NICE Guidelines Technical Support Unit

Meta-Analysis of Continuous Outcomes

Guideline Methodology Document 2

Version 1 (January 2021)

Caitlin Daly[1], Nicky J Welton[1], Sofia Dias[2], Sumayya Anwer[2], AE Ades[1]

[1] Population Health Sciences, Bristol Medical School, University of Bristol

[2] Centre for Reviews and Dissemination and Centre for Health Economics, University of York

About the NICE Guidelines Technical Support Unit

The NICE Guidelines Technical Support Unit (TSU) is a collaboration between the Universities of Bristol, Sheffield, York and Leicester. The TSU is commissioned by the Centre for Guidelines at the National Institute for Health and Clinical Excellence (NICE) to provide rapid-response technical support, methodology training, and methods research, in the context of guideline development. Please see this website for further information http://www.bristol.ac.uk/population-health-sciences/centres/cresyda/mpes/nice/

About the Guideline Methodology Document series

This series of Guideline Methodology Documents (GMDs) complements the Guide to the Methods of Technology Appraisal (1), the Guidelines Manual (2), and the NICE Decision Support Unit (DSU) Technical Support Documents (TSDs) (3-9).

The aim of the GMDs is to assist all those involved in guideline development, including guideline developers, guideline committee members, those commenting on draft guidelines during the consultation period, manufacturers, and stakeholders.

There is, of course, already a wealth of tutorial material on how to conduct systematic review and meta-analysis (10-12). The GMDs are in agreement with virtually all this material, although there are some significant differences in the way that meta-analytic methods are used.

The GMDs take the particular perspective of the guideline developer. They therefore go beyond standard treatments in which systematic review and meta-analysis tend to be seen as methods for producing "pooled" analyses that "summarise the literature". The decision context requires a focus on patients at specific points in their disease progression, methods that have particular properties regarding coherence and complete use of evidence, and procedures that are compatible with decision making under conditions of uncertainty.

The GMDs are aimed at a basic and introductory level: more advanced topics are indicated with an asterisk (*), and readers are referred elsewhere.

There are several areas of methodological uncertainty, controversy or rapid change. These are indicated in the GMDs. GMDs are extensively peer reviewed prior to publication (see acknowledgements). However, the responsibility for each GMD lies with the authors, who welcome any constructive feedback on the content, suggestions for updates and further guides. Readers should be aware that while the TSU is funded by NICE, these documents do not constitute formal NICE guidance or policy.

Acknowledgements

The TSU thanks the NICE Centre for Guidelines Methods and Economic Team, their NMA Working Group and Guidelines Methodology Group, for their substantial contribution to this document. The joint editors for the GMD series are Nicky Welton (University of Bristol) and Sofia Dias (University of York). The production of this document was funded by the National Institute for Health and Clinical Excellence (NICE) through the NICE Guidelines Technical Support Unit. We are especially grateful to the external reviewers: Julian Higgins (University of Bristol), Alex Sutton (University of Leicester), Tom Trikalinos (Brown University). The views, and any errors or omissions, expressed in the Guideline Methodology Documents are those of the authors only. NICE and NICE Guideline Developers may take account of any part of this document, but they are not bound to do so.

Contents

1. INTRODUCTION

1.1. PURPOSE AND SCOPE OF THIS DOCUMENT

The purpose of this document is to make recommendations about methods for meta-analyses of trials that report continuous outcomes. This guidance should be read alongside GMD1 Meta-analysis.

Among the topics covered will be:

1. Different forms of analysis depending on whether the treatment is believed to have an additive effect, or a proportional or multiplicative effect.
2. How to carry out meta-analysis when trials report outcomes on different scales measuring similar constructs
3. How to include data on more than one outcome from each trial

There is an abundance of excellent tutorial material on systematic review and meta-analysis of continuous data from RCTs (10, 11, 13, 14). As far as basic methodology is concerned, our recommendations do not depart greatly from previous work. However, there are some major differences in how the basic methodology is used, and these are documented and explained.

1.2. OUTLINE OF THIS DOCUMENT

We begin with a summary of the recommendations for easy reference (Section 2). Section 3 gives a brief overview of how a meta-analytic model is developed in the context of guideline development and describes the key decisions that guideline developers must make, before data extraction can begin. These will be based in part on a scoping review.

Section 4 provides details on the preferred summary statistics that should be extracted from each trial, and alerts readers to the need for data conversions that are needed if the data are not reported in the form required for the meta-analytic model.

Section 6 provides some worked examples to illustrate recommended procedures for standardisation.

The guidance on what should be reported in Guidelines (Section 7) is specific to continuous outcomes: more general guidance is available in GMD1. We close with a brief section on research recommendations (Section 8).

An accompanying *Data Conversion Workbook (GMD-2 Data Conversion Workbook.xlsx)* is available to assist in calculating the correct summary statistics for data input into Review Manager (RevMan) version 5.3, and the *metafor* package version 2.4-0 in R (version 3.6.3). The use of these packages is illustrated in a Software Appendix. Mathematical notation is avoided throughout, although the formulae used in the *GMD2 Data Conversion Workbook* are set out in Appendix C for reference.

2. SUMMARY OF RECOMMENDATIONS

Recommendation 1. Use multiplicative models (additive on the log scale) for outcomes that are commonly analysed after log transformation (Section 5.1)

Recommendation 2. For outcomes commonly reported on the natural scale, but where there is convincing evidence that treatment acts in a proportional way, an analysis using RoM is preferred (Section 5.1)

Recommendation 3. In a meta-analysis of mean differences (MDs), the order of preference for the estimates to be extracted is:

(i) MD from analysis of covariance (ANCOVA) with baseline score as a covariate
(ii) MD based on means of change-from-baseline scores

(iii) MD based on means of post-treatment scores

Combinations of each type may be pooled together in a meta-analysis. When data permit, convert post-treatment MD to change-from-baseline MD (5.2)

Recommendation 4. If meta-analyses include trials with different outcomes of the same construct and an additive model is assumed, they should be mapped to a common scale, using one of the following methods, in this order or preference:

(i) Division by an SD from an external reference population (Section 5.3.1)
(ii) Division by an SD derived internally from the trials included in the meta-analysis**,** taking care to adjust for differences in the direction of effect (Section 5.3.2)

SMD meta-analysis (standardisation by dividing the MD in each trial by the same trial's sample SD) should not be used (Section 5.3.3)

Recommendation 5. If trials report more than one outcome on the same construct, a within-trial synthesis should be conducted taking account of the correlations between outcomes (Section 5.4)

3. PRELIMINARY STEPS AND OVERVIEW

3.1. STRATEGIC DECISIONS BASED ON A SCOPING REVIEW

Before data extraction is begun in earnest, Guideline Developers need to make a series of strategic decisions, based partly on a scoping review of the literature. In the specific case of continuous outcomes, a series of decisions must be made before full data extraction and meta-analysis can begin. These relate to:

- Choice of an additive or multiplicative model of the treatment effect (Sections 3.2, 5.1)
- Continuous outcomes or proportions (Section 3.2.4)
- Contrast-based vs Arm-based data (Section 3.6)
- How to conduct meta-analysis of different outcomes measuring similar constructs (Section 5.3)
- Within-trial synthesis of multiple outcomes measuring similar constructs (Section 5.4)

3.2. DOES THE TREATMENT WORK IN AN ADDITIVE OR MULTIPLICATIVE WAY?

This is a fundamental decision, representing alternative models of the data. Different models can easily result in different recommendations based on the same set of data. Four kinds of model are considered, three of which are for multiplicative or proportional effects:

- Standard *Additive* model – additive treatment effect
- Multiplicative or proportional treatment effects
    a. Additive models for *log-transformed data* – multiplicative treatment effect
    b. *Ratio of Means* (RoM) – proportional treatment effect
    c. *Percent change from baseline* – proportional treatment effect

The additive effect model is generally regarded as the standard approach.

3.2.1.Multiplicative effects: log-transformed data

Multiplicative treatment effects are most often seen in trials reporting laboratory outcomes. Usually, the observations are skewed to the left, and are log transformed in primary data analysis. Results may be reported as geometric means. The strategy in these cases should be to conduct the meta-analysis on the log-transformed scale. On this scale, the treatment effect is additive. The pooled results should be converted back onto to original scale and reported as Geometric Means with 95% CIs.

3.2.2.Ratio of means

A second kind of "multiplicative" model assumes that the treatment effect acts in a *proportional* way, but the log transformation is applied not to the observations themselves, but to their means. If observations tend to be skewed but are analysed on the natural scale, for example outcomes measured in hours, days or weeks, RoM analysis should be preferred. Patient Reported or Clinician Reported Outcomes based on multi-item questionnaires should also probably be analysed as RoM. Criteria for deciding whether an additive or RoM analysis is most appropriate are presented in Section 5.1. RoM is also a method for combining data from trials that report similar outcomes measured on different scales (Section 3.4).

3.2.3.Percent change from baseline

There are some specific outcomes, such as HbA1c in diabetes studies, that tend to be reported as mean percent change from baseline. This also reflects a belief that treatment acts in a proportional way.

3.2.4.Continuous outcomes or proportions

Within the same meta-analysis, some trials may report results as means and standard deviations, while others may report the proportion of "responders" – for example the proportion who improve by more than 50% from baseline, or whose score exceeds a specified threshold. Then a decision must be made as to whether the meta-analysis will be on the continuous or on the binomial outcome. The choice can be made on several grounds, for example: which is the most commonly reported; or which is required for use in a cost-effectiveness analysis. Like RoM, proportion responding is a convenient way of pooling across trials that report similar outcomes on different scales.

3.3. CHOICE OF MODEL AND STRATEGY FOR EXTRACTION, DATA CONVERSION AND META-ANALYSIS

Guideline developers need to make a decision as to which model of the data should be adopted. The decision should be based, as much as possible, on what seems to be the most appropriate model for the data: different models may lead to different recommendations, and will have a major impact on the guideline development process (See Section 4). But, at the same time, the decision can be tempered by the way in which trials tend to report data, and on the feasibility of converting the summary statistics reported in trials into a format suitable for input into the chosen model.

Table 4.1 shows which sets of summary statistics can be converted for use in each model, and which conversions are available in the *GMD2 Data Conversion Workbook.* These data conversions make it possible to incorporate trials reported in several different ways into the same meta-analysis. Among the salient points to note:

- It is feasible to convert summary statistics calculated on the natural scale into a form suitable for meta-analysis of log-transformed data.
- Summary statistics calculated on the natural scale, or on the log scale, can be used in an RoM meta-analysis only if arm-based summaries are available.
- Percent change from baseline cannot easily be combined with any other form of summary. Trials reported this way must be analysed separately.
- Binomial outcomes can be converted to a continuous form.

3.4. DIFFERENT OUTCOMES MEASURING SIMILAR CONSTRUCTS

For continuous outcomes it is common for different studies to report different outcome measures that attempt to capture the same underlying construct. Well known examples are the various scales used in trials of treatments for depression, such as the Beck Depression Inventory (BDI), and the

Hamilton Depression Scale (HamD); or, for social anxiety, the Liebovitz Social Anxiety Scale (LSAS), the Social Phobia Inventory (SPIN), and so on. To carry out a meta-analysis it will be necessary to either convert them to a common scale, or to carry out an RoM analysis. The scoping review will establish which scales are used in the literature and a decision will need to be made on which ones are considered sufficiently reliable and sufficiently similar to synthesise together. While guideline developers should aim to be as inclusive as possible, in practice the decision will be influenced by: how frequently each outcome is reported, and the availability of external data to inform the standardisation (see Section 5.3.1).

3.5. MULTIPLE OUTCOMES WITHIN THE SAME TRIAL

The scoping review will also establish how many trials report the outcome on more than one scale. The recommended approach in these cases is to carry out a "within-trial" synthesis of these outcomes before proceeding with the meta-analysis (see Section 5.4).

3.6. ARM-BASED AND CONTRAST-BASED DATA

Data can be extracted and analysed in two forms, either *Arm-Based* (arm means, standard deviations (SDs) and sample sizes, or *Contrast-Based* (mean difference and its standard error (SE). Analyses of *Arm-based and Contrast-based forms will give identical results, so long as the same adjustments (if any) are applied to both forms of data.*

Examples of how each kind of data are entered into software can be found in the Software Appendix for RevMan (*Examples 1 and 2*) and the *metafor* package (version 2.4-0) in R (*Examples 3 and 4*).

None of these software options allow one to combine Arm- and Contrast-Based data in the same meta-analysis. Therefore, in every meta-analysis where there is mixed data, it is necessary to convert all the arm-based data into contrast-based data. These conversions can all be carried out in the *Data Conversion Workbook.* There is also facility for this in RevMan (*Example 5* in the Software Appendix). R scripts can also be written to facilitate this process.

[*] Different forms of data can be combined if a Bayesian simulation approach is taken, for example in WinBUGS (see TSD2 Example 8 under the heading "Shared Parameter Models" (4)). Several of the data conversions not available in the *GMD2 Data Conversion Workbook* can be carried out by modelling in a Bayesian framework.

4. DATA EXTRACTION AND PROCESSING

Once the model has been chosen - additive, RoM, log-transformed, or percent CFB – data extraction and processing can begin. The order of preference regarding which summary statistics should be extracted from each study is shown in Table 4.1, which is essentially the same regardless of the model.

It is relatively unusual for every trial to report results in exactly the same way. In most cases there is a standard formula which "converts" the extracted summary statistics into a form that is appropriate for the treatment effect model that has been chosen. These formulae, which are implemented in the *GMD2 Data Conversion Workbook* are set out in Appendix C for reference.

[*] Advanced material that may be skipped.

Data Extraction Procedure

1. *Order of preference*. Table 4.1 shows several sets of summary statistics that might be reported. Extract the set that is highest on the list, and enter them into a spreadsheet, using the *Data Conversion Workbook* as necessary, before entering them into the meta-analysis package.

2. *Change from baseline*. Change from baseline (listed 3rd-4th in Table 4.1) is preferred to follow-up (listed 5th-6th) *even when the SE or SD of CFB is not reported* (Recommendation 3, Section 5.2)

3. *SD and SE*. Distinguish between SD (a measure of variation), and the SE of the mean (a measure of uncertainty). The SE is sometimes referred to as the "SD of the mean", and papers may report a mean "+/- SD": this is usually an SE. The two are related: for a single sample $SD = SE\sqrt{n}$, where n is the sample size. When converting between the two, it is important to ensure that the sample size used is the correct one used for either measurement's calculation.

4. If the relevant mean, SD, SE_D or SD_{pooled} is not available, it may be possible to infer it from other statistics such as the median, IQR, confidence intervals, z-statistic, t-statistic, or p-value. See Appendix A, otherwise contact the authors.

5. If standardisation by internal SDs is intended, extract the pooled baseline SDs from every trial, or extract the arm SDs and calculate it from that (Section 6.2).

6. If the paper reports none of the statistics in Table 4.1, contact the authors.

Table 4.1. Order of preference which summary statistics to extract, and feasibility of conversion to form required for each meta-analytic model.

		Available synthesis models			
			Multiplicative		
	Data to be extracted, in order of preference	**Additive**	**RoM**	**Logged data**	**% change**
	Summary statistics on natural scale				
1	MD, SE(MD) from ANCOVA	Y	-	Y[e]	-
2	Change-from-baseline arm means, CFB SDs, N	Y	-	-	-
3	Arm Means, SDs at baseline and follow-up, N, correlation [b]	Y	Y	Y	-
4	MD, SE(MD) change from baseline	Y	-	Y[e]	-
5	Arm means, SD, N at follow-up	Y	Y	Y	-
6	MD, SE(MD) at follow-up	Y	-	Y[e]	-
7	Geometric arm means, confidence intervals, N	[a]	Y	Y	-
8	Ratio of geometric means, confidence intervals	[a]	-	Y	-
9	Mean percent change in each arm, SD of % change, and N	-	-	-	Y
10	MD of % change, and its SE	-	-	-	Y
	Summary statistics on log scale				
11	MD, SE(MD) from ANCOVA	[a]	-	Y	-
12	Arm means, SD, at baseline and follow-up, N [b]	[a]	Y	Y	-
13	MD, SE(MD) change from baseline	[a]	-	Y	-
14	Arm means, SD, N at follow-up	[a]	Y	Y	-
15	MD, SE(MD) at follow-up	[a]	-	Y	-
	Summary statistics as probabilities				-
16	Pr(Response), N in each arm. Also extract SD [c]	Y	Y	Y	-
17	Log odds ratio, and its SE. Also extract SD [c]	Y	-	Y	-
	Less common outcomes				
18	SMD, and SE. Also extract SD [d] and samples sizes	Y	-	Y	-
20	% change based on mean scores, SE		Y		-
21	Any statistic listed in Point 4 of the Data Extraction Procedure above.				

Abbreviations: ANCOVA – analysis of covariance; CfB – Change from baseline; MD – mean difference; N – sample size; RoM – ratio of means; SD – standard deviation; SE – standard error; SMD – standardised mean difference.

[a] These conversions are possible but very unlikely to be useful as their presentation suggests a multiplicative model should be fitted. They are not currently available in the *GMD2 Data Conversion Workbook*

[b] Extract the correlation between baseline and follow-up scores if available

[c] The SD required here is the pooled SD at Follow-up in that trial

[d] The SD required here is the SD that was used to produce the SMD in that trial

[e] To transform MDs on the raw scale to MDs on the log-transformed scale, also extract the overall mean response averaged over both arms

5. ISSUES IN META-ANALYSIS OF CONTINUOUS OUTCOMES

This section looks at some particular topics in meta-analysis of continuous outcomes, focussing on those where clarification of existing methods is required, or where our proposals differ from currently accepted approaches. The issues we review are

- Multiplicative vs additive treatment effects (5.1)
- Preference for change-from-baseline data (5.2)
- Standardisation and RoM analysis (5.3)
- Within-trial synthesis of multiple outcomes (5.4)

5.1. MODELLING ADDITIVE VS. MULTIPLICATIVE EFFECTS

The effect of a treatment is considered to be additive when it adds to or subtracts from a value on a scale. Difference measures are, then, the natural way to quantify the additive effect on an outcome.

Laboratory outcomes are often skewed to the right, and statistical analysis of trial data may be carried out on log-transformed observations. In these trials it is implicitly assumed that the treatment effect is multiplicative: its effect is to multiply or divide a value on a scale. In these cases meta-analysis is carried out on the log scale. Some complications arise when trials report log-transformed data analysis on the natural scale, but in most cases standard conversions are available to convert summary statistics on the raw scale into summary statistics on the log-scale, or vice versa (15). These conversions can be carried out in the *Data Conversion Workbook*. Meta-analysis should be performed on the log-transformed data, and converted to that form if necessary. The pooled results should be converted back to the natural scale, in exactly the same way that Odds Ratios meta-analysis is carried out on log-transformed data, and the pooled result converted back to ORs.

Recommendation 1: Use multiplicative models (additive on the log scale) for outcomes that are commonly analysed after log transformation

A different analysis of "proportional treatment effects" is possible using Ratio of Means (RoM) meta-analysis. This is applied to summary statistics on the original raw scale, where the original scores are Patient, or Clinical Reported Outcomes (PROs, CROs) based on the sum of multiple ratings. Examples are the BDI scales for depression, the PANSS score for Schizophrenia, or the Liebowitz Social Anxiety Scale (LSAS).

It is emphasised that additive and RoM models represent different models of how treatments work, and they can lead to different treatment recommendations for the same data.

Literature addressing how to choose between additive and RoM methods is inconclusive. Simulation studies (16) suggest that RoM and MD have comparable performance in most scenarios, although this assessment was limited to simulated data in which effects were always additive. A study in which results of RoM and MD meta-analysis were compared on a large number of reviews from the Cochrane database (17) revealed similar treatment effects and assessments of heterogeneity. However, the relative merits of additive and multiplicative effect models should ideally be considered separately for each condition. Ideally the choice "should be determined by the biological effect of the treatment" (17). This is an area requiring further research.

In considering whether PROs and CROs should be analysed using additive or RoM methods, Guideline Developers should take account of the following "signs" that treatment effects are likely to be proportional:

- If there is a positive correlation between the groups mean and SD: this might be established by a plot of the arm SDs against the arm means, over all the trials.

- Similarly, if RoM shows less heterogeneity than mean difference, when plotted against baseline mean.
- If trials tend to report proportion of patients responding, where "response" is defined as a % change from baseline.
- If trials tend to report mean % change from baseline as a continuous outcome.
- If there is evidence that the size of the treatment effect is greater (or less) in more severe disease: i.e. a positive (negative) interaction between treatment effect and baseline severity.
- Most PRO and CRO total scores are sums of scores on correlated items. These would be expected to be log distributed on *a priori* grounds, suggesting RoM is more appropriate.

A further consideration in favour of RoM analysis is that it provides a solution to the issue of different trials reporting outcomes on different scales that measure essentially the same construct. Although RoM offers a convenient and very general solution to this problem, which avoids "standardisation" (see Section 5.3), RoM modelling should only be done if the assumption of proportional treatment effects is reasonable.

Against adoption of an RoM model, it should be born in mind that only trials that report arm-based summary data can be included in an RoM analysis (Table 4.1).

Even when there is every indication that a treatment effect is proportional, and that RoM analysis would be preferred, Guideline Developers may prefer to carry out separate analyses on an additive scale, looking at "Mild", "Moderate" and "Severe" disease separately. Another option, consistent with a belief that treatment effects are proportional, is to carry out a meta-analysis on a proportion responding to treatment outcome.

Recommendation 2. For outcomes commonly reported on the natural scale, but where there is convincing evidence that treatment acts in a proportional way, a RoM analysis is preferred

5.2. ORDER OF PREFERENCE: ANCOVA, CHANGE FROM BASELINE, OR POST-TREATMENT SCORE

In most trials, patients are enrolled, allocated to a treatment at random, treated, then followed up, at which point a *post-treatment* observation is made. In some cases, a *baseline* observation is also recorded: this may be before randomisation if it is used to determine eligibility, or after. If a baseline observation is available, there is a choice of outcome measure: either the mean post-treatment observation in each treatment arm, or the mean of the difference between the post-treatment score and the baseline, in each arm: the *change from baseline* (CFB).

5.2.1.Basic order of preference

The preferred method for analysing change-from-baseline data is by analysis of covariance (ANCOVA) (18) in which the patients' post-treatment scores are regressed against their pre-treatment scores, and possibly against other covariates. This both corrects for any baseline imbalance in the covariates included, and results in more precise estimates of the treatment effect, by removing variation due to covariates.

A simple change-from-baseline analysis may adjust or partially adjust for imbalance in baseline outcome scores (18). Baseline imbalance in the outcome score can come about in two ways.

The first is random variation, which affects both baseline and post-treatment scores. However, if change-from-baseline has a lower variance, or (equivalently) the pre-post correlation is greater than 0.5, then we expect there to be an advantage in using the CFB MD in preference to the post-treatment MD, just on the grounds of greater precision. Empirical studies have reported a median correlation of 0.59 for tests with reasonable test-retest reliability (19). Therefore, if precision was the only consideration, CFB would be expected to have an advantage.

A second cause of baseline imbalance is bias: for example, if baseline testing takes place before randomisation, failure to conceal allocation could lead to a selective inclusion of more severe patients in the more active intervention arm. Use of the change-from-baseline MD may reduce this, and other, biases. For these reasons, we strongly prefer CFB to post-treatment MD, and, unlike some authorities (14) we do not propose any explicit criteria for baseline imbalance.

Recommendation 3. In a meta-analysis of mean differences (MDs), the order of preference for the estimates to be used is:

(i) MD from analysis of covariance (ANCOVA) with baseline score as a covariate
(ii) MD based on means of change-from-baseline scores
(iii) MD based on means of post-treatment scores

Combinations of each type may be pooled together in a meta-analysis. When data permit, convert post-treatment MD to change-from-baseline MD.

The same recommendation applies when the summary statistics have been calculated on log-transformed data.

In the case of RoM analysis, change from baseline expressed as a *difference* cannot be converted to an RoM summary. Instead change from baseline must be expressed as a *ratio*. Thus it is the ratio of the ratios of follow-up mean to baseline mean that is used. If arm baseline and follow-up means and their SDs are available, an RoM analysis should be based on a ratio of follow-up-to baseline ratios. As with an additive analysis, the correlation between baseline and follow-up scores must be extracted, or assumed. This option is available in the *GMD2 Data Conversion Workbook*.

5.2.2.Procedure when the change from baseline standard deviation is missing

The change-from-baseline standard error or standard deviation is generally only reported in papers which have carried out a change-from-baseline analysis. Our recommendation, which differs from advice in the literature, is that change-from-baseline should be used in preference to post-treatment analysis, whenever possible, *even when the change from baseline SD is not reported.*

The change-from-baseline SD can be calculated from the pre- and post-treatment SDs, if the correlation is known (*Section 5.2*). However, trials seldom report the correlation, in which case one may either assume a correlation of 0.5, which is on average slightly conservative (19, 20), or a correlation may be sourced from literature on similar outcomes, or from an appropriate area of medicine. Balk (2012) (19) is a useful reference, and a summary of findings from that study appear in Appendix B.

Various kinds of imputation have been discussed (14), including calculation of correlations in trials where it is can be estimated, and then use of the average to impute values for trials where the correlation cannot be calculated. We do not recommend this for routine use as it requires additional data extraction, is time consuming, and because correlations calculated this way are highly variable (estimates of +1 or -1 are not uncommon (19, 20)) so the process can be hazardous even with large samples.

Conversion of arm-based change-from-baseline data to a change-from-baseline mean difference and SE is available on the *GMD2 Data Conversion Workbook*: users need to specify the pre-post correlation.

5.3. MAPPING OUTCOMES TO A COMMON SCALE, STANDARDISATION AND ROM

When trials report what are essentially the same or similar outcomes measured on different scales, there is a need to convert them to a common scale, so that the results of each trial can be compared and pooled.

Conversion of feet and inches into centimeters, Fahrenheit to Centigrade, or pounds into kilograms, are trivial examples of "mapping to a common scale", and we assume that this is done at the time of data extraction, and make no further comment except that extreme care is required to harmonise concentrations reported in different standard units (milli-, micro-, pico-, nano-grams *per* mm^3, litre, or decilitre), with concentrations reported in molecular units.

A slightly less trivial example is pain scores. Pain is generally either measured on an 11-point Numerical Rating Scale (NRS) (0,1,2….10), or on a Visual Analog Scale (VAS) from zero to 100. There is a literature comparing these scales, and it is accepted that they give closely similar results (21, 22). Results should therefore be converted to the same units.

Where there is no suitable "mapping" between outcome scales, for example different questionnaires aimed at measuring the same underlying construct such as anxiety, two approaches have been advocated. The first is standardisation, in which all responses are mapped onto a common unit SD scale by dividing by an SD. However, the traditional "standardisation", in which each trial's MD is divided by the sample SD in the same trial, is not recommended (*Section 5.3.3*). Instead we suggest two ways of mapping to a common scale that are applicable especially to patient reported outcomes (PROs) and clinician reported outcomes (CROs). Both are forms of "standardisation" in that they map data onto an SD scale, but this is done by dividing the summary statistics by a *scale-specific* (not trial-specific) SD. The preferred method is to use SDs derived from external reference data (Section 5.3.1); the other option is to derive the SDs "internally" from the trials included in the meta-analysis (Section 5.3.2).

A major advantage of this approach is its simplicity, as it reduces the problem to the same kind of calculation needed to convert between inches and centimeters, and also avoids having to undertake complex calculations to take account of difference in the SDs generated by different study designs (23, 24).

A second possibility is RoM meta-analysis: ratios of mean results on treatment and control arms constitute a "mapping" onto a common, unit-less scale (Section 5.1).

5.3.1. Standardisation using an external reference SD

This method requires a large cohort or cross-sectional study, in an external reference population that matches the target population in the trial data in the meta-analysis. The subjects in the cohort or cross-sectional study must have been tested on all the measurement scales of interest and the SDs on each scale must be reported. Each MD and its SE, *or* each arm mean and its SD, is then divided by the relevant scale SD.

The resulting standardised means and mean differences can be considered comparable because the SDs used for the mapping were based on the same population. The process is illustrated in *Section* 6.1.

When using this method, guideline developers should choose a reference population that is as similar as possible to the target population for the decision, represented in the trials included in the meta-analysis. Note this method does not take into account the sampling uncertainty in the reference SD. More research is needed to investigate the degree of heterogeneity in SDs and ratios of SDs of different scales, in large observational datasets, so that this can be incorporating into the method in future.

5.3.2. Standardisation using an internal reference SD

External reference SDs can only be used if there is a dataset that provides SDs on all the scales of interest.

An alternative that can always be used is to create a set of *internal* reference SDs, using the sample SDs reported in the trials included in the meta-analysis. The simplest approach is to use the mean of the SDs on each scale. For example, if there are 15 trials, with 10 reporting on scale "A" and 5 on scale "B", we standardise the Scale "A" MDs and SEs (or the arm means and SDs) by the mean of the 10 pooled Scale "A" SDs, and the Scale "B" results by the mean of the 5 pooled Scale "B" SDs.

The SDs at baseline, pooled over all the treatment arms, should be used for this purpose as the baseline SD most closely approximates that of the trial population. The calculations are illustrated in Section 6.2.

If there are trials where the baseline SD is not reported, the analysis can proceed using the mean SD based on the other trials. If several baseline SDs are unreported, and post-treatment SDs are reported instead, the best approach is to regress all the baseline SDs against the post-treatment SDs. The best estimate of the average pooled baseline SD is now the average of all the fitted baseline SDs. This is illustrated in Section 6.3.

Methods for deriving reference SDs for a set of similar outcome scales, either from observational studies, trials, or both require further research (see Section 8).

5.3.3. Sample-based standardization: the traditional SMD

The Standardised Mean Difference (SMD) of a trial, also sometimes called the standardised effect size, or simply "the effect size", is the mean difference divided by the sample standard deviation (SD). The SMD, realised as Cohen's *d* (25) or Hedges *g* (26) is an apparently simple and very frequently used solution to the problems created by having similar outcomes reported on different scales (27), but *it must be regarded as a flawed procedure that should be avoided wherever possible*.

The fundamental problem is that the SMD is created by dividing the MD by what is essentially an arbitrary number. If one trial reports an MD on the Beck Depression Inventory (BDI) of 4.5, and another reports and MD of 3.0, then no one would dispute that the treatment effect in the first trial has clearly reported a larger effect. This fact cannot be changed by the SDs. And yet, if the pooled SDs happened to be 9 and 5 respectively, then the ranking of trials on the "SMD scale" would be reversed: 0.50 and 0.60. The fact that the SDs were 9 and 5 was no more than a coincidence: it would have been equally probable that they happened to be 5 and 9 respectively. In this case we would have SMDs of 0.90 and 0.33, representing a massively greater effect in Trial 1.

The origin of sample-based standardisation appears to have been in sample size calculations. The sample size required to declare a given standardised effect statistically significant at a stated level is independent of the SD (25). Thus, *within the same trial,* SMDs on two outcome scales should be the same, within measurement error and sampling variation. However, the same model which gives us these results also tells us that, in the presence of sampling variation or variation in population SDs, no such equivalence can exist *between* trials.

The problematic nature of SMDs in this respect has in fact been widely recognised for many years. Leading epidemiologists (28-30) have described SMDs as "non-comparable and useless for meta-analysis". The Cochrane Handbook (2019) notes that the use of SMDs assumes that "between-study variation in standard deviations reflects *only* [our emphasis] differences in measurement scales and not differences in the reliability of outcome measures or variability among study populations" (31) (p252). The use of SMDs is therefore not compatible with *any* degree of between-trial variation in SDs, let alone the high level of variation that can be routinely observed in systematic reviews using the same scale. It should be added that the requirement is not that the SDs in different studies should be "similar" (32): the requirement is that they should be the same.

Hunter and Schmidt (33) express the same concerns at what they term "range variation" and "range restriction", whereby the study SD is typically restricted by the study inclusion criteria. They regarded

these as artefacts which need to be removed to reveal the true treatment effect. As a solution they advocate that a correction is applied to the sample SD, so that the denominator SD becomes, in effect, that of a *reference population, and therefore fixed*. This is exactly equivalent to the use of a single reference population to define a set of fixed reference SDs for each measurement scale, as recommended in Section 5.3.1.

Interestingly, this same idea underlies the suggestion that the pooled summary SMD can be "back-transformed" onto the scale of the original measurement instrument by using a reference SD, derived either from a representative observational study (Section 5.3.1), or based on the weighted average of the study SDs that measure the outcome on the scale of interest (14). Once again, this requires a population reference SD, although this is needed for only one outcome. The notion of a weighted average of the study SDs is exactly our proposal if no external one can be found: use an internal reference SD (see Recommendation 4).

Besides the logical flaws in SMDs flowing from between-trial variation in the population SDs, SMDs meta-analysis is exposed to several further unwelcome sources of heterogeneity:

- *Sampling error*: The standard deviation is based on the trial sample, and will therefore vary across studies by chance alone. For example, if the true SD is 1, the 95% limits on what would be observed in a study of with a sample size of 20 in each treatment arm is (0.68, 1.32) and with 50 in each arm it is (0.80, 1.20).
- *Variation in measurement error and construct validity:* Measurement scales vary in both precision, and in the extent to which they reflect the underlying construct, for example "depression" or "anxiety", to which the treatment is directed. Both measurement error and construct validity will affect the *responsiveness* to the effects of treatment, so that two scales which are identical in every other way will produce different SMDs
- *Risk of deliberate manipulation.* It is good experimental practice to minimise unwanted sources of variation. Patient selection criteria that limit trial populations to a narrow range therefore represent good practice, and allow investigators to obtain a given power with a smaller number of patients. However, minimising unwanted variation will lower the population SD, and will then generate treatment effects that are higher on the SMD scale, without being any different on the MD scale.

Standardisation based on an internal reference SD, using an average of the baseline SDs in the meta-analysis (Section 5.3.2) is a compromise, somewhere between the external reference SD and the traditional SMD dividing each trial by the sample SD. Because it takes an *average* of the trial-specific SDs *it is always less variable than the SMD, and thus introduces less heterogeneity.*

Recommendation 4. If meta-analyses include trials with different outcomes of the same construct and an additive model is assumed, they should be mapped to a common scale, using one of the following methods, in this order or preference:

(i) Division by an SD from an external reference population

(ii) Division by an SD derived internally from the trials included in the meta-analysis, taking care to adjust for differences in the direction of effect.

SMD meta-analysis (standardisation by dividing the MD in each trial by the same trial's sample SD) should not be used.

5.4. SYNTHESIS WITHIN AS WELL AS BETWEEN TRIALS

When a trial reports more than one "similar" outcome, for example the BDI and the HAMD depression scores, the usual advice is to select a single outcome from each study (14). This is because the measures are correlated (as they are measured on the same patients) and including both is a form of "double counting". A common practice is to propose a preference hierarchy, for example: choose scale A if available, otherwise scale B, otherwise… and so on. But this raises the question: if evidence from scale B is acceptable when it is the only one reported, why should it be excluded from trials also

reporting scale A? It seems strange to go to such lengths to include every trial, if we then discard what could be a high percentage of the information we have found.

Once the scores are converted to a common SD scale using internal or external standardisation, it is preferable to include all relevant information, taking the correlations into account. Table 5.1 shows how much the standard error of a trial rescaled MD is reduced if we include 2, 3, 4 or 5 measures, given a range of correlations. The correlations cited here are the between-score correlations between individuals that would be recorded in cross-sectional studies of patients with the condition under study. Methods and formulae for pooling mean treatment differences within trials, taking account of correlations, are found in (34, 35), Appendix C10-12, and may be implemented in the *GMD2 Data Conversion Workbook* (see Appendix D).

Table 5.1 Benefits of within-study synthesis of multiple outcomes when using rescaled MDs. The cell entries represent the proportional decrease in the SE of the composite rescaled Mean Difference. The reduction is less as the correlation increases, and is independent of sample size.

Proportional decrease in the SE of the composite rescaled Mean Difference		Number of outcomes				
		1	2	3	4	5
Correlation	0.60	1.000	0.894	0.856	0.837	0.825
	0.65	1.000	0.908	0.876	0.859	0.849
	0.70	1.000	0.922	0.894	0.880	0.872
	0.75	1.000	0.935	0.913	0.901	0.894
	0.80	1.000	0.949	0.931	0.922	0.917

To implement this, estimates of the cross-sectional correlations are required. Typically, for the kinds of patient reported outcomes (PROs) and clinician reported outcomes (CROs) where multiple outcomes are reported, correlations of around 0.65 to 0.75 between outcome measures are observed in the general psychometric literature (36-39). The true correlations are likely to be higher (close to 0.95), as the observed variability in PROs and CROs consists of both between-patient variation in the underlying scores and measurement error, which is reflected by a degree of attenuation in the observed correlations due to measurement error (33). Guideline developers should consult literature in the relevant field to obtain ranges for correlation.

If an estimate of the correlation between two outcomes is not available in the literature, a plausible range may be elicited from clinical experts based on unpublished datasets. For example, a reasonable estimate for depression and anxiety scales may be 0.70, and the impact of this imputation may be assessed with a sensitivity analysis using 0.60 and 0.80. If there is doubt about which value is appropriate, or if there is variation between correlations reported in the literature, a conservative approach is to choose a higher value as this will lead to less benefit of within-trial synthesis.

The arguments for within-trial synthesis apply equally to RoM meta-analysis and the necessary calculations are available in the *GMD2 Data Conversion Workbook.*

Recommendation 5. If trials report more than one outcome on the same construct, a within-trial synthesis should be conducted taking account of the correlations between outcomes.

GMD1 includes some discussion and references to other methods for synthesis of multiple outcomes.

6. WORKED EXAMPLES FOR PROCEDURES FOR STANDARDISATION

6.1. STANDARDISATION USING SDS FROM AN EXTERNAL REFERENCE POPULATION

If the trials included in the meta-analysis have reported the outcome on various scales that aim to measure the same underlying construct, the analyst will have to rescale the outcomes across trials to a common scale. The preferred approach to standardise the arm-based means and SDs or contrast-based MDs and SEs by dividing them by a SD on the same scale from an external reference population (Recommendation 4) that matches the patient populations in the included trials (Section 5.3.1).

Worked Example

A systematic review compared Donepezil to placebo for dementia due to Alzheimer's disease (AD) (40). The included studies reported change from baseline in cognition scores on the Alzheimer's Disease Assessment Scale – cognitive subscale (ADAS-cog), or the Mini-Mental State Examination (MMSE) scale in individuals with mild-to-moderate AD are listed in Table 6.1.

The observational ICTUS study (41) was considered to represent the patient population of the trials. It included patients with mild-to-moderate AD from 12 European countries, like the review's target population, and reports the baseline SDs in 973 individuals on the ADAS-cog and MMSE scales (41), shown as the reference SDs in Table 6.1.

Table 6.1. Studies reporting change in cognition scores from baseline on one of two scales in patients with mild-to-moderate Alzheimer's disease. First six data columns: original data on two scales; last four data columns: data mapped to a common SD scale using external reference SDs.

	Data on two scales, ADAS-Cog and MMSE							Reference SD	Data mapped to a common SD scale			
	Placebo			Donepezil (10 mg/day)					Placebo		Donepezil (10 mg/day)	
Study	$\overline{Y}_1$	S_1	n_1	$\overline{Y}_2$	S_2	n_2	Scale		$\overline{Y}_1$	S_1	$\overline{Y}_2$	S_2
Burns 1999	1.66	5.5	264	-1.26	5.5	254	ADAS-Cog	9.0	0.18	0.61	-0.14	0.61
Maher-Edwards 2011	-0.3	6.26	61	-1.5	6	67	ADAS-Cog	9.0	-0.03	0.70	-0.17	0.67
Moraes 2006b	3.8	26.3	18	-7.3	18.4	17	ADAS-Cog	9.0	0.42	2.92	-0.81	2.04
Rogers 1998b	1.82	5.43	153	-1.06	5.43	150	ADAS-Cog	9.0	0.20	0.60	-0.12	0.60
Seltzer 2004	0.69	4.61	55	-1.64	4.69	91	ADAS-Cog	9.0	0.08	0.51	-0.18	0.52
Tariot 2001	-0.81	4.03	102	-0.1	4.05	103	MMSE	3.8	0.21	1.06	-0.03	1.07
Winblad 2006	0.1	3.3	120	1.1	3.3	120	MMSE	3.8	-0.03	0.87	0.29	0.87

We map the scores to a common scale by dividing the means and SDs in all studies reporting on the ADAS-cog scale by 9.0 and the means and SDs in all studies reporting on the MMSE scale by 3.8. For example, in the placebo arm of Burns 1999:

$$\overline{Y}_1 = {}^{1.66}\!/_{9.0} = 0.184,\ S_1 = {}^{5.5}\!/_{9.0} = 0.611$$

Note that because a high score means greater cognitive impairment on ADAS-cog but lower on MMSE we have multiplied the mean MMSE scores by -1, so that they can then be pooled as (rescaled) mean differences.

6.2. STANDARDISATION USING AN INTERNAL REFERENCE STANDARD: THE AVERAGE OF POOLED SDS AT BASELINE

If the outcome has been reported on different scales across trials included in the meta-analysis, and an external set of reference SDs is not available (Section 5.3.1), the preferred rescaling option is then to standardise the arm-based means and SDs or contrast-based MDs and SEs by an internal reference SD (Recommendation 4). That is, for each scale, the average of the pooled baseline SDs reported by trials included in the meta-analysis. This is further explained in Section 5.3.2.

Worked Example

The example data come from 17 parallel RCTs comparing atomoxetine to placebo in patients with attention deficit hyperactivity disorder (ADHD) symptoms and the efficacy outcome has been reported as the mean change from baseline on one of two scales (i.e., Swanson, Nolan, and Pelham, Version IV (SNAP-IV) or ADHD Rating Scale-IV (ADHD-IV)) (42).

Table 6.2. Baseline $\left(\overline{Y}_B, S_B\right)$ and change from baseline $\left(\overline{Y}_C, S_C\right)$ measurements from studies comparing atomoxetine and placebo in ADHD patients [25]

	Placebo					Atomoxetine					
Study	$\overline{Y}_{B_1}$	S_{B_1}	$\overline{Y}_{C_1}$	S_{C_1}	n_1	$\overline{Y}_{B_2}$	S_{B_2}	$\overline{Y}_{C_2}$	S_{C_2}	n_2	Scale
Bangs 2008	45.3	5.7	-4.4	8.4	68	44.7	6.4	-9.6	11.4	153	SNAP-IV
Dell'Agnello 2009	41.5	6.9	-2.0	4.7	32	42.7	6.2	-8.1	9.2	105	SNAP-IV
Dittmann 2011	36.4	9.3	-6.7	10.5	59	37.6	9.7	-14.7	10.3	60	SNAP-IV
Gau 2007	37.1	6.4	-9.3	13.2	34	36.7	6.7	-17.3	10.6	72	ADHD-IV
Kaplan 2004	41.9	8.0	-7.5	11.4	44	42.2	8.3	-17.0	13.9	52	ADHD-IV
Kelsey 2004	42.3	7.1	-7.0	10.8	60	42.1	9.2	-16.7	14.5	126	ADHD-IV
Kratochvil 2011	37.6	7.0	-5.8	8.4	49	38.9	6.6	-13.2	11.3	44	ADHD-IV
Marenyi 2009	37.0	7.5	-11.4	8.0	33	38.1	7.3	-15.8	7.6	72	ADHD-IV
Michelson 2001	38.3	8.9	-5.8	10.9	83	39.2	9.2	-13.6	14.0	84	ADHD-IV
Michelson 2002	36.7	8.8	-5.0	10.4	83	37.6	9.4	-12.8	12.4	84	ADHD-IV
Montoya 2009	39.5	9.0	-4.7	7.4	50	39.1	9.0	-12.8	9.3	99	ADHD-IV
Newcorn 2008	41.7	8.5	-7.3	11.5	74	40.9	8.8	-14.4	12.7	222	ADHD-IV
Spencer 2002a	37.6	8.0	-5.9	13.0	60	37.8	7.9	-14.4	13.0	63	ADHD-IV
Spencer 2002b	41.4	7.9	-5.5	11.6	61	41.2	8.9	-15.6	13.7	64	ADHD-IV
Svanborg 2009	39.5	6.7	-6.3	10.6	50	38.9	7.7	-19.0	10.5	49	ADHD-IV
Takahasi 2009	32.3	9.6	-8.1	7.1	61	33.3	8.7	-10.8	6.8	58	ADHD-IV
Weiss 2005	36.7	8.4	-7.2	9.7	51	38.9	7.2	-14.5	12.3	100	ADHD-IV

In the absence of an external reference standard, an internal reference standard is constructed using the averages of the pooled SDs at baseline. For example, the pooled SD in Bangs 2008 is calculated based on the formula provided in Appendix C1:

$$S_{pooled} = \sqrt{\frac{(68-1)5.7^2 + (153-1)6.4^2}{68+153-2}} = 6.1942$$

For the SNAP-IV scale, the internal reference SD is calculated as the average of pooled baseline SDs reported in Bangs 2008, Dell'Agnello 2009, and Dittmann 2011:

$$\text{Reference } SD_{SNAP-IV} = \frac{6.1942 + 6.3676 + 9.5308}{3} = 7.3552$$

The change from baseline means and SDs on the SNAP-IV scale are then rescaled to a common scale by dividing them by this reference SD. For example, in the placebo group of Bangs 2008:

$$\bar{Y}_{C_1} = -4.4 \big/ 7.3552 = -0.598,\ S_{C_1} = 8.4 \big/ 7.3552 = 1.142$$

A similar procedure follows for studies reporting the outcome on the ADHD-IV scale, resulting in the rescaled values provided in Table 6.3 that can be in turn pooled as [rescaled] mean differences.

Table 6.3. Change from baseline scores mapped to a common SD scale using an internal reference. The pooled baseline SDs on each scale are averaged to form the Reference SD, and then the Means and Standard deviations are divided by the reference SDs

Study	S_{pooled}	Scale	Reference SD	Placebo			Atomoxetine		
				$\bar{Y}_{C_1}$	S_{C_1}	n_1	$\bar{Y}_{C_2}$	S_{C_2}	n_2
Bangs 2008	6.1942	SNAP-IV	7.3552	-0.598	1.142	68	-1.305	1.550	153
Dell'Agnello 2009	6.3676	SNAP-IV	7.3552	-0.272	0.639	32	-1.101	1.251	105
Dittmann 2011	9.5038	SNAP-IV	7.3552	-0.911	1.428	59	-1.999	1.400	60
Gau 2007	6.6063	ADHD-IV	8.1286	-1.144	1.624	34	-2.128	1.304	72
Kaplan 2004	8.1641	ADHD-IV	8.1286	-0.923	1.402	44	-2.091	1.710	52
Kelsey 2004	8.5828	ADHD-IV	8.1286	-0.861	1.329	60	-2.054	1.784	126
Kratochvil 2011	6.8139	ADHD-IV	8.1286	-0.714	1.033	49	-1.624	1.390	44
Marenyi 2009	7.3627	ADHD-IV	8.1286	-1.402	0.984	33	-1.944	0.935	72
Michelson 2001	9.0522	ADHD-IV	8.1286	-0.714	1.341	83	-1.673	1.722	84
Michelson 2002	9.1068	ADHD-IV	8.1286	-0.615	1.279	83	-1.575	1.525	84
Montoya 2009	9.0000	ADHD-IV	8.1286	-0.578	0.910	50	-1.575	1.144	99
Newcorn 2008	8.7265	ADHD-IV	8.1286	-0.898	1.415	74	-1.772	1.562	222
Spencer 2002a	7.9489	ADHD-IV	8.1286	-0.726	1.599	60	-1.772	1.599	63
Spencer 2002b	8.4270	ADHD-IV	8.1286	-0.677	1.427	61	-1.919	1.685	64
Svanborg 2009	7.2122	ADHD-IV	8.1286	-0.775	1.304	50	-2.337	1.292	49
Takahasi 2009	9.1726	ADHD-IV	8.1286	-0.996	0.873	61	-1.329	0.837	58
Weiss 2005	7.6238	ADHD-IV	8.1286	-0.886	1.193	51	-1.784	1.513	100

6.3. PREDICTING BASELINE SD IN STUDIES REPORTING POST-TREATMENT SD ONLY

If a trial did not report the baseline SDs, but reported the post-treatment SDs, its pooled baseline SD may be imputed from the studies reporting both baseline and post-treatment SDs on the same scale through geometric regression.

Worked Example

The trials in Table 6.4 come from the meta-analysis presented in Table 6.3. These trials report both the baseline and post-treatment SDs on the ADHD-IV scale. However, for illustrative purposes, suppose Kaplan 2004 did not report the baseline SDs. We will impute its pooled baseline SD using information from the remaining trials reporting both SDs on the ADHD-IV scale.

First, we calculate the pooled baseline SD and pooled post-treatment SD in all trials that report both (Table 6.4). We then calculate a factor, β, as follows:

$$\beta = \sqrt{\frac{SD_{S_{pooled_B}}}{SD_{S_{pooled_F}}}}$$

(β is in fact the geometric mean regression slope). For example, in Excel, we compute the standard deviations of the pooled baseline and post-treatment SDs in each arm using the stdev.S formula. For example, for the pooled baseline SDs, we would input =STDEV.S(8.58,6.83,9.00,9.17) in an Excel cell (Table 6.5).

β is then:

$$\beta = \sqrt{1.07 / 2.41} = 0.6679$$

The pooled pre-treatment SD for Kaplan 2004 would then be computed as

$$S_{pooled_B} = \beta \times S_{pooled_F} = 0.6679 \times 14.01 = 9.36$$

Table 6.4. Baseline and post-treatment SDs and sample sizes in trials reporting both on the ADHD-IV scale

	Sample Size		Baseline SD			Post-treatment SD		
	Placebo	Atom-oxetine	Placebo	Atom-oxetine		Placebo	Atom-oxetine	
Study	n_1	n_2	S_{B_1}	S_{B_2}	Pooled	S_{F_1}	S_{F_2}	Pooled
Kaplan 2004	44	52	NR	NR	NR	13.4	14.5	14.01
Kelsey 2004	60	126	7.1	9.2	8.58	12.3	14.3	13.69
Kratochvil 2011	49	44	7	6.63	6.83	9.8	5.80	8.16
Montoya 2009	50	99	9	9	9.00	12.3	12.7	12.57
Takahasi 2009	61	58	9.6	8.7	9.17	11.4	10.3	10.88

Table 6.5. Standard deviations of the pooled baseline and post- treatment SDs of trials reporting both

	Pooled baseline SDs	Pooled post-treatment SDs
SD	1.07	2.41

7. REPORTING GUIDANCE

This section makes recommendations about additional items to be reported in the case of continuous outcomes, over and above those mentioned in GMD1.

Guideline documents should provide an account of the strategic decisions that were made, concerning:

1. Which continuous outcomes were available, and which were extracted, and which not: how was this decided?
2. The reasons for choosing an additive or multiplicative model, referring to the issues set out in Section 5.1.
3. If different but related outcomes were reported in different trials, within the framework of an additive model, the method for mapping to a common scale, or standardisation should be explained. If external reference SDs were used, what was the source of data to inform the values, and how was this determined.
4. If multiple outcomes reflecting the same construct were reported in some trials, whether within-trial synthesis was used, and if so, what was the source of information about the correlation coefficient(s).

8. RESEARCH RECOMMENDATIONS

8.1. IS ANCOVA SUPERIOR TO CHANGE-FROM-BASELINE?

It is widely considered that the best way to analyse RCT data is the analysis of covariance using the patient's baseline score as the covariate, and the estimate of the relative treatment effect computed this way is widely seen as the optimal choice in meta-analysis. However, the core research papers in this area (43-45) have consistently employed the standard ANCOVA model which assumes that the baseline score is measured with no error. This assumption is irrational because the post-treatment score, based on the exact same test instrument as the baseline score, *is* assumed to be measured with error. Thus, the preference for ANCOVA is based on a false assumption, and research is needed to re-evaluate its role in trial analysis and meta-analysis.

8.2. METHODS FOR STANDARDISATION AND MAPPING TO A COMMON SCALE

In Section 5 we recommend two simple ways to derive sets of reference SDs, either from external observational studies, or internally from sets of trials. More research is required to develop these methods, with the objective of generating sets of "off the shelf" reference SDs for use in meta-analyse of treatments for different disorders, and "off the shelf" correlations for use in within-trial synthesis These methods assume that ratios of SDs of different instruments are relatively stable, with minimal variation across studies. This needs to be assessed. At the same time there is a need to develop methods for pooling data on ratios of SDs across observational studies, and possibly to combine data from trials and observational studies.

At the same time, the additivity or proportionality of treatment effects measured on these outcomes needs to be assessed, so that more definitive guidance on the use of RoM could be made available

A final possibility is the use of "cross-walking" between outcome scales (46, 47). Also known as test equating, aligning, or linking. The objective of cross-walks is to determine, for each point on one scale, what is the corresponding point on the other. Methods include Item Response Theory (IRT) (48-51) and equi-percentile matching. There is a widely accepted set of properties that these mapping must have, in order to be valid (46). Originally designed for educational psychology applications, cross-walk data are being seen increasingly in psychiatry, especially depression (52-56) and schizophrenia (57, 58), using either equi-percentile or IRT methods. Research is needed, in the first instance, on whether, and how, published cross-walk tables can help map group means and standard deviations between scales. If they have a role, further work will be required to develop cross-walks for a wider set of disorders.

APPENDICES

APPENDIX A

Indirect estimation of the standard error of the mean difference

Methods for deriving the pooled standard deviation, SD_{pooled}, and the standard error of the mean difference, SE_D, from available statistics, when it is not reported directly.

Finding SD_{pooled} from SE_D	$SD_{pooled} = \frac{SE_D}{\sqrt{\frac{n_1+n_2}{n_1 n_2}}}$
Finding SE_D from 95% confidence interval of mean difference	$SE_D = \frac{\text{upper limit - lower limit}}{3.92}$
Finding SE_D from mean difference, D, and *z*-statistic	$SE_D = \frac{\lvert D \rvert}{z}$
Finding SE_D from mean difference, D, and *t*-statistic	$SE_D = \frac{\lvert D \rvert}{t}$
Finding *z*-statistic from one sided *p*-value, p, corresponding to z-test	$z = \Phi^{-1}(1-p)$, where Φ^{-1} is the inverse of the standard normal cumulative distribution function
Finding *z*-statistic from two sided *p*-value, p, corresponding to z-test	$z = \Phi^{-1}\left(1-{}^{p}\!/_{2}\right)$
Finding *t*-statistic from one sided *p*-value, p, corresponding to t-test	$t = \pm t^{-1}(p, df), df = n_1 + n_2 - 2$, where t^{-1} is the inverse of the t distribution, n_1 and n_2 are the number of patients in arm 1 and arm 2, respectively.
Finding *t*-statistic from two sided *p*-value, p, corresponding to t-test	$t = \pm t^{-1}\left({}^{p}\!/_{2}, df\right), df = n_1 + n_2 - 2$

Alternatively, the RevMan calculator may assist in these calculations. An Excel File containing the RevMan calculator may be obtained from:

http://training.cochrane.org/resource/revman-calculator

Indirect estimation of the mean and SD from the median and interquartile range (IQR)

Where the distribution of the individual measurements are approximately normal, which may be the case when the data has been log-transformed, the median of the log-transformed measurements, $\tilde{X}_j$, is approximately equal to the mean, $\bar{X}_j$, and the standard deviation of the log-transformed measurements, S_{X_j}, may be approximated as $S_{X_j} \approx {}^{IQR}\!/_{1.35}$ [6].

APPENDIX B

Empirical correlations between baseline and post-treatment scores

Table B.1: Empirical correlations of baseline and final values based on a collection of trials, overall and broken down by different characteristics (19)

Characteristic	**Median**	**25th, 75th Percentile**
All studies	0.59	0.40, 0.81
Treatment type		
Active treatment	0.54	0.37, 0.77
Inactive treatment	0.73	0.53, 0.87
Outcome types		
Device measure	0.83	0.61, 0.94
Lab	0.63	0.39, 0.81
Sign	0.51	0.37, 0.72
Questionnaire/Score	0.51	0.34, 0.68
Symptoms	0.44	0.38, 0.50
Other	0.78	0.71, 0.87
Clinical domains		
Nephrology	0.61	0.44, 0.82
Cardiovascular medicine	0.59	0.35, 0.86
Pulmonary medicine	0.77	0.54, 0.94
Diabetology	0.65	0.44, 0.76
Internal medicine/Geriatrics/Primary care	0.73	0.56, 0.83
Gastroenterology/Hepatology	0.44	0.23, 0.55
Psychiatry	0.36	0.22, 0.58
Neonatology/Pediatrics	0.60	0.46, 0.84
Ophthalmology	0.38	0.26, 0.54
Critical care	0.56	0.38, 0.70
Others	0.52	0.34, 0.77

APPENDIX C

Formulae used in the GMD2 Data Conversion Workbook

Note that in all calculations involving a log-transformation, we use the natural logarithm, denoted by $\ln$.

Notation

Let $Y_{i,jk}$ denote the outcome observed on individual $i = 1, 2, ..., n_j$ in treatment group $j = 1, 2$, where k indicates the outcome is the baseline measurement, b, follow-up measurement, f, or the change from baseline measurement, c, such that $Y_{i,jc} = Y_{i,jf} - Y_{i,jb}$. $\overline{Y}_{1b}, \overline{Y}_{1f}$ and $\overline{Y}_{2b}, \overline{Y}_{2f}$ are the mean responses in treatment groups 1 and 2 at baseline and follow-up. For example, $\overline{Y}_{jf} = \frac{1}{n_j}\sum_{i=1}^{n_j} Y_{i,jf}$ is the average over the individuals in treatment group j at follow-up, with standard deviation $S_{jf} = \sqrt{\frac{1}{(n_j - 1)}\sum_{i=1}^{n_j}\left(Y_{i,jf} - \overline{Y}_{jf}\right)^2}$. The mean change from baseline (CFB) in treatment group j is: $\overline{Y}_{jc} = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(Y_{i,jf} - Y_{i,jb}\right) = \overline{Y}_{jf} - \overline{Y}_{jb}$.

Alternatively, the mean response in group j may be reported as a mean percent change from baseline, $\overline{P}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\frac{\left(Y_{i,jf} - Y_{i,jb}\right)}{Y_{i,jb}}$, and it's standard error $S_{\overline{P}_j}$.

Note that in the calculation of relative effects described below, Y_{ij} may refer to either the post-treatment values, $Y_{i,jf}$, or to changes from baseline, $Y_{i,jc}$, so that $\overline{Y}_j$ may refer to the post-treatment mean, $\overline{Y}_{jf}$, or the mean change from baseline, $\overline{Y}_{j,c}$, in group j.

In addition, note the data may be recorded on the log-scale, i.e., $X_{i,j} = \ln\left(Y_{i,j}\right)$. The mean and standard deviation of these log-transformed measurements in group j are denoted as $\overline{X}_j$ and S_{X_j}, respectively. The geometric mean, $\overline{G}_j = \exp(\overline{X}_j)$, may be reported on the natural scale with a corresponding confidence interval $\left(L_{\overline{G}_j}, U_{\overline{G}_j}\right)$.

C1. Converting arm-based data into contrast-based data: mean differences (two-arm trial)

The mean difference, D, in a trial is calculated as the difference between the means $\overline{Y}_1$, $\overline{Y}_2$ of the outcome in two treatment groups, and has variance V_D :

$$D = \bar{Y}_2 - \bar{Y}_1, \quad SE_D = \sqrt{V_D}$$

$$V_D = \frac{n_1 + n_2}{n_1 n_2} S^2_{pooled}, \quad S^2_{pooled} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{df}, \quad df = n_1 + n_2 - 2$$

C2. Converting arm-based data into ratio of means

The log of a ratio of means, RoM, in a trial is calculated as (16)

$$\log(RoM) = \ln\left(\frac{\bar{Y}_{2f}}{\bar{Y}_{1f}}\right) = \ln\left(\bar{Y}_{2f}\right) - \ln\left(\bar{Y}_{1f}\right)$$

$$SE_{\log(RoM)} = \sqrt{V_{\log(RoM)}}, \quad V_{\log(RoM)} = \frac{1}{n_1}\left(\frac{S_{1f}}{\bar{Y}_{1f}}\right)^2 + \frac{1}{n_2}\left(\frac{S_{2f}}{\bar{Y}_{2f}}\right)^2$$

where this variance is based on a Taylor series approximation (16). Note that the $\log\left(RoM\right)$ may only be calculated if the means are positive. If sufficient statistics are not available to directly calculate $V_{\log(RoM)}$ or $SE_{\log(RoM)}$, they may be derived using other statistics reported in a trial (see Appendix A).

C3. Converting baseline and follow-up arm-based data into a ratio of ratios of means

A ratio of ratio of means, $RoRoM$, i.e., the ratio of ratio of follow-up to baseline means, is computed as

$$RoRoM = \frac{\bar{Y}_{2f} \big/ \bar{Y}_{2b}}{\bar{Y}_{1f} \big/ \bar{Y}_{1b}},$$

where these ratios are pooled on the log-scale:

$$\log\left(RoRoM\right) = \ln\left(\bar{Y}_{2f} \big/ \bar{Y}_{2b}\right) - \ln\left(\bar{Y}_{1f} \big/ \bar{Y}_{1b}\right)$$

and the standard error is calculated as

$$SE_{\log(RoRoM)} = \sqrt{V_{\log(RoRoM)}},$$

$$V_{\log(RoRoM)} = \frac{1}{n_1}\left[\left(\frac{S_{1b}}{\bar{Y}_{1b}}\right)^2 + \left(\frac{S_{1f}}{\bar{Y}_{1f}}\right)^2\right] - 2\ln\left(\frac{\rho_1 S_{1b} S_{1f}}{n_1 \bar{Y}_{1b} \bar{Y}_{1f}} + 1\right) + \frac{1}{n_2}\left[\left(\frac{S_{2b}}{\bar{Y}_{2b}}\right)^2 + \left(\frac{S_{2f}}{\bar{Y}_{2f}}\right)^2\right] - 2\ln\left(\frac{\rho_2 S_{2b} S_{2f}}{n_2 \bar{Y}_{2b} \bar{Y}_{2f}} + 1\right)$$

where ρ_j is the correlation between baseline and follow-up measurements in group j. Note that this variance is based on a multivariate Taylor series approximation.

C4. Calculating CFB and its SE in studies reporting baseline and post-treatment means

$$\bar{Y}_{jc} = \bar{Y}_{jf} - \bar{Y}_{jb}, \quad S_{jc} = \sqrt{S^2_{jb} + S^2_{jf} - 2r_j S_{jb} S_{jf}}$$

where r_j is the correlation between the baseline and post-treatment measures in arm j.

C5. Working out a pooled SD in a multi-arm trial

In a *M*-arm trial, S_{pooled} is the pooled standard deviation, calculated as the square root of a weighted average of the sample variances S_j^2, $j = 1, 2, ..., M$

$$S_{pooled} = \sqrt{\frac{\sum_{j=1}^{M}\left(n_j - 1\right)S_j^2}{df}}, \quad df = \sum_{j=1}^{M} n_j - M$$

where $n_j, j = 1, 2, ..., M$ are the sample sizes of each treatment group j.

C6. Converting means and standard deviations of raw data to that of log-transformed data

Assuming the individual observations on the natural scale, Y_{ij}, are log-normally distributed, the arithmetic mean, $\overline{X}_j$, and standard deviation, S_{X_j}, of the measurements on the log-scale may be calculated based on the arithmetic mean, $\overline{Y}_j$, and standard deviation, S_j, of the measurements on the natural scale, using (15)

$$\overline{X}_j = \ln\left(\frac{\overline{Y}_j}{\sqrt{1 + \frac{S_j^2}{\overline{Y}_j^2}}}\right), \quad S_{X_j} = \sqrt{\ln\left(1 + \frac{S_j^2}{\overline{Y}_j^2}\right)}.$$

Note that if we are converting mean change from baselines, then the assumption behind this conversion is that the individual's change from baseline on the natural scale are log-normally distributed.

C7. Converting geometric means and confidence intervals to log-transformed data

A geometric mean of raw values in a particular treatment group, $\overline{G}_j$, may be converted to an arithmetic mean of log-transformed values, $\overline{X}_j$, using (15)

$\overline{X}_j = \ln\left(\overline{Y}_{G_j}\right)$.

A $100(1 - \alpha/2)\%$ confidence interval of the geometric mean, $\left(L_{\overline{G}_j}, U_{\overline{G}_j}\right)$, may be used to obtain a measure of the standard deviation of the log-transformed values, S_{X_j}, using

$$S_{X_j} = \frac{\ln\left(U_{\overline{G}_j}\right) - \ln\left(L_{\overline{G}_j}\right)}{2z_{1-\alpha/2}} \times \sqrt{n_j}$$

where n_j is the sample size in treatment group j and $z_{1-\alpha/2}$ may be determined in an Excel worksheet by inputting $= \text{norm.inv}(1 - \alpha/2, 0, 1)$ into a cell. In the case of a 95% confidence interval, $z_{1-\alpha/2} = 1.96$

C8. Converting MD and its SE calculated on the natural scale to MD and SE calculated on the log-transformed scale

If the effect size is small, and the distributions of the individual observations on the natural scale, Y_{ij} , are similar for both treatment groups, a mean difference reported on the natural scale may be converted to a mean difference on the log scale, so long as the overall (or grand) arithmetic mean response across both treatment arms, $\overline{Y}$, is available (15):

$$D_{\log} = \frac{D}{\overline{Y}}, \quad SE_{D_{\log}} = \frac{SE_D}{\overline{Y}}$$

C9. Converting ratio of geometric means to MD and SE of log-transformed data

A ratio of geometric means, $RoGM$, based on raw data may be converted to a mean difference based on log-transformed data by taking the log of the ratio, $D_{\log} = \ln\left(RoGM\right)$.

A 100(1-$\alpha/2$)% confidence interval of the ratio of geometric means, $\left(L_{RoGM}, U_{RoGM}\right)$, may be converted to an approximate standard error of the MD based on log-transformed values, $SE_{D_{\log}}$, using

$$SE_{D_{\log}} = \frac{\ln\left(U_{RoGM}\right) - \ln\left(L_{RoGM}\right)}{2z_{1-\alpha/2}}$$

where $z_{1-\alpha/2}$ may be determined in an Excel worksheet by inputting $=\mathrm{norm.inv}(1-\alpha/2, 0, 1)$ into a cell. In the case of a 95% confidence interval, $z_{1-\alpha/2} = 1.96$.

C10. Within-trial synthesis of a continuous outcome with additive effects

A composite MD, $\overline{D}^*$, in a trial may be calculated as (35)

$$\overline{D}^* = \frac{1}{K}\sum_{k=1}^{K} D_k^* \ ,$$

where $D_k^* = \dfrac{D_k}{\text{Reference } SD_k}$, the rescaled MD on the k^{th} scale reported in a particular trial, $k = 1, \ldots, K$. Its variance is

$$V_{\overline{D}^*} = \frac{1}{K^2}\left(\sum_{k=1}^{K} V_{D_k^*} + \sum_{a \neq b} \mathrm{cov}\left(D_a^*, D_b^*\right)\right).$$

If we assume that the correlation between the measurements on each pair of scales a, b is ρ_{ab}, and all participants in treatment group j were measured on both scales, i.e., $n_{ja} = n_{jb} = n_j$, then (34)

$$\mathrm{cov}\left(D_a^*, D_b^*\right) = \frac{1}{\left(\text{Reference } SD_a\right)\left(\text{Reference } SD_b\right)}\left(\frac{1}{n_1}\rho_{ab} S_{1a} S_{1b} + \frac{1}{n_j}\rho_{ab} S_{ja} S_{jb}\right),$$

where S_{ja} and S_{jb} are the standard deviations of the measurements recorded on scales a and b, respectively, for treatment group j.

C11. Within-trial synthesis of a continuous outcome in a log-transformed MD analysis

A composite MD in terms of the log-transformed measurements, $\overline{D}_{\log}$, in a trial may be calculated as (35)

$$\overline{D}_{\log} = \frac{1}{K}\sum_{k=1}^{K} D_{\log_k} ,$$

Where $D_{\log_k} = \overline{X}_{jk} - \overline{X}_{1k}$, the difference between arithmetic means of log-transformed values, $\overline{X}_{jk}$ in treatment groups 1 and j on the k^{th} scale reported in a particular trial, $k = 1,...,K$. Its variance is

$$V_{\overline{D}_{\log}} = \frac{1}{K^2}\left(\sum_{k=1}^{K} V_{D_{\log_k}} + \sum_{a\neq b}\operatorname{cov}\left(D_{\log_a}, D_{\log_b}\right)\right).$$

If we assume that the correlation between the log-transformed measurements on each pair of scales a,b is $\rho_{\log_{ab}}$, and all participants in treatment group j were measured on both scales, i.e., $n_{ja} = n_{jb} = n_j$, then (34)

$$\operatorname{cov}\left(D_{\log_a}, D_{\log_b}\right) = \frac{1}{n_1}\rho_{\log_{ab}} S_{X_{1a}} S_{X_{1b}} + \frac{1}{n_j}\rho_{\log_{ab}} S_{X_{ja}} S_{X_{jb}} ,$$

where $S_{X_{ja}}$ and $S_{X_{jb}}$ are the standard deviations of the log-transformed measurements recorded on scales a and b, respectively, for treatment group j. Note that if a correlation between the raw measurements on each pair of scales a,b, ρ_{ab}, is only available, then

$$\operatorname{cov}\left(D_{\log_a}, D_{\log_b}\right) = \frac{1}{n_1}\ln\left(\frac{\rho_{ab} S_{1a} S_{1b}}{\overline{Y}_{1a}\overline{Y}_{1b}} + 1\right) + \frac{1}{n_j}\ln\left(\frac{\rho_{ab} S_{ja} S_{jb}}{\overline{Y}_{ja}\overline{Y}_{jb}} + 1\right),$$

where $\overline{Y}_{ja}, \overline{Y}_{jb}$ and S_{ja}, S_{jb} are the means and standard deviations of the raw measurements recorded on scales a and b, respectively, for treatment group j.

C12. Within-trial synthesis of a continuous outcome in an RoM analysis

A composite log-RoM, $\overline{\log(RoM)}$, in a trial may be calculated as (35)

$$\overline{\log(RoM)} = \frac{1}{K}\sum_{k=1}^{K} \log\left(RoM_k\right),$$

where $\log\left(RoM_k\right)$ is the log-RoM measured on the k^{th} scale reported in a particular trial, $k = 1,...,K$. Its variance is

$$V_{\overline{\log(RoM)}} = \frac{1}{K^2}\left(\sum_{k=1}^{K} V_{\log(RoM_k)} + \sum_{a\neq b}\operatorname{cov}\left(\log\left(RoM_a\right), \log\left(RoM_b\right)\right)\right).$$

If we assume that the correlation between the measurements on each pair of scales a,b is ρ_{ab}, and all participants in treatment group j were measured on both scales, i.e., $n_{ja} = n_{jb} = n_j$, then (34)

$$\operatorname{cov}\left(\log\left(RoM_a\right), \log\left(RoM_b\right)\right) = \frac{\rho_{ab} S_{1a} S_{1b}}{n_1 \overline{Y}_{1a}\overline{Y}_{1b}} + \frac{\rho_{ab} S_{ja} S_{jb}}{n_j \overline{Y}_{ja}\overline{Y}_{jb}},$$

where $\overline{Y}_{ja}, \overline{Y}_{jb}$ and S_{ja}, S_{jb} are the means and standard deviations of the measurements recorded on scales a and b, respectively, for treatment group j. Note that this covariance has been derived using a multivariate Taylor series approximation (34).

C13. Converting log odds ratio, or proportions responding on each arm to an MD and its SE for an additive model

Trials reporting the proportion of responders in each arm can be combined with other trials reporting additive effects. The following procedure assumes that the underlying individual responses have a logistic distribution, and they start with the log OR, map to a "standardised mean difference", and then rescale this to the scale of measurement, using a reference SD.

If p_1, p_2 are the proportions of responders on arms 1 and 2, n_1, n_2 are the total number of patients on each arm, and $\text{Reference } SD$ is an appropriate reference SD for the desired scale, then the treatment effect and SE in that trial can be found as follows:

$$LOR = \ln\left(\frac{p_2(1-p_1)}{p_1(1-p_2)}\right); \quad Var(LOR) = \frac{1}{p_1(1-p_1)n_1} + \frac{1}{p_2(1-p_2)n_s}, \quad SMD = \frac{\sqrt{3}}{\pi} LOR$$

$$D = SMD\left(\text{Reference } SD\right); \qquad SE_D = \frac{\sqrt{3}}{\pi}\left(\text{Reference } SD\right)\sqrt{Var(LOR)}$$

The final step maps the "standardised effect" into the units of the mean difference.

If either the arm probabilities or the LOR are reported, the mean difference and its SE may be further converted into a form suitable for log transformed analyses, provided sufficient statistics are reported (see Appendix C7).

If the probabilities are reported, they can be further converted to a form suitable for RoM analysis.

C14. Converting arm-based data into contrast-based data: Differences in mean percentage change from baseline

We note here that if all studies report mean percentage change from baseline, $\overline{P}_j$, and its standard error, $S_{\overline{P}_j}$, then the data may be directly pooled as mean differences (Appendix C1). It is not possible to combine this type of data with any other data formats.

APPENDIX D

General guidance on GMD2 Data Conversion Workbook

This workbook should allow the user to input one of the prioritised sets of statistics listed in Table 4-1 for each trial. To keep track of data conversions, we suggest saving a workbook for each review. Data may be stored for multiple trials within the worksheets for this purpose.

The data should be outputted in the format required for synthesis, whether that be arm-level or contrast-level. Outputted data from some worksheets may be copied and pasted as inputted data in other worksheets. For example:

- The outputted mean CFB, SD, and n from the CFB Calculation worksheet may be inputted into:
    - the External or Internal Rescaling worksheets for standardisation, or
    - the Additive worksheet to be converted into a MD, or
    - the Log worksheet to be converted into a MD of the log-transformed CFB data,
- The outputted mean differences and SE from the Additive worksheet may be inputted into the Within-trial synthesis_MDs worksheet.

Table D1. Reference to conversions in GMD2 Data Conversion Workbook

		Data Input		Data Output	
Procedure	**Worksheet**	**Statistics**	**Column(s)**	**Statistics**	**Column(s)**
Calculate mean change from baseline (CFB) based on baseline and post-treatment means	CFB Calculation	Sample size (n)	B; G	Mean CFB, SD, sample size for both treatment groups	N-S
		Baseline mean, SD	C-D; H-I		
		Post-treatment mean, SD	E-F; J-K		
		Correlation between baseline and post-treatment measurements	L		
Standardising using an externally sourced reference SD	External Rescaling	Scale name, code[a], and external reference SD	A-C	<u>If arm-level data inputted:</u> Rescaled group means and SDs, sample size <u>If contrast-level data inputted:</u> Rescaled MD and SE	<u>If arm-level data inputted:</u> Q-V <u>If contrast-level data inputted:</u> W-X
		Scale code[a] reported in trial	F		
		CFB or Post-treatment mean, SD, n	H-M		
		Mean difference, SE	N-O		

Procedure	Worksheet	Data Input		Data Output	
		Statistics	**Column(s)**	**Statistics**	**Column(s)**
Standardising using an internally sourced reference SD	Internal Rescaling	Baseline SDs and sample size from all included trials reporting data on the same scale	B-E	Trial-specific pooled baseline SD; Average pooled baseline SD	F; H
		Scale name, code[a], and internal reference SD[b] from column H	J-L	<u>If arm-level data inputted:</u> Rescaled group means and SDs, sample size <u>If contrast-level data inputted:</u> Rescaled MD and SE	<u>If arm-level data inputted:</u> Z-AE <u>If contrast-level data inputted:</u> AF-AG
		Scale code[a] reported in trial	O		
		CFB or Post-treatment mean, SD, n	Q-V		
		Mean difference, SE	W-X		
Additive - MD	Additive	CFB mean, SD, n	B-D; H-J	Mean difference and SE	V-W
		Post-Treatment mean, SD, n	E-G, K-M		
		Mean difference, SE	N-O		
Within-trial synthesis of Rescaled MDs	Within-trial synthesis_MDs	Correlation between scales	A-E	Composite mean difference and SE	BC; BE
		Rescaled MD, SE, Reference SD for each scale	H-J; O-Q; V-X; AC-AE; AJ-AL		
		Group SDs, n for each scale	K-N; R-U; Y-AB; AF-AI; AM-AP		
Multiplicative – Calculate the MD for log-transformed measurements	Log	Sample size (n)	B; O	Mean difference and SE for log-transformed measurements	BZ; CA
		CFB mean, SD of log-transformed data	C-D; P-Q		
		Post-treatment mean, SD of log-transformed data	E-F; R-S		
		CFB mean, SD of raw data	G-H; T-U		
		Post-treatment mean, SD of raw data	I-J; V-W		
		Geometric mean and confidence interval limits[d]	K-N; X-AA		
		Mean difference, SE based on log-transformed data	AB-AC		
		Overall (grand) mean based on raw data	AD		
		Mean difference, SE based on raw data	AE-AF		

Procedure	Worksheet	Data Input		Data Output	
		Statistics	Column(s)	Statistics	Column(s)
Within-trial synthesis of MDs of log transformed measurements	Within-trial synthesis_MDs_log	Correlation between raw or log-transformed measurements on scales	A-E	Composite mean difference of log-transformed measurements and SE	BH; BJ
		MD of log-transformed measurements, SE for each scale	H-I; P-Q; X-Y; AF-AG; AN-AO		
		Group means of raw measurements (if correlations are between raw measurements)	J, M, R, U, Z, AC, AH, AK, AP, AS		
		Group SDs of raw measurements (if correlations are between raw measurements) OR log-transformed measurements (if correlations are between log-transformed measurements)	K, N, S, V, AA, AD, AI, AL, AQ, AT		
		Group sample size (n) for each scale	L, O, T, W, AB, AE, AJ, AM, AR, AU		
Multiplicative - Calculate the log(RoM) and SE in each trial	RoM	Correlation between baseline and post-treatment measurements[c]	B	<u>If arm-level CFB data inputted:</u> $\log(RoM_{F:B})$ and SE <u>Otherwise:</u> log(RoM) and SE	U-V
		Sample size (n)	C; H		
		Baseline mean, SD	D-E; I-J		
		Post-treatment mean, SD	F-G; K-L		
		log(RoM), SE	M-N		
Within-trial synthesis of RoMs	Within-trial synthesis_RoMs	Correlation between scales	A-E	Composite log(RoM) and SE	BH; BJ
		log(RoMs), SE	H-I; P-Q; X-Y; AF-AG; AN-AO		
		Group means, SDs, n for each scale	J-O; R-W; Z-AE; AH-AM; AP-AU		
Multiplicative – Calculate ratios of mean percentage change from baseline data	% change	Mean percentage CFB, SD, n	B-G	log(RoM) and SE	I-J

Procedure	Worksheet	Data Input		Data Output	
		Statistics	Column(s)	Statistics	Column(s)
Converting log odds ratio, or proportions responding on each arm, to an MD and its SE	Proportion of Responders	Proportion of responders, n	B-E	Rescaled MD and SE	P-Q
		log odds ratio, variance	F-G		
		Scale name and reference SD	I-J		

[a]Code as determined by user. This will prompt the worksheet to look up the reference SD listed in column C (External Rescaling worksheet) or column P (Internal Rescaling worksheet).

[b]The internal reference SDs will have to be inputted one-scale-at-a-time, as the user iteratively inputs baseline SDs and sample size from all included trials that report data on the same scale into columns B-E.

[c]Only required if inputting baseline and post-treatment data.

[d]Default confidence level is 0.95, but can be changed by user.

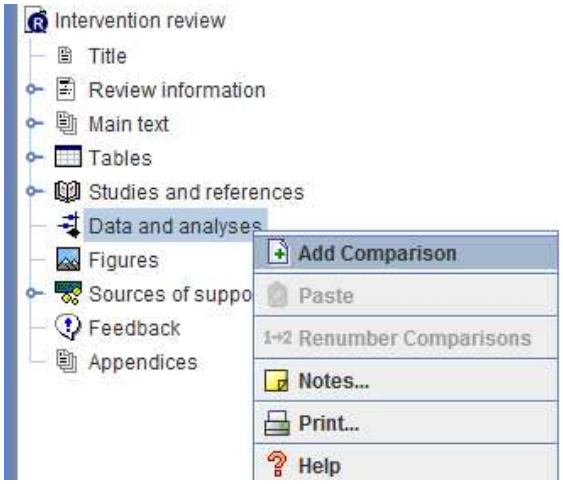
Abbreviations: CFB – change from baseline, RoM – ratio of means, $RoM_{F:B}$ – ratio of the post-treatment to baseline mean ratios, SD – standard deviation, SE – standard error

SOFTWARE APPENDIX

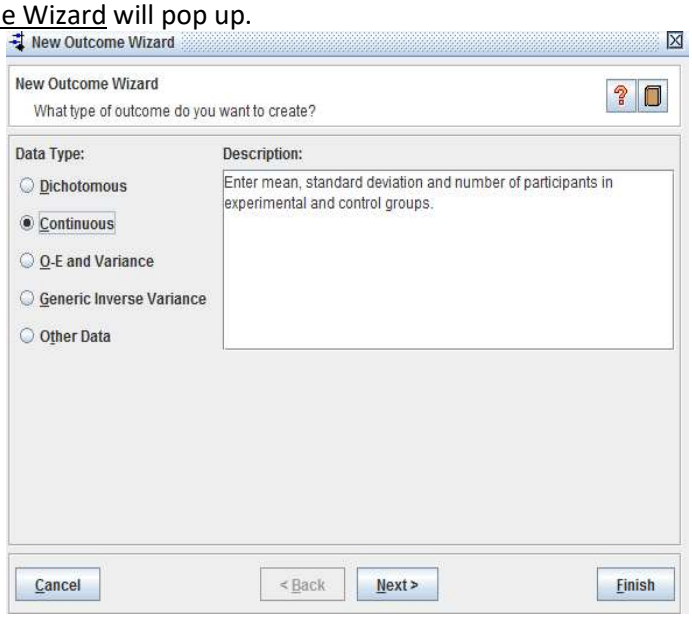
1. Inputting arm-based continuous data into Review Manager (RevMan) 5.3

Steps to import data:

- Right click 'Data and analyses'.
- Select 'Add Comparison'.



- The New Comparison Wizard will pop up.
  i. Under 'What name should the comparison have?' enter the treatments being compared (e.g., "LABA/LAMA vs. LAMA").
  ii. Click 'Next >'.
- Under 'What do you want to do after the wizard is closed?', select 'Add an outcome under the new comparison'. Then click 'Continue'.
- The New Outcome Wizard will pop up.

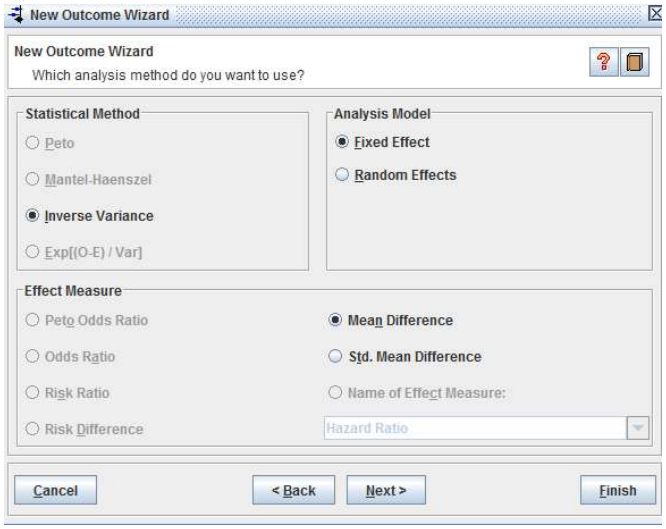


  i. Select the appropriate Data Type.
     - In this example, we are entering means and standard deviations (SDs) for each arm, so we will select 'Continuous'.
  ii. Click 'Next >'.
  iii. For 'Name', enter the outcome.
       For example: 'Change from baseline in FEV1 at 3 months'.
  iv. Enter the treatment names. E.g.:

- Group Label 1: LABA/LAMA
- Group Label 2: LAMA

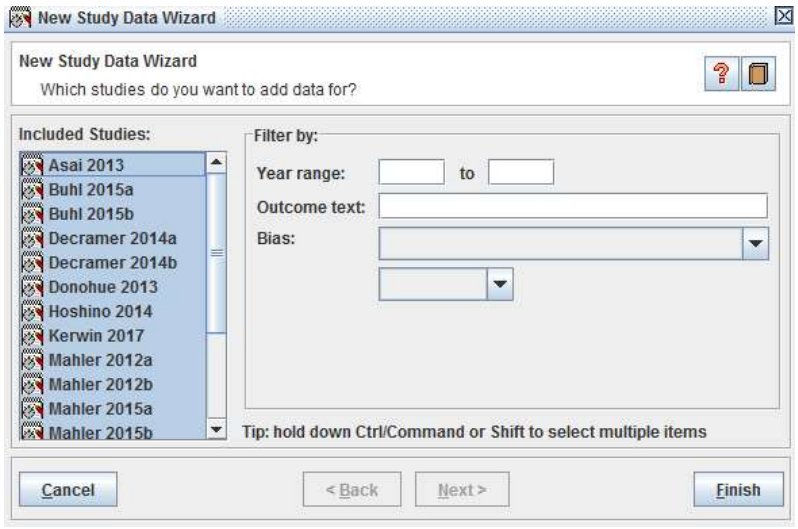
v. Click ‘Next >’.

- The Wizard will prompt you to specify which analysis method you would like to use.
  - Different methods will be available depending on the data type you specified previously. For our example, the default selections are what we want.

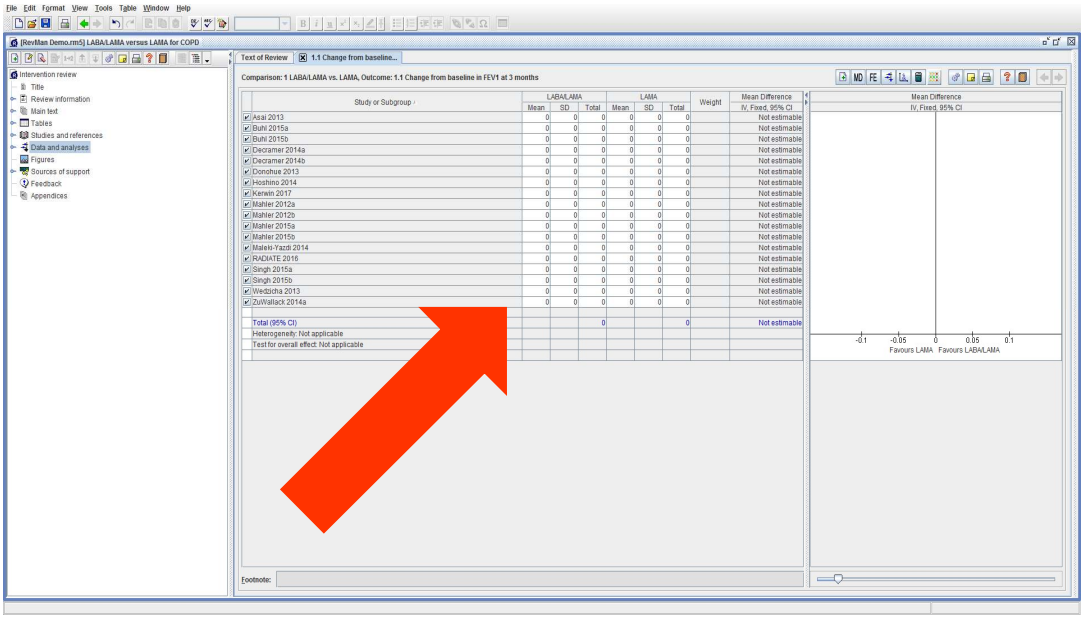


- Click ‘Next >’.
- Under ‘Which analysis details do you want to use?’, select desired options.
- Click ‘Next >’.
- Under ‘Which graph details do you want to use?’, enter:
  i. Left Graph Label: E.g., ‘Favours LAMA’
  ii. Right Graph Label: E.g., ‘Favours LABA/LAMA’

  (Note this depends on whether the outcome is desirable or not.)
- Click ‘Next >’.
- Under ‘What do you want to do after the wizard is closed?’, select ‘Add study data for the new outcome’.
- Click ‘Continue’.
- Select all of the studies under ‘Included Studies’
  - Tip: Ctrl + A
- Click ‘Finish’.



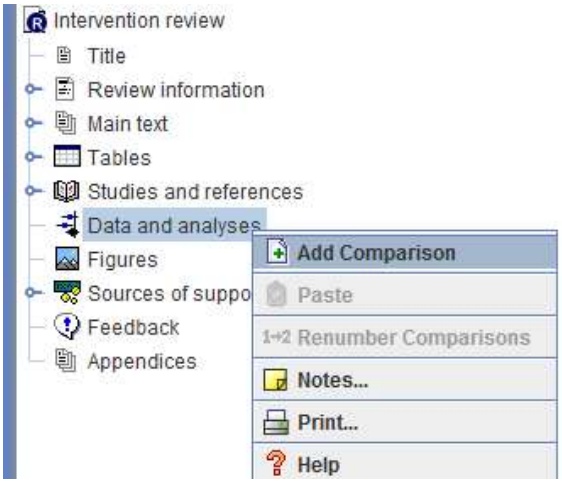
- Copy the data from Excel and paste it into RevMan.
  - Tip: Make sure data columns between the programs match!


File Edit Format View Tools Table Window Help
[RevMan Demo.rm5] LABA/LAMA versus LAMA for COPD
Intervention review
Title
Review information
Main text
Tables
Studies and references
Data and analyses
Figures
Sources of support
Feedback
Appendices
Text of Review
1.1 Change from baseline...
Comparison: 1 LABA/LAMA vs. LAMA, Outcome: 1.1 Change from baseline in FEV1 at 3 months
Study or Subgroup
LABA/LAMA
LAMA
Weight
Mean Difference
IV, Fixed, 95% CI
Mean
SD
Total
Asai 2013
Buhl 2015a
Buhl 2015b
Decramer 2014a
Decramer 2014b
Donohue 2013
Hoshino 2014
Kerwin 2017
Mahler 2012a
Mahler 2012b
Mahler 2015a
Mahler 2015b
Maleki-Yazdi 2014
RADIATE 2016
Singh 2015a
Singh 2015b
Wedzicha 2013
ZuWallack 2014a
Not estimable
Total (95% CI)
Heterogeneity: Not applicable
Test for overall effect: Not applicable
-0.1
-0.05
0
0.05
0.1
Favours LAMA
Favours LABA/LAMA
Footnote:

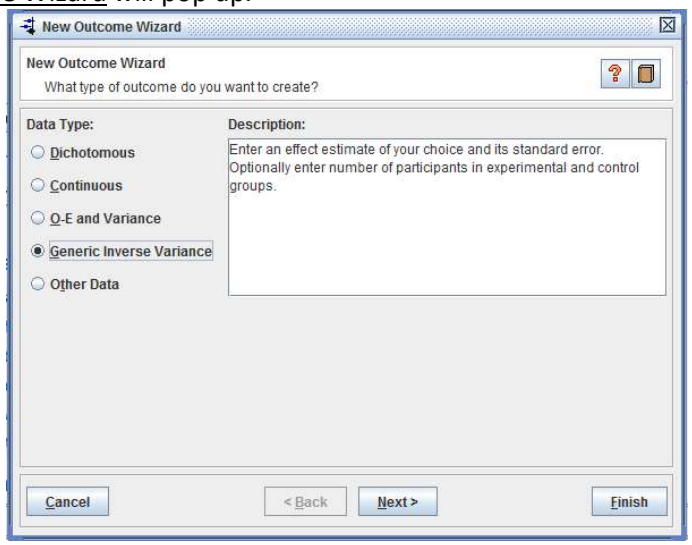
2. Inputting contrast-based continuous data into Review Manager (RevMan) 5.3

Steps to import data:

1. Right click 'Data and analyses'.
2. Select 'Add Comparison'.



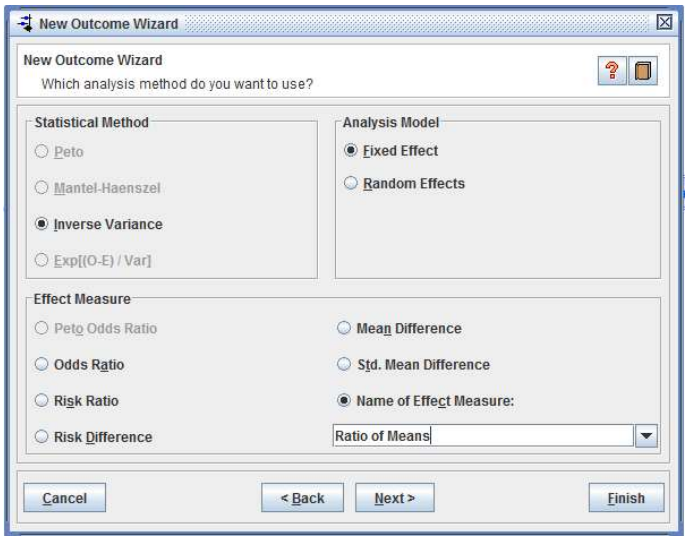
3. The New Comparison Wizard will pop up.
   iii. Under 'What name should the comparison have?' enter the treatments being compared (e.g., "LABA/LAMA vs. LAMA").
   iv. Click 'Next >'.
4. Under 'What do you want to do after the wizard is closed?', select 'Add an outcome under the new comparison'. Then click 'Continue'.
5. The New Outcome Wizard will pop up.



   i. Select the appropriate Data Type.
      - In this example, we are entering log(ratio of means) and standard errors (SEs) for each study, so we will select 'Generic Inverse Variance'.
   ii. Click 'Next >'.
   iii. For 'Name', enter the outcome.
      For example: 'Change from baseline in FEV1 at 3 months'.
   vi. Enter the treatment names. E.g.:
      - Group Label 1: LABA/LAMA
      - Group Label 2: LAMA
   vii. Click 'Next >'.

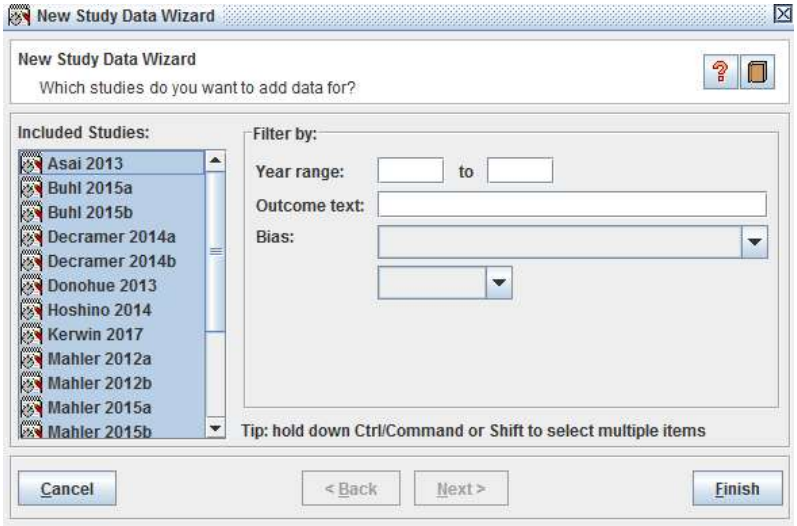
- The Wizard will prompt you to specify which analysis method you would like to use.

- Different methods will be available depending on the data type you specified previously. For our example, the default selections are what we want, and we specify the effect measure to be “Ratio of Means”

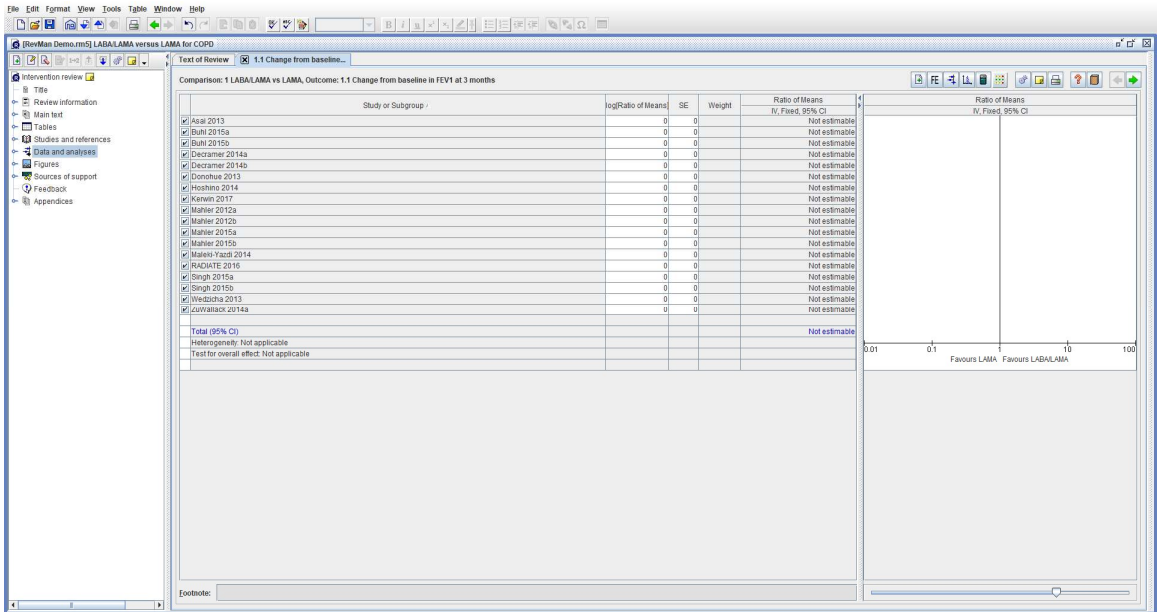


- Click ‘Next >’.
- Under ‘Which analysis details do you want to use?’, check ‘Entered data are on log scale (Generic Inverse Variance only)’ if appropriate (e.g., if entering log(RoM)).
- Click ‘Next >’.
- Under ‘Which graph details do you want to use?’, enter:
    i. Left Graph Label: E.g., ‘Favours LAMA’
    ii. Right Graph Label: E.g., ‘Favours LABA/LAMA’

  (Note this depends on whether the outcome is desirable or not.)
- Click ‘Next >’.
- Under ‘What do you want to do after the wizard is closed?’, select ‘Add study data for the new outcome’.
- Click ‘Continue’.
- Select all of the studies under ‘Included Studies’
    - Tip: Ctrl + A
- Click ‘Finish’.



- Copy the data from Excel and paste it into RevMan.
    - Tip: Make sure data columns between the programs match!


File Edit Format View Tools Table Window Help
[RevMan Demo.rm5] LABA/LAMA versus LAMA for COPD
Intervention review
Title
Review information
Main text
Tables
Studies and references
Data and analyses
Figures
Sources of support
Feedback
Appendices
Text of Review
1.1 Change from baseline...
Comparison: 1 LABA/LAMA vs LAMA, Outcome: 1.1 Change from baseline in FEV1 at 3 months
Study or Subgroup
log[Ratio of Means]
SE
Weight
Ratio of Means
IV, Fixed, 95% CI
Asai 2013 0 0 Not estimable
Buhl 2015a 0 0 Not estimable
Buhl 2015b 0 0 Not estimable
Decramer 2014a 0 0 Not estimable
Decramer 2014b 0 0 Not estimable
Donohue 2013 0 0 Not estimable
Hoshino 2014 0 0 Not estimable
Kerwin 2017 0 0 Not estimable
Mahler 2012a 0 0 Not estimable
Mahler 2012b 0 0 Not estimable
Mahler 2015a 0 0 Not estimable
Mahler 2015b 0 0 Not estimable
Maleki-Yazdi 2014 0 0 Not estimable
RADIATE 2016 0 0 Not estimable
Singh 2015a 0 0 Not estimable
Singh 2015b 0 0 Not estimable
Wedzicha 2013 0 0 Not estimable
Zuwallack 2014a 0 0 Not estimable
Total (95% CI) Not estimable
Heterogeneity: Not applicable
Test for overall effect: Not applicable
Ratio of Means
IV, Fixed, 95% CI
0.01
0.1
1.0
100
Favours LAMA
Favours LABA/LAMA
Footnote:

3. Inputting arm-based continuous data into R

The following code delivers a fixed effect meta-analysis on the mean difference scale where data are inputted as an Excel file where the columns contain the study IDs ("studyid"), treatment group 1's mean ("m1"), standard deviation ("sd1"), and sample size ("n1"), and treatment group 2's mean ("m2"), standard deviation ("sd2"), and sample size ("n2") (Figure 1). Output from R is displayed in blue. The forest plot produced by the code applied to the data in Figure 1 is given in Figure 2.

```
#Set working directory to be folder that contains relevant files
setwd(c:/ '')

##load metafor package
library(metafor)
##load xlsx package
library(xlsx)

########################
#### Arm-based data ######
########################

# Load data from the "Depression (BDI)_arm" worksheet in an Excel file entitled "metafor example data.xlsx"
dat<-read.xlsx(file="metafor example data.xlsx",sheetName="Depression (BDI)_arm", header=TRUE)

# Preview a maximum of the first 6 rows of the data
head(dat)

# Fit a Fixed Effect model
model.arm <- rma(m1i=m1, m2i=m2, # specify group means
        sd1i=sd1, sd2i=sd2,  # specify group SDs
        n1=n1, n2=n2,  # specify group sample sizes
        measure="MD",  # specify effect measure (MD = mean difference, SMD = standardised mean difference (Hedges' g))
        data=dat,     # specify data
        slab=studyid, # specify column name containing study labels
        method="FE")  # specify you want to fit a fixed effect ("FE")

# Display the MA results
model.arm

 # R displays a summary of the meta-analysis model. It first notes that a fixed effect model was fitted, where k = 3 studies were included.
 # The results of a Q-test for heterogeneity are displayed, followed by the pooled summary results (estimate = 6.92, CI = (4.58, 9.25)).
 # Fixed-Effects Model (k = 3)
 #
 # I^2 (total heterogeneity / total variability):   0.00%
 # H^2 (total variability / sampling variability):  0.65
 #
 # Test for Heterogeneity:
 #   Q(df = 2) = 1.3081, p-val = 0.5199
 #
 # Model Results:
 #
 #  estimate     se    zval    pval   ci.lb   ci.ub
 #    6.9162  1.1908  5.8082  <.0001  4.5823  9.2501  ***
 #
 #  ---
 #  Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Create a forest plot
forest(x=model.arm,              # specify label given to fitted MA model (e.g., "model")
     xlim=c(-40,35),        # specify horizontal limits of plot
     alim=c(-10,15),          # specify actual x-axis limits
     showweights=TRUE,       # Tell R to include study weights
     at=c(-10,-5,0,5,10,15),   # Specify tick marks on x-axis
     ilab=round(cbind(dat$m1,dat$sd1,dat$n1,dat$m2,dat$sd2,dat$n2),1),        #specify column names of study data, and round values to 1 decimal place
     ilab.xpos=c(-28,-25,-22,-19,-16,-13),    # specify location of m1, sd1, n1, m2, sd2, n2
     digits=1,               # specify number of decimal places of x-axis labels and effect measures
     refline=0)              # specify null value to draw reference line at

# Add titles to forest plot
text(x=c(-28,-25,-22,-19,-16,-13),   # horizontal position of labels
    y=4.25,                  # vertical position of labels
    c("m1","sd1","n1","m2","sd2","n2"))   # labels
text(x=-40,y=4.25,"Study",pos=4)    # pos=4 indicates to position text to right of x
```

```
text(x=35,y=4.25,"MD [95% CI]",pos=2)  # pos=2 indicates to position text to left of x
text(x=26,y=4.25,"Weight",pos=2)
```

	A	B	C	D	E	F	G	H	I
1	studyid	m1	sd1	n1	m2	sd2	n2		
2	Hassiotis 2013	18.46	12.96	13	16.79	12.38	14		
3	McCabe 2006	12.80	4.23	15	5.71	4.54	34		
4	McGillivray 2008	16.51	13.81	27	8.45	6.69	20		
5									

Figure 1. Excel spreadsheet containing arm-level data for three studies comparing CBT intervention to a control in terms of reducing depression scores on the Beck Depression Index (BDI) scale. This structure of the data is required to run the R code provided above.

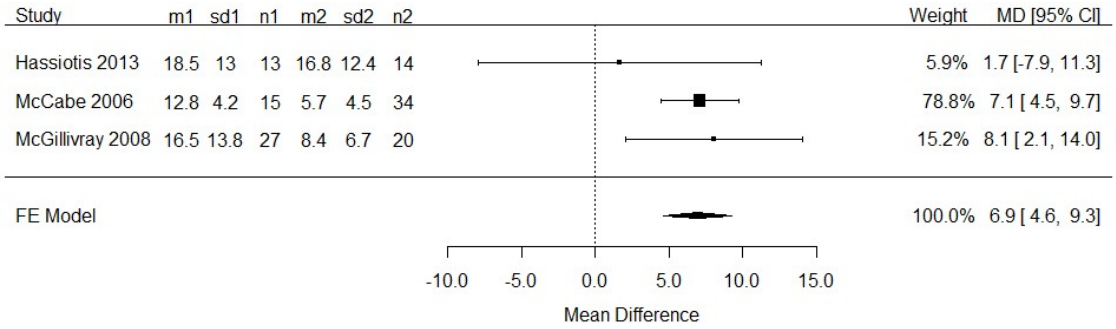


Figure 2. Meta-analysis of arm-level data from studies using the same scale to measure the outcome: CBT vs. control for depressive symptoms (measured with Beck Depression Index, BDI). This forest plot was produced using the R code provided above.

4. Inputting contrast-based continuous data into R

The following code delivers a fixed effect meta-analyis on the mean difference scale where data are inputted as an Excel file where the columns contain the study IDs ("studyid"), the mean differences in each study ("md") and their corresponding variance ("v") and standard error ("se") (Figure 3). Output from R is displayed in blue. The forest plot produced by the code applied to the data in Figure 3 is given in Figure 4.

```
#Set working directory to be folder that contains relevant files
setwd("C:/")

##load metafor package
library(metafor)
##load xlsx package
library(xlsx)

##############################
##### Contrast-based data #####
##############################

# Load data from the "Depression (BDI)_contrast" worksheet in an Excel file entitled "metafor example data.xlsx"
dat <- read.xlsx(file="metafor example data.xlsx",sheetName="Depression (BDI)_contrast", header=TRUE)

# Preview a maximum of the first 6 rows of the data
head(dat)

# Fit a Fixed Effect model
model<-rma(yi=md,          # specify mean differences
        sei=se,          # specify standard error of mean difference
        measure="MD",    # specify effect measure (MD = mean difference, SMD = standardised mean difference (Hedges' g))
        data=dat,        # specify data
        slab=studyid,    # specify column name containing study labels
        method="FE")     # specify you want to fit a fixed effect ("FE")

# MA results
model

 # R displays a summary of the meta-analysis model. It first notes that a fixed effect model was fitted, where k = 3 studies were included.
 # The results of a Q-test for heterogeneity are displayed, followed by the pooled summary results (estimate = 6.92, CI = (4.58, 9.25)).
 # Fixed-Effects Model (k = 3)
 #
 # I^2 (total heterogeneity / total variability):   0.00%
 # H^2 (total variability / sampling variability):  0.65
 #
 # Test for Heterogeneity:
 #   Q(df = 2) = 1.3081, p-val = 0.5199
 #
 # Model Results:
 #
 #  estimate     se    zval    pval   ci.lb   ci.ub
 #    6.9162  1.1908  5.8081  <.0001  4.5823  9.2501  ***
 #
 # ---
 # Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

# Create forest plot
forest(x=model,              # specify label given to fitted MA model (e.g., "model")
      xlim=c(-20,30),        # specify horizontal limits of plot
      alim=c(-10,15),        # specify actual x-axis limits
      showweights=TRUE,      # tell R to include study weights
      at=c(-10,-5,0,5,10,15),  # specify tick marks on x-axis
      digits=1,              # specify number of decimal places of x-axis labels and effect measures
      refline=0)             # specify null value to draw reference line at

# Add titles to forest plot
text(x=-20,y=4.25,"Study",pos=4)       # pos=4 indicates to position text to right of x
text(x=30,y=4.25,"MD [95% CI]",pos=2)  # pos=2 indicates to position text to left of x
text(x=23.5,y=4.25,"Weight",pos=2)
```

	A	B	C	D	E	F
1	studyid	md	v	se		
2	Hassiotis 2013	1.68	23.87	4.89		
3	McCabe 2006	7.09	1.80	1.34		
4	McGillivray 2008	8.06	9.30	3.05		
5						
6						

Figure 3. Excel spreadsheet containing contrast-level data for three studies comparing CBT intervention to a control in terms of reducing depression scores on the Beck Depression Index (BDI) scale. This structure of the data is required to run the R code provided above.

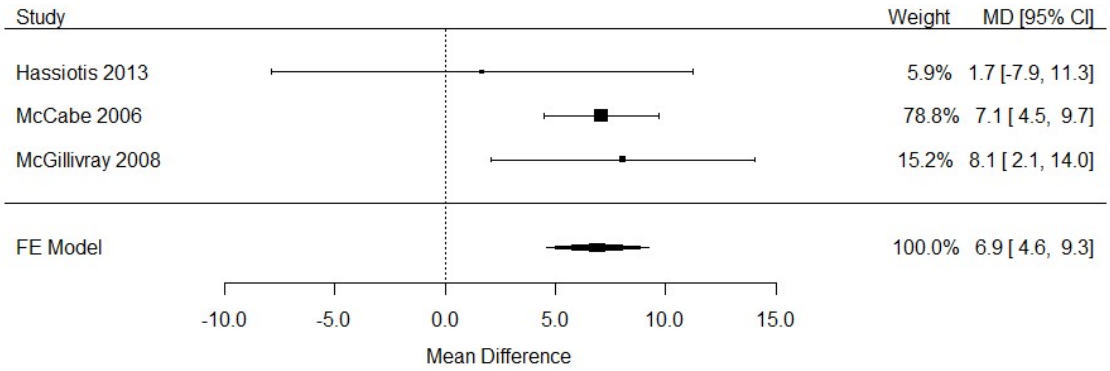


Figure 4. Meta-analysis of contrast-level data from studies using the same scale to measure the outcome: CBT vs. control for depressive symptoms (measured with Beck Depression Index, BDI). This forest plot was produced using the R code provided in above.

5. RevMan calculator for calculating contrast-based data from arm-based data

The mean difference and standard error for a Contrast-Based meta-analysis may be computed from Arm-Based data using the built-in calculator in RevMan, as illustrated for McGillivray 2008 (59) in Figure 5.

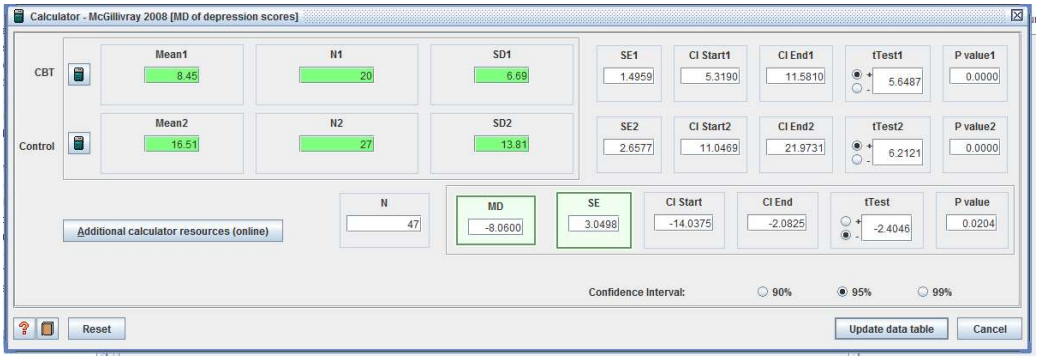


Figure 5. Calculator in RevMan for continuous outcomes

REFERENCES

1.National Institute for Health and Care Excellence. Guide to the methods of technology appraisal. London; 2013.

2.National Institute for Health and Care Excellence. Developing NICE guidelines: the manual. London; 2014.

3.Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. 2011.

4.Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 2: A generalised linear modelling framework for pair-wise and network meta-analysis of randomised controlled trials. 2011.

5.Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011.

6.Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. NICE DSU Technical Support Document 4: Inconsistency in networks of evidence based on randomised controlled trials. 2011.

7.Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 5: Evidence synthesis in the baseline natural history model. 2011.

8.Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 6: Embedding evidence synthesis in probabilistic cost-effectiveness analysis: Software choices. 2011.

9.Ades AE, Caldwell DM, Reken S, Welton NJ, Sutton AJ, Dias S. NICE DSU Technical Support Document 7: Evidence synthesis of treatment efficacy in decision making: a reviewer's checklist. 2012.

10.Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. Chichester: Wiley; 2009.

11.Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al. Cochrane handbook for systematic reviews of interventions. 2nd ed. Chichester: John Wiley & Sons; 2019.

12.Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for meta-analysis in medical research. London: Wiley; 2000.

13.Borenstein M. Effect sizes for continuous data. In: Cooper H, Hedges LV, Valentine JC, editors. The handbook of research synthesis and meta-analysis. 2nd ed. New York, NY, US: Russell Sage Foundation; 2009. p. 221-35.

14.Fu R, Vandermeer BW, Shamliyan TA, O'Neil ME, Yazdi F, Fox SH, et al. Handling continuous outcomes in quantitative synthesis. Rockville: Agency for Healthcare Research and Quality; 2013.

15.Higgins JPT, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: Combining results reported on log-transformed or raw scales. Statistics in Medicine. 2008;27(29):6072-92.

16.Friedrich JO, Adhikari NK, 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 Medical Research Methodology. 2008;8:32.

17.Friedrich JO, Adhikari NKJ, Beyene J. Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. Journal of Clinical Epidemiology. 2011;64(5):556-64.

18.Vickers AJ, Altman DG. Analysing controlled trials with baseline and follow up measurements. BMJ. 2001;323:1123-4.

19.Balk EM, Earley A, Patel K, Trikalinos TA, Dahabreh IJ. Empirical assessment of within-arm correlation imputation in trials of continuous outcomes. Rockville: Agency for Healthcare Research and Quality; 2012.

20.Fu R, Holmer HK. Change score or followup score? An empirical evaluation of the impact of choice of mean difference estimates. Rockville: Agency for Healthcare Research and Quality; 2015.

21.Price DD, Bush FM, Long S, Harkins SW. A comparison of pain measurement characteristics of mechanical visual analogue and simple numerical rating scales. Pain. 1994;56(2):217-26.

22.Hjermstad MJ, Fayers PM, Haugen DF, Caraceni A, Hanks GW, Loge JH, et al. Studies comparing numerical rating scales, verbal rating scales, and visual analogue scales for assessment of pain intensity in adults: A systematic literature review. Journal of pain and symptom management. 2011;41(6):1073-93.

23.White IR, Thomas J. Standardized mean differences in individually-randomized and cluster-randomized trials, with applications to meta-analysis. Clinical Trials. 2005;2(2):141-51.

24.Elbourne DR, Altman DG, Higgins JPT, Curtin F, Worthington HV, Vail A. Meta-analyses involving cross-over trials: Methodological issues. Internation Journal of Epidemiology. 2002;31:140-9.

25.Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic Press; 1969 1969.

26.Hedges LV. Distribution theory for Glass's estimator of effect size and related estimators. Journal of Educational Statistics. 1981;6:107-28.

27.Light RJ, Pillemer DB. Summing up: the science of reviewing research. Cambridge Mass: Harvard University Press; 1984.

28.Greenland S, Schlesselman J, Criqui M. The fallacy of employing standardized coefficients and correlations as measures of effect. American Journal of Epidemiology. 1986;123:203-8.

29.Greenland S, Maclure M, Schlesselman J, Poole C, Morgenstern H. Standardized regression coefficients: a further critique and review of some alternatives. Epidemiology and Infection. 1991;2(5):387-92.

30.Rothman KJ, Greenland S, Lash TL. Modern epidemiology. 3rd ed. Philadelphia: Lippincott, Williams & Wilkins; 2012.

31.Deeks JJ, Higgins JPT, Altman DG (editors). Chapter 10: Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, et al., editors. Cochrane handbook for systematic reviews of interventions. 2nd ed. Chichester: John Wiley & Sons; 2019. p. 241-84.

32.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. Research Synthesis Methods. 2011;2:188-203.

33.Hunter JE, Schmidt FL. Methods of meta-analysis: correcting error and bias in research findings. 2nd ed. London: Sage Publications; 2004.

34.Wei Y, Higgins J. Estimating within-study covariances in multivariate meta-analysis with multiple outcomes. Statistics in Medicine. 2013;32:1191-205.

35.Borenstein M, Hedges LV, Higgins JP, Rothstein HR. Multiple outcomes or time-points within a study. In: Borenstein M, Hedges LV, Higgins JPT, Rothstein HR, editors. Introduction to meta-analysis. Chichester: Wiley; 2009. p. 225-38.

36.Bowling A. Measuring health: A review of quality of life measurement scales. 3rd ed. Maidenhead, UK: Open University; 2004. 224 p.

37.Bowling A. Measuring disease: A review of disease-specific quality of life measurement scales. 2nd ed. Buckingham, UK: Open University; 2001. 390 p.

38.McDowell I. Measuring health: A guide to rating scales and questionnaires. 3rd ed. New York: Oxford University Press; 2006.

39.Baer L, Blais MA. Handbook of clinical rating scales and assessment in psychiatry and mental health. New York: Humana Press; 2010.

40.Birks JS, Harvey RJ. Donepezil for dementia due to Alzheimer's disease. Cochrane Database of Systematic Reviews. 2018;6.

41.Kelaiditi E, Andrieu S, Cantet C, Vellas B, Cesari M. Frailty index and incident mortality, hospitalization, and institutionalization in Alzheimer's disease: Data from the ICTUS study. The Journals of Gerontology (A). 2016;71(4):543-8.

42.Stuhec M, Munda B, Svab V, Locatelli I. Comparative efficacy and acceptability of atomoxetine, lisdexamfetamine, bupropion and methylphenidate in treatment of attention deficit hyperactivity disorder in children and adolescents: A meta-analysis with focus on bupropion. Journal of affective disorders. 2015;178:149-59.

43.Wei L, Zhang J. Analysis of data with imbalance in the baseline outcome variable for randomized clinical trials. Drug Information Journal. 2001;35(4):1201-14.

44.Senn S. Meta-analysis. In: Senn S, editor. Statistical issues in drug development. Chichester: John Wiley & Sons, Ltd.; 2007. p. 251-72.

45.McKenzie JE, Herbison GP, Deeks JJ. Impact of analysing continuous outcomes using final values, change scores and analysis of covariance on the performance of meta-analytic methods: A simulation study. Research Synthesis Methods. 2016;7(4):371-86.

46.Dorans NJ, Pommerich M, Holland PW, editors. Linking and aligning scores and scales. New York: Springer; 2007.

47.Kolen MJ, Brennan RL. Test equating, scaling and linking: methods and practices. New York: Springer; 1994.

48.Hambleton R, Swamninathan H, Rogers HJ. Fundamentals of item response theory. Newbury Park, CA: Sage Press; 1991.

49.Reeve BB. Item response theory modeling in health outcomes measurement. Expert review of pharmacoeconomics & outcomes research. 2003;3(2):131-45.

50.Hays RD, Morales LS, Reise SP. Item response theory and health outcomes measurement in the 21st century. Medical care. 2000;38(9 Suppl):II28-42.

51.Guilera G, Gomez J. Item response theory test equating in health sciences education. Advances in Health Sciences Education. 2008;13(1):3-10.

52.Wahl I, Lowe B, Bjorner JB, Fischer F, Langs G, Voderholzer U, et al. Standardization of depression measurement: A common metric was developed for 11 self-report depression measures. Journal of Clinical Epidemiology. 2014;67(1):73-86.

53.Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, 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. Biological psychiatry. 2003;54(5):573-83.

54.Uher R, Farmer A, Maier W, Rietschel M, Hauser J, Marusic A, et al. Measuring depression: comparison and integration of three scales in the GENDEP study. Psychological medicine. 2008;38(2):289-300.

55.Carmody TJ, Rush AJ, Bernstein I, Warden D, Brannan S, Burnham D, et al. The Montgomery Asberg and the Hamilton ratings of depression: a comparison of measures. European Neuropsychopharmacology. 2006;16(8):601-11.

56.Leucht S, Fennema H, Engel RR, Kaspers-Janssen M, Szegedi A. Translating the HAM-D into the MADRS and vice versa with equipercentile linking. Journal of affective disorders. 2018;226:326-31.

57.Leucht S, Rothe P, Davis JM, Engel RR. Equipercentile linking of the BPRS and the PANSS. European Neuropsychopharmacology. 2013;23(8):956-9.

58.Samara MT, Engel RR, Millier A, Kandenwein J, Toumi M, Leucht S. Equipercentile linking of scales measuring functioning and symptoms: examining the GAF, SOFAS, CGI-S, and PANSS. European Neuropsychopharmacology. 2014;24(11):1767-72.

59.McGillivray JA, McCabe MP, Kershaw MM. Depression in people with intellectual disability: An evaluation of a staff-administered treatment program. Research in Developmental Disabilities. 2008;29(6):524-36.