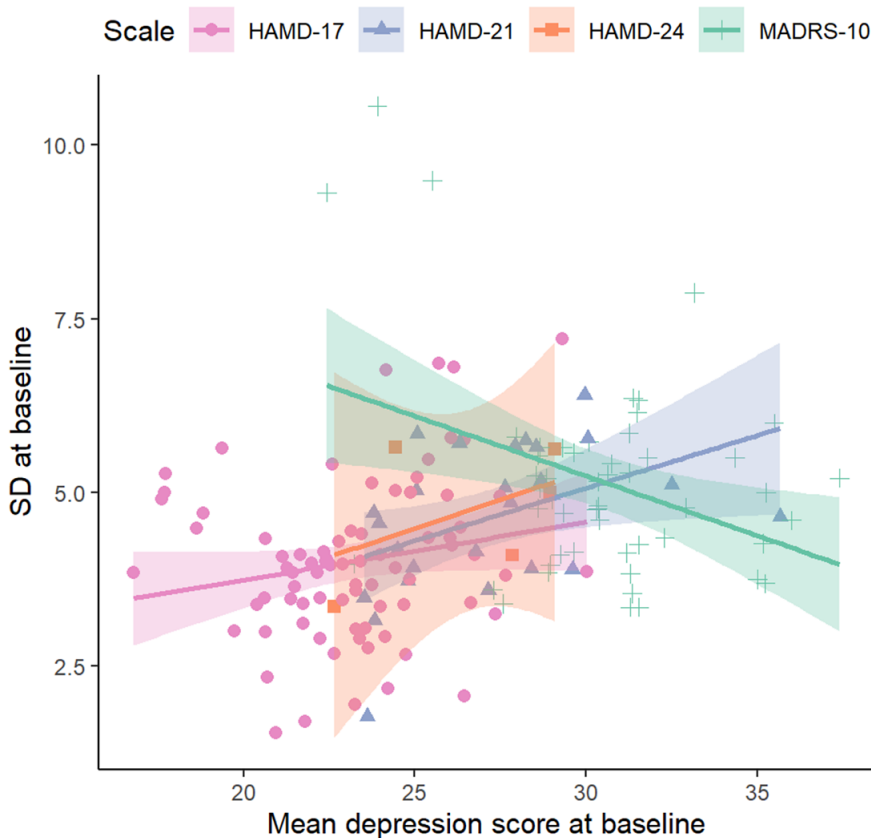
#### 3.1. Descriptive analysis

Bartlett tests<sup>20</sup> of the null hypothesis of equal variance in studies reporting on the same scale indicated statistically large differences between the baseline standard deviations (SDs) from each study. The ratio of maximum to minimum baseline SDs varied from 1.7 (HAMD-24) to 4.7 (HAMD-17) (Table 1), with a weighted average of 4.0.

For the three HAMD scales, mean depression scores at baseline increased with baseline SD, but this was not observed with the MADRS scale (Table 1, Figure 2).

#### 3.2. Model fit, heterogeneity, relative effects, and their precision

Study-SMD and Study-SMD-MR models have a slightly lower posterior mean residual deviance (Table 2), indicating a better fit compared to the other measures. The closer fit is the result of greater model complexity, reflected in the higher  $pD$ , and larger heterogeneity in treatment effects, as seen in the between-study SD. RoM has the best fit given model complexity with the lowest DIC. Scale-SMD and Scale-SMD-MR models have a similar fit to RoM, but higher DIC reflecting the higher effective number of parameters. These results on DIC reflect the greater shrinkage (reduced heterogeneity) seen with RoM (20.5%), with an intermediate degree of shrinkage with the Scale-SMD (18.2%), and



**Figure 2.** The relationship between the baseline pooled standard deviation and mean depression score at baseline. Results are shown by scale for HAMD-17, HAMD-21, HAMD-24 and MADRS scales. BDI-I could not be plotted because it was reported in a single study.

**Table 2.** *Model fit statistics and heterogeneity estimated within each of the five models.*

Model	Dbar	pD	DIC	Regression coefficient mean (95% CrI)	% shrinkage	Between-study SD Scale: linear predictor Median (95% CrI)	Between-study SD Scale: HAMD-17 Median (95% CrI)
Study SMD	625.3	538.2	1163.6	–	16.7	0.34 (0.27, 0.41)	1.35 (1.10, 1.65)
Study SMD (MR)	625.3	537.5	1162.8	–0.168 (–0.333, –0.002)	17.0	0.33 (0.27, 0.40)	1.33 (1.08, 1.62)
Scale SMD	630.6	531.7	1162.3	–	18.2	0.29 (0.23, 0.36)	1.18 (0.93, 1.45)
Scale SMD (MR)	631.3	531.7	1163.0	–0.121 (–0.272, 0.030)	18.3	0.29 (0.23, 0.36)	1.17 (0.92, 1.44)
Ratio of means	629.0	521.9	1150.9	–	20.5	0.09 (0.07, 0.12)	Not calculable

*Note:* Fit statistics: posterior mean residual deviance (relative to 611 data points), DIC, and percentage shrinkage. Coefficient of meta-regression (MR) on baseline severity. The posterior median of between-study SD with the 95% credible interval (CrI) on both the linear predictor scale and, for SMD models, on the HAMD-17 scale.

lower shrinkage (higher heterogeneity) with Study-SMD (16.7%). The degree of heterogeneity, as measured by the between-study SD, was 13% lower in the Scale-SMD than in the Study-SMD models (Table 2).

The regression coefficients in the meta-regression models were negative (Table 2), indicating a stronger treatment effect in trials whose study populations were more severely depressed at the outset. This effect was somewhat stronger with the Study-SMD outcome measure. The global fit and shrinkage statistics of the SMD meta-regression models were no different from their non-regression counterparts.

The effects of all treatments relative to pill placebo on the HAMD-17 scale appear in Table 3 and in Figure 3 in the form of a forest plot as median estimates with credible and predictive intervals. Predictive intervals depict where the treatment effect from a new study might lie and are generated from the between-study SD, therefore they reflect the heterogeneity within the modelled data. The predictive intervals were narrower in the Scale-SMD models than in the Study-SMD models.

Effects of active treatments relative to the best treatment on each scale appear in Supplementary Table S2.

### 3.3. Treatment recommendations

The treatment recommendations generated by the different models are presented in Table 4, using five different decision rules. The single best treatment (highest expected efficacy) was Amitriptyline for the RoM model and Scale-SMD models, and Mirtazapine for the Study-SMD models. The second decision rule selected all treatments with evidence of effect “significantly” better than pill placebo (that is where the 95%CrI did not cross zero). All active treatments qualified on all models.

If decision-makers were to recommend all treatments that were better than placebo by at least 1 HAMD-17 unit, the third decision rule, all five models pick out the same 11 treatments, and exclude trazodone, nortriptyline, and lofepramine.

The fourth approach selects from the above 11 treatments all those that are also not inferior to the best treatment. In this case, RoM picks out only seven treatments besides the best, Scale-SMD models a further eight, and Study-SMD models a further nine. Our fifth decision rule, based on GRADE recommendations, identifies a set of treatments that is inferior to no other treatment, again using a 95% credible limit. In this case, all five models picked out the same three treatments. Unlike the first three decision rules, which are driven by differences between active treatments and placebo, the fourth and fifth depend on differences between the active treatments. It seems that discrimination between active treatments is better in RoM models than in Scale-SMD, and that Scale-SMD models discriminate better than Study-SMD.

### 3.4. Sensitivity analyses

The results from the sensitivity analysis using a different prioritisation of data format are given in Supplementary Materials, Appendix 3. The findings were generally similar (Supplementary Materials, Appendix 3, Supplementary Table S3) to those from the main analysis (Table 2), although the benefits of the RoM model compared to the Scale-SMD models were minimal. All five models selected the same treatment as in the main analysis, and the results with the different decision rules differed only slightly from the main analysis (Supplementary Materials Appendix 3). Better discrimination between active treatments on the HAMD-17 scale was observed for the RoM model in both analyses.

A sensitivity analysis setting the before-after correlation at 0.5, as opposed to our base-case 0.3, had negligible effects on shrinkage. Shrinkage was lowered from 17% to 16% for Study-SMD, from 18% to 17% for Scale-SMD, and remained at 21% for RoM.

**Table 3.** The mean difference of treatments relative to pill placebo, presented as units on the HAMD-17 scale, with their 95% CrI.

SMD by study SD	SMD by study SD (MR on baseline severity)		SMD by scale SD		SMD by scale SD (MR on baseline severity)		RoM		
Mirtazapine	-3.63	Mirtazapine	-3.71	Amitriptyline	-3.30	Amitriptyline	-3.37	Amitriptyline	-3.83
	(-4.79, -2.46)		(-4.87, -2.56)		(-4.13, -2.49)		(-4.19, -2.56)		(-4.63, -3.00)
Amitriptyline	-3.52	Amitriptyline	-3.64	Mirtazapine	-3.29	Mirtazapine	-3.34	Mirtazapine	-3.64
	(-4.38, -2.67)		(-4.50, -2.79)		(-4.38, -2.23)		(-4.42, -2.27)		(-4.70, -2.51)
Clomipramine	-3.34	Clomipramine	-3.42	Clomipramine	-3.10	Clomipramine	-3.14	Lofepramine	-3.52
	(-5.17, -1.52)		(-5.25, -1.60)		(-5.01, -1.25)		(-4.98, -1.27)		(-5.86, -0.82)
Venlafaxine	-3.18	Lofepramine	-3.28	Venlafaxine	-2.96	Venlafaxine	-2.99	Venlafaxine	-3.30
	(-3.96, -2.41)		(-5.69, -0.87)		(-3.68, -2.26)		(-3.69, -2.30)		(-4.05, -2.53)
Lofepramine	-3.16	Venlafaxine	-3.23	Lofepramine	-2.86	Lofepramine	-2.95	Clomipramine	-3.17
	(-5.63, -0.73)		(-4.01, -2.46)		(-5.30, -0.38)		(-5.40, -0.49)		(-5.12, -1.03)
Paroxetine	-3.03	Paroxetine	-3.12	Paroxetine	-2.72	Paroxetine	-2.77	Paroxetine	-3.00
	(-3.89, -2.20)		(-3.96, -2.30)		(-3.46, -1.97)		(-3.52, -2.03)		(-3.82, -2.15)
Escitalopram	-2.93	Escitalopram	-2.96	Imipramine	-2.69	Duloxetine	-2.71	Duloxetine	-2.99
	(-3.62, -2.23)		(-3.64, -2.27)		(-3.71, -1.68)		(-3.42, -2.01)		(-3.71, -2.25)
Imipramine	-2.68	Duloxetine	-2.89	Escitalopram	-2.63	Imipramine	-2.69	Escitalopram	-2.83
	(-3.75, -1.60)		(-3.66, -2.11)		(-3.23, -2.02)		(-3.71, -1.67)		(-3.46, -2.17)
Duloxetine	-2.68	Imipramine	-2.69	Duloxetine	-2.58	Escitalopram	-2.65	Imipramine	-2.80
	(-3.44, -1.93)		(-3.74, -1.64)		(-3.25, -1.89)		(-3.26, -2.04)		(-3.90, -1.61)
Sertraline	-2.59	Sertraline	-2.68	Sertraline	-2.29	Sertraline	-2.35	Sertraline	-2.61
	(-3.36, -1.81)		(-3.46, -1.90)		(-2.97, -1.59)		(-3.03, -1.67)		(-3.34, -1.83)
Citalopram	-2.43	Citalopram	-2.51	Citalopram	-2.28	Citalopram	-2.31	Citalopram	-2.58
	(-3.47, -1.38)		(-3.55, -1.47)		(-3.21, -1.34)		(-3.25, -1.37)		(-3.64, -1.48)
Fluoxetine	-2.40	Fluoxetine	-2.48	Fluoxetine	-2.18	Fluoxetine	-2.23	Nortriptyline	-2.57
	(-3.04, -1.78)		(-3.11, -1.85)		(-2.75, -1.60)		(-2.80, -1.65)		(-4.22, -0.76)
No treatment	-2.31	Nortriptyline	-2.10	Nortriptyline	-2.03	Nortriptyline	-2.07	Fluoxetine	-2.31
	(-5.32, 0.72)		(-3.61, -0.61)		(-3.50, -0.59)		(-3.52, -0.64)		(-2.94, -1.67)
Nortriptyline	-2.09	Trazodone	-1.82	Trazodone	-1.75	Trazodone	-1.83	Trazodone	-1.70
	(-3.63, -0.56)		(-3.13, -0.52)		(-2.88, -0.62)		(-2.94, -0.71)		(-3.07, -0.21)
Trazodone	-1.72	No treatment	-1.61	No treatment	-1.23	No treatment	-0.73	No treatment	-1.03
	(-3.04, -0.40)		(-4.68, 1.46)		(-4.62, 2.11)		(-4.12, 2.72)		(-4.78, 3.62)

Note: Estimates are presented from models of depression score standardised as SMDs by study-level and scale-level pooled SD, with and without meta-regression (MR) on baseline severity, and a model of the ratio of means (RoM).

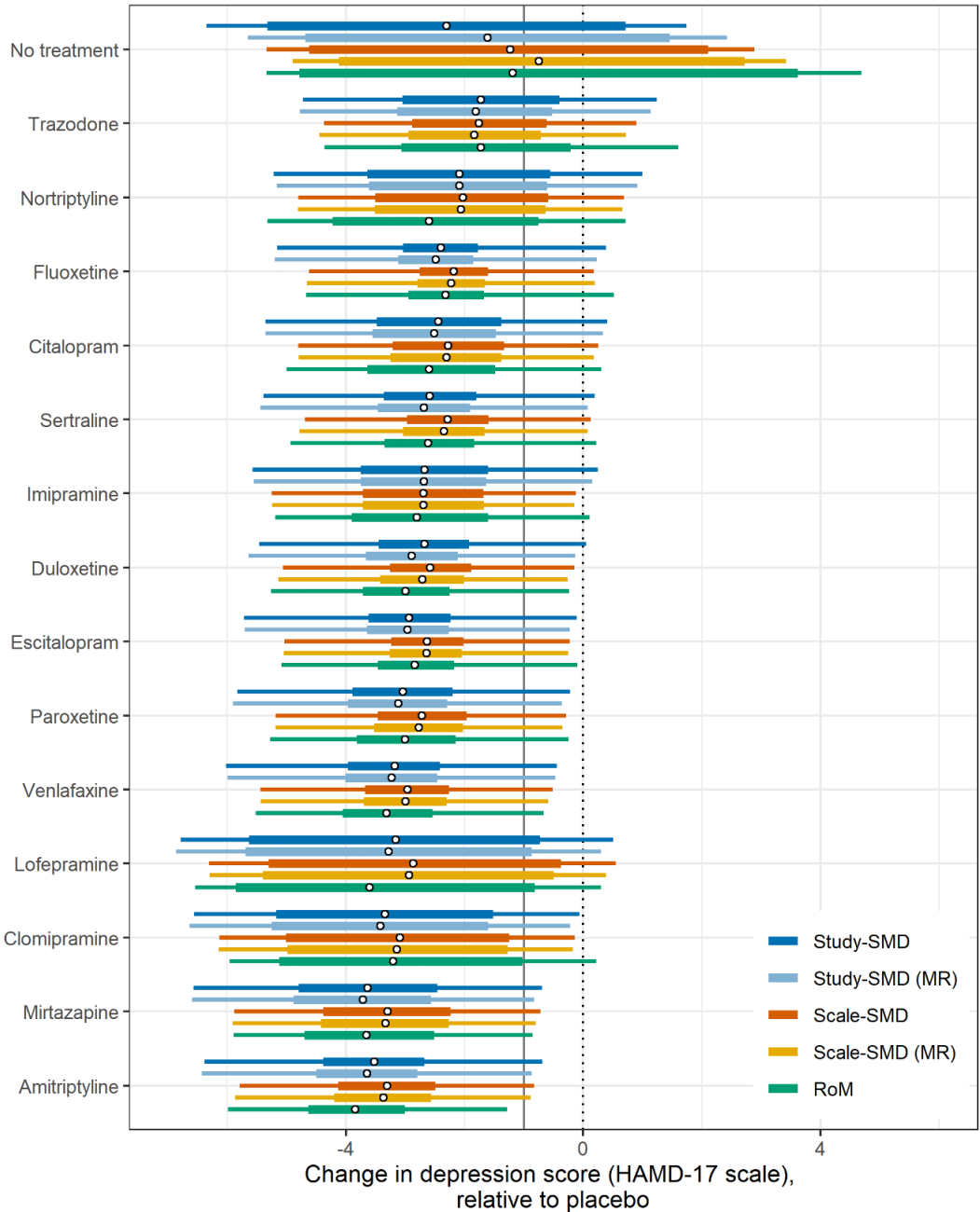
**Table 4.** Treatment recommendations based on each model, ranked by efficacy, according to five decision rules.

Decision rule	Study SMD	Study SMD Meta-regression	Scale-SMD	Scale-SMD Meta-regression	RoM
1. Single best treatment	Mirtazapine	Mirtazapine	Amitriptyline	Amitriptyline	Amitriptyline
2. Treatments better than placebo <sup>a</sup>	All active treatments	All active treatments	All active treatments	All active treatments	All active treatments
3. Treatments better than placebo by 1 HAMD-17 <sup>b</sup>	Mirtazapine	Mirtazapine	Amitriptyline	Amitriptyline	Amitriptyline
	Amitriptyline	Amitriptyline	Mirtazapine	Mirtazapine	Mirtazapine
	Clomipramine	Clomipramine	Clomipramine	Clomipramine	Venlafaxine
	Venlafaxine	Venlafaxine	Venlafaxine	Venlafaxine	Clomipramine
	Paroxetine	Paroxetine	Paroxetine	Paroxetine	Paroxetine
	Escitalopram	Escitalopram	Imipramine	Duloxetine	Duloxetine
	Imipramine	Duloxetine	Escitalopram	Imipramine	Escitalopram
	Duloxetine	Imipramine	Duloxetine	Escitalopram	Imipramine
	Sertraline	Sertraline	Sertraline	Sertraline	Sertraline
	Citalopram	Citalopram	Citalopram	Citalopram	Citalopram
4. Treatments better than placebo by 1 HAMD-17, <sup>b</sup> AND no worse than the best treatment <sup>c</sup>	Fluoxetine	Fluoxetine	Fluoxetine	Fluoxetine	Fluoxetine
	Amitriptyline	Amitriptyline	Mirtazapine	Mirtazapine	Mirtazapine
	Clomipramine	Clomipramine	Clomipramine	Clomipramine	Venlafaxine
	Venlafaxine	Venlafaxine	Venlafaxine	Venlafaxine	Clomipramine
	Paroxetine	Paroxetine	Imipramine	Imipramine	Duloxetine
	Escitalopram	Escitalopram	Paroxetine	Duloxetine	Paroxetine
	Imipramine	Duloxetine	Escitalopram	Paroxetine	Imipramine
	Duloxetine	Imipramine	Duloxetine	Escitalopram	Citalopram
	Sertraline	Sertraline	Citalopram	Citalopram	
	Citalopram	Citalopram			
5. GRADE, with a threshold zero. Treatments that are superior to placebo by 1 HAMD-17 <sup>b</sup> and inferior to no other treatment. (Number of treatments satisfying criterion (3) that they were superior to)	Amitriptyline (2)	Amitriptyline (3)	Amitriptyline (3)	Amitriptyline (3)	Amitriptyline (4)
	Mirtazapine (2)	Mirtazapine (2)	Mirtazapine (2)	Mirtazapine (2)	Mirtazapine (2)
	Venlafaxine (2)	Venlafaxine (1)	Venlafaxine (2)	Venlafaxine (2)	Venlafaxine (2)

<sup>a</sup> X better than Y means that the 95%CrI on the (X-Y) difference did not include zero.

<sup>b</sup> X better than Y by more than 1 HAMD-17 unit means that the 95% CrI on the (X-Y) difference did not include -1.0.

<sup>c</sup> X no worse than Y means that the 95%CrI on the (X-Y) difference did not include zero.



**Figure 3.** Treatment effect vs placebo as change in depression score on the HAMD-17 scale by model. In each case, circular points indicate the median estimate, thick bars indicate the 95% credible interval (CrI) and thin bars indicate the 95% prediction interval, for each treatment vs placebo. Treatments are ordered by median treatment effect under the RoM model. The vertical grey line indicates one unit on the HAMD-17 scale.

## 4. Discussion

### 4.1. Summary of findings

In this dataset, which was typical of pharmaceutical trials for more severe depression, the RoM model performed better than any SMD model. This assessment was based on lower DIC and greater shrinkage of estimates. Furthermore, examination of differences between active treatments' effects on the HAMDD-17 scale produced evidence that treatment effects were estimated with greater precision in the RoM models than in the Scale-SMD model. The Scale-SMD model gave a similar fit to RoM but had higher heterogeneity.

Scale-SMD models performed better than Study-SMD, giving lower heterogeneity and more precise estimates, by about 6%. A core assumption of the Study-SMD method is that all the studies that report on the same scale should have the same SD.<sup>15</sup> Hunter and Schmidt<sup>7</sup> regarded the changes in SMDs due to different study variances as “range variation” artefacts that required removing, and they advocated a form of Scale-SMD approach. In our dataset, range variation, which is generated by sampling variation as well as between-study population differences, was statistically and materially highly significant, with study SDs reported on the same scale varying over an average 4.0-fold range. The effect of the Study-SD method is therefore to multiply relative treatment effects by an arbitrary number between 1 and 4. Similar findings have been reported in trials reporting on the Liebowitz Social Anxiety Scale.<sup>29</sup>

Scale-SMD models represent an improvement on Study-SMDs because they avoid range variation by using a single normalising SD for each scale. To create a reference SD we used the average SD over the available studies using that scale. This is an imperfect solution, as it risks introducing random error when there are small numbers of studies. The ideal would be to use reference SDs from a single large population study. However, large population studies looking at the full range of scales are uncommon. Alternatively, it may be possible to derive reference SDs by modelling large collections of trials or observational studies.

### 4.2. Criteria employed in this study, and robustness of findings

The criteria used here to compare models were: global model fit; percentage shrinkage (a scale-independent measure of heterogeneity); DIC, which is model fit penalised for complexity; and precision of relative effects. These are in line with previous literature on the choice of scale in meta-analysis, which has focused on between-study heterogeneity measured by the Q-statistic,<sup>9,30,31</sup> or on both Q-statistics and precision of estimates.<sup>32,33</sup> Another measure that has been used is the percentage of study effects within the 95% confidence interval of the pooled effect.<sup>33</sup> A study comparing five different effect measures in an NMA of trials reporting binary data used global fit statistics and DIC.<sup>34</sup>

A limitation of reliance on aggregated data is that assumptions have to be made about correlations between baseline and follow-up scores. Our results assumed a correlation of 0.3, based on the evidence available; a sensitivity analysis with a correlation of 0.5 showed a greater advantage of RoM models. A second sensitivity analysis examined the priority order for data extraction. The base-case analysis prioritised (i) CFB, (ii) baseline and follow-up, and (iii) follow-up only. Results with an alternative ordering (ii, iii, and i) produced similar findings although there was no longer any difference between Scale-SMD and RoM methods. However, the better discrimination between treatments observed in the base-case analysis was also evident in this sensitivity analysis. It should be noted that this analysis was limited by the available data (which had been extracted following the priorities for SMDs).

### 4.3. Generality of findings

A limitation of the study is that it is restricted to a single dataset of trials, which includes only pharmaceutical interventions, for a particular severity range—more severe—of a specific disorder, depression. It is therefore unclear how far one can generalise from the present results to other PCROs consisting of correlated sums of sub-scale scores, let alone to other types of continuous outcomes.



Although studies comparing RoM with Study-SMD in large collections of pair-wise meta-analyses have reported more heterogeneity in SMD models,<sup>9</sup> this finding may be due to the poor performance of Study-SMD, and the comparison may yield different results if Scale-SMD were used instead. We would in any case urge caution before adopting any “one-size-fits-all” solution, and recommend that further studies are conducted, along the lines of the present study, using large networks of trials to examine commonly used PCROs. Separate investigations are required in each clinical area of psychiatric, neurologic, and other fields where PCROs are used.

#### **4.4. Minimally important difference**

Another approach, attracting increasing attention, is to standardise by dividing mean treatment effects in each trial by the MID, generating treatment effects in MID units. Some researchers regard MIDs as more easily interpretable than either SMDs or the units of the original scales.<sup>1,3,16</sup> Interestingly, in parallel to SMDs, both scale-specific<sup>1,16,35</sup> and study-specific<sup>36</sup> MIDs have been proposed, the latter for studies using scales for which no MID estimate is available. MIDs have also been expressed as percent improvement, indicating a proportional rather than additive effect, most notably in studies of depression.<sup>37,38</sup> The adoption of a MID approach does not in itself, therefore, contribute directly to the issues addressed in this paper: whether treatment effects are additive or multiplicative, and the choice of scale-specific or study-specific standardisation. Researchers using MID face several additional challenges: multiple ways of constructing and/or presenting the MID,<sup>39,40</sup> the extreme variability of estimates,<sup>41–43</sup> and between-patient variation.<sup>38,44</sup> Before choosing a solution to the problems posed by outcomes reported on multiple scales, we believe the first priority must be to determine whether treatment effects are multiplicative or additive.

#### **4.5. Impact of choice of scale on treatment recommendations**

Although there were differences between the five models in treatment rankings, including differences regarding which treatment was best, there was a high level of concordance. A decision rule similar to the one proposed by the GRADE Working Group<sup>27</sup> picked out the same three treatments under all five models. The relative insensitivity of treatment rankings to the choice of scale has been observed in previous NMAs of binomial data.<sup>34</sup> One should not conclude from this that the choice of scale does not matter, as it can have a substantial impact on the estimates themselves and their precision.<sup>34</sup> The formulation of decision rules for recommending treatments following an NMA is a topic calling for further research, particularly decision rules that take uncertainty into account.<sup>45</sup> Rules such as those proposed by GRADE,<sup>27</sup> because they can recommend treatments that are not “significantly” different from the best treatment, may have the unwanted effect of privileging treatments with more uncertain evidence.

#### **4.6. Recommendations regarding the choice of scale**

A decision about whether treatment effects are multiplicative or additive should be based on all the evidence available, including clinical opinion.<sup>30</sup> Table 5 summarises the kinds of evidence that would support the conclusion that effects are multiplicative. Depression scales score on the first four of the five criteria. Thus, although the empirical evidence on model fit and shrinkage in this paper is not strong, in combination with other evidence, we would regard the overall evidence for multiplicative effects as strong enough to favour RoM as the default.

For measurement scales where additivity can be assumed, we would recommend scale-specific standardisation by SMD or, if reliable estimates exist, MID. This is supported by the results reported here and is based on well-established arguments in textbooks and tutorial papers.<sup>6,15,17</sup> It is interesting to note that measurement scales in educational and psychological research, where meta-analytic methods including SMDs originated,<sup>2,14</sup> were generally constructed to be additive. This may explain the durability of additive models in meta-analysis.

**Table 5.** Criteria supporting the assumption of multiplicative treatment effects.

Criteria
1 Superior fit, greater shrinkage, and lower heterogeneity in models with a log link compared to an identity link
2 Increasing SD with increasing baseline severity (heteroscedasticity)
3 Scale constructed from the sum of correlated Poisson variables
4 Individual patient data shows that the treatment effect increases with baseline severity
5 Scores commonly reported and analysed following a log transformation

#### 4.7. Future research directions

The additive SMD and proportional RoM approaches assume, respectively, uniform linearity or uniform proportionality of treatment effects on the underlying depression scale. Rather than enquiring into the relation between the measurement scales and the underlying severity of depression, we should instead consider the relationships between the scales themselves. For this, we need to turn to methods for test equating and linking.<sup>46,47</sup> While proportionality appears to be closer to the truth for the depression scales studied here, based on the Table 5 criteria, test-linking studies show that the assumption of any simple uniform relation can only be an approximation. Studies using three distinct methodologies, factor analysis,<sup>48</sup> Item Response Theory,<sup>49–52</sup> and equi-percentile linking,<sup>53</sup> have all shown that each scale's ability to discriminate different degrees of depression varies unevenly across the severity spectrum and that each scale has a unique sensitivity profile. This is confirmed by recent work on the MID of depression scales, showing that while MID generally increases with baseline severity, the relationship is uneven and scale-dependent.<sup>38</sup>

Further research is required to find ways of leveraging the one-to-one mapping information generated by test-linking studies, to drive algorithms that map between aggregated results (mean and SD) reported on different scales.<sup>54</sup> Methods of this type would be considerably more flexible than either RoM or SMD, as they would allow synthesis even when scales are not monotonically related.

**Acknowledgements.** We thank Larry Hedges, Ian Shrier, and Alex Sutton for helpful discussions on this work.

**Author contributions.** B.C.D.: data curation [lead]; formal analysis [lead]; original draft [equal]; visualisation [lead]; software [lead]; N.J.W.: methodology [equal]; conceptualisation [equal]; original draft [equal]; supervision [equal]; software [equal]; visualisation [supporting]; funding [lead]; H.P.: data curation [supporting]; review of draft [equal]; I.M.: data curation [supporting]; review of draft [equal]; O.M.V.: data curation [supporting]; review of draft [equal]; A.E.A.: conceptualisation [lead]; methodology [equal]; formal analysis [supporting]; original draft [equal]; supervision [equal]; visualisation [equal].

**Competing interest.** The authors declare that no competing interests exist.

**Data availability statement.** The data and code files that support the findings of this study are available from OSF: <https://osf.io/2jyf7>. The WinBUGS code is also included in the Supplementary Materials (Appendix 2).

**Funding statement.** B.C.D., N.J.W., H.P., and T.A. were supported by the Guidelines Technical Support Unit at the University of Bristol, funded by the National Institute of Health and Care Excellence. I.M. and O.M.V. received support from NICE. The views expressed in this publication are those of the authors and not necessarily those of NICE.

**Supplementary material.** To view supplementary material for this article, please visit <http://doi.org/10.1017/rsm.2025.7>.

## References

- [1] Thorlund K, Walter SD, Johnston BC, Furukawa TA, Guyatt GH. Pooling health-related quality of life outcomes in meta-analysis—a tutorial and review of methods for enhancing interpretability. *Res Synth Methods*. 2011;2: 188–203.
- [2] Glass GV. Primary, secondary, and meta-analysis of research. *Educ Res*. 1976;5(10): 1–8.
- [3] Johnston BC, Thorlund K, Schunemann H, et al. Improving the interpretation of quality of life evidence in meta-analysis: the application of minimally important difference units. *BMC Health Quality Life Outcomes*. 2010;8(116).
- [4] Murad MH, Wang Z, Chu H, Lin L. When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation. *Br Med J*. 2019; 364.

- [5] Friedrich JO, Adhikari NKJ, Beyene J. The ratio of means method as an alternative to mean differences for analyzing continuous outcome variables in meta-analysis: a simulation study. *BMC Med Res Methodol.* 2008;8(32).
- [6] Guyatt GH, Thorlund K, Oxman AD, et al. GRADE guidelines: 13. preparing summary of findings tables and evidence profiles—continuous outcomes. *J Clin Epidemiol.* 2013;66(2): 173–183.
- [7] Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. *Introduction to Meta-Analysis.* Wiley; 2009.
- [8] Morton SC, Murad MH, O'Connor E, et al. *Quantitative Synthesis—An Update. Methods Guide for Comparative Effectiveness Reviews.* Agency for Healthcare Quality and Research; 2018.
- [9] Friedrich JO, Adhikari KJ, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *J Clin Epidemiol.* 2011; 64(5): 556–564.
- [10] Fournier JC, DeRubeis RJ, Hollon SD, Dimidjian S, Amsterdam JD, Shelton RC, Fawcett J. Antidepressant drug effects and depression severity: a patient-level meta-analysis. *J Am Med Assoc.* 2010; 303(1): 47–53.
- [11] Bower P, Kontopantelis E, Sutton A, et al. Influence of initial severity of depression on effectiveness of low intensity interventions: meta-analysis of individual patient data. *Br Med J.* 2013; 346.
- [12] Hieronymus F, Lisinski A, Nilsson S, Eriksson E. Influence of baseline severity on the effects of SSRIs in depression: an item-based, patient-level post-hoc analysis. *Lancet Psychiatry.* 2019;6: 745–752.
- [13] Khan A, Levanthal RM, Khan SR, Brown WA. Severity of depression and response to antidepressants and placebo: an analysis of the Food and Drug Administration database. *J Clin Pharmacol.* 2002;22: 40–45.
- [14] Hedges LV, Olkin I. *Statistical Methods for Meta-analysis.* Academic Press; 1985.
- [15] Higgins JPT, Li T, Deeks JJ. Choosing effect measures and computing estimates of effect. In: Higgins JPT, Thomas J, eds. *Cochrane Handbook for Systematic Reviews of Interventions.* 2nd ed. The Cochrane Collaboration; 2019.
- [16] Jaeschke R, Singer J, Guyatt G. Measurement of health status: Ascertaining the minimal clinically important difference. *Control Clin Trials.* 1989;10(4): 407–415.
- [17] Hunter JE, Schmidt FL. *Methods of Meta-analysis: Correcting Error and Bias in Research Findings.* 2nd ed. Sage Publications; 2004.
- [18] Daly C, Welton SJ, Dias S, Anwer S, Ades AE. *Meta-Analysis of Continuous Outcomes. Guideline Methodology Document 2: NICE Guidelines Technical Support Unit.* Decision Support Unit, University of Sheffield; 2021.
- [19] National Institute for Health and Care Excellence. *Depression in Adults: Treatment and Management. NICE Guideline [NG 222].* London. National Institute for Health and Care Excellence; 2022.
- [20] Snedecor GW, Cochran WG. *Statistical Methods.* 7th ed. Iowa State University Press; 1980.
- [21] Cohen J. *Statistical Power Analysis for the Behavioral Sciences.* Academic Press; 1969.
- [22] Lu G, Ades AE. Assessing evidence consistency in mixed treatment comparisons. *J Am Stat Assoc.* 2006;101: 447–459.
- [23] Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: A generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Med Decis Making.* 2013;33: 607–617.
- [24] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J Royal Stat Soc (B).* 2002;64(4): 583–616.
- [25] Tollefson GD, Holman SL. Analysis of the Hamilton Depression Rating Scale factors from a double-blind, placebo-controlled trial of fluoxetine in geriatric major depression. *Int Clin Psychopharmacol.* 1993;8(4): 253–260.
- [26] Tollefson GD, Bosomworth JC, Heiligenstein JH, Potvin JH, Holman S. A double-blind, placebo-controlled clinical trial of fluoxetine in geriatric patients with major depression. *Int Psychogeriatr.* 1995;7(1): 89–104.
- [27] Brignardello-Petersen R, Florez ID, Izcovich A, et al. GRADE approach to drawing conclusions from a network meta-analysis using a minimally contextualised framework. *BMJ.* 2020;371: m3900.
- [28] Spiegelhalter D, Thomas A, Best N, Lunn D. WinBUGS User Manual Version 1.4 January 2003. Upgraded to Version 1.4.32007; 2007. <https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/>. Accessed November 2022.
- [29] Ades AE, Lu G, Dias S, Mayo-Wilson E, Kounali D. Simultaneous synthesis of treatment effects and mapping to a common scale: an alternative to standardisation. *Res Synth Methods.* 2015;6: 96–107.
- [30] Deeks JJ. Issues on the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Stat Med.* 2002;21: 1575–1600.
- [31] Engels EA, Schmid CH, Terrin N, Olkin I, Lau J. Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Stat Med.* 2000;19: 1707–1728.
- [32] Friedrich JO, Adhikari KJ, Beyene J. Ratio of geometric means to analyze continuous outcomes in meta-analysis: comparison to mean differences and ratio of arithmetic means using empiric data and simulation. *Stat Med.* 2012;31(17): 1857–1886.
- [33] Takeshima N, Sozu T, Tajika A, Ogawa Y, Hayasaka Y, Furukawa TA. Which is more generalizable, powerful and interpretable in meta-analyses, mean difference or standardized mean difference? *BMC Med Res Methodol.* 2014;14(30).
- [34] Caldwell DM, Welton NJ, Dias S, Ades AE. Selecting the best scale for measuring treatment effect in a network meta-analysis: a case study in childhood nocturnal enuresis. *Res Synth Methods.* 2012;3: 126–141.
- [35] Schunemann HJ, Vist GE, Higgins JPT, et al. Interpreting results and drawing conclusions. In: Higgins JPT, Thomas J, eds. *Cochrane Handbook for Systematic Reviews of Interventions.* 2nd ed. The Cochrane Collaboration; 2019.
- [36] Johnston BC, Thorlund K, Da Costa BR, Furukawa TA, Guyatt GH. New methods can extend the use of minimal important difference units in meta-analyses of continuous outcome measures. *J Clin Epidemiol.* 2012;65: 817–826.

- [37] Button KS, Kounali D, Thomas L, et al. Minimal clinically important difference on the Beck Depression Inventory—II according to the patient's perspective. *Psychol Med*. 2015;45(15): 3269–3279.
- [38] Bauer-Staeb C, Kounali D-Z, Welton NJ, et al. Effective dose 50 method as the minimal clinically important difference: evidence from depression trials. *J Clin Epidemiol*. 2021;137: 200–208.
- [39] Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC. Understanding the minimum clinically important difference: a review of concepts and methods. *Spine J*. 2007;7: 541–546.
- [40] Johnston BC, Patrick DL, Thorlund K, et al. Patient-reported outcomes in meta-analysis—part 2: methods for improving interpretability for decision makers. *Health Qual Life Outcomes*. 2013;11: 211.
- [41] Chung AS, Copay AG, Olmscheid N, Campbell D, Walker JB, Chutkan N. Minimum clinically important difference: current trends in the spine literature. *Spine*. 2017;42(14): 1096–1105.
- [42] Hengartner MP, Plöderl M. Estimates of the minimal important difference to evaluate the clinical significance of antidepressants in the acute treatment of moderate-to-severe depression. *BMJ Evid-Based Med*. 2022;27: 69–73.
- [43] Kolin DA, Moverman MA, Pagani NR, et al. Substantial inconsistency and variability exists among minimum clinically important differences for shoulder arthroplasty outcomes: a systematic review. *Clin Orthop Relat Res*. 2022;470(7): 1371–1383.
- [44] Shrier I, Christensen R, Juhl C, Beyene J. Meta-analysis on continuous outcomes in minimal important difference units: an application with appropriate variance calculations. *J Clin Epidemiol*. 2016;80: 57–67.
- [45] Ades AE, Pedder H, Davies AL, et al. Treatment recommendations based on network meta-analysis: rules for risk-averse decision makers. *MedRxiv*. 2024; <https://doi.org/10.1101/2024.07.01.24309758>.
- [46] Dorans NJ, Pommerich M, Holland PW, eds. *Linking and Aligning Scores and Scales*. Springer; 2007.
- [47] Kolen MJ, Brennan RL. *Test Equating, Scaling and Linking: Methods and Practices*. Springer; 1994.
- [48] Uher R, Maier W, Hauser J, et al. Measuring depression: comparison and integration of three scales in the GENDEP study. *Psychol Med*. 2008;38: 289–300.
- [49] Rush AJ, Madhumar H, Trivedi H, et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), Clinician Rating (QIDS-C), and Self Report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. *Biol Psychiatry*. 2003;54: 573–583.
- [50] 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: 73–86.
- [51] Carmody TJ, Rush AJ, Bernstein I, et al. The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. *Eur Neuropsychopharmacol*. 2006;16: 601–611.
- [52] Olin TM, Yu L, Klein DN, et al. Measuring depression using item response theory: an examination of three measures of depressive symptomatology. *Int J Methods Psychiatr Res*. 2012;21(1): 76–85.
- [53] Leucht S, Fennema H, Engel RR, Kaspers-Janssen M, Szegedi A. Translating the HAM-D into the MADRS and vice versa with equipercntile linking. *J Affect Disord*. 2018;226: 326–331.
- [54] Davies AL, Ades AE, Higgins JPT. Mapping between measurement scales in meta-analysis, with application to measures of body mass index in children. *Res Synth Methods*. 2024;15(6): 1072–1093. <https://doi.org/10.1002/jrsm.1758>.

---

**Cite this article:** Downing BC, Welton NJ, Pedder H, Mavranzeouli I, Megnin-Viggars O, Ades A. Synthesis of depression outcomes reported on different scales: A comparison of methods for modelling mean differences. *Research Synthesis Methods*. 2025: 1–19. <https://doi.org/10.1017/rsm.2025.7>