When continuous outcomes are measured using different scales: guide for meta-analysis and interpretation

Mohammad Hassan Murad,1 Zhen Wang,1 Haitao Chu,2 Lifeng Lin3

It is common to measure continuous outcomes using different scales (eg, quality of life, severity of anxiety or depression), therefore these outcomes need to be standardized before pooling in a meta-analysis. Common methods of standardization include using the standardized mean difference, the odds ratio derived from continuous data, the minimally important difference, and the ratio of means. Other ways of making data more meaningful to end users include transforming standardized effects back to original scales and transforming odds ratios to absolute effects using an assumed baseline risk. For these methods to be valid, the scales or instruments being combined across studies need to have assessed the same or a similar construct.

Clinical scenario
A child and her parent present to the clinic to discuss anxiety symptoms that the child has had for over a year. The therapist talks with the parent and child about the possibility of starting a selective serotonin reuptake inhibitor (SSRI). A systematic review comparing SSRIs with placebo has shown that SSRIs reduce anxiety symptoms by a standardized mean difference (SMD) of −0.65 (95% confidence interval −1.10 to −0.21).1 2 The therapist finds these results difficult to interpret and not easy to explain to the parent and child.

SUMMARY POINTS
• When an outcome is measured using several scales (eg, quality of life or severity of anxiety or depression), it requires standardization to be pooled in a meta-analysis
• Common methods of standardization include using the standardized mean difference, converting continuous data to binary relative and absolute association measures, the minimally important difference, the ratio of means, and transforming standardized effects back to original scales
• The underlying assumption in all these methods is that the different scales measure the same construct

The problem
Outcomes of importance to patients such as quality of life and severity of anxiety or depression are often measured using different scales. These scales can have different signaling questions, units, or direction. For example, when comparing the effect of two cancer treatments on quality of life, trials can present their results using the short form health survey 36, the short form health survey 12, the European quality of life five dimensions, or others. Trials may also present their results as binary outcomes (proportion of patients who had improved quality of life in each trial arm). Decision makers need to know the best estimate of the impact of interventions on quality of life. The best estimate for decision makers is usually the pooled estimate (that is, from a meta-analysis), which has the highest precision (narrower confidence intervals).

Pooling outcomes across studies is challenging because they are measured using different scales. Pooling the results of each scale independently is undesirable because it does not allow all the available evidence to be included and can lead to imprecise estimates (only a few studies would be included in each analysis, leading to an overall small sample size and wide confidence intervals). As long as the different scales represent the same construct (eg, severity of anxiety), pooling outcomes across studies is needed.

In this guide, we describe several approaches for meta-analyzing outcomes measured using multiple scales. The methods used can be applied before the meta-analysis (to individual study estimates that are then meta-analyzed), after the meta-analysis and generation of the SMD, or they can be based on individual trial summary statistics and established minimally important differences (MIDs) for all instruments.3 We present a simplified approach focused on the general concepts of the SMD, the ratio of means (ROM), the MID, and conversion to relative and absolute binary measures.

For each approach, we describe the method used and the associated assumptions (fig 1). We apply these methods to a dataset of five randomized trials comparing SSRIs with placebo (table 1). These trials used different anxiety scales and one trial presented its results as a binary outcome. We use this dataset to show the common approaches described in this guide and how the clinical scenario was addressed by providing an interpretation (a narrative) to convey the results to end users such as clinicians and patients.

SMD
A common approach for combining outcomes from studies that used different scales is to standardize the outcomes (that is, express outcomes in multiples of standard deviations), which makes the outcomes
Standardized mean difference

- Cohen’s \( d \):
  \[
  d = \frac{\bar{X}_1 - \bar{X}_2}{s_p}
  \]

- Hedges’ \( g \):
  \[
  g = d \left( 1 - \frac{3}{4(n_1+n_2)-9} \right)
  \]

- SE \( g \):
  \[
  SE_g = \sqrt{\frac{1}{n_1} + \frac{1}{n_2} + \frac{d^2}{2(n_1+n_2-2)}}
  \]

Conversion of standardized mean difference to odds ratio

- The cumulative distribution functions of the logistic distributions of the two groups are:
  \[
  \Phi(c\mu_1 - c\mu_2) = \frac{1}{1 + e^{-c(\mu_1 - \mu_2)}}
  \]

- The variance of each logistic distribution is:
  \[
  \sigma^2 = \frac{n^2}{3}s^2 = \frac{\pi^2}{6}
  \]

- If we obtain counts in a 2 x 2 table based on a specific cut-off point (say \( c \)) and express the results as odds ratios from this 2 x 2 table, the properties of the logistic distribution imply that the log odds for the two groups are:

  \[
  \frac{\mu_1 - \mu_2}{3} \quad \text{and} \quad \frac{\mu_1 - \mu_2}{3}
  \]

Therefore, the log OR is their difference, that is,

\[
\frac{\mu_1 - \mu_2}{3} \quad \text{and} \quad \frac{\mu_1 - \mu_2}{3}
\]

Therefore, the log OR is their difference, that is,

\[
\frac{\mu_1 - \mu_2}{3} \quad \text{and} \quad \frac{\mu_1 - \mu_2}{3}
\]

Final equation: In OR = \( \frac{\pi}{\sqrt{3}} \times SMD \)

Expressing the difference in means using minimally important difference units

- Effect size in MID units = \( \frac{MD}{MID} \)
- SE \( \text{in MID} = \frac{SE_{MD}}{MID} \)

Expressing the results as ratio of means

- \( \text{ROM} = \frac{X_1}{X_0} \)
- \( SE_{\text{p.ROM}} = \sqrt{\frac{\Sigma_1^2}{n_1X_1^2} + \frac{\Sigma_0^2}{n_0X_0^2}} \)

Meta-analysis is done using a log scale

Calculating risk difference (also called absolute risk reduction) from odds ratio

Risk difference = \( \text{CER} - \frac{\text{OR} \times \text{CER}}{1 - \text{CER} + \text{OR} \times \text{CER}} \)

The upper and lower boundaries of the OR can be translated using the same equation to produce the upper and lower boundaries for the risk difference

- OR = odds ratio; \( d \) = standardized mean difference; \( \bar{X} \) = mean of each group; \( n \) = sample size in each group; \( s \) = standard deviation in each group; \( S_p \) = pooled standard deviation from the two groups; \( SE_g \) = standard error of Cohen’s \( d \); \( g = \) Hedges’ \( g \); \( f \) = bias correction factor for small sample size; \( SE_g \) = standard error of Hedges’ \( g \); MD = mean difference; MID = minimally important difference of a scale; ROM = ratio of means; \( CER = \) control event rate (that is, baseline risk); SMD = standardized mean difference

Assumptions of converting SMD to OR: the two groups being compared follow normal distributions with population means \( \mu_0 \) and \( \mu_1 \) and equal variance \( \sigma^2 \) for continuous outcomes.

The normal distribution can be approximated by the logistic distribution with the same means and the same variance

The equation of risk difference can be written in reverse (multiplied by \(-1\))

Fig 1 | Calculations for the different methods used to standardize outcomes measured using different scales

This approach consists of dividing the mean difference between the intervention and control in each study by the study’s pooled standard deviation of the two groups, which allows the studies to be combined and a pooled SMD to be generated. The SMD is justified based on one of the following two arguments. Firstly, outcome measures across studies may be interpreted as linear transformations of each other. Secondly, the SMD may be considered to be the difference between two distributions of distinct clusters of scores, even if these distributions did not measure exactly the same outcomes. Therefore, for the standard deviation to be used as a scaling factor, between-study variation in standard deviations is assumed to only reflect differences in measurement scales and not differences in the reliability of outcome measures or variability among study populations. However, this assumption cannot always be met. For example, when a meta-analysis includes pragmatic and explanatory trials, pragmatic trials are expected to have more variation in the study population and higher standard deviations.

Figure 1 shows two commonly used methods to derive the SMD: Cohen’s \( d \) and Hedges’ \( g \). Hedges’ \( g \) includes a correction for small sample size. Small sample size can lead to biased overestimation of the SMD. The SMD method can be complemented by three additional approaches.

Provide a judgment about size of effect

Meta-analysts can provide end users with the commonly used arbitrary cut-off points for the magnitude of a standardized effect. SMD cut-off points of 0.20, 0.50, and 0.80 can be considered to represent a small, moderate, and large effect, respectively.

Transform SMD to odds ratio

Continuous outcome measures such as the SMD can be converted to odds ratios. Although several approaches are available, the most commonly used method is to multiply the SMD by \( \pi/\sqrt{3} \) (about 1.81) to produce the natural logarithm of the odds ratio. This conversion from the SMD to the odds ratio can be performed by some statistical software packages. The main advantage of this approach is the ability to combine studies that present the outcome in a binary fashion (that is, number of responders) with studies that present the results on a continuous scale. Figure 1 presents the assumptions and an explanation for this approach.

Interpretation of this odds ratio is challenging. The Cochrane Handbook for Systematic Reviews of Interventions implies that the odds ratio refers to an improvement by some unspecified amount. Based on the characteristics of logistic distribution, which indicate that the calculated odds ratio is invariant to the cut-off point (fig 1), we propose that this odds ratio can be interpreted as follows: the ratio of the odds of patients with a measure higher than any specific cut-off point to those with a lower measure. Therefore, this odds ratio applies to any cut-off point of the continuous data. The cut-off point defining the magnitude of improvement on the various anxiety scales can be determined by practitioners to represent a meaningful change.

Back transform SMD to an original scale

SMDs can be made more clinically relevant by translating them back to scales with which clinicians are more familiar. This rescaling is done by simply multiplying the SMD generated from the meta-analysis
by the standard deviation of the specific scale. The results are then given in the natural units of a scale, which allows a more intuitive interpretation by end users. The standard deviation used here is the pooled standard deviation of baseline scores in one of the included trials (the largest or most representative) or the average value from several of the trials, or from a more representative observational study.

It is also possible to perform this rescaling on the results of each individual trial before conducting the meta-analysis; the meta-analysis can then be performed using the transformed values.

Box 1 shows the SMD based methods applied to the example of anxiety in children.

MID
The MID is defined as “the smallest difference in score in the outcome of interest that informed patients or informed proxies perceive as important, either beneficial or harmful, and which would lead the patient or clinician to consider a change in the management.” Meta-analysts might consider expressing the outcomes of each study using MID units and then pooling the results (which now have the same unit, the MID) in the meta-analysis. Figure 1 shows the formula for this expression.

One advantage of the MID approach is to reduce heterogeneity (often a lower I² value, which is the proportion of heterogeneity not attributable to chance). Such heterogeneity observed with the SMD approach would have been caused by variability in the standard deviation across studies. A second advantage is that a more intuitive interpretation can be made by clinicians and patients. This approach requires the availability of published MID values for the scales used in the various studies. MIDs are determined on the basis of prior research or expert consensus.

SMD
Method: the first four trials in table 1 provide the mean, standard deviation, and sample size for each study arm; however, the trials used two different scales. Data from each trial are standardized by dividing the difference in means by the pooled standard deviation (pooled from the intervention and control groups). The odds ratio from the fifth trial is 2.00 (95% confidence interval 1.44 to 2.78) by using the equation ln(odds ratio)=π/√3 (fig 1) and multiplying by –1 (because the odds ratio is for improvement whereas the scales measure anxiety symptoms, and a higher score suggests worsening of symptoms). This odds ratio is converted to an SMD of −0.38 (95% confidence interval −0.56 to −0.20). SMDs of all five trials were pooled in a random effects meta-analysis to give a final SMD of −0.97 (95% confidence interval −1.34 to −0.59).

Interpretation:
• Compared with no treatment, SSRIs reduce anxiety symptoms by 0.97 standard deviations of anxiety scales
• Compared with no treatment, the reduction in anxiety symptoms associated with SSRIs is consistent with a large effect

Odds ratio derived from SMD
Method: this pooled SMD of the five trials can also be expressed as an odds ratio using the equation ln(odds ratio)=π/√3 (fig 1); that is, an odds ratio of 5.75 (95% confidence interval 2.90 to 11.35) is obtained. Interpretation: the odds of improvement in anxiety symptoms after taking SSRIs are approximately six times higher compared with not taking SSRIs

Transformation to natural units
Method: the SMD can be transformed back to the natural units of the Pediatric Anxiety Rating Scale by multiplying it by the pooled standard deviation (pooled from the intervention and control groups in a trial that used this scale). This standard deviation can be obtained from the largest trial or as an average of the pooled standard deviations of the two trials, which here is 2.91. This multiplication gives a mean reduction of −2.81 (95% confidence interval −3.90 to −1.71).

Interpretation: compared with no treatment, SSRIs reduce anxiety symptoms by approximately three points on the Pediatric Anxiety Rating Scale.
Box 2 | Minimally important difference (MID) method applied to the example of selective serotonin reuptake inhibitors (SSRIs) for anxiety in children*

Method: assuming the smallest change a patient can feel on the Pediatric Anxiety Rating Scale and on the Screen for Child Anxiety Related Emotional Disorders is 5 and 10 points, respectively, the mean difference in each study is divided by the corresponding MID to obtain the difference between the two groups in MID units. The standard error of the difference in MID units is then calculated (fig 1). The differences in MID units from each study are meta-analyzed using the random effects model to give a difference of –0.98 (95% confidence interval –1.27 to –0.69).

Interpretation: compared with no treatment, the reduction in anxiety symptoms associated with SSRI use is 0.98 of the minimal amount of improvement that a patient can feel.

*All analyses use the DerSimonian-Laird random effect model (presuming that the assumptions of this model are met). For simplicity, the end-of-trial means in the two groups are compared (rather than comparing the change in means in the two groups).

Box 3 | Ratio of means (ROM) method applied to the example of selective serotonin reuptake inhibitors (SSRIs) for anxiety in children

Method: in each study, the mean of anxiety symptoms in the group that received SSRIs is divided by the mean in the placebo group, giving a ROM. The standard error of the ROM is then calculated (fig 1). The natural logarithms of ROMs from each study are meta-analyzed using the random effects model and then exponentiated to give a pooled ROM of 0.66 (95% confidence interval 0.61 to 0.70).

Interpretation: the average scores on anxiety symptom scales for patients who used SSRIs are 66% of the average symptom scores for patients who did not use SSRIs (thus better).

*All analyses use the DerSimonian-Laird random effect model (presuming that the assumptions of this model are met). For simplicity, the end-of-trial means in the two groups are compared (rather than comparing the change in means in the two groups).
treatment choices. Meta-analysis of these outcomes provides more precise estimates for decision making but is challenged when individual studies use multiple instruments with different scales and units. Several methods are available to deal with this issue and include using the SMD, back transformation of the SMD to natural units, converting the SMD to an odds ratio, using MID units, using the ROM, or converting continuous outcomes to absolute effects using a baseline risk appropriate for the target population.

Each of these methods has statistical or conceptual limitations. The SMD is often associated with heterogeneity because of variation in the standard deviations across trials and has also been reported to be biased towards the null.\textsuperscript{12} 22 23 The variance of the SMD, which impacts meta-analysis weights, is not independent of the magnitude of the SMD, as can be seen in the equation shown in figure 1. Larger SMDs tend to have larger variances and thus lower weights in inverse variance weighted meta-analysis, which may be another limitation.\textsuperscript{12} 22 23

By using the MID, some of these statistical challenges may be reduced; however, the MID is not always known for many scales. When the SMD is converted to an odds ratio, empirical evaluation shows that at least four of five available methods have performed well and were consistent with each other (intraclass correlation coefficients were ≥0.90).\textsuperscript{17} Nevertheless, the assumptions of these methods vary and may not always be met. When the effect size is extreme, the conversion to an odds ratio may be poor. Some conversion methods can be considered exact methods using the normal distribution, which makes the resultant odds ratio dependent on the cut-off point\textsuperscript{17} and further complicates intuitive interpretation.

When results are converted to natural units (to the scale most familiar to end users), linear transformation may not be valid when the instruments have different measurement scales.\textsuperscript{3} Although the ROM has reasonable statistical properties,\textsuperscript{22} its assumptions are not always met (such as having outcome measures with natural units and natural zero values).\textsuperscript{4} The ROM cannot be used with change data, which can have a negative value.\textsuperscript{7} The ROM is also criticized for having a multiplicative nature, which is appealing for clinicians and patients because treatments are often discussed in these terms, but this interpretation may not be appropriate.\textsuperscript{24} The limitation of having different ROMs calculated in studies with similar absolute change is shared with other relative association measures (such as odds ratio and relative risk). Although methods for establishing certainty in baseline risk have been proposed, they have not been widely used.\textsuperscript{25}

Not all of the described methods have been implemented in the commonly used meta-analysis software packages and may require statistical coding. It is important to reiterate that for any of these methods to be valid, the scales or instruments being combined across studies need to have assessed the same or a similar construct.

Contributors: All authors contributed to the design of the manuscript and interpretation of the data. MHM developed the first draft and ZW, HC, and LL critically revised the manuscript and approved the final version. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. MHM is the guarantor.

Funding: This research did not receive any specific grant from any funding agency in the public, commercial, or not-for-profit sector. HC is supported in part by the National Library of Medicine (R21 LM012197, R21 LM012744), and the National Institute of Diabetes and Digestive and Kidney Diseases (U01 DK106786). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Competing interests: All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: no support from any organisation for the submitted work; no financial relationships with any organisations that might have an interested in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Provenance and peer review: Not commissioned; externally peer reviewed.

This is an Open Access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

Box 4 | Absolute effect generation

Method: the odds ratio derived from the standardized mean difference in previous steps was used as the relative effect of selective serotonin reuptake inhibitors (SSRIs) (other ways to derive the odds ratio can also be used). A baseline risk (here, the likelihood of symptom improvement without SSRIs) is obtained from the placebo arm of the fifth trial: 100/300 = 0.33 (this baseline risk can also be derived from a better source that shows the clinical course of anxiety in children without treatment). By using the odds ratio and the baseline risk (fig 1), the resultant risk difference is 0.41.

Interpretation: in 100 patients with anxiety who do not receive treatment, 33 will improve. However, when 100 patients with anxiety receive SSRIs, 74 will improve (difference of 41 attributable to treatment with SSRIs).


