Draft Methods Guide for Comparative Effectiveness Reviews

Number XX

Handling Continuous Outcomes in Quantitative Synthesis

Prepared for:
Agency for Healthcare Research and Quality
U.S. Department of Health and Human Services
540 Gaither Road
Rockville, MD 20850
www.ahrq.gov

Contract No. xxx-xx-xxxx

Prepared by:
xxxx

Investigators:
[Deleted for external review]

AHRQ Publication No. xx-EHCxxx
<Month Year>

Comparative Effectiveness Reviews are systematic reviews of existing research on the effectiveness, comparative effectiveness, and harms of different health care interventions. They provide syntheses of relevant evidence to inform real-world health care decisions for patients, providers, and policymakers. Strong methodologic approaches to systematic review improve the transparency, consistency, and scientific rigor of these reports. Through a collaborative effort of the Effective Health Care (EHC) Program, the Agency for Healthcare Research and Quality (AHRQ), the EHC Program Scientific Resource Center, and the AHRQ Evidence-based Practice Centers have developed a Methods Guide for Comparative Effectiveness Reviews. This Guide presents issues key to the development of Comparative Effectiveness Reviews and describes recommended approaches for addressing difficult, frequently encountered methodological issues.

The Methods Guide for Comparative Effectiveness Reviews is a living document, and will be updated as further empiric evidence develops and our understanding of better methods improves. Comments and suggestions on the Methods Guide for Comparative Effectiveness Reviews and the Effective Health Care Program can be made at http://www.effectivehealthcare.ahrq.gov/.

Persons using assistive technology may not be able to fully access information in this report. For assistance contact info@ahrq.gov.

None of the investigators has any affiliations or financial involvement that conflicts with the material presented in this report.

Suggested citation: <Authors>. <Topic>. Methods research report. No. <#>. (Prepared by <EPC Name> under Contract No. <##>.) Rockville, MD: Agency for Healthcare Research and Quality. <Month, Year>. www.effectivehealthcare.ahrq.gov/reports/final.cf. Accessed: <Date>. <URL>.

Preface

The Agency for Healthcare Research and Quality (AHRQ), through its Evidence-based Practice Centers (EPCs), sponsors the development of evidence reports and technology assessments to assist public- and private-sector organizations in their efforts to improve the quality of health care in the United States. The reports and assessments provide organizations with comprehensive, science-based information on common, costly medical conditions and new health care technologies and strategies. The EPCs systematically review the relevant scientific literature on topics assigned to them by AHRQ and conduct additional analyses when appropriate prior to developing their reports and assessments.

To improve the scientific rigor of these evidence reports, AHRQ supports empiric research by the EPCs to help understand or improve complex methodologic issues in systematic reviews. These methods research projects are intended to contribute to the research base in and be used to improve the science of systematic reviews. They are not intended to be guidance to the EPC program, although may be considered by EPCs along with other scientific research when determining EPC program methods guidance.

AHRQ expects that the EPC evidence reports and technology assessments will inform individual health plans, providers, and purchasers as well as the health care system as a whole by providing important information to help improve health care quality. The reports undergo peer review prior to their release as a final report.

We welcome comments on this Methods Research Project. They may be sent by mail to the Task Order Officer named below at: Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850, or by e-mail to epc@ahrq.hhs.gov.

Carolyn M. Clancy, M.D.
Director
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Director
Evidence-based Practice Program
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Jean Slutsky, P.A., M.S.P.H.
Director, Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Stephanie Chang, M.D., M.P.H.
Task Order Officer
Center for Outcomes and Evidence
Agency for Healthcare Research and Quality

Acknowledgments

The authors gratefully acknowledge the following individuals for their contributions to this project: <Acknowledgments>.

Key Informants

<Name>
<Place>
<City>, <ST>

Technical Expert Panel

<Name>
<Place>
<City>, <ST>

<Name>
<Place>
<City>, <ST>

Peer Reviewers

<Name>
<Place>
<City>, <ST>

<Name>
<Place>
<City>, <ST>

Contents

Tables

Figures

Appendixes

1. Introduction

In quantitative synthesis of randomized clinical trials for a comparative effectiveness review, continuous outcomes are usually less straightforward to analyze than are binary outcomes. The continuous outcomes are often measured at both baseline and followup time points. Results of continuous data are reported in many different forms as means, or mean differences or differences in change score from baseline, and measures of precision are reported as standard deviation (SD) or standard error (SE) or confidence intervals. The distribution of the data is not always symmetric and journal publications may not report all information that is required for a meta-analysis.

The original quantitative synthesis chapter[1] of the *Methods Guide for Effectiveness and Comparative Effectiveness Reviews* has a very brief section on continuous outcomes. It provides limited guidance on using mean difference versus standardized mean difference, but does not provide guidance of a number of issues relating to meta-analysis of continuous outcome. Evidence-based Practice Center investigators who use the *Guide* have noted the need for additional guidance in this area. Therefore the objective of this report is to update the section for quantitative synthesis of continuous outcomes.

This report addresses the following topics including effect measures of continuous outcomes; choice of estimates for mean difference and baseline imbalance; calculation of SD and SE; how to handle missing data and skewed data; use and interpretation of the standardized mean difference (SMD) and of the ratio of means (RoM) as an alternative measure; and dichotomization of continuous outcomes in meta-analyses.

For each of these topics, we searched the Effective Health Program’s methods database and as well as Ovid Medline, Current Index to Statistics, and Scopus for methodological papers (Appendix A). Recommendations for each topic were developed based on the current knowledge in the literature and group discussion and consensus.

2. Effect Measures for Continuous Outcomes

The two most often used measures for continuous outcomes are mean difference and standardized effect sizes. The choice of effect measure is determined primarily by the scale of the available data. Investigators can combine mean differences if multiple trials report results using the same or similar scales. Standardized mean difference is typically used when the outcome is measured using different scales. A recently proposed measure, ratio of means,[2, 3] is an alternative to SMD for outcomes measured using different scales and has an interpretation of evaluating percentage change of a continuous outcome. This report will focus mainly on estimates of mean difference and related statistics, and SMD and ratio of mean are discussed in detail in Section 7 and Section 8 of this report.

In randomized clinical trials, continuous outcomes are usually measured at both baseline and followup. There are several ways to calculate mean difference:

1. Use followup score only to calculate a mean difference between intervention groups;
2. Calculate the change score from baseline to followup for each intervention group and use the difference in change scores between the intervention groups as the effect measure;
3. Use the followup score as the dependent variable in an analysis of covariance model (ANCOVA) with the baseline score as an covariate and the estimated intervention difference from the model is used as the effect measure;

4. Use the change score from baseline to followup as the dependent variable in an ANCOVA with the baseline score as a covariate, and the estimated intervention difference from the model is used as the effect measure.

Options 3 and 4 are equivalent statistically in terms of estimating the effect measure, that is, the difference between the intervention groups. When the variance of the baseline score equals the variance of the followup score, an ANCOVA estimate is a weighted sum of the two estimates from options 1 and 2, and the weight is the correlation between baseline and followup score.[4] If the correlation is greater than 0.5, difference in change in score from option 2 has more weight; otherwise, difference between followup score has more weight. It is possible that the observed variance at baseline is very different from the variance of the followup score, and an ANCOVA estimate is not exactly a weighted sum of the two measures, however, the ANCOVA estimate usually lies between the estimates from options 1 and 2.

3. Choice of Estimate for Mean Difference and Baseline Imbalance

For a well-randomized randomized controlled trial (RCT), distribution of baseline characteristics should be similar between treatment groups. However, baseline imbalance often occurs for one or more baseline characteristic. This imbalance could be due to chance, especially in small trials,[6] though Senn (1989)[7] argued that balance does not improve with sample size. Selection bias due to inadequate randomization concealment is another reason for baseline imbalance.[8]

The imbalance of baseline scores is usually considered as part of quality rating but little attention has been paid to this in quantitative synthesis. A meta-analysis may have different results depending on whether or not we adjust for baseline imbalance.[9] Here we distinguish between two types of baseline variables. The first type of variables are the usual patient characteristics and important prognostic factors for the medical condition under study and the second type of variables are the baseline measurements of continuous variables that are specified as outcomes. Both types of variables should be incorporated in quality rating, and in terms of quantitative synthesis, the latter is more relevant.

Assessment of Baseline Balance

Should Investigators Assess Baseline Balance of Included Studies in Quantitative Synthesis?

As mentioned above, in most systematic reviews assessment of the baseline balance for both types of variables is a quality rating criterion. Imbalance of important prognostic factors and outcome variables may imply inadequate randomization and allocation concealment and lead to biased results. Quality should be downgraded if the imbalance of important prognostic factors and outcome variables is not achieved and addressed in the included studies.

In addition to quality rating, for the second type of variables, investigators should also assess the baseline balance for each continuous outcome and take any imbalance into consideration when conducting quantitative synthesis.

How to Assess Whether the Baseline Scores are Balanced?

Though opinion is divided,[10] use of statistical testing for baseline difference, for both types of variables, is generally not recommended for individual studies.[7, 11-15] It is argued that "it is a test of a null hypothesis that is known to be true,"[15] and it "assesses the probability of something having occurred by chance when we know that it did occur by chance."[13] Imbalance of important prognostic factors could have an impact on results and the unadjusted estimates could be biased, even if the statistical tests are not significant.

Current practices vary. In a study of published RCTs in leading medical journals, unadjusted estimates of treatment effects were reported more frequently than adjusted estimates.[16] Of the 110 included RCTs, 42 used statistical testing to compare baseline difference. In a systematic review, if an included study reported tests of homogeneity for baseline continuous outcomes, investigators should not judge the baseline distribution based on the p-value of these tests. Small trials often lack the power to detect a significant difference and large trials pick up many unimportant differences. While it is difficult to put down concrete criteria to determine balanced versus imbalanced distribution, the actual differences between baseline measurements, clinically important differences, and the direction of the imbalance are important considerations. If an imbalance favors the control group, the consequence of this imbalance may be less serious than an imbalance favoring the treatment group. The decision could be subjective and we recommend the conservative decision of imbalanced baseline scores when the decision is not readily clear cut.

If the baseline scores of the continuous outcome are not part of the baseline characteristics reported between groups, and the reported baseline characteristics were deemed to be comparable by the investigators, investigators should not automatically assume that the baseline scores of the continuous outcome are comparable. If possible, investigators should also consider how attrition may impact the baseline imbalances for the second type of variables in quantitative synthesis. For studies with high attrition, the baseline balance may not be maintained in the subsample with outcome data,[17] affecting the choice of estimates of mean difference (see discussion below). If baseline scores are not reported adequately to judge whether they are comparable, don't assume that they are and the study's quality should be downgraded.

If the baseline score imbalance is only by chance, meta-analysis of baseline score differences between treatment groups of included studies should provide a combined estimate close to zero (given no publication bias).[9] Investigators are encouraged to do such an analysis.

Choice of Estimate for Mean Difference

When the baseline scores are balanced, options 1, 2, or 3 would provide unbiased estimates and the ANCOVA approach (option 3) provides a more efficient estimator with more precision.[12, 18, 19] When the baseline scores are imbalanced, options 1 and 2 produce biased effect estimates. Option 1 simply ignores baseline imbalance; option 2, contrary to the common misconception, does not control for the baseline imbalance. The change score is negatively associated with the baseline score and patients with worse baseline score are more likely to experience a high change score (regression to the mean). Suppose that a trial has an intervention and a placebo group and the intervention group has worse baseline score. The treatment effect from the intervention will be underestimated using option 1) and overestimated using option 2).[20] The ANCOVA has been shown to be a better method to control for this imbalance and the estimates from ANCOVA are less biased. When baseline scores are correlated to followup scores, adjusting for baseline using ANCOVA has been shown to remove conditional bias in treatment group comparisons due to chance imbalances[7] and improve efficiency over unadjusted comparisons.[7, 19]

Choice of Estimate for Mean Difference When There is No or Only Minimal Baseline Imbalance

When there is no or only minimal baseline imbalance, we provide the following recommendations for the choice of estimates for mean difference:

1. Use an ANCOVA estimate if reported. It is an unbiased and more efficient estimator.

   When a study did not report ANCOVA estimates, it is possible to calculate them when the studies reported enough information including: 1) means and SDs at baseline and followup for both the intervention and the control groups; 2) means and SDs of change for both the intervention and the control groups; 3) sample size in both the intervention and the control groups. It is rare for the studies to report such detailed data, so this is usually not a practical choice.
2. If an ANCOVA estimate is not reported and the study directly reported or reported enough data to calculate mean difference based on both options 1) and 2), use the estimate with a smaller SE.

   Option 2), difference in change score produces a small SE when correlation between baseline and post treatment is high (> 0.5 when variance is equal at baseline and post intervention). Otherwise, Option 1), difference between post score produces a small SE. There is evidence to show that the correlation between baseline and post score is often greater than 0.5 [http://www.effectivehealthcare.ahrq.gov/ehc/products/344/1087/Correlation_Draft-Report_20120515.pdf].
3. If the study did not reported or reported enough data to calculate the mean difference based on both options 1) and 2), use the reported estimate or whichever estimate can be calculated from the reported data. Sometimes important data to allow the study to be included in the meta-analysis could be missing but may be imputed. Section 5 provides more guidance on handling such situations.
4. Since all options provide unbiased estimates, it is also appropriate if the investigators choose to use the same estimate across studies. Since ANCOVA estimates are usually not consistently reported, practically this applies to options 1) and options 2). In such cases, some assumptions about missing data are usually needed to obtain an estimate of the same effect measure for all studies. For example, if change score between baseline and followup needs to be calculated, the correlation between baseline and the followup score is often not known and an assumption about the correlation is needed to calculate the SE of change score. See Section 5 for more information about handling missing data.

Choice of Estimate for Mean Difference When There is Baseline Imbalance

When there is baseline imbalance, ANCOVA estimates are preferred as they provide the least unbiased estimate with more precision. Options 1) and 2) would provide biased estimates. However, studies that are otherwise appropriate for inclusion but lack ANCOVA estimates should be not excluded from the quantitative synthesis. Such bias is usually not as serious as the bias caused by excluding these studies from the quantitative synthesis. For the choice of estimates for mean difference for each study, we provide the following recommendations:

1. Use ANCOVA estimates if reported (more precision and less bias)
2. If ANCOVA estimates were not reported, between options 1) and 2), the investigators choose an estimate based on the magnitude of correlation in the primary analysis: if the correlation is > 0.5, option 2) may be more likely to provide an estimate close to the ANCOVA estimate; otherwise, use option 1). Then a sensitivity analysis must be conducted using the other estimate. If the results from the two estimates don't agree, the investigators should present both combined estimates and clearly explain that the combined estimates are sensitive to the choice of estimate for mean difference. A meta-regression approach[9] was also suggested to adjust for baseline imbalance though its performance has not been fully studied. The investigators may choose this approach as another sensitivity analysis.

4. Calculating Standard Deviation and Standard Error When They Are Not Directly Reported

Commonly used meta-analysis packages (e.g. RevMan, Stata) require three parameters from each of the intervention groups to calculate a weighted mean difference: the mean, the SD, and the sample size. The mean could be the mean change score from baseline or the mean score at followup based on choice for calculating the estimate for mean difference. If any of these are missing, the study will be omitted from the meta-analysis.

Alternatively, investigators could also use the mean difference between the intervention groups and its associated SE directly. It is important for investigators to recognize that the results from continuous outcomes are reported in many different forms, especially for the precision parameters such as SD and SE. If SD and SE are not directly reported, they can often be calculated from other reported information. Investigators should always look for reported data that could be used to conduct exact algebraic calculation.

In this section, we present formulas for calculating SD and SE using other reported information. We also briefly discuss the issue of incorporating correlation into calculation of SD for crossover and cluster randomization trials.

Calculation of Standard Deviation and Standard Error Using Available Data

When SD is not directly reported, it can be computed (assuming both mean and sample size are given) from other reported data: SEs, confidence intervals, z or t statistics, or exact parametric p-values using available formulas.[21] These other reported data could be available for the mean between baseline and followup from each intervention group, or for the mean difference between two intervention groups.

Available Data for one Intervention Group

In this section, all calculations apply to obtaining the SD for the mean between baseline and followup from any one intervention group, and are particularly pertinent to the situation that investigators conduct a meta-analysis using three parameters from each intervention group.

If given a SE of the mean of one intervention group in a trial of sample size *n*, the SD for that group can be computed as:

$$SD = SE\sqrt{n} \qquad (1)$$

If given a 95% normal confidence interval in the form (lower confidence bound [LCB], upper confidence bound [UCB]) around the mean, we can compute the SE using the formula:

$$SE = \frac{UCB - LCB}{3.92} \qquad (2)$$

Formula (1) can then be used to compute SD. If a 90% confidence interval is given, rather than a 95% confidence interval the divisor in formula (2) should be changed to 3.29.
If given a z-statistic or a t-statistic, usually for the situation of change score from baseline in each intervention group, the SE can be computed using the change score:

$$SE = \frac{|\,change\ score\,|}{z} \quad \text{or} \quad SE = \frac{|\,change\ score\,|}{t} \qquad (3)$$

Again, formula (2) can then be used to determine the SD.

If given an exact p-value, again usually for the situation of change score from baseline in each intervention group, it can be converted to a z-statistic first, using the inverse normal value. The easiest way to obtain this is by typing in any cell in Microsoft excel "= normsinv(1-*p*/2)", where p is the reported *p*-value. For example, if the given p-value was 0.03 we would type in excel "=normsinv(0.985)" which obtains the z-stat 2.17. If the sample size is small and the study obtained the p-value using a paired t-test, then *t*-statistic could be obtained by tying in Microsoft excel "=tinv (p,df)" where p is the reported p-value, and df is the degree of freedom for the t-test and equals *n*-1, where *n* is the sample size of the intervention group. Then formulas (3) and (2) could be used to calculate SD.

If an upper-bound p-value (e.g. p<0.05) is given, then the same formulas can be used to obtain a conservative estimate of the SD.

For a change score, if the SD at baseline (SD_b) and followup (SD_f) are reported, SD for the change score could also be calculated as:

$$SD = \sqrt{SD_b^2 + SD_f^2 - 2 * r * SD_b * SD_f} \qquad (4)$$

where *r* is the correlation between baseline and followup score. Information about ρ is often not available and needs to be imputed. See Section 5 for more information on handling missing data for ρ.

Available Data for the Mean Difference Between Two Groups

If a confidence interval, a z-statistic, or a t-statistic is given for the difference of means between two intervention groups, formulas (2) and (3) apply in a similar way to calculate the SE, for the mean difference between groups in this case. For formula (3), replace change score with the mean difference. If given an exact p-value of for a mean difference, it can be converted to a z-statistic using the same Excel "normsinv(1-*p*/2)" function. If the sample size is small and the study obtained the p-value using a two-sample t-test, then *t*-statistic could be obtained by using the same Excel function "tinv (p,df)" where p is the reported p-value, but df equals $n_1 + n_2$ -2 in this case, where n_1 and n_2 are the sample size of each intervention group. If an upper-bound p-value (e.g., p<0.05) is given, then the same formulas can be used to obtain a conservative estimate of the SE of mean difference.

In some cases, the SD for each intervention group (SD_1 and SD_2) is reported, SE for the mean difference could be calculated as:

$$SE = \sqrt{\frac{SD_1^2}{n_1} + \frac{SD_2^2}{n_2}}, \qquad (5)$$

where n_1 and n_2 are sample sizes of the two intervention groups. If the estimates of SD_1 and SD_2 are similar, one could also use:

$$SE = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \ (6)$$

Unlike formula (4), there is no need to consider correlation since the intervention groups are independent in a parallel design.

With the SE for mean difference calculated, if investigators choose to conduct a meta-analysis using three parameters (the mean, the SD, and the sample size) from each intervention group, the simplifying assumption could be made that treatment SD is equal to the control SD. This assumption will not affect the final result and the computed SD can then be used for both the intervention and control group. The common SD can be estimated as:

$$SD = SE\sqrt{\frac{n_1 n_2}{n_1 + n_2}}. \qquad (7)$$

Alternatively, investigators could directly use the SE of the difference between groups (and the mean difference) in the meta-analysis. Both methods will provide the same results. Usually the choice of method depends on the type of data reported in the included studies (hence one method involves less calculation) and the meta-analysis package one is using.

A Working Example

A parallel study with 15 patients in each group reports the following: “The mean systolic blood pressure in treatment was 122.4 mmHG while in control it was 134.5 mmHG. This difference was not statistically significant (p=0.24).” How would we go about computing the SD?

- Mean difference = 134.5 – 122.4 = 12.1.
- 1-p/2 = 1-0.24/2 = 0.88. Typing “=normsinv(0.88)” in excel gives a gives z-stat of 1.175. If assuming a t-test, then the t-stat = tinv(0.24, 28) = 1.201 where 28 = 15+15-2.
- SE = 12.1/1.175 = 10.298. This number could be used directly in the meta-analysis, or if one is using a software that requires the SD in each group, it can be computed from this SE:

$$SD = SE\sqrt{\frac{n_1 n_2}{n_1 + n_2}} = 10.298\sqrt{\frac{15*15}{15+15}} = 28.2$$

- This SD can be entered for *both* treatment and control.

Crossover Trials

For trials with a parallel design, the intervention groups are independent from each other. There is no need to consider correlation between intervention groups when calculating SE for mean difference. When a crossover trial is to be included in a meta-analysis, using the methods of a parallel design to calculate SE for mean difference will usually underestimate the precision of the crossover trial in most cases. This is because that the positive correlation associated with using the same patients in both the treatment and control groups lowers the variance around the estimate. The formula to use to compute the pooled SE for a crossover trial is:

$$SE_d = \sqrt{SE_T^2 + SE_C^2 - 2rSE_TSE_C}$$

where *r* is the within patient correlation coefficient and SE_d, SE_T and SE_C are the difference, treatment, and control SE respectively. For a parallel trial the value of r is always 0, thus the last term becomes 0. For a crossover study, however, the value of r is usually not reported from the trial and needs to be estimated in order to properly compute the correct SE. See Section 5 on missing data for methods for estimating or imputing r.

Cluster Randomized Trials

Cluster randomized trials are similar to crossover trials in that using the methods of a parallel design to estimate SE for mean difference will produce incorrect results. Data among patients within a cluster are usually positively correlated. However, unlike crossover trials, ignoring this correlation in cluster randomized trials will overestimate the precision of the mean difference. If a cluster randomized trial reported a SE that failed to account for this correlation, the simplest way to account for this discrepancy is to compute a design effect (DE) as:

$$DE = 1 + (m - 1)ICC$$

where m is the average cluster size and ICC is the intra-class correlation coefficient. The ICC is defined as the proportion of the total variance (the within cluster variance plus the between cluster variance) that is attributed to the between cluster variance. The design effect can then be multiplied by the variance of the mean difference computed as if it were parallel. This new adjusted variance will appropriately reflect the loss of precision to the cluster randomization design.

The ICC is generally be quite low (less than 0.1), but it can still have a fairly large effect on the trial variance, particularly when the average cluster size is quite large. Usually this ICC is not reported from trials and the investigators need to assume a plausible value to calculate the SE. Investigators should always conduct sensitivity analysis by assuming several values of ICC and check how robust the results to the assumed ICC values.

5. Dealing With Missing Data

Missing data is a common issue when conducting meta-analysis and often leads to biased estimates. Missing data can take many forms: missing studies, missing outcomes, missing summary data, missing individuals, and missing study-level characteristics. Missing studies and missing outcomes are complex issues that are not specific to continuous data and will not be discussed here. The issue of missing summary data is most relevant to continuous data and the focus of this section. The issue of missing individuals and missing study-level data will be discussed briefly.

How Are the Missing Data Distributed?

All missing data can be categorized into one of three types: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). Data are said to be MCAR if being missing does not depend on observed or unobserved measurements. MAR means that the reason data are missing is unrelated to their actual values. Data are MNAR if they are neither MCAR nor MAR. Missing data that are MCAR or the more reasonably acceptable MAR are considered ignorable, because there is no bias in simply performing the meta-analysis without the missing data, and the combined estimate only suffers from less precision. Unfortunately, missing data are usually MNAR and thus the consequences of how we deal with it must be considered. Simply omitting studies with missing data that are MNAR will lead to biased results.[22]

Missing Summary Data

If a study is missing data elements that are required in a meta-analysis and could not be calculated from reported data, it is often a good idea to contact the authors to obtain the missing values first. If unsuccessful, either we need to exclude the study or we need to impute the missing data in some way. Both omitting a study and imputing for missing values can result in bias and under-precision. It is important that the investigators select the method that gives the best estimate of combined effect size.

Standard deviation is the most commonly missing parameter. We recommend that studies missing only SDs should not be excluded as this will often lead to a biased combined estimate. Methods for imputing SDs will be discussed below and we will also address the issue of missing correlation between baseline and followup.

Imputation of Standard Deviation

If the data are not available in an alternative form that allow direct calculation, imputation of the missing values is an often recommended alternative as shown in simulation studies.[23] Several simple methods have been suggested for directly imputing missing SDs, including direct substitution using the largest SD of the included studies, arithmetic means,[24] linear regression,[25] coefficient of variation,[26] and imputation from correlation.[23] We demonstrate some of these methods using the following example which is taken from a review comparing asthma patients using long-acting beta agonist (LABA)/inhaled corticosteroid (ICS) combination versus using ICS alone. The outcome is pulmonary function in L/min.

The studies labeled Strand and SAM40036 are missing their SD and are not counted in the final meta-analysis (Figure 1). If we do a direct substitution of the largest SD we can see that the largest SD in the LABA/ICS group is 52.14 and in the ICS group is 49.64 (Figure 2).

Figure 1. Results of meta-analysis of pulmonary function without including studies with missing data

	LABA/ICS			ICS				Mean Difference	
Study or Subgroup	Mean [L/min]	SD [L/min]	Total	Mean [L/min]	SD [L/min]	Total	Weight	IV, Random, 95% CI [L/min]	Year
Nelson 2003	51.5	46.2	95	29.9	49.64	97	6.3%	21.60 [8.04, 35.16]	2003
Chuchalin 2004	55.2	52.14	111	33.6	46.3	114	6.9%	21.60 [8.70, 34.50]	2004
SAS30015 2004	45.6	45.8	74	26.7	37.5	75	6.4%	18.90 [5.45, 32.35]	2004
Strand 2004	40	0	78	14	0	72		Not estimable	2004
SAM40034 2004	51	43.4	75	27.7	43.3	79	6.1%	23.30 [9.60, 37.00]	2004
Murray 2004	51	50.66	88	30.4	45.28	89	5.7%	20.60 [6.44, 34.76]	2004
SAM40036 2004	51	0	288	40.1	0	289		Not estimable	2004
SAS30039 2005	64.4	48.83	179	42.9	49.64	180	11.1%	21.50 [11.31, 31.69]	2005
SAS40068 2005	42.3	41.83	251	27.3	41.44	262	22.2%	15.00 [7.79, 22.21]	2005
Boonsawat 2008	37.5	38.09	151	17.7	37.35	155	16.1%	19.80 [11.35, 28.25]	2008
Kerwin 2008	48.7	40.58	210	27.9	40.77	212	19.1%	20.80 [13.04, 28.56]	2008
Total (95% CI)			**1600**			**1624**	**100.0%**	**19.56 [16.16, 22.95]**	

Heterogeneity: Tau² = 0.00; Chi² = 2.28, df = 8 (P = 0.97); I² = 0%
Test for overall effect: Z = 11.29 (P < 0.00001)

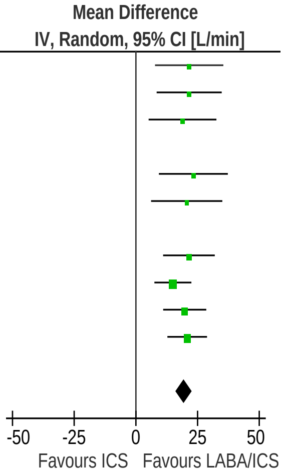
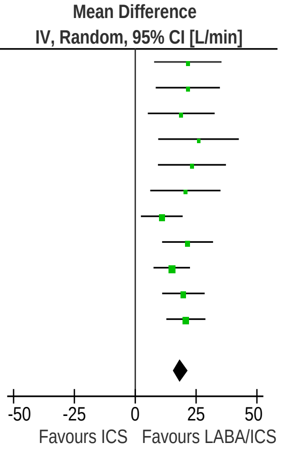


Figure 2. Results of meta-analysis of pulmonary function with imputing missed data using direct substitution

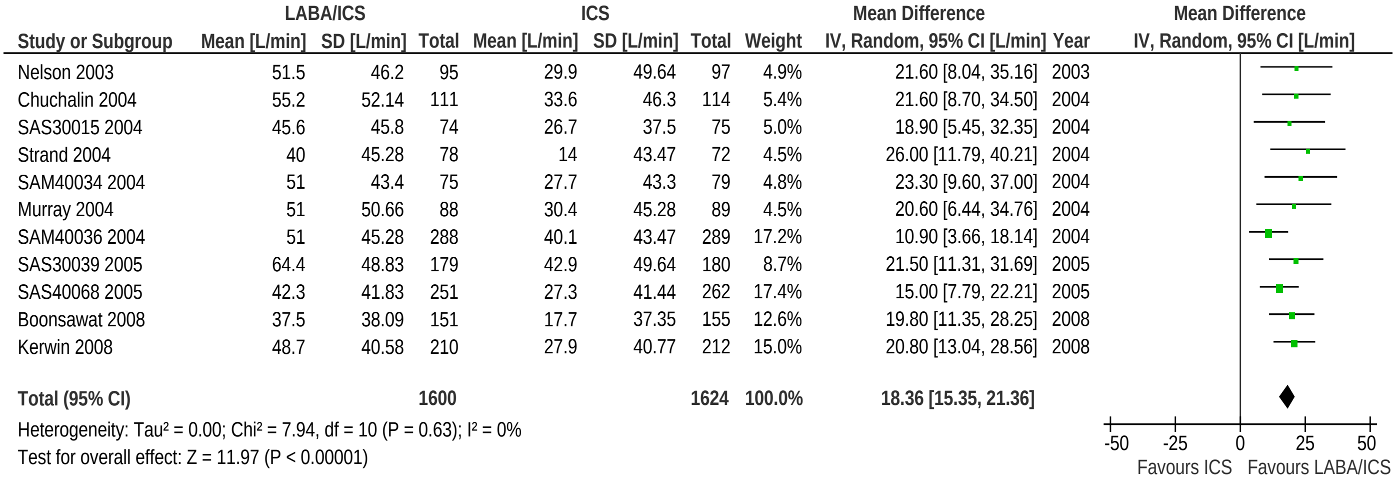
	LABA/ICS			ICS				Mean Difference	
Study or Subgroup	Mean [L/min]	SD [L/min]	Total	Mean [L/min]	SD [L/min]	Total	Weight	IV, Random, 95% CI [L/min]	Year
Nelson 2003	51.5	46.2	95	29.9	49.64	97	5.2%	21.60 [8.04, 35.16]	2003
Chuchalin 2004	55.2	52.14	111	33.6	46.3	114	5.7%	21.60 [8.70, 34.50]	2004
SAS30015 2004	45.6	45.8	74	26.7	37.5	75	5.3%	18.90 [5.45, 32.35]	2004
Strand 2004	40	52.14	78	14	49.64	72	3.6%	26.00 [9.71, 42.29]	2004
SAM40034 2004	51	43.4	75	27.7	43.3	79	5.1%	23.30 [9.60, 37.00]	2004
Murray 2004	51	50.66	88	30.4	45.28	89	4.7%	20.60 [6.44, 34.76]	2004
SAM40036 2004	51	52.14	288	40.1	49.64	289	13.8%	10.90 [2.59, 19.21]	2004
SAS30039 2005	64.4	48.83	179	42.9	49.64	180	9.2%	21.50 [11.31, 31.69]	2005
SAS40068 2005	42.3	41.83	251	27.3	41.44	262	18.3%	15.00 [7.79, 22.21]	2005
Boonsawat 2008	37.5	38.09	151	17.7	37.35	155	13.3%	19.80 [11.35, 28.25]	2008
Kerwin 2008	48.7	40.58	210	27.9	40.77	212	15.8%	20.80 [13.04, 28.56]	2008
Total (95% CI)			**1600**			**1624**	**100.0%**	**18.59 [15.51, 21.68]**	

Heterogeneity: Tau² = 0.00; Chi² = 6.68, df = 10 (P = 0.76); I² = 0%
Test for overall effect: Z = 11.81 (P < 0.00001)



Alternatively, we could use the arithmetic means of the SDs in each group. That is, for the LABA/ICS group we take (46.2 + 51.14 + 45.8 + … + 40.58)/9 = 45.28. For the ICS group we get 43.47. Using these values for our two missing studies yields similar results to imputing using the maximum (Figure 3).

Figure 3. Results of meta-analysis of pulmonary function with imputing missed data using arithmetic means



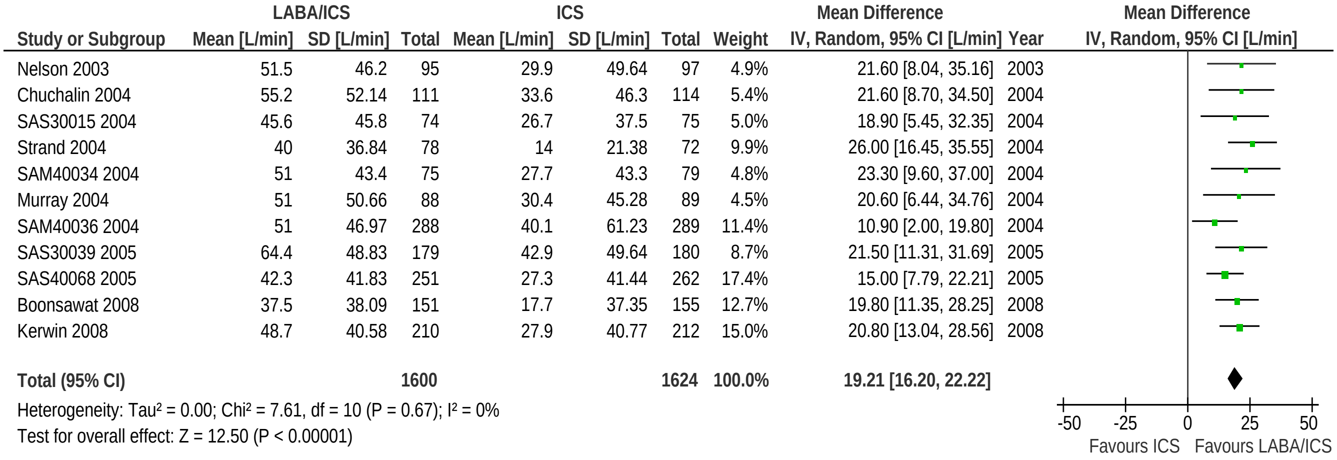
	LABA/ICS			ICS				Mean Difference	
Study or Subgroup	Mean [L/min]	SD [L/min]	Total	Mean [L/min]	SD [L/min]	Total	Weight	IV, Random, 95% CI [L/min]	Year
Nelson 2003	51.5	46.2	95	29.9	49.64	97	4.9%	21.60 [8.04, 35.16]	2003
Chuchalin 2004	55.2	52.14	111	33.6	46.3	114	5.4%	21.60 [8.70, 34.50]	2004
SAS30015 2004	45.6	45.8	74	26.7	37.5	75	5.0%	18.90 [5.45, 32.35]	2004
Strand 2004	40	45.28	78	14	43.47	72	4.5%	26.00 [11.79, 40.21]	2004
SAM40034 2004	51	43.4	75	27.7	43.3	79	4.8%	23.30 [9.60, 37.00]	2004
Murray 2004	51	50.66	88	30.4	45.28	89	4.5%	20.60 [6.44, 34.76]	2004
SAM40036 2004	51	45.28	288	40.1	43.47	289	17.2%	10.90 [3.66, 18.14]	2004
SAS30039 2005	64.4	48.83	179	42.9	49.64	180	8.7%	21.50 [11.31, 31.69]	2005
SAS40068 2005	42.3	41.83	251	27.3	41.44	262	17.4%	15.00 [7.79, 22.21]	2005
Boonsawat 2008	37.5	38.09	151	17.7	37.35	155	12.6%	19.80 [11.35, 28.25]	2008
Kerwin 2008	48.7	40.58	210	27.9	40.77	212	15.0%	20.80 [13.04, 28.56]	2008
Total (95% CI)			**1600**			**1624**	**100.0%**	**18.36 [15.35, 21.36]**	

Heterogeneity: Tau² = 0.00; Chi² = 7.94, df = 10 (P = 0.63); I² = 0%

Test for overall effect: Z = 11.97 (P < 0.00001)

If we wish to use average coefficient of variation (CV) to impute, we need to first calculate a CV for each study. CV is defined as SD/mean. For example, for the Nelson study, CV=46.2/51.5 = 0.897. Computing CV for each study and then taking the average gives us 0.921 for the LABA/ICS group and 1.527 for the ICS group. We now use these values and the formula SD = CV*mean to estimate the SD for studies with a missing SD. For example, for the Strand study, in the LABA/ICS group, the mean is 40, and so we estimate SD as 40*0.921 = 36.84. Using this method gives us similar results to the previous two methods (Figure 4).

Figure 4. Results of meta-analysis of pulmonary function with imputing missed data using coefficient of variation



	LABA/ICS			ICS				Mean Difference	
Study or Subgroup	Mean [L/min]	SD [L/min]	Total	Mean [L/min]	SD [L/min]	Total	Weight	IV, Random, 95% CI [L/min]	Year
Nelson 2003	51.5	46.2	95	29.9	49.64	97	4.9%	21.60 [8.04, 35.16]	2003
Chuchalin 2004	55.2	52.14	111	33.6	46.3	114	5.4%	21.60 [8.70, 34.50]	2004
SAS30015 2004	45.6	45.8	74	26.7	37.5	75	5.0%	18.90 [5.45, 32.35]	2004
Strand 2004	40	36.84	78	14	21.38	72	9.9%	26.00 [16.45, 35.55]	2004
SAM40034 2004	51	43.4	75	27.7	43.3	79	4.8%	23.30 [9.60, 37.00]	2004
Murray 2004	51	50.66	88	30.4	45.28	89	4.5%	20.60 [6.44, 34.76]	2004
SAM40036 2004	51	46.97	288	40.1	61.23	289	11.4%	10.90 [2.00, 19.80]	2004
SAS30039 2005	64.4	48.83	179	42.9	49.64	180	8.7%	21.50 [11.31, 31.69]	2005
SAS40068 2005	42.3	41.83	251	27.3	41.44	262	17.4%	15.00 [7.79, 22.21]	2005
Boonsawat 2008	37.5	38.09	151	17.7	37.35	155	12.7%	19.80 [11.35, 28.25]	2008
Kerwin 2008	48.7	40.58	210	27.9	40.77	212	15.0%	20.80 [13.04, 28.56]	2008
Total (95% CI)			**1600**			**1624**	**100.0%**	**19.21 [16.20, 22.22]**	

Heterogeneity: Tau² = 0.00; Chi² = 7.61, df = 10 (P = 0.67); I² = 0%
Test for overall effect: Z = 12.50 (P < 0.00001)

More complex methods have been suggested for imputing a weighted mean difference directly in the presence of missing SD data: these include sample size weights,[27] bootstrap methods,[28] multiple imputation methods,[29, 30] interval method,[31] and prognostic method.[31] These methods have the disadvantage of complexity and the inability to present the standard forest plot that readers are accustomed to seeing. The advantage to these methods is a likely more accurate accounting of the true variance in the meta-analysis. Some work has also been done taking into account the uncertainty of the SD when it is imputed.[32, 33] A full accounting of these methods is beyond the scope of this paper and the investigators are encouraged to look more into each of these methods themselves. There is no enough evidence to indicate their relative performance yet, though there is some evidence that the method that one chooses for imputation may not make a huge difference in the final meta-analysis.[21, 34]

To summarize the recommendations for missing SD, investigators should always try to contact authors to request exact estimates. Studies missing only SDs should not be excluded as this will often lead to a biased combined estimate. If exact estimates could not be obtained, imputation using one of the various methods listed should be done. Direct substitution using the largest SD is the simplest method and most likely lead to a conservative estimate, but if one is comfortable with one of the more complex methods listed, this may lead to a more accurate estimate of precision parameter and is encouraged. Investigators should use alternative imputation method in a sensitivity analysis to determine how robust the results are to the different imputation methods.

Missing Correlations

When meta-analyzing change from baseline score or data from crossover studies, to calculate the SD for change from baseline, the correlation between baseline and followup scores is required in addition to the SDs for baseline and followup score. This information is often not available from trials and has to be imputed.

The first option of imputation is to use estimates of correlation from other similar studies included in the same meta-analysis. If a study gives the SDs for both individual scores as well as for the change score, one can compute the correlation (r) using the following formula:

$$r = \frac{SD_1^2 + SD_2^2 - SD_C^2}{2SD_1SD_2}$$

where SD_1, SD_2, and SD_C represent the SD for baseline, followup, and change score, respectively. This correlation can be used as an estimate of the correlation in studies where the SD for change score is not available but the SDs for baseline and followup score are available.

If it is not possible to compute a correlation from any of the included studies, one can either estimate it from historical data or use an approximate value. In the case of the latter, the most common value to use is 0.5.[24] A recent study (http://www.effectivehealthcare.ahrq.gov/ehc/products/344/1087/Correlation_Draft-Report_20120515.pdf) showed that the median correlation for change from baseline among trials included in systematic reviews was 0.59 (IQR: 0.40, 0.81). As in the case of missing SDs, investigators should always conduct sensitivity analysis by assuming several values of correlation.

Missing Individuals and Missing Study Level Characteristics

The issue of individuals missing from a study either due to withdrawals, or other reasons is more an issue at the study level than the meta-analysis level. Nevertheless, three methods have been proposed to account for missing patient data: reweighting by completion rate, incorporating completion rate into a Bayesian random-effects model, and inference based on a Bayesian shared-parameter model (including the completion rate).[35]

When study level characteristics are missing, it will not affect the primary meta-analysis but can have an effect or even prevent one from performing sub-group analysis and meta-regression. Bayesian methods have been suggested for combating this problem when doing a meta-regression.[36]

Both of these issues are complex and beyond the scope of this report as they do not specifically pertain to continuous data. We don’t have any particular recommendations for using these methods, and investigators may try these methods for exploratory purpose.

6. Dealing with Skewed Data

Most meta-analytic techniques for continuous data are based on the mean of the variable of interest, for example, a clinical outcome and a measure of dispersion. If the variable’s distribution is asymmetric, then the data are classified as skewed. Meta-analytic methods based on means provide correct inference when the individual studies have sufficiently large sample size regardless of the variable’s distribution itself due to the Central Limit Theorem, or if the variable of interest is at least approximately normally distributed.[37] However, if neither the sample size is sufficient nor the variable of interest is approximately normal, ignoring variable skewness or treating skewness inadequately can result in misleading conclusions. For example, Ziguras et al. (2002)[38] compared two meta-analyses of interventions to reduce alcohol consumption, one of which excluded skewed data and one of which did not. The difference in handling skewed data was discussed as one of the reasons that the two analyses produced different results. Shen et al. (2007)[5] provided an example regarding the relationship between hospital ownership and financial performance in which disregarding skewness produced misleading results.

Several possible scenarios exist with respect to skewed data in meta-analysis. First, an individual study may report nonparametric summaries such as the median and interquartile range for a variable of interest. Second, the variable of interest may be suspected to be skewed and yet an individual study will report parametric summaries, that is, the mean and SD (or SE or variance). Third, an individual study may transform the data and present summary statistics on the transformed scale or different statistics, for example, the geometric mean, on the raw (original) scale.

Using Nonparametric Summaries Assuming Symmetry

If symmetry could be assumed, nonparametric statistics like medians, ranges, and inter-quartile ranges can be used to estimate both means and SDs. These nonparametric summaries are only estimates of the true parameters, not direct calculations like Section 4. Different methods using non-parametric summaries have been used to obtain both means from medians and SDs from ranges and inter-quartile ranges depending on sample size.[21, 39]

The median is similar to the mean when the variable distribution is symmetric. Thus, if an individual study reports the median for a variable of interest, the median could be used in place of the mean to calculate the mean difference. Most past analyses have used a simple direct substitution of median, but there has been a recent study[39] showing that if the range (i.e. the minimum [a] and maximum [b] values) are given, a better estimate of the mean for sample sizes less than 25 is:

$$\bar{x} = \frac{a + 2m + b}{4}$$

while the median itself remains the best estimator for sample sizes greater than 25.
For estimating SD, the most common practice has been to simply compute it from the range or inter-quartile range (IQR). IQR indicates the length of the interval in which the central 50 percent of the sample values of variable lie between the 25th percentile and 75th percentile. In these situations, SD can be estimated as IQR/1.35 or as range/4. Hozo[39] has suggested that range/4 should be used for sample sizes between 15 and 70, while range/6 should be used for sample sizes greater than 70. For sample sizes smaller than 15, the formula below can be used to calculate SD:

$$SD = \sqrt{\frac{1}{12}\left\{\frac{(a - 2m + b)^2}{4} + (b - a)^2\right\}}$$

Since range is inherently dependent upon sample size, Wiebe[21] has suggested that the table below (taken from Pearson[40]) should be used to impute SD from range. The SD can be determined simply by dividing the range by the given divisor (which represents the percentage limit for the distribution of the range in a normal population).

Table 1. Percentage limits for the distribution of range in samples from a normal population

Sample Size	Divisor	Sample Size	Divisor	Sample Size	Divisor	Sample Size	Divisor
2	1.128	13	3.336	24	3.895	55	4.572
3	1.693	14	3.407	25	3.931	60	4.639
4	2.059	15	3.472	26	3.964	65	4.699
5	2.326	16	3.532	27	3.997	70	4.755
6	2.534	17	3.588	28	4.027	75	4.806
7	2.704	18	3.640	29	4.057	80	4.854
8	2.847	19	3.689	30	4.086	85	4.898
9	2.970	20	3.735	35	4.213	90	4.939
10	3.078	21	3.778	40	4.322	95	4.978
11	3.173	22	3.819	45	4.415	100	5.015
12	3.258	23	3.858	50	4.498		

To use this table, simply look up your sample size and use the given divisor. For example if the sample size is 22, then SD could be estimated as range/3.819. It should be noted that the above table is assuming a normal distribution on the data. Investigators should use it only when the distribution of data is at least symmetric.

Assessing Skewness

The fact that nonparametric summaries have been reported in individual studies is an indication that the study authors have evidence of skewness in the data. Thus, prior to beginning analysis, we recommend that the meta-analyst carefully consider the distribution of each variable of interest and ascertain whether the distribution is skewed. This assessment should be based on substantive knowledge of the variable and prior data if available. For example, utilization and cost variables are often skewed due to a subpopulation of users with no use, and thus no cost, and a few individuals with very high use and hence high cost. When median (or mean) with IQR or range are reported, some idea about the distribution usually could be gained. The two end points of IQR and range are not symmetric around median (or mean) if the distribution of the data is skewed. Altman and Bland (1996)[41] also provided two useful tricks for checking skewness. If the mean is smaller than twice the SD in each intervention group, the data are likely to be skewed. If there are data from several groups of individuals, and the SD increases as the mean increases, it is a good indication that the data are positively skewed.

Dealing With Skewness

If skewness is suspected, and individual studies present nonparametric summaries, one can estimate the mean and SD and proceed with usual meta-analysis methods using the resulting estimates. This could work if the degree of skewness is at most moderate, for example, when the variable of interest has a symmetric distribution in most included studies but shows some skewness in others. We recommend, however, in the case of significant skewness, transforming the data to reduce skew. An additional advantage of such a transformation can be increased clinical interpretability (Higgins et al. 2008).[37] Generally a logarithmic transformation is used, particularly when the data are economic in nature. Some studies may report summaries on the logarithmic scale. An alternative approach when the data have been log-transformed is to present the geometric mean on the raw (original) scale and it's SD. One cannot combine summaries on the raw and transformed scales together. Higgins et al. (2008)[37] present methods for transforming between different scales. This ability to transform allows the meta-analyst to determine whether to conduct the meta-analysis on the raw scale or on the log-transformed scale as appropriate. Issues to take into consideration when choosing the scale include, for example, which scale was most commonly used across the individual studies. Investigators are encouraged to employ these methods.

Some research has recently focused on conducting nonparametric meta-analysis. For example, Ma et al. (2011)[31] discuss a nonparametric method that utilizes U-statistic theory. Nonparametric approaches would obviate the need for distributional assumptions, be they Normality or symmetry, but may be statistically inefficient. Other authors have proposed using a ratio of geometric means to analyze skewed continuous data.[42] However, the lack of clinician experience with geometric means may make such methods difficult to implement. Investigators may choose to explore these methods and see how they compare to their primary analysis.

7. Standardized Mean Difference

For continuous outcomes, it is often the case that different studies in a meta-analysis use a variety of measures to assess the same outcome. For example, included trials might use the Beck Depression Inventory, the Geriatric Depression Scale, and the Center for Epidemiologic Studies Depression scale to measure depression. If these measures are sufficiently similar to suggest that they are truly measuring the same outcome, standardized mean difference (SMD), a measure of effect size, could be used to combine the studies using different scales.

Choice of Standardized Mean Difference

Commonly used SMD includes Cohen’s d, Hedges’ g, and Glass’Δ. (Card, 2012).[43] These measures are all calculated similarly by dividing mean difference by the SD. The differences lie in the denominator: Cohen’s d divides by the estimate of the pooled population SD, which is calculated as:

$$d = \frac{\overline{X}_1 - \overline{X}_2}{S_P} \text{ where } S_P = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2}}$$

where $\bar{X}_1 - \bar{X}_2$ is the mean difference between the two intervention groups and SD_1 and SD_2 are the standards deviation of the two intervention groups.

Hedges’ g uses the pooled sample SD, which is calculated as:

$$g = \frac{\overline{X}_1 - \overline{X}_2}{S_{Pooled}} \text{ where } S_{Pooled} = \sqrt{\frac{(n_1 - 1)SD_1^2 + (n_2 - 1)SD_2^2}{n_1 + n_2 - 2}}$$

Glass’ Δ uses the estimate of the SD from the control group:

$$\Delta = \frac{\bar{X}_1 - \bar{X}_2}{sd_{control}}$$

All three parameters are biased and the bias can be more than trivial when the sample sizes of both intervention groups are small. Durlak (2009)[44] suggests that the positive bias “amounts to a

4% reduction in effect when the total sample size is 20 and around 2% when *N* = 50." Hedges and Olkin (1985)[45] provided a correction formula to correct for this small sample bias and the corrected version serves as an unbiased estimator of the population SMD. Under the equal variance assumption, Cohen's d and Hedges' g are more precise estimators than Glass' Δ and Hedges' g has smaller sample variance than Cohen's d.

Hedges' unbiased estimator should be used whenever possible, especially when the sample sizes are smaller than 20. Otherwise, Hedges' g is generally preferred over Cohen's d or Glass' Δ. When sample size is large, difference between Hedges' g and Cohen's d is small and they could be used interchangeably. When variance across the groups differs and the control group may be a more accurate estimate of true population variance, Glass' Δ is preferable. Sensitivity analyses are recommended to check how the results differ between using Hedges' g versus Glass' Δ.

Interpreting Values of Standard Mean Difference

In theory, SMD can be any number, positive or negative. SMDs of 0.3, 0.5, and 0.8 are suggested corresponding to small, medium, and large referents (Cohen 1988)[46] and are widely used, although they were not anchored in meaningful clinical context. Conclusions about clinical importance of the differences are often not clear using SMDs.

We recommend that Investigators consider back transforming the pooled SMD to the original scale to facilitate assessing the clinical importance of combined SMDs and aid decisionmaking. Back transforming could be done by multiplying the SMDs with the among-person SD of the original scale derived from the population representative studies. The back transformed mean difference should be evaluated for clinical importance according to evidence based definitions of minimum clinically important differences from published studies and evidence-based reports.

Caveats of using Standard Mean Difference

Sample variance heterogeneity. Some studies have identified bias associated with using SMD in heterogeneous studies and studies with large SD (Van Den Noortgate et al., 2003).[47] Because the SMD is greatly influenced by the SD, factors affecting the SD will affect the SMD. Therefore, if there are meaningful differences in variance across studies due to factors such as different inclusion criteria (e.g., one study includes only severely depressed participants, while another includes participants with mild, moderate, and severe depression), then these differences in variance due to populations will affect the SMD. However, the bias associated with the use of SMD is small when the true variance is small (Van Den Noortgate et al., 2003).[47]

Investigators should examine sample variance heterogeneity when combining SMDs across studies and evaluate how these differences could affect the meta-analysis results by doing subgroup analyses based on the magnitude of the variance. In each subgroup, only SMDs from more homogeneous populations with similar variance estimates should be combined. If subgroup analyses suggest that results differ, then SMDs should not be combined across all studies with heterogeneous sample variance.

Covariates. Studies may account for the effect of covariates. When combining SMDs, SMDs calculated using the unadjusted mean difference (Nakagawa & Cuthill, 2007)[48] are not recommended to be combined with SMDs adjusted for covariates if there is heterogeneity between the two sets of the SMDs. For SMDs calculated from mean difference adjusted for covariates, investigators should consider only combining results with similar degree of adjustment to ensure comparable effect size across studies. Otherwise, the combined estimate may be biased. If a study uses balanced groups based on important covariates, and another study adjusts for covariates, these two studies could be considered as having similar degree of adjustment and could be combined in a meta-analysis.

Directionality. Note that the direction of the scale must be consistent across the scales used in the included studies. For example, if in one study a higher score indicates depression and in another study a low score indicates depression, then one of the scores must be reverse-coded to account for scale direction differences. Investigators should assure that scales are converted to a consistent direction of effect across all studies when calculating SMD.

Missing standard deviation. Calculation of SMD needs information from SD. When SD is missing, investigators could use imputed SD. One study showed that imputed SD produced similar results when comparing studies using imputed and known SD values. More information is provided in Section 5 for imputing SD and Furukawa et al. (2006)[49] provided discussion on how imputing SD applies to SMD.

Multiplicity of data. Studies often report data from outcomes based on multiple measures from multiple time points, an important source of possible bias in meta-analysis (Tendal et al., 2011).[50] For example, one trial may assess an outcome using five measures assessed at three time points and report results in four published articles. Investigators should establish *a priori* inclusion criteria on which outcomes and time points should be used in a meta-analysis and make sure that all outcome measures meeting inclusion criteria are included. Outcome measures should not be excluded on the basis of statistical significance or other types of selection bias. Investigators must also make sure that only one outcome measure is included in the same meta-analysis. Sensitivity analyses may be conducted to assess the impact of the different measures (for the same outcome) on the combined estimate.

8. Ratio of Means

Mean difference or SMD has been the most commonly used measure in meta-analysis for continuous outcomes. Recently, RoM[2, 3] was proposed as an alternative measure of mean difference and standardized mean difference for meta-analyzing continuous outcome. This measure offers the advantage that it could be used regardless of the units used in the individual trials. As SMD, it can be used to combine outcomes that are measured using different scales. Mathematically, it is equivalent to use the percentage change of the intervention group from the control group.

The RoM is calculated by dividing the mean outcome value from the intervention group ($\bar{X}_1$) by the mean outcome value from the control group ($\bar{X}_2$). For meta-analysis, the natural logarithm of each trial's RoM and its SE is calculated using the mean values, number of participants (n), and SD in each group [2] as:

$$\mathrm{Iog}\;(\mathrm{RoM}) = \mathrm{Iog}\left(\frac{\bar{X}_1}{\bar{X}_2}\right)$$

$$SE[\mathrm{Iog}(\mathrm{RoM})] = \sqrt{\frac{1}{n_1}\left(\frac{SD_1}{\bar{X}_1}\right)^2 + \frac{1}{n_2}\left(\frac{SD_2}{\bar{X}_2}\right)^2}$$

Then the natural logarithm transformed ratios are combined across studies using the standard inverse variance method. A combined ratio and its 95% confidence interval could be obtained by back transforming the combined log-transformed ratio and its 95% confidence interval:

$$\text{RoM} \; = \exp(\text{Iog}(\text{RoM})_{pooled})$$
$$95\% \text{ Confidence Interval} = \exp\{\text{Iog}(\text{RoM})_{pooled} \pm 1.96 \times \text{SE}[\text{In}(\text{RoM})]_{pooled}\}$$

This method can be employed using a free meta-analysis software package called COMPARE2.[5] RoM has a straightforward interpretation and expresses the percentage change in the mean value of the intervention group relative to the control group. The results are in a relative form similar to the risk ratio. For example, if the combined RoM is 1.15, it means that the mean of the intervention group is 15 percent higher than the control group; if the combined RoM is 0.85, then the mean of the intervention group is 15 percent lower than the control group.

Based on simulation studies,[2] RoM showed comparable statistical performance to mean difference methods in terms of bias, coverage probability, and statistical power. Overall, the data suggested that RoM is a reasonable alternative. Further data from an empirical analysis of 232 clinically diverse published meta-analyses[3] confirmed the findings of simulated data and suggested that, on average, RoM produced similar combined effect estimates. SMDs of 0.2, 0.5, and 0.8 corresponded to increases in mean of 8, 22, and 37 percent, respectively. There was less heterogeneity in meta-analyses using RoM compared with mean difference but more compared with SMD.

Several meta-analyses have used the RoM when faced with the limitation of the data in the original studies prohibiting the use of traditional difference methods.[51-54] For example, one study (Peng et al., 2007)[53] utilized the RoM method when included studies reported various units of dosing for analgesics for a meta-analysis of total analgesic used within a post operative period. Traditional method would need standardizing all analgesic doses (i.e. conversion to “morphine equivalent”) which was not possible in all cases since not all analgesics have a reliable equivalent ratio. The treatment effect of cumulative analgesics used was therefore expressed as RoM in the experimental versus the control groups.

In summary, RoM appears to be a reasonable alternative to the traditional effect measures of continuous outcome based on empirical evidence. When the outcome is assessed using different scales, it may be preferred to SMD by clinicians due to the ease of interpretation. RoM has no units and allows for pooling of the studies expressed in different units and facilitates comparisons regarding relative effect sizes across different interventions. On the other hand, investigators should note that RoM can only be used in scenarios when the mean values of the intervention and control groups are both positive or both negative. Caution is warranted when RoM is used for small trials with large SDs and large effect sizes. Similar to the limitation of SMD for small trials, the combined estimate of RoM biases towards no effect and this bias is accentuated by high heterogeneity.

9. Dichotomizing Continuous Outcomes in Meta-Analyses

For some continuous outcomes, a meaningful clinically important change is often defined and patients achieving such change are considered as “responders.”[55] Understanding the relationship between continuous effect measures and proportion of “response” is not straightforward and the assumptions used to assessing such relationship are usually difficult to verify. Further research is necessary and we currently recommend against inferring response rate from a combined mean difference.

References

1. Fu R, Gartlehner G, Grant M, et al. Conducting quantitative synthesis when comparing medical interventions: AHRQ and the Effective Health Care Program. Journal of Clinical Epidemiology. 2011;64(11):1187-97.

2. 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 Med Res Methodol. 2008;8:32. PMID: 18492289.

3. 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.

4. Senn S. Baseline distribution and conditional size. J Biopharm Stat. 1993 Sep;3(2):265-76. PMID: 8220409.

5. Shen YC, Eggleston K, Lau J, et al. Hospital ownership and financial performance: what explains the different findings in the empirical literature? Inquiry. 2007 Spring;44(1):41-68. PMID: 17583261.

6. Rosenberger W LJ, ed Randomization in Clinical Trials: theory and practice: New York: Wiley; 2002.

7. Senn SJ. Covariate imbalance and random allocation in clinical trials. Stat Med. 1989 Apr;8(4):467-75. PMID: 2727470.

8. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. The Lancet. 2002;359(9306):614-8.

9. Trowman R, Dumville JC, Torgerson DJ, et al. The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study. J Clin Epidemiol. 2007 Dec;60(12):1229-33. PMID: 17998076.

10. Berger VW, Weinstein S. Ensuring the comparability of comparison groups: is randomization enough? Control Clin Trials. 2004 Oct;25(5):515-24. PMID: 15465620.

11. Roberts C, Torgerson DJ. Understanding controlled trials: baseline imbalance in randomised controlled trials. BMJ. 1999 Jul 17;319(7203):185. PMID: 10406763.

12. Senn S. Testing for baseline balance in clinical trials. Stat Med. 1994 Sep 15;13(17):1715-26. PMID: 7997705.

13. Altman DG. Comparability of randomised groups. Statistician. 1985;34:125-36.

14. Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. The Lancet. 1990;335(8682):149-53.

15. Begg CB. Suspended judgment. Significance tests of covariate imbalance in clinical trials. Controlled Clinical Trials. 1990(11):223-5.

16. Austin PC, Manca A, Zwarenstein M, et al. A substantial and confusing variation exists in handling of baseline covariates in randomized controlled trials: a review of trials published in leading medical journals. J Clin Epidemiol. 2010 Feb;63(2):142-53. PMID: 19716262.

17. Hewitt CE, Kumaravel B, Dumville JC, et al. Assessing the impact of attrition in randomized controlled trials. J Clin Epidemiol. 2010 Nov;63(11):1264-70. PMID: 20573482.

18. Senn S. Change from baseline and analysis of covariance revisited. Stat Med. 2006 Dec 30;25(24):4334-44. PMID: 16921578.

19. Crager MR. Analysis of covariance in parallel-group clinical trials with pretreatment baselines. Biometrics. 1987 Dec;43(4):895-901. PMID: 3427174.

20. Vickers AJ, Altman, D. G. Statistics notes: Analysing controlled trials with baseline and follow up measurements. BMJ. 2001;323:1123-4.

21. Wiebe N, Vandermeer B, Platt RW, et al. A systematic review identifies a lack of standardization in methods for handling missing variance data. Journal of Clinical Epidemiology. 2006 Apr;59(4):342-53. PMID: 16549255.

22. Little RJA, Rubin DB. Statistical analysis with missing data. 2 ed. Hoboken, N.J.: Wiley; 1987.

23. Idris N, Robertson C. The Effects of Imputing the Missing Standard Deviations on the Standard Error of Meta Analysis Estimates. Communications in Statistics - Simulation and Computation. 2009;38(3):513-26.

24. Follmann D, Elliott P, Suh I, et al. Variance imputation for overviews of clinical trials with continuous response. Journal of Clinical Epidemiology. 1992 Jul;45(7):769-73. PMID: 1619456.

25. Pigott T. *Methods for handling missing data in research synthesis*. In: H. C, LV. H, eds. The Handbook of Research Synthesis. New York: Sage Publications Inc.; 1994:163-76.

26. Bracken M. *Statistical methods for analysis of effects of treatment in overviews of randomized tirals*. In: JC. S, MB. B, eds. *Effective care of the newborn infant*. Oxford: Oxford University Press; 1992:13-20.

27. Sanchez-Meca J, Marin-Martinez F. Weighting by inverse variance or by sample size in meta-analysis: A simulation study. Educational and Psychological Measurement. 1998 Apr;58(2):211-20. PMID: Peer Reviewed Journal: 1998-01441-004.

28. Zhu W. Making bootstrap statistical inferences: a tutorial. Research Quarterly for Exercise & Sport. 1997 Mar;68(1):44-55. PMID: 9094762.

29. Rubin DB, Schenker N. Multiple imputation in health-care databases: an overview and some applications. Statistics in Medicine. 1991 Apr;10(4):585-98. PMID: 2057657.

30. Stevens JW, Stevens JW. A note on dealing with missing standard errors in meta-analyses of continuous outcome measures in WinBUGS. Pharmaceutical Statistics. 2011 Jul-Aug;10(4):374-8. PMID: 21394888.

31. Ma Y, Mazumdar M. Multivariate meta-analysis: a robust approach based on the theory of U-statistic. Stat Med. 2011 Oct 30;30(24):2911-29. PMID: 21830230.

32. White IR, Higgins JP, Wood AM, et al. Allowing for uncertainty due to missing data in meta-analysis--part 1: two-stage methods. Statistics in Medicine. 2008 Feb 28;27(5):711-27. PMID: 17703496.

33. White IR, Welton NJ, Wood AM, et al. Allowing for uncertainty due to missing data in meta-analysis--part 2: hierarchical models. Statistics in Medicine. 2008 Feb 28;27(5):728-45. PMID: 17703502.

34. Thiessen Philbrook H, Barrowman N, Garg AX. Imputing variance estimates do not alter the conclusions of a meta-analysis with continuous outcomes: a case study of changes in renal function after living kidney donation. Journal of Clinical Epidemiology. 2007 Mar;60(3):228-40. PMID: 17292016.

35. Yuan Y, Little RJ, Yuan Y, et al. Meta-analysis of studies with missing data. Biometrics. 2009 Jun;65(2):487-96. PMID: 18565168.

36. Hemming K, Hutton JL, Maguire MG, et al. Meta-regression with partial information on summary trial or patient characteristics. Statistics in Medicine. 2010 May 30;29(12):1312-24. PMID: 20087842.

37. Higgins JP, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. Stat Med. 2008 Dec 20;27(29):6072-92. PMID: 18800342.

38. Ziguras SJ, Stuart GW, Jackson AC. Assessing the evidence on case management. Br J Psychiatry. 2002 Jul;181:17-21. PMID: 12091258.

39. Hozo SP, Djulbegovic B, Hozo I, et al. Estimating the mean and variance from the median, range, and the size of a sample. BMC Medical Research Methodology. 2005;5:13. PMID: 15840177.

40. Pearson E. The percentage limits for the distribution of range in samples from a normal population (N less than 100). Biometrika. 1932;24:404-17. PMID: Peer-Reviewed Status-Unknown: 1933-01620-001.

41. Altman DG, Bland JM. Detecting skewness from summary information. BMJ. 1996 Nov 9;313(7066):1200. PMID: 8916759.

42. Friedrich JO, Adhikari NK, Beyene J, et al. Ratio of geometric means to analyze continuous outcomes in meta-analysis: comparison to mean differences and ratio of arithmetic means using empiric data and simulation. Statistics in Medicine. 2012 Jul 30;31(17):1857-86. PMID: 22438170.

43. Card NA. Applied meta-analysis for social science research. First ed. New York, NY: The Guilford Press; 2012.

44. Durlak JA, Durlak JA. How to select, calculate, and interpret effect sizes. Journal of Pediatric Psychology. 2009 Oct;34(9):917-28. PMID: 19223279.

45. Hedges LV, Olkin I. Statistical Methods for Meta-Analysis. 1st ed. Waltham, MA: Academic Press; 1985.

46. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.; 1988.

47. Van Den Noortgate W, Onghena P. Estimating the mean effect size in meta-analysis: bias, precision, and mean squared error of different weighting methods. Behav Res Methods Instrum Comput. 2003 Nov;35(4):504-11. PMID: 14748494.

48. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc. 2007 Nov;82(4):591-605. PMID: 17944619.

49. Furukawa TA, Barbui C, Cipriani A, et al. Imputing missing standard deviations in meta-analyses can provide accurate results. J Clin Epidemiol. 2006 Jan;59(1):7-10. PMID: 16360555.

50. Tendal B, Nuesch E, Higgins JP, et al. Multiplicity of data in trial reports and the reliability of meta-analyses: empirical study. BMJ. 2011;343:d4829. PMID: 21878462.

51. Adhikari NK, Burns KE, Friedrich JO, et al. Effect of nitric oxide on oxygenation and mortality in acute lung injury: systematic review and meta-analysis. BMJ. 2007 Apr 14;334(7597):779. PMID: 17383982.

52. Kunz R, Friedrich C, Wolbers M, et al. Meta-analysis: effect of monotherapy and combination therapy with inhibitors of the renin angiotensin system on proteinuria in renal disease. Annals of Internal Medicine. 2008 Jan 1;148(1):30-48. PMID: 17984482.

53. Peng PW, Wijeysundera DN, Li CC, et al. Use of gabapentin for perioperative pain control -- a meta-analysis. Pain Research & Management. 2007;12(2):85-92. PMID: 17505569.

54. Sud S, Sud M, Friedrich JO, et al. High frequency oscillation in patients with acute lung injury and acute respiratory distress syndrome (ARDS): systematic review and meta-analysis. BMJ. 2010;340:c2327. PMID: 20483951.

55. Tubach F, Ravaud P, Baron G, et al. Evaluation of clinically relevant states in patient reported outcomes in knee and hip osteoarthritis: the patient acceptable symptom state. Annals of the Rheumatic Diseases. 2005 January 1, 2005;64(1):34-7.

Appendix A. Search Strategies

Standardized Mean Difference

Ovid Medline (Date Searched 3/8/2012)

1	(standardized adj1 mean adj1 difference).ti,ab.	532
2	meta-analysis as topic/	12130
3	meta-analys$.ti,ab.	41814
4	exp statistics as topic/	1697404
5	meta-analysis.sh.	33853
6	2 or 3 or 5	59871
7	1 and 4 and 6	79

Current Index to Statistics (Date Searched 2/22/2012)

Keyword search using combinations of standardized mean difference

Baseline Imbalances

Ovid Medline (Date Searched 2/22/2012)

1	((imbalance* or balance* or distribution) and (pre-treatment or pretreatment or baseline or pre-intervention or preintervention or covariat*)).ti,ab.	18981
2	exp clinical trials as topic/	255550
3	meta-analysis as topic/	12130
4	"review literature as topic"/	4314
5	exp "bias (epidemiology)"/	45684
6	exp "analysis of variance"/	237153
7	((analys$ adj3 covarian$) or ANCOVA).ti,ab.	8690
8	data interpretation, statistical/	42335
9	3 or 4 or 5 or 6 or 7 or 8	338233

10	1 and 2 and 9	210

Current Index to Statistics (Date Searched 2/22/2012)

Keyword search using combinations of (imbalance* or balance* or distribution) and (pre-treatment or pretreatment or baseline or pre-intervention or preintervention or covariat*)

Scopus

Pearling search to identify additional relevant citations from relevant articles already identified.

Meta-analysis of Skewed Data

Ovid Medline (Date Searched: 3/8-20/2012), Current Index to Statistics, Scopus

Took Higgins et al article (Higgins, White and Anzures-Cabrera, "Meta-analysis of skewed data: combining results reported on log-transformed or raw scales." Stats in Med 2008; 27:6072-6092.) as a starting point but was unable to define a subject search that worked, so did a combination of keyword and pearling searches in Ovid Medline, Current Index to Statistics, and Scopus.

Means Ratios in Pooled Analyses and Categorizing for Continuous Outcomes

We searched Ovid MEDLINE(R) <1946 to January Week 4 2012> and PubMed on March 1st 2012 for (Dichotomis* or Dichotomiz*) limited to: Humans, Meta-Analysis, and English. We searched Web of Science for articles citing either of 2 known studies [1,2] in combination with a known author/expert (Friedrich, JO). Experts and reviewers also recommended references based on experience and reference list checking.