Research White Paper

Change Score or Followup Score?

An Empirical Evaluation of the Impact of Choice of Mean Difference Estimates

DEPARTMENT OF HEALTH & HUMAN SERVICES • USA

AHRQ
Agency for Healthcare Research and Quality
Advancing Excellence in Health Care • www.ahrq.gov

Research White Paper

Change Score or Followup Score?

An Empirical Evaluation of the Impact of Choice of Mean Difference Estimates

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. 290-2007-10057-I

Prepared by:
Oregon Evidence-based Practice Center
Portland, OR

Investigators:
Rongwei Fu, Ph.D.
Haley K. Holmer, M.P.H.

AHRQ Publication No. 15-EHC016-EF
April 2015

This report is based on research conducted by the Oregon Evidence-based Practice Center (EPC) under contract to the Agency for Healthcare Research and Quality (AHRQ), Rockville, MD (Contract No. 290-2007-10057-I). The findings and conclusions in this document are those of the authors, who are responsible for its contents; the findings and conclusions do not necessarily represent the views of AHRQ. Therefore, no statement in this report should be construed as an official position of AHRQ or of the U.S. Department of Health and Human Services.

The information in this report is intended to help health care decision makers—patients and clinicians, health system leaders, and policymakers, among others—make well-informed decisions and thereby improve the quality of health care services. This report is not intended to be a substitute for the application of clinical judgment. Anyone who makes decisions concerning the provision of clinical care should consider this report in the same way as any medical reference and in conjunction with all other pertinent information, i.e., in the context of available resources and circumstances presented by individual patients.

AHRQ or U.S. Department of Health and Human Services endorsement of any derivative products that may be developed from this report, such as clinical practice guidelines, other quality enhancement tools, or reimbursement or coverage policies may not be stated or implied.

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

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

Suggested citation: Fu R, Holmer HK. Change Score or Followup Score? An Empirical Evaluation of the Impact of Choice of Mean Difference Estimates. Research White Paper. (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-2007-10057-I.) AHRQ Publication No. 15-EHC016-EF. Rockville, MD: Agency for Healthcare Research and Quality. April 2015. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

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 Research White Paper. 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 email to epc@ahrq.hhs.gov.

Richard G. Kronick, Ph.D.
Director
Agency for Healthcare Research and Quality

David Meyers, M.D.
Acting Director
Center for Evidence and Practice Improvement
Agency for Healthcare Research and Quality

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

Acknowledgments

The authors gratefully acknowledge Elaine Graham, M.L.S., and Amber L. Laurie, M.S., for their contributions to this project.

Peer Reviewers

Prior to publication of the final report, the EPC sought input from independent Peer Reviewers without financial conflicts of interest. However, the conclusions and synthesis of the scientific literature presented in this report does not necessarily represent the views of individual reviewers.

Peer Reviewers must disclose any financial conflicts of interest greater than $10,000 and any other relevant business or professional conflicts of interest. Because of their unique clinical or content expertise, individuals with potential non-financial conflicts may be retained. The TOO and the EPC work to balance, manage, or mitigate any potential non-financial conflicts of interest identified.

The list of Peer Reviewers follows:

Doug Altman
Director, Centre for Statistics in Medicine
and Cancer Research UK Medical Statistics Group
University of Oxford
Oxford, United Kingdom

Joseph C. Cappelleri, Ph.D., M.P.H., M.S.
Pfizer Inc
Groton, CT

Bruno R. da Costa, Bsc.P.T., M.Sc., Ph.D.
Assistant Professor, Department of Physical Therapy
Florida International University
Miami, FL

Steven H. Fox, M.D., S.M., M.P.H.
Fulbright Senior Research Scholar
National Taiwan University
Taipei, Taiwan

George Papandonatos, Ph.D.
Associate Professor of Biostatistics
Brown University
Providence, RI

Change Score or Followup Score? An Empirical Evaluation of the Impact of Choice of Mean Difference Estimates

Structured Abstract

Background. In randomized controlled clinical trials, continuous outcomes are typically measured at both baseline and followup time points, and mean difference is analyzed as the effect measure. There are multiple ways to estimate the mean difference: using the change score from the baseline, using the followup scores, or estimating the mean difference using the analysis of covariance (ANCOVA) model. Use of the ANCOVA model is generally preferable to using the change scores from the baseline or using the followup scores. When the baseline scores are imbalanced, using either the change score from the baseline or the followup scores would produce biased effect estimates of mean difference, while the ANCOVA model provides the least biased estimates. Nonetheless, individual studies often report results incompletely, and investigators have to summarize results across studies that are not optimally reported. The impact of using the change versus the followup score on meta-analysis has not been well studied.

Methods. We selected six comparative effectiveness reviews published by the Agency for Healthcare Research and Quality that included at least one meta-analysis for continuous outcomes using mean difference. Data were abstracted from a total of 63 meta-analyses (156 trials) to evaluate differences in baseline scores and how the choice of using the change score or the followup score impacted the combined mean difference using a random effects model and discrepancy in conclusions. Discrepancy in conclusion occurs when one estimate (e.g., change score) shows significant difference and the other estimate (e.g., followup score) does not. We also evaluated whether the impact qualitatively varied by the comparator and alternative random effect estimates.

Results. Based on the Dersimonian-Laird (DL) method, using the change score versus the followup score led to 5 out of the 63 meta-analyses (7.9%) showing discrepancy in conclusions; based on the profile likelihood (PL) method, 9 (14.3%) showed discrepancy in conclusions. Using the change score was more likely to show a significant difference in effects between interventions (4 out of 5 using the DL method, and 7 out of 9 using the PL method). The impact of the change score versus the followup score using the maximum likelihood was similar to using the DL method, and the impact using the restricted maximum likelihood method was similar to using the PL method. Using the Knapp-Hartung modification of random effect estimate led to most (10) meta-analyses showing discrepancy in conclusions. A significant difference in baseline scores did not necessarily lead to discrepancy in conclusion. Finally, among the 10 meta-analyses that compared active intervention versus control or usual care, using the change score versus the followup score led to one discrepancy in conclusion using the DL or the PL method (10%) but using change scores consistently produced larger intervention effects in nine meta-analyses. Among the other 53 meta-analyses comparing different interventions, there were 4 discrepancies using the DL method (7.5%) and 8 discrepancies using the PL method (15.1%).

Conclusions. This study of 63 meta-analysis indicated that using the change score versus the followup score to estimate mean difference could lead to important discrepancies in conclusions. Using the change score is more likely to produce significant results when there are discrepancies in conclusions; using the followup score is more likely to produce more conservative results. Sensitivity analyses using both change scores and final values should be conducted to check the robustness of results to the choice of mean difference estimates.

Contents

Tables

Appendixes

Introduction

In randomized controlled clinical trials (RCTs), continuous outcomes are typically measured at both baseline and followup time points, and mean difference is analyzed as the effect measure. There are multiple ways to estimate the mean difference: using the change score from the baseline, using the followup scores, or estimating the mean difference using the analysis of covariance (ANCOVA) model with the baseline score as a covariate. When using the ANCOVA analysis, either the followup score or the change score can be used as the dependent variable, and they are equivalent in estimating the mean difference. All of these estimates provide unbiased estimates of mean difference when the clinical trials are adequately randomized, and the ANCOVA estimate provides a more efficient estimator with more precision.[1-3] Comparing using the change versus followup score, using the change score will provide a more precise estimate if the correlation between baseline and followup is high (> 0.5 if the standard deviations at baseline and followup are the same); otherwise, using the followup score will provide a more precise estimate.

The distribution of the baseline outcome scores is similar for adequately randomized RCTs. The distribution of the baseline outcome scores could become imbalanced in inadequately randomized trials, for example, due to chance, especially in small trials,[4] or due to selection bias, often caused by inadequate randomization concealment.[5] In addition, even for trials without baseline imbalance, systematic attrition linked to outcome is a concern, which may cause baseline imbalance (among patients with followup scores).[6] Baseline imbalance occurs quite commonly in clinical trials. Hartling et al.[7] reported that 35 percent of RCTs at high/unclear risk of bias in child health had imbalanced baseline distribution when this was subjectively evaluated as one of the methodological characteristics.

When the baseline scores are imbalanced, using either the change scores from the baseline or the followup scores would produce biased effect estimates of mean difference. Using the followup scores simply ignores baseline imbalance, and using the change score from the baseline, contrary to common belief, does not address the issue of the baseline imbalance. Instead, the change score is negatively associated with the baseline score, and patients with a worse baseline score are more likely to experience a high change score (regression to the mean). For instance, suppose that a trial has an intervention group and a placebo group, and the intervention group has a worse baseline score; then the treatment effect size from the intervention will be underestimated using the followup score and overestimated using the change score from baseline.[8] When baseline imbalance occurs by chance, the ANCOVA method removes conditional bias in treatment group comparisons and improves efficiency over unadjusted comparisons.[1-3] While the issue of baseline imbalance could be attenuated by using the ANCOVA model to adjust for baseline imbalance in individual studies, its impact on meta-analysis has not been well studied. In a meta-epidemiological study of osteoarthritis trials with pain scores as the clinical outcomes, Da Costa et al.[9] found no evidence for systematic differences between standardized mean differences (SMD) derived from followup and change scores, and concluded that it is valid to combine followup and change scores. However, the study did not consider whether the baseline pain scores were balanced in the included studies. Other studies have emphasized the importance of considering the impact of baseline imbalance in meta-analyses[10-12] and the use of ANCOVA estimates whenever possible.[10,12] Nonetheless, for meta-analyses using study-level estimates, which is still the dominant approach to conducting meta-analyses, the choice of mean difference estimates has to depend on the reported data. Individual studies don’t always report ANCOVA estimates, and they rarely report all the necessary data to calculate the ANCOVA estimates. Estimates using the change scores or followup scores for mean difference often become the practical choice.

Therefore, in this meta-epidemiological study, we empirically evaluated how the choice of using change scores or followup scores to estimate the mean difference impacted the meta-analyses, and whether the impact qualitatively varied by the comparator (whether the intervention was compared with a control group or multiple interventions were compared to each other), or linked with differences in baseline scores. For this paper, we specifically looked at mean difference for continuous outcomes measured in the same scale. In addition, we evaluated how different random effect models might affect the impact of the choice of mean difference estimates. A random effects model is generally recommended for combining continuous outcomes, and recently there has been a call to use alternative random-effects estimates to replace the universal use of Dersimonian-Laird random effects model.[13]

Methods

Selection and Abstraction of Data

We first reviewed all systematic reviews the Agency for Healthcare Research and Quality (AHRQ) published from 1999 to June 2012[14], including the general reviews conducted from 1999 to 2011 and the comparative effectiveness reviews (CERs) the AHRQ Evidence-based Practice Center Program conducted from 2005 to June 2012. For each review, we abstracted data on whether continuous outcome and meta-analysis were used, the effect measure of continuous outcomes (mean difference [MD] vs. standardized mean difference [SMD]), and meta-analysis methods. The use of MD versus SMD typically depends on the scales of the included outcomes. When the reported outcomes use the same scale, MD could be used; otherwise, SMD should be used. Additional abstracted information included whether baseline imbalance was addressed and other issues of analyzing continuous outcomes (e.g., how to handling missing data and skewed data).

Next, we selected six CERs to be used to evaluate the impact of using the change score or the followup score to estimate the MD. For this analysis, we only selected CERs conducted from 2005 onward, since most general systematic reviews were conducted earlier and CERs were more recent and relevant. To be included, the CER had to include at least one meta-analysis for continuous outcomes using mean difference. Only meta-analyses of RCTs were included in this study, and we specifically evaluated MD only. The CERs were selected to cover a wide range of commonly used continuous outcomes in health research (Table 1). We avoided including more than one CER that evaluated the same continuous outcome (duplicate outcome, e.g., if two CERs examined lipid variables or HbA1C, we only picked one of them). Also, we only considered the updated review if there was an earlier review and updated version (duplicate review). The medical conditions included hip fracture, obstructive sleep apnea, depression, diabetes, and alcohol abuse, and the interventions included pharmacologic, behavioral, and other therapies.

Table 1. Included comparative effectiveness reviews and continuous outcomes

Comparative Effectiveness Review Title	Publication Year	Continuous Outcomes	Number of Meta-Analyses
Pain Management Interventions for Hip Fracture[15]	2011	Acute pain	1
Diagnosis and Treatment of Obstructive Sleep Apnea in Adults[16]	2011	Apnea-Hypopnea Index (AHI), Epworth Sleepiness Scale (ESS)	6
Nonpharmacologic Interventions for Treatment-Resistant Depression in Adults[17]	2011	Depressive severity	3
Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression: An Update of the 2007 Comparative Effectiveness Review[18]	2011	Montgomery–Åsberg Depression Rating Scale (MADRS)	1
Oral Diabetes Medications for Adults With Type 2 Diabetes: An Update of the 2007 Report[19]	2011	Hemglobin A1C, Weight, LDL, HDL, Triglycerides	50
Screening, Behavioral Counseling, and Referral in Primary Care To Reduce Alcohol Misuse[20]	2012	Weekly alcohol use	2

HDL = high-density lipoprotein; LDL = low-density lipoprotein.

A total of 63 meta-analyses were evaluated in this study; each meta-analysis needed to include at least 3 studies in order to be included. For each meta-analysis, we identified all original publications. One investigator abstracted data on outcomes, comparison groups, the analysis method and all data from the publication at baseline, followup, and any change data including estimates from analysis of covariance (ANCOVA) model, and a second investigator reviewed data abstraction for accuracy. We abstracted data based on the comparisons and time points included in each meta-analysis in the CER but did not use data from the CERs in our evaluation, as the data from the CERs were not adequate for the purpose of this study.

Statistical Analysis

The characteristics of the systematic reviews were summarized descriptively. For each meta-analysis, we calculated two mean differences and the associated standard errors based on change score and followup score. When the standard deviation (SD) for baseline or followup score was missing, it was imputed using the mean SD from studies with reported SDs in that meta-analysis. The standard deviation of the change score, when not reported, was calculated from baseline and followup SDs by assuming that the correlation between baseline and followup was 0.5. When mean difference could not be calculated based on change score or followup score due to inadequately reported data, we used the ANCOVA estimate, or other estimate of mean difference available or reported in the publication (e.g., an estimate from a mixed effects model). When studies reported geometric mean and its standard deviation, we converted them to the mean and standard deviation on the raw scale.[21]

The mean difference estimates based on the change score or the followup score were separately combined using random effects models. We also assessed whether the baseline scores imbalance across the included studies was due to chance by meta-analyzing the baseline score differences between treatment groups.[11] The patterns of baseline scores imbalance across the included studies could vary. However, if the baseline scores imbalance occurred by chance, the combined overall baseline score difference between treatment groups should be close to zero, in particular when the number of studies and the total number of patients randomized are large. This is a different issue from evaluating baseline scores imbalance within a single RCT.[1,22,23]

Given that different random effect models might affect the results of using different mean difference estimates, we evaluated alternative random effect estimators, and each meta-analysis was conducted using six random effects estimates: the Dersimonian-Laird (DL) method, the profile likelihood (PL) method, the maximum likelihood (ML) method, the restricted maximum likelihood (REML) method, the permutation (PE) method, and the Knapp-Hartung modification (KH) of random effect estimate.[24] The primary analyses focused on results from the DL and PL methods as they are the methods with better performance.[25,26]

We assessed the presence of statistical heterogeneity among the studies by using the standard Cochran's chi-square test, and the magnitude of heterogeneity by using the I^2 statistic.[27] In addition, we conducted sensitivity analysis by using ANCOVA estimates whenever they were available (not only when the mean difference could not be calculated based on change score and followup score) and by assuming different values of correlation between baseline and followup. We did not use the ANCOVA estimates in the primary analyses as we aimed to focus on the comparison between using the change versus the followup score. For each random effect method, we qualitatively compared the combined estimates using the change score and followup score to see whether there was discrepancy in conclusion. *Discrepancy in conclusion* means one estimate shows statistically significant difference and the other estimate does not (i.e., the combined estimate using the mean difference based on the change score shows a statistically significant difference, while using the followup score does not; or vice versa). Further, *qualitative difference* means that two estimates show a difference in the magnitude of effect, but no discrepancy in conclusion.

All analyses were performed using Stata/IC 13.1 (StataCorp, College Station, TX).

Results

Use of Continuous Outcomes in AHRQ Systematic Reviews

Between 1999 and June 2012, a total of 263 AHRQ systematic reviews were conducted and 247 reviews evaluated at least one continuous outcome. Among these reviews, 133 (50.6%) included a meta-analysis and 74 (28.1%) included a meta-analysis of continuous outcomes, which was 55.6 percent among those with a meta-analysis conducted. However, only four reviews explicitly mentioned the issue of baseline imbalance. In particular, one review used meta-regression to adjust for baseline imbalance; one review stated that meta-regression was not used, citing concerns for ecological fallacy. The other two reviews only combined studies in the presence of baseline imbalance when mean change score from baseline could be used in the meta-analysis, though they failed to recognize that using change score did not address the problem of baseline imbalance.

Among the 74 reviews including a meta-analysis of continuous outcomes, 41 (55.4%) used MD exclusively, 21 (28.4%) used SMD exclusively, 9 (12.2%) used both MD and SMD, 2 (2.7%) used percent change and 1 (1.4%) used both MD and percent change. In addition, 16 of the 74 reviews (21.6%) provided some description on how to handling missing standard deviation, 5 (6.8%) described how to handle missing correlation between baseline and followup point, and 4 reviews (5.4%) mentioned using median to approximate mean values (for skewed data).

A total of 63 CERs were conducted between 2005 to June 2012, and 19 of them conducted a meta-analysis using mean difference. Six CERs were selected for evaluation of use of the change score versus the followup score, and other CERs were excluded due to duplicate reviews, duplicate outcomes, small number of studies (<3), inclusion of variables that were not measured at both baseline and followup (e.g., birth weight, length of stay), or lack of forest plots to identify which studies and data were included in the meta-analyses. A total of 63 meta-analyses from six CERs[15-20] were included in the following evaluation.

Impact of the Mean Difference Estimates

One CER[15] reported using both change score and followup score to estimate mean difference and the others[16-20] reported using change score. These meta-analyses included 156 trials, among which 58 trials (37.2%) reported using an ANCOVA model, though only 16 trials reported at least one ANCOVA estimate, less than 30 percent of trials that used an ANCOVA model. Comparisons of results using change score versus followup score were shown in Table 2 and Table 3, though Table 2 and Table 3 only included results based on Dersimonian-Laird (DL) method and profile likelihood (PL) method and Table 3 only presented meta-analyses with discrepancy in conclusion or significant baseline difference. The complete results for all analyses based on all six random effects estimates were shown in appendixes A and B.

Based on the DL method, using change score versus followup score led to 5 out of the 63 meta-analysis (7.9%) showing discrepancy in conclusions; and based on the PL method, 9 (14.3%) showed discrepancy in conclusions (see bolded values in tables 2 and 3; only discrepancies based on DL and PL methods are bolded). In general, using the change score is more likely to show a significant difference in effects between interventions. For the five meta-analyses showing discrepancies using DL method, four showed significant differences when using change score, and one showed significant result when using the followup score. For the nine meta-analyses showing discrepancies using PL method, seven showed significant differences when using change score and two showed significant differences when using the followup score. Therefore, though not necessarily more accurate, using the followup score is more likely to produce conservative results based on these comparisons.

Compared with the DL method, estimates based on the maximum likelihood (ML) and restricted maximum likelihood (REML) methods tended to have narrower 95 percent confidence intervals (CIs), though the REML method sometimes provided more conservative 95 percent CIs. The permutation (PE) method worked only when there were at least six studies in a meta-analysis and provided much wider, and often unrealistically wide 95 percent CIs. In addition, the PE 95 percent CIs are often practically invalid with lower bounds like -4.8e+15 when there were only six studies. The Knapp-Hartung modification (KH) produced narrower 95 percent CIs compared to the PE method, but wider 95 percent CIs than others. Using ML method led to discrepancy in conclusions in six meta-analyses and the impact of change score versus followup score using the ML method is largely similar to using the DL method. Interestingly, the impact using the REML method is similar to using the PL method with eight meta-analyses showing discrepancy in conclusions (see italicized values on appendixes A and B). Estimates from PE method are not available in many cases, and the 95 percent CIs of PE method, when present, are too wide to make meaningful comparisons. Lastly, using the KH modification led to a discrepancy in conclusions in ten meta-analyses, the highest among the different methods, seven showed significant differences when using the change score, and three showed significant differences when using the followup score.

Sensitivity analyses using ANCOVA estimates (whenever they were available) did not change the results, which may be partly because only a few ANCOVA estimates were reported. Sensitivity analyses assuming different values of correlation between baseline and followup from 0.3 to 0.7 also generally provided the same results.

Additional details of the results are provided below, with results stratified by the comparison of intervention versus control and the comparison between interventions.

Meta-Analysis of Intervention Versus Control

Ten meta-analyses from four CERs[15-17,20] compared active intervention versus control or usual care on pain, Apnea-Hypopnea Index, Epworth Sleepiness Scale, depression, and alcohol use. These meta-analyses included 5 to 13 studies and 218 to 4,100 patients (Table 2).

When the baseline imbalance occurs by chance, the combined baseline difference across included studies should be close to zero. The combined baseline differences indicated significant imbalance (different from zero) in 1 of the 10 meta-analyses (Apnea-Hypopnea Index, continuous positive airway pressure [CPAP] vs. sham CPAP), and the combined mean differences using change score versus endpoint also led to discrepancy in conclusions in another meta-analysis (depressive severity, repetitive transcranial magnetic stimulation vs. sham for condition = Tier1, major depressive disorder). Therefore the significant baseline imbalance does not necessarily coincide with the discrepancy in conclusions. For this particular case (Apnea-Hypopnea Index, CPAP vs. sham CPAP), the magnitude of the mean difference is large and so the baseline difference only led to qualitative differences in the combined mean difference.

For the one meta-analysis showing discrepancy in conclusions, using the change score showed significant results based on both PL and DL methods (as well as REML and the KH modification method, Appendix A). However, more interestingly, except for one meta-analysis (acute pain, skin traction vs. no traction), the combined mean difference using change score consistently showed a larger intervention effect than the combined mean difference using endpoint score, by a considerable amount. Compared with the combined mean difference using change score, the combined mean difference using endpoint score showed an intervention effect of about 20 percent smaller on average, ranging from 5 percent to more than 40 percent.

For these comparisons, the magnitude of heterogeneity and the results of testing heterogeneity were generally similar between the two estimates.

Meta-Analysis of Comparison Between Interventions

Fifty-three meta-analyses from three CERs[16,18,19] compared outcomes between active interventions. Most (50) of these are from one CER[19] comparing the various oral diabetes medications for adults with type 2 diabetes on weight, A1C, and lipid variables. The other three meta-analyses compared sleepiness[16] and depression[18] variables. These meta-analyses included 3 to 17 studies and 215 to 3,252 patients (Table 3).

Based on the DL method, using the change score versus the followup score led to 4 out of the 53 meta-analyses (7.5%) showing discrepancy in conclusions, and 3 showed significant results when using the change score. At the same time, significant baseline differences occurred in six meta-analyses (highlighted in Table 3), and three meta-analyses showed both significant baseline difference and discrepancy in conclusions. Based on the PL method, eight (15.1%) showed discrepancy in conclusions, with six showing significant results when using the change score and two showing significant results using the followup score. Only two meta-analyses showed significant baseline differences using the PL method, and one is associated with discrepancy in conclusions. On the other hand, not all studies in each meta-analysis provided data to meta-analyze baseline difference (though most did), so the results of baseline difference may not fully reflect the baseline distribution of the included studies. We did not calculate the percent difference between the two combined mean differences since there was no clearly defined (active) control for these comparisons.

For most meta-analyses, the magnitude of heterogeneity and the results of testing heterogeneity were comparable between the two measures. When heterogeneity showed a difference between the two estimates, there was no clear pattern of one estimate having more heterogeneity than the other. In the presence of a discrepancy in conclusions, the estimate with significant results did not necessarily have less heterogeneity among studies (lower I^2 values). In addition, the number of individual studies included in meta-analyses with discrepant conclusions varied from 3 to 14, the total number of patients varied from a few hundred to a few thousand, and the studies reported various outcomes (depressive severity, Montgomery-Asberg Depression Rating Scale, HbA1C, weight, lipid variables). These meta-analyses revealed no clear common characteristics.

Table 2. Comparison of combined mean differences using change score and endpoints (treatment versus control) based on PL and DL methods

Comparative Effectiveness Review	Outcome: Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min–Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)	Percentage Difference in Combined Estimates
Pain Management Interventions for Hip Fracture[15]	Acute pain: skin traction vs. no traction	13 studies #1: 568 (30–166) #2: 662 (34–151)	PL: -0.300 (-0.812, 0.260) DL: -0.281 (-0.843, 0.280) I^2 = 77.2%, P < 0.001	PL: 0.432 (-0.137, 0.997) DL: 0.431 (-0.125, 0.987) I^2 = 76.5%, P < 0.001	PL: 0.115 (-0.332, 0.617) DL: 0.122 (-0.316, 0.561) I^2 = 63.2%, P = 0.008	PL: 73% DL: 72%
Diagnosis and Treatment of Obstructive Sleep Apnea in Adults[16]	Apnea-Hypopnea Index: CPAP vs. control	6 studies #1: 177 (12–66) #2: 159 (12–59)	PL: 2.208 (-3.442, 9.009) DL: 2.532 (-2.798, 7.862) I^2 = 64.7%, P = 0.015	PL: -27.022 (-41.194, -13.631) DL: -27.051 (-38.909, -15.193) I^2 = 93.3%, P < 0.001	PL: -24.176 (-36.050, -13.153) DL: -24.134 (-33.599, -14.669) I^2 = 91.9%, P < 0.001	PL: -11% DL: -11%
	Epworth Sleepiness Scale: CPAP vs. control	7 studies #1: 448 (19–178) #2: 398 (21–181)	PL: -0.015 (-0.523, 0.494) DL: -0.015 (-0.435, 0.406) I^2 = 0%, P = 0.805	PL: -2.684 (-4.311, -1.169) DL: -2.714 (-4.274, -1.155) I^2 = 84.3%, P < 0.001	PL: -2.325 (-3.645, -1.220) DL: -2.371 -3.415, -1.327) I^2 = 54.2%, P = 0.042	PL: -13% DL: -13%
	Apnea-Hypopnea Index: CPAP vs. sham CPAP	8 studies #1: 163 (15–27) #2: 149 (10–29)	**PL: 7.161 (1.241, 13.088)**[a] **DL: 7.161 (1.260, 13.063)**[a] I^2 = 0%, P = 0.920	PL: -45.634 (-58.285, -34.180) DL: -45.891 (-57.534, -34.249) I^2 = 70.9%, P = 0.001	PL: -40.495 (-52.287, -29.403) DL: -40.690 (-52.073, -29.307) I^2 = 82.4%, P < 0.001	PL: -11% DL: -11%
	Epworth Sleepiness Scale: CPAP vs. sham CPAP	11 studies #1: 293 (16–52) #2: 291 16–49)	PL: 0.309 (-0.263, 0.876) DL: 0.309 (-0.258, 0.876) I^2 = 0%, P = 0.920	PL: -2.684 (-4.345, -1.033) DL: -2.688 (-4.399, -0.977) I^2 = 83.4%, P < 0.001	PL: -2.555 (-4.197, -0.918) DL: -2.556 (-4.209, -0.904) I^2 = 82.2%, P < 0.001	DL: -5% PL: -5%
Nonpharmacologic Interventions for Treatment-Resistant Depression in Adults[17]	Depressive severity: rTMS vs. sham (condition = Tier 1, MDD)	8 studies #1: 116 (7–32) #2: 102 (5–31)	PL: 0.638 (-0.530, 2.671) DL: 0.638 (-0.460, 1.736) I^2 = 0%, P = 0.520	**PL: -5.442 (-7.792, -2.672)**[a] **DL: -5.394 (-7.764, -3.025)**[a] I^2 = 43.9%, P = 0.086	**PL: -3.068 (-6.171, 0.371)**[a] **DL: -2.898 (-6.508, 0.711)**[a] I^2 = 79.2%, P < 0.001	PL: -44% DL: -46%
	Depressive severity: rTMS vs. sham (condition = Tier 1)	11 studies #1: 197 (7–36) #2: 149 (5–31)	PL: 0.680 (-0.374, 2.276) DL: 0.680 (-0.276, 1.636) I^2 = 0%, P = 0.452	PL: -5.690 (-7.813, -3.478) DL: -5.672 (-7.784, -3.561) I^2 = 54.4%, P = 0.015	PL: -3.906 (-6.284, -1.401) DL: -3.841 (-6.482, -1.199) I^2 = 73.6%, P < 0.001	PL: -31% DL: 32%
	Depressive severity: rTMS vs. sham (condition = Tier 1&2, MDD)	12 studies #1: 374 (7–155) #2: 364 (5–146)	PL: -0.013 (-0.595, 0.764) DL: -0.013 (-0.569, 0.544) I^2 = 0%, P = 0.688	PL: -4.719 (-7.097, -2.319) DL: -4.714 (-7.126, -2.301) I^2 = 80.4%, P < 0.001	PL: -3.441 (-5.943, -0.802) DL: -3.436 (-5.83, -1.041) I^2 = 79.3%, P < 0.001	PL: -27% DL: -27%
Screening, Behavioral Counseling, and Referral in Primary Care To Reduce Alcohol Misuse[20]	Drinks/week: BCI vs. control (adults, 6 months)	11 studies #1: 1547 (39-353) #2: 1556 (32-376)	PL: 0.613 (-0.247, 1.606) DL: 0.613 (-0.229, 1.456) I^2 = 0%, P = 0.490	PL: -3.121 (-4.255, -2.310) DL: -3.228 (-4.214, -2.242) I^2 =13.9%, P = 0.311	PL: -2.504 (-4.037, -1.210) DL: -2.593 (-4.063, -1.123) I^2 =46.3%, P = 0.046	PL: -20% DL: -20%
	Drinks/week: BCI vs. control (adults, 12 months)	13 studies #1: 2088 (33-371) #2: 2012 (39-381)	PL: 0.455 (-0.403, 1.282) DL: 0.455 (-0.364, 1.274) I^2 = 0%, P = 0.649	PL: -3.700 (-4.641, -2.805) DL: -3.718 (-4.760, -2.677) I^2 =16.9%, P = 0.274	PL: -2.920 (-4.912, -0.800) DL: -2.936 (-4.601, -1.272) I^2 =57.2%, P = 0.005	PL: -21% DL: -21%

[a] Significant baseline difference or discrepancy in conclusion.

BCI = behavioral counseling intervention; CI = confidence interval; CPAP = continuous positive airway pressure; DL = DerSimonian-Laird random-effects method; MDD = major depressive disorderPL = profile likelihood method; rTMS = repetitive transcranial magnetic stimulation.

Table 3. Discrepancy in comparison of combined mean differences using change score and endpoints (comparison between different treatments) based on PL and DL methods

Comparative Effectiveness Review	**Outcome** **Comparison (Group #1 vs. Group #2)**	**Number of Studies (N)** **Group #1 and Group #2 Total Sample Size (Min-Max)**	**Baseline Difference (95% CI)**	**Difference in Change Score (95% CI)**	**Difference in Endpoint Score (95% CI)**
Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression: An Update of the 2007 Comparative Effectiveness Review[18]	MADRS: citalopram vs. escitalopram	6 studies #1: 939 (125–214) #2: 932 (108–241)	PL: 0.178 (-0.341, 0.646) DL: 0.171 (-0.283, 0.625) I^2 = 12.7%, P = 0.333	**PL: 2.108 (0.078, 4.008)**[a] DL: 2.077 (0.174, 3.979) I^2 = 67.8%, P = 0.008	**PL: 2.258 (-0.073, 4.430)**[a] DL: 2.221 (0.060, 4.383) I^2 = 68.7%, P = 0.007

Table 3. Discrepancy in comparison of combined mean differences using change score and endpoints (comparison between different treatments) based on PL and DL methods, (continued)

Comparative Effectiveness Review	Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
Oral Diabetes Medications for Adults With Type 2 Diabetes: An Update of the 2007 Report[19]	HbA1c: metformin vs. thiazolidinediones	14 studies #1: 1132 (13–501) #2: 1127 (14–499)	PL: -0.017 (-0.097, 0.062) DL: -0.017 (-0.109, 0.074) I^2 = 9.7%, P = 0.347	**PL: -0.105 (-0.188, -0.025)**[a] DL: -0.106 (-0.214, 0.003) I^2 = 24.9%, P = 0.186	**PL: -0.054 (-0.186, 0.103)**[a] DL: -0.053 (-0.182, 0.077) I^2 = 58.0%, P = 0.003
	HbA1c: metformin vs. metformin and thiazolidinediones	11 studies #1: 1428 (34–277) #2: 1688 (60–296)	**PL: -0.160 (-0.283, -0.034)**[a] **DL: -0.158 (-0.290, -0.027)**[a] I^2 = 69.6%, P = 0.001	PL: 0.633 (0.432, 0.854) DL: 0.635 (0.437, 0.832) I^2 = 85.5%, P < 0.001	PL: 0.503 (0.280, 0.733) DL: 0.506 (0.268, 0.744) I^2 = 93.4%, P < 0.001
	HbA1c: metformin and Sulfonylureas vs. Thiazolidinediones and Sulfonylureas	6 studies #1: 847 (37–320) #2: 871 (34–319)	PL: -0.123 (-0.238, 0.008) **DL: -0.126 (-0.217, -0.034)**[a] I^2 = 8.3%, P = 0.363	**PL: -0.050 (-0.151, 0.069)**[a] **DL: -0.050 (-0.148, 0.047)**[a] I^2 = 0%, P = 0.628	**PL: -0.165 (-0.262, -0.061)**[a] **DL: -0.165 (-0.262, -0.068)**[a] I^2 = 30.7%, P = 0.205
	Weight: metformin vs. sulfonylureas (studies < 24 weeks in duration)	8 studies #1: 718 (21–164) #2: 797 (18–161)	**PL: 1.669 (0.115, 3.208)**[a] **DL: 1.669 (0.163, 3.175)**[a] I^2 = 0%, P = 0.939	**PL: -2.240 (-2.747, -1.792)**[a] **DL: -2.234 (-2.640, -1.828)**[a] I^2 = 2.8%, P = 0.408	**PL: -0.634 (-2.144, 0.852)**[a] **DL: -0.634 (-2.118, 0.850)**[a] I^2 = 0%, P = 0.883
	Weight: metformin and sulfonylureas vs. combination thiazolidinediones and sulfonylureas.	4 studies #1: 653 (37–320) #2: 646 (34–319)	PL: 1.746 (-2.856, 6.117) DL: 1.672 (-2.815, 6.160) I^2 = 87.5%, P < 0.001	**PL: -2.689** (**-3.840**, **-1.550**)[a] **DL: -2.689** (**-3.747**, **-1.631**)[a] I^2 = 78.1%, P = 0.003	**PL: -0.935** (**-6.105**, **3.972**)[a] **DL: -1.038** (**-6.510**, **4.434**)[a] I^2 = 90.8%, P < 0.001
	LDL: metformin vs. rosiglitazone	6 studies #1: 198 (9–117) #2: 213 (14–128)	PL: 2.100 (-4.188, 8.886) DL: 2.100 (-3.960, 8.159) I^2 = 0.0%, P = 0.682	**PL: -12.535 (-22.237, -5.876)**[a] DL: -13.263 (-20.553, -5.974) I^2 = 58.0%, P = 0.036	**PL: -14.009 (-29.491, 2.268)**[a] DL: -13.944 (-27.561, -0.327) I^2 = 65.7%, P = 0.012
	HDL: metformin vs. rosiglitazone	6 studies #1: 198 (9–117) #2: 213 (14–128)	PL: Does not converge **DL: -2.408 (-4.400, -0.416)**[a] I^2 = 0.0%, P = 0.673	PL: -0.595 (-1.658, 2.319) DL: -0.052 (-1.877, 1.774) I^2 = 42.6%, P = 0.121	PL: 0.445 (-2.983, 3.832) DL: 0.445 (-2.941, 3.831) I^2 = 0%, P = 0.922
	HDL: metformin vs. DPP-4 inhibitors	3 studies #1: 913 (241–427) #2: 791 (95–440)	PL: 0.711 (-0.271, 1.673) DL: 0.711 (-0.217, 1.639) I^2 = 0.0%, P = 0. 787	**PL: 1.432 (-0.004, 3.251)**[a] DL: 1.514 (0.132, 2.897) I^2 = 34.1%, P = 0.219	**PL: 2.112 (0.705, 3.899)**[a] DL: 2.199 (0.814, 3.584) I^2 = 33.9%, P = 0.220

Table 3. Discrepancy in comparison of combined mean differences using change score and endpoints (comparison between different treatments) based on PL and DL methods, (continued)

Comparative Effectiveness Review	Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
	HDL: pioglitazone vs. sulfonylurea	6 studies #1: 304 (17–91) #2: 311 (18–109)	PL: 1.375 (-0.020, 2.695) **DL: 1.375 (0.074, 2.675)**[a] I^2 = 0.0%, P = 0. 706	PL: 5.295 (3.480, 6.964) DL: 5.278 (3.472, 7.084) I^2 = 43.3%, P = 0.117	PL: 6.369 (3.528, 8.853) DL: 6.325 (4.021, 8.630) I^2 = 60.5%, P = 0.027
	Triglycerides: metformin vs. sulfonylureas	11 studies #1: 812 (19–210) #2: 858 (18–209)	PL: 17.186 (-2.503, 23.406) **DL: 17.186 (10.965, 23.406)**[a] I^2 = 0.0%, P = 0. 500	**PL: -17.217 (-28.683, -4.876)**[a] **DL: -17.217 (-28.684, -5.750)**[a] I^2 = 0.0%, P = 0.669	**PL: -5.847 (-23.606, 10.612)**[a] **DL: -6.444 (-24.536, 11.648)**[a] I^2 = 67.9%, P = 0.001

[a] Significant baseline difference or discrepancy in conclusion.

CI = confidence interval; DL = DerSimonian-Laird random-effects method; DPP-4 = Dipeptidyl peptidase-4; HbA1c = hemoglobin A1c/glycated hemoglobin; HDL = high-density lipoprotein; LDL = low-density lipoprotein; MADRS = Montgomery-Asberg Depression Rating Scale; PL = profile likelihood method.

Discussion

The issue of baseline imbalance was inadequately addressed in systematic reviews and comparative effectiveness reviews sponsored by AHRQ. To our knowledge, this is the first empirical evaluation of how using the change versus the followup score would affect the combined mean differences. Discrepancy in conclusions was shown in 5 out of the 63 meta-analysis (7.9%) using the DL method and 9 (14.3%) using the PL method. In general, the mean difference based on the change score was more likely to show significant results. While the conclusions were consistent in most cases, it was concerning to see that up to 14.3 percent of results were not. Such discrepancy emphasizes the need for careful selection of mean difference estimates, in particular, given the recent call for using alternative random-effects estimates, like the PL estimate, to replace the universal use of the DL random effects model.[13] The common misconception was that using change scores accounts for baseline difference, and five of the six CERs used the change score to estimate the mean difference, even though using the change score does not address the baseline imbalance. These results support the current AHRQ guidance that advises review authors to conduct sensitivity analyses using both scores to assess the robustness of results to the choice of mean difference estimates.[12]

It was interesting that in the meta-analyses of intervention versus controls, the combined mean differences using the change score consistently showed larger treatment effect in all but one meta-analysis, and the relative difference could be as large as more than 40 percent higher. Since the change score was negatively associated with the baseline score, such results would occur when the baseline scores in the intervention groups were systematically worse. This was also roughly shown by the combined baseline differences and suggested potential bias in randomization that may not de due to chance (patients with a more severe condition were more likely to be randomized to the intervention group). Baseline imbalance due to systematic bias poses a more serious problem than baseline imbalance due to chance, which would not be as effectively adjusted for by using ANCOVA models. This finding warrants further study to evaluate its prevalence and impact in the literature.

Whenever there is baseline imbalance, using the change versus followup score to estimate mean difference would lead to some qualitative difference in the combined mean difference. Nevertheless, significant baseline difference does not necessarily lead to discrepancy in conclusions and discrepancy in conclusions does not necessarily occur when there is significant baseline difference. No other common characteristics relating to the size and number of included studies, type of outcomes, and between-study heterogeneity among discrepant meta-analyses were identified. When the correlation between baseline and followup is large, the standard error of the mean difference based on the change score will be smaller than the mean difference based on the followup score. This may play a role why the mean difference based on the change score was more likely to show significant results.

Both qualitative difference and discrepancy in conclusions are important. Sometimes, discrepancy in conclusion only reflects small shifts in numerical values. However, such shifts could be critical to consider and may have potential health policy implications since it is unavoidable to use the cutoff points of *P*-values in the current scientific world and some health decisions could be affected by such shifts. While this study is not powered to look at the association between significant baseline difference and discrepancy in conclusions, in theory, discrepancy in conclusions should be a function of the magnitude of baseline difference, magnitude of effect, and heterogeneity between (and within) studies. When the significance of effect is closer to borderline, significant baseline difference may be more likely to lead to discrepancy in conclusions. Whether or not significant baseline difference is directly linked with discrepancy in conclusions, it highlights the importance to evaluate specifically baseline imbalance for each meta-analysis.

Results on discrepancy in conclusions based on the ML method are similar to the DL method and those based on the REML method are similar to the PL method. Further, an additional purpose of comparing these methods was to provide some empirical examples and practical sense of how much the estimates differed among the methods. The results were consistent with simulation studies in that the DL and PL methods generally provided wider CIs and better coverage probability than ML and REML methods, though the PL method did not converge in a small proportion of meta-analyses, and the 95 percent CIs based on the PE method are too wide and conservative for general use.[25,26] Interestingly, the KH modification produced the most discrepancy, and it generally produced a wider CI than the DL and PL methods. Though the KH modification has been shown to result in more adequate error rates than the DL method,[28] its overall performance has been shown not to be better than the DL and PL methods.[25]

We did not use the Trowman's method[11] or the modified Trowman's method[10] to adjust summary baseline score using meta-regression because it has been shown that such methods could provide misleading results.[10] There is general agreement within the literature that estimates from the ANCOVA model are least biased to estimate mean difference in the presence of baseline imbalance due to chance. When the variance of the baseline score equals the variance of the followup score, the ANCOVA estimate is the weighted sum of the two estimates from change score and followup score, and the weight is the correlation between baseline and followup score.[29] However, meta-regression adjusts the summary baseline score only in the study level. It is an ANCOVA-analogue analysis using study level data and ecological fallacy could play a role here. Results from Riley et al.[10] showed that the adjustment using ANCOVA model and individual patient-level data (IPD) could not be achieved by using summary baseline score in meta-regression, but were limited to a couple of examples. More research is needed to determine whether adjusting the summary baseline score in a meta-regression is useful at all.

In the presence of baseline imbalance, the ANCOVA estimates should be used in a meta-analysis whenever reported.[10,12] Unfortunately, in this study, we found that the reporting of ANCOVA estimates has been poor, even though many trials used an ANCOVA model in their analyses. The importance of adequate reporting of such estimates needs to be better recognized among the authors and journal editors. The actual ANCOVA estimates should be reported when such analyses were conducted so they could be properly used in evidence synthesis and potential health policy decision making. On the other hand, this is one situation where availability of IPD would provide the best solution, where baseline imbalance could be consistently adjusted using ANCOVA model across studies. IPD could also address the concern that systematic attrition may lead to baseline imbalance (among patients with followup scores),[6] and provide the means to better handle missing data in general.

We tried to evaluate a wide range of outcomes in this analysis, but choice of outcomes was limited by the outcomes studied in CERs published by AHRQ during the study period. Most outcomes were related to diabetes and came from one review[19] with multiple studies included in more than one meta-analysis. This would affect the generalizability the results. However, these meta-analyses did address multiple key questions, outcomes, interventions and comparators, and it was valuable to study the choice of estimates for mean difference in such a complex CER. Many other commonly used continuous outcomes such as pain, quality of life, and functional status were not assessed or not adequately assessed. They will be evaluated in future research.

We specifically focused only on MD in this study without looking at SMD. Baseline imbalance directly affects the MD, but the choice of change score versus followup score has further implications on estimating SMD. The standard deviation is also involved in calculating SMD and the standard deviations are calculated differently when using change score versus followup score. Da Costa et al.[9] found no evidence for systematic differences between SMDs derived from followup and change pain scores without evaluating the impact of baseline scores, but as MD, the overall evidence is still limited.

Only RCTs were included in this study and RCT is the most commonly used study design of the included studies in CERs to evaluate effectiveness. Other designs, like cohort or other observational studies were not considered. The issue of baseline imbalance in observational studies was different from RCTs since baseline balance was not expected and Lord's paradox could further complicate the analysis.[30]

In summary, using the change score versus the followup score to estimate MD could lead to important discrepancy in conclusions. Using the change score is more likely to produce significant results when there are discrepancies in conclusions based on this study, though the overall evidence is limited; and using the followup score is more likely produce more conservative results. Sensitivity analyses using both change scores and final values should be conducted to check the robustness of results to the choice of mean difference estimates.

References

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

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

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

4. Rosenberger W, Lachin J, eds. Randomization in Clinical Trials: Theory and Practice. New York: Wiley; 2002.

5. Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet. 2002;359(9306):614-8. PMID: 11867132.

6. 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.

7. Hartling L, Hamm MP, Fernandes RM, et al. Quantifying bias in randomized controlled trials in child health: a meta-epidemiological study. PLoS One. 2014;9(2):e88008. PMID: 24505351.

8. Vickers AJ, Altman, D. G. Statistics notes: Analysing controlled trials with baseline and follow up measurements. BMJ. 2001;323:1123-4. PMID: 11701584.

9. da Costa BR, Nuesch E, Rutjes AW, et al. Combining follow-up and change data is valid in meta-analyses of continuous outcomes: a meta-epidemiological study. J Clin Epidemiol. 2013 Aug;66(8):847-55. PMID: 23747228.

10. Riley RD, Kauser I, Bland M, et al. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. Stat Med. 2013 Jul 20;32(16):2747-66. PMID: 23303608.

11. 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.

12. Fu R, Vandermeer BW, Shamliyan TA, et al. Handling Continuous Outcomes in Quantitative Synthesis. Methods Guide for Comparative Effectiveness Reviews. (Prepared by the Oregon Evidence-based Practice Center under Contract No. 290-2007-10057-I.) AHRQ Publication No. 13-EHC103-EF. Rockville, MD: Agency for Healthcare Research and Quality; July 2013. PMID: 24006546. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

13. Cornell JE, Mulrow CD, Localio R, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Ann Intern Med. 2014 Feb 18;160(4):267-70. PMID: 24727843.

14. Evidence-based Practice Center Reports by Year. Rockville, MD: Agency for Healthcare Research and Quality; 2014. www.ahrq.gov/research/findings/evidence-based-reports/year/index.html. Accessed April 1, 2015.

15. Abou-Setta AM, Beaupre LA, Jones CA, et al. Pain Management Interventions for Hip Fracture. Comparative Effectiveness Review No. 30. (Prepared by the University of Alberta Evidence-based Practice Center under Contract No. 290-02-0023.) AHRQ Publication No. 11-EHC022-EF. Rockville, MD: Agency for Healthcare Research and Quality; May 2011. PMID: 21938799. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

16. Balk EM, Moorthy D, Obadan NO, et al. Diagnosis and Treatment of Obstructive Sleep Apnea in Adults. Comparative Effectiveness Review No. 32. (Prepared by Tufts Evidence-based Practice Center under Contract No. 290-2007-10055-1). AHRQ Publication No. 11-EHC052-EF. Rockville, MD: Agency for Healthcare Research and Quality; July 2011. PMID: 21977519. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

17. Gaynes BN, Lux L, Lloyd S, et al. Nonpharmacologic Interventions for Treatment-Resistant Depression in Adults. Comparative Effectiveness Review No. 33. (Prepared by RTI International-University of North Carolina (RTI-UNC) Evidence-based Practice Center under Contract No. 290-02-0016I.) AHRQ Publication No. 11-EHC056-EF. Rockville, MD: Agency for Healthcare Research and Quality; September 2011. PMID: 22091472. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

18. Gartlehner G, Hansen RA, Morgan LC, et al. Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression: An Update of the 2007 Comparative Effectiveness Review. (Prepared by the RTI International–University of North Carolina Evidence-based Practice Center, Contract No. 290-2007-10056-I.) AHRQ Publication No. 12-EHC012-EF. Rockville, MD: Agency for Healthcare Research and Quality; December 2011. PMID: 22299185. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

19. Bennett WL, Wilson LM, Bolen S, et al. Oral Diabetes Medications for Adults With Type 2 Diabetes: An Update. Comparative Effectiveness Review No. 27. (Prepared by Johns Hopkins University Evidence-based Practice Center under Contract No. 290-02-0018.) AHRQ Publication No. 11-EHC038-EF. Rockville, MD: Agency for Healthcare Research and Quality; March 2011. PMID: 21735563. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

20. Jonas DE, Garbutt JC, Brown JM, et al. Screening, Behavioral Counseling, and Referral in Primary Care to Reduce Alcohol Misuse. Comparative Effectiveness Review No. 64. Prepared by the RTI International–University of North Carolina Evidence-based Practice Center under Contract No. 290-2007-10056-I.) AHRQ Publication No. 12-EHC055-EF. Rockville, MD: Agency for Healthcare Research and Quality; July 2012. PMID: 22876371. www.effectivehealthcare.ahrq.gov/reports/final.cfm.

21. 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.

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

23. Begg CB. Suspended judgment. Significance tests of covariate imbalance in clinical trials. Control Clin Trials. 1990 Aug;11(4):223-5. PMID: 2171874.

24. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med. 2003 Sep 15;22(17):2693-710. PMID: 12939780.

25. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. Stat Methods Med Res. 2012 Aug;21(4):409-26. PMID: 21148194.

26. Kontopantelis E, Reeves D. Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: a comparison between DerSimonian-Laird and restricted maximum likelihood. Stat Methods Med Res. 2012 Dec;21(6):657-9. PMID: 23171971.

27. Higgins JP, Thompson SG, Deeks JJ, et al. Measuring inconsistency in meta-analyses. BMJ. 2003 Sep 6;327(7414):557-60. PMID: 12958120.

28. IntHout J, Ioannidis JP, Borm GF. The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol. 2014;14:25. PMID: 24548571.

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

30. Wright DB. Comparing groups in a before-after design: when t test and ANCOVA produce different results. Br J Educ Psychol. 2006 Sep;76(Pt 3):663-75. PMID: 16953968.

Abbreviations and Acronyms

AHRQ	Agency for Healthcare Research and Quality
ANCOVA	analysis of covariance
CER	comparative effectiveness review
CI	confidence interval
CPAP	continuous positive airway pressure
DL	Dersimonian-Laird
IPD	individual patient-level data
KH	Knapp-Hartung
MD	mean difference
ML	maximum likelihood
PE	permutation
PL	profile likelihood
RCT	randomized controlled clinical trial
REML	restricted maximum likelihood
SMD	standardized mean difference
TMS	transcranial magnetic stimulation

Appendix A. Comparison of Combined Mean Differences Using Change Score and Endpoints (Treatment Versus Control)

Appendix A. Comparison of Combined Mean Differences Using Change Score and Endpoints (Treatment Versus Control)

Comparative Effectiveness Review	Outcome: Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min–Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)	Percentage Difference in Combined Estimates
Pain Management Interventions for Hip Fracture[15]	Acute pain: skin traction vs. no traction	13 studies #1: 568 (30–166) #2: 662 (34–151)	PL: -0.300 (-0.812, 0.260) DL: -0.281 (-0.843, 0.280) REML: -0.293 (-0.802, 0.216) ML: -0.300 (-0.782, 0.181 PE: -0.281 (-2.097, 0.062) KH: -0.293 (-0.907, 0.321) $I^2 = 77.2\%$, $P < 0.001$	PL: 0.432 (-0.137, 0.997) DL: 0.431 (-0.125, 0.987) REML: 0.431 (-0.106, 0.968) ML: 0.432 (-0.073, 0.938) PE: 0.431 (-0.411, 0.743) KH: 0.431 (-0.217, 1.08) $I^2 = 76.5\%$, $P < 0.001$	PL: 0.115 (-0.332, 0.617) DL: 0.122 (-0.316, 0.561) REML: 0.123 (-0.318, 0.565) ML: 0.115 (-0.290, 0.519) PE: 0.122 (-1.211, 0.344) KH: 0.123 (-0.427, 0.673) $I^2 = 63.2\%$, $P = 0.008$	PL: 73% DL: 72% REML: 71% ML: 73% PE: 72% KH:71%

Appendix A. Comparison of Combined Mean Differences Using Change Score and Endpoints (Treatment Versus Control) (continued)

Comparative Effectiveness Review	Outcome: Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min–Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)	Percentage Difference in Combined Estimates
Diagnosis and Treatment of Obstructive Sleep Apnea in Adults[16]	Apnea-Hypopnea Index: CPAP vs. control	6 studies #1: 177 (12–66) #2: 159 (12–59)	PL: 2.208 (-3.442, 9.009) DL: 2.532 (-2.798, 7.862) REML: 2.592 (-3.029, 8.212) ML: 2.208 (-2.134, 6.550) PE: 2.532 (-7.179, 9.793) KH: 2.592 (-5.711, 10.895) I^2 = 64.7%, P = 0.015	PL: -27.022 (-41.194, -13.631) DL: -27.051 (-38.909, -15.193) REML: -27.161 (-39.922, -14.401) ML: -27.022 (-38.680, -15.365) PE: -27.051 (-157.441, 0.000) KH: -27.161 (-43.898, -10.425) I^2 = 93.3%, P < 0.001	PL: -24.176 (-36.050, -13.153) DL: -24.134 (-33.599, -14.669) REML: -24.325 (-34.925, -13.724) ML: -24.176 (-33.865, -14.488) PE: -24.134 (-136.260, -0.000) KH: -24.325 (-38.227, -10.422) I^2 = 91.9%, P < 0.001	PL: -11% DL: -11% REML: -10% ML: -11% PE: -11% KH: -10%
	Epworth Sleepiness Scale: CPAP vs. control	7 studies #1: 448 (19–178) #2: 398 (21–181)	PL: -0.015 (-0.523, 0.494) DL: -0.015 (-0.435, 0.406) REML: -0.015, (-0.435, 0.406) ML: -0.015 (-0.435, 0.406) PE: -0.015 (-2.242, 0.336) KH: -0.015 (-0.540, 0.510) I^2 = 0%, P = 0.805	PL: -2.684 (-4.311, -1.169) DL: -2.714 (-4.274, -1.155) REML: -2.702 (-4.179, -1.225) ML: -2.684 (-4.064, -1.304) PE: -2.714 (-18.518, -0.186) KH: -2.702 (-4.546, -0.858) I^2 = 84.3%, P < 0.001	PL: -2.325 (-3.645, -1.220) DL: -2.371 -3.415, -1.327) REML: -2.387 (-3.458, -1.317) ML: -2.325 (-3.305, -1.346) PE: -2.371 (-18.241, -0.161) KH: -2.387 (-3.724, -1.051) I^2 = 54.2%, P = 0.042	PL: -13% DL: -12% REML: -12% ML: -13% PE: -13% KH: -12%
	Apnea-Hypopnea Index: CPAP vs. sham CPAP	8 studies #1: 163 (15–27) #2: 149 (10–29)	**PL: 7.161 (1.241, 13.088)**[a] **DL: 7.161 (1.260, 13.063)**[a] *REML: 7.161 (1.260, 13.063)* *ML: 7.161 (1.260, 13.063)* PE: 7.161 (0.780, 9.416) KH: 7.161 (0.041, 14.281) I^2 = 0%, P = 0.920	PL: -45.634 (-58.285, -34.180) DL: -45.891 (-57.534, -34.249) REML: -45.826 (-57.209, -34.444) ML: -45.634 (-56.334, -34.933) PE: -45.891 (-321.166, -7.516) KH: -45.826 (-59.559, -32.094) I^2 = 70.9%, P = 0.001	PL: -40.495 (-52.287, -29.403) DL: -40.690 (-52.073, -29.307) REML: -40.608 (-51.424, -29.791) ML: -40.495 (-50.659, -30.330) PE: -40.690 (-280.617, -6.796) KH: -40.608 (-53.657, -27.558) I^2 = 82.4%, P < 0.001	PL: -11% DL: -11% REML: -11% ML: -11% PE: -11% KH: -11%
	Epworth Sleepiness Scale: CPAP vs. sham CPAP	11 studies #1: 293 (16–52) #2: 291 16–49)	PL: 0.309 (-0.263, 0.876) DL: 0.309 (-0.258, 0.876) REML: 0.309 (-0.258, 0.876) ML: 0.309 (-0.258, 0.876) PE: 0.309 (-0.192, 0.644) KH: 0.309 (-0.336, 0.954) I^2 = 0%, P = 0.920	PL: -2.684 (-4.345, -1.033) DL: -2.688 (-4.399, -0.977) REML: -2.686 (-4.277, -1.095) ML: -2.684 (-4.206, -1.162) PE: -2.688 (-10.240, -0.329) KH: -2.686 (-4.495, -0.877) I^2 = 83.4%, P < 0.001	PL: -2.555 (-4.197, -0.918) DL: -2.556 (-4.209, -0.904) REML: -2.555 (-4.130, -0.980) ML: -2.555 (-4.061, -1.048) PE: -2.556 (-9.499, -0.309) KH: -2.555 (-4.346, -0.765) I^2 = 82.2%, P < 0.001	DL: -5% PL: -5% REML: -5% ML: -5% PE: -5% KH: -5%

Appendix A. Comparison of Combined Mean Differences Using Change Score and Endpoints (Treatment Versus Control) (continued)

Comparative Effectiveness Review	Outcome: Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min–Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)	Percentage Difference in Combined Estimates
Nonpharmacologic Interventions for Treatment-Resistant Depression in Adults[17]	Depressive severity: rTMS vs. sham (condition = Tier 1, MDD)	8 studies #1: 116 (7–32) #2: 102 (5–31)	PL: 0.638 (-0.530, 2.671) DL: 0.638 (-0.460, 1.736) REML: 0.638 (-0.460, 1.736) ML: 0.638 (-0.460, 1.736) PE: 0.638 (0.110, 3.733) KH: 0.638 (-0.687, 1.963) I^2 = 0%, P = 0.520	**PL: -5.442 (-7.792, -2.672)**[a] **DL: -5.394 (-7.764, -3.025)**[a] *REML: -5.372 (-7.787, -2.958)*[b] ML: -5.442 (-7.726, -3.158) PE: -5.394 (-33.449, -0.607) *KH: -5.372 (-8.285, -2.459)*[b] I^2 = 43.9%, P = 0.086	**PL: -3.068 (-6.171, 0.371)**[a] **DL: -2.898 (-6.508, 0.711)**[a] *REML: -3.023 (-6.114, 0.068)*[b] ML: -3.068 (-5.990, -0.146) PE: -2.898 (-23.845, 0.351) *KH: -3.023 (-6.752, 0.706)*[b] I^2 = 79.2%, P < 0.001	PL: -44% DL: -45% REML: -44% ML: -44% PE: -46% KH: -44%
	Depressive severity: rTMS vs. sham (condition = Tier 1)	11 studies #1: 197 (7–36) #2: 149 (5–31)	PL: 0.680 (-0.374, 2.276) DL: 0.680 (-0.276, 1.636) REML: 0.680 (-0.276, 1.636) ML: 0.680 (-0.276, 1.636) PE: 0.680 (0.039 1.849) KH: 0.680 (-0.407, 1.767) I^2 = 0%, P = 0.452	PL: -5.690 (-7.813, -3.478) DL: -5.672 (-7.784, -3.561) REML: -5.676 (-7.760, -3.593) ML: -5.690 (-7.684, -3.696 PE: -5.672 (-22.270, -1.136) KH: -5.676 (-8.045, -3.308) I^2 = 54.4%, P = 0.015	PL: -3.906 (-6.284, -1.401) DL: -3.841 (-6.482, -1.199) REML: -3.891 (-6.236, -1.547) ML: -3.906 (-6.156, -1.657) PE: -3.841 (-16.307, -0.396) KH: -3.891 (-6.557, -1.226) I^2 = 73.6%, P < 0.001	PL: -31% DL: -32% REML: -31% ML: -3% PE: -32% KH: -31%
	Depressive severity: rTMS vs. sham (condition = Tier 1 & Tier 2, MDD)	12 studies #1: 374 (7–155) #2: 364 (5–146)	PL: -0.013 (-0.595, 0.764) DL: -0.013 (-0.569, 0.544) REML: -0.013 (-0.569, 0.544) ML: -0.013 (-0.569, 0.544) PE: -0.013 (-2.420, 0.169) KH: -0.013 (-0.637, 0.612) I^2 = 0%, P = 0.688	PL: -4.719 (-7.097, -2.319) DL: -4.714 (-7.126, -2.301) REML: -4.717 (-7.024, -2.410) ML: -4.719 (-6.937, -2.501) PE: -4.714 (-17.932, -0.890) KH: -4.717 (-7.307, -2.126) I^2 = 80.4%, P < 0.001	PL: -3.441 (-5.943, -0.802) DL: -3.436 (-5.83, -1.041) REML: -3.424 (-5.893, -0.955) ML: -3.441 (-5.805, -1.077) PE:-3.436 (-14.074, -0.313) KH: -3.424 (-6.197, -0.651) I^2 = 79.3%, P < 0.001	PL: -27% DL: -27% REML: -27% ML: -27% PE: -27% KH: -27%
Screening, Behavioral Counseling, and Referral in Primary Care to Reduce Alcohol Misuse[20]	Drinks/week: BCI vs. control (adults, 6 months)	11 studies #1: 1547 (39-353) #2: 1556 (32-376)	PL: 0.613 (-0.247, 1.606) DL: 0.613 (-0.229, 1.456) REML: 0.613 (-0.229, 1.456) ML: 0.613 (-0.229, 1.456) PE: 0.613 (-0.312, 1.561) KH: 0.613 (-0.345, 1.571) I^2 = 0%, P = 0.490	PL: -3.121 (-4.255, -2.310) DL: -3.228 (-4.214, -2.242) REML: -3.121 (-3.931, -2.310) ML: -3.121 (-3.931, -2.310) PE: -3.228 (-17.773, -0.355) KH: -3.121 (-4.114, -2.127) I^2 =13.9%, P = 0.311	PL: -2.504 (-4.037, -1.210) DL: -2.593 (-4.063, -1.123) REML: -2.561 (-3.839, -1.284) ML: -2.504 (-3.466, -1.542) PE: -2.593 (-13.018, -0.138) KH: -2.561 (-4.322, -0.800) I^2 =46.3%, P = 0.046	PL: -20% DL: -20% REML: -18% ML: -20% PE: -20% KH: -18%
	Drinks/week: BCI vs. control (adults, 12 months)	13 studies #1: 2088 (33-371) #2: 2012 (39-381)	PL: 0.455 (-0.403, 1.282) DL: 0.455 (-0.364, 1.274) REML: 0.455 (-0.364, 1.274) ML: 0.455 (-0.364, 1.274) PE: 0.455 (-0.903, 1.079) KH: 0.455 (-0.455, 1.365) I^2 = 0%, P = 0.649	PL: -3.700 (-4.641, -2.805) DL: -3.718 (-4.760, -2.677) REML: -3.700 (-4.529, -2.871) ML: -3.700 (-4.529, -2.871) PE: -3.718 (-15.177, -0.711) KH: -3.700 (-4.711, -2.689) I^2 =16.9%, P = 0.274	PL: -2.920 (-4.912, -0.800) DL: -2.936 (-4.601, -1.272) REML: -2.900 (-4.849, -0.951) ML: -2.920 (-4.718 -1.121) PE: -2.936 (-12.488, -0.215) KH: -2.900 (-5.182, -0.618) I^2 =57.2%, P = 0.005	PL: -21% DL: -21% REML: -22% ML: -21% PE: -21% KH: -22%

[a] Significant baseline difference or discrepancy in conclusion using DL or PL methods.

[b] Significant baseline difference or discrepancy in conclusion using REML, ML, PE or KH methods.

BCI = behavioral counseling intervention; CI = confidence interval; CPAP = continuous positive airway pressure; DL = DerSimonian-Laird random-effects method; MDD = major depressive disorder; ML = maximum likelihood method; PE = permutations method; PL = profile likelihood method; REML = restricted maximum likelihood method; rTMS = repetitive transcranial magnetic stimulation.

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments)

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments)

Comparative Effectiveness Review	Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
Diagnosis and Treatment of Obstructive Sleep Apnea in Adults[16]	Apnea-Hypopnea Index: AutoCPAP vs. CPAP	6 studies #1: 115 (10–50) #2: 116 (10–50)	PL: 1.004 (-3.766, 5.783) DL: 1.004 (-3.767, 5.775) REML: 1.004 (-3.767, 5.775) ML: 1.004 (-3.767, 5.775) PE: 1.004 (-8.0e+15, 2.058) KH: 1.004 (-5.253, 7.262) I^2 = 0%, P = 0.938	PL: -0.844 (-5.359, 3.646) DL: -0.844 (-5.330, 3.643) REML: -0.844 (-5.330, 3.643) ML: -0.844 (-5.330, 3.643) PE: -0.844 (-35.933, 1.3e+16) KH: -0.844 (-6.728, 5.040) 0.0%, P = 0.975	PL: -0.326 (-2.710, 2.058) DL: -0.448 (-2.519, 1.928) REML: -0.402 (-2.461, 1.658) ML: -0.448 (-2.351, 1.455) PE: -0.326 (-1.7e+16, 1.058) KH: -0.402 (-3.102, 2.299) I^2 = 80.8%, P < 0.001
	Epworth Sleepiness Scale: AutoCPAP vs. CPAP	5 studies #1: 112 (10–50) #2: 103 10–50)	PL: -0.880 (-2.840, 0.568) DL: -0.959 (-2.467, 0.548) REML: -0.977 (-2.518, 0.564) ML: -0.880 (-2.261, 0.501) PE: Not calculable[a] KH: 1.004 (-5.253, 7.262) I^2 = 26.3%, P = 0.246	PL: 0.959 (-0.313, 2.628) DL: 0.959 (-0.299, 2.216) REML: 0.995 (-0.316, 2.306) ML: 0.959 (-0.299, 2.216) PE: Not calculable[a] KH: 0.995 (-0.862, 2.852) I2 = 0%, P = 0.452	PL: -0.085 (-1.176, 1.028) DL: -0.085 (-1.155, 0.984) REML: -0.085 (-1.155, 0.984) ML: -0.085 (-1.155, 0.984) PE: Not calculable[a] KH: -0.085 (-1.600, 1.429) I^2 = 0%, P = 0.785
Second-Generation Antidepressants in the Pharmacologic Treatment of Adult Depression: An Update of the 2007 Comparative Effectiveness Review[18]	MADRS: citalopram vs. escitalopram	6 studies #1: 939 (125–214) #2: 932 (108–241)	PL: 0.178 (-0.341, 0.646) DL: 0.171 (-0.283, 0.625) REML: 0.171 (-0.285, 0.626) ML: 0.178 (-0.245, 0.601) PE: 0.171, -0.798, 0.443) KH: 0.171 (-0.426, 0.768) I^2 = 12.7%, P = 0.333	**PL: 2.108 (0.078, 4.008)**[b] DL: 2.077 (0.174, 3.979) REML: 2.086 (0.258, 3.914) ML: 2.108 (0.418, 3.798) PE: 2.077 (0.000, 6.109) KH: 2.042 (-0.316, 4.400) I^2 = 67.8%, P = 0.008	**PL: 2.258 (-0.073, 4.430)**[b] DL: 2.221 (0.060, 4.383) REML: 2.231 (0.139, 4.323) ML: 2.258 (0.330, 4.186) PE: 2.221 (-0.000, 6.552) KH: 2.231 (-0.513, 4.975) I^2 = 68.7%, P = 0.007

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Comparative Effectiveness Review	Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
Oral Diabetes Medications for Adults With Type 2 Diabetes: An Update of the 2007 Report[19]	HbA1c: metformin vs. thiazolidinediones	14 studies #1: 1132 (13–501) #2: 1127 (14–499)	PL: -0.017 (-0.097, 0.062) DL: -0.017 (-0.109, 0.074) REML: -0.017 (-0.094, 0.060) ML: -0.017 (-0.094, 0.060) PE: -0.017 (-0.226, 0.034) KH: -0.017 (-0.107, 0.072) I^2 = 9.7%, P = 0.347	**PL: -0.105 (-0.188, -0.025)**[b] DL: -0.106 (-0.214, 0.003) *REML: -0.105 (-0.183, -0.027)*[c] *ML: -0.105 (-0.183, -0.027)*[c] PE: -0.106 (-0.550, 0.005) *KH: -0.105 (-0.204, -0.005)*[c] I^2 = 24.9%, P = 0.186	**PL: -0.054 (-0.186, 0.103)**[b] DL: -0.053 (-0.182, 0.077) *REML: -0.050 (-0.186, 0.087)*[c] *ML: -0.054 (-0.181, 0.074)*[c] PE: -0.053 (-0.452, 0.058) *KH: -0.050 (-0.220, 0.121)*[c] I^2 = 58.0%, P = 0.003
	HbA1c: metformin vs. sulfonylureas	17 studies #1: 1153 (16–210) #2: 1229 (17–209)	PL: 0.042 (-0.087, 0.169) DL: 0.044 (-0.066, 0.153) REML: 0.042 (-0.082, 0.166) ML: 0.042 (-0.078, 0.162) PE: 0.044 (-0.208, 0.119) KH: 0.042 (-0.093, 0.176) I^2 = 14.7%, P = 0.282	PL: 0.060 (-0.113., 0.228) DL: 0.058 (-0.115, 0.232) *REML: 0.059 (-0.107, 0.225)*[c] *ML: 0.060 (-0.101, 0.220)*[c] PE: 0.058 (-0.237, 0.159) KH: 0.059 (-0.12, 0.238) I^2 = 62.4%, P < 0.001	PL: 0.116 (-0.005, 0.233) DL: 0.115 (-0.008, 0.237) *REML: 0.115 (0.001, 0.230)*[c] *ML: 0.116 (0.005, 0.226)*[c] PE: 0.115 (-0.027, 0.180) KH: 0.115 (-0.018, 0.249) I^2 = 51.8%, P = 0.007
	HbA1c: metformin vs. DPP-4 inhibitors	3 studies #1: 1037 (243–439) #2: 855 (175–455)	PL: -0.039 (-0.210, 0.047) DL: -0.066 (-0.178, 0.047) REML: -0.066 (-0.180, 0.048) ML: -0.039 (-0.110, 0.032) PE: Not calculable[a] KH: -0.066 (-0.316, 0.183) I^2 = 43.9%, P = 0.168	PL: -0.184 (-0.387, -0.049) DL: -0.203 (-0.354, -0.052) REML: -0.201 (-0.348, -0.055) ML: -0.184 (-0.299, -0.069) PE: Not calculable[a] KH: -0.201 (-0.523, 0.12) I^2 = 62.7%, P = 0.068	PL: -0.301 (-0.606, -0.028) DL: -0.312 (-0.593, -0.030) REML: -0.308 (-0.558, -0.058) ML: -0.301 (-0.509, -0.093) PE: Not calculable[a] KH: -0.308 (-0.857, 0.241) I^2 = 87.7%, P < 0.001
	HbA1c: metformin vs. metformin and thiazolidinediones	11 studies #1: 1428 (34–277) #2: 1688 (60–296)	**PL: -0.160 (-0.283, -0.034)**[b] **DL: -0.158 (-0.290, -0.027)**[b] *REML: -0.160 (-0.280, -0.039)*[c] *ML: -0.160 (-0.276, -0.044)*[c] PE: -0.158 (-0.624, -0.013) KH: -0.160 (-0.296, -0.023) I^2 = 69.6%, P = 0.001	PL: 0.633 (0.432, 0.854) DL: 0.635 (0.437, 0.832) REML: 0.636 (0.435, 0.837) ML: 0.633 (0.442, 0.824) PE: 0.635 (0.399, 0.866) KH: 0.636 (0.406, 0.866) I^2 = 85.5%, P < 0.001	PL: 0.503 (0.280, 0.733) DL: 0.506 (0.268, 0.744) REML: 0.504 (0.287, 0.721) ML: 0.503 (0.295, 0.710) PE: 0.506 (0.313, 0.659) KH: 0.504 (0.257, 0.751) I^2 = 93.4%, P < 0.001
	HbA1c: metformin vs. metformin and sulfonylureas	14 studies #1: 1188 (16–210) #2: 1949 (23–323)	PL: 0.043 (-0.085, 0.147) DL: 0.033 (-0.090, 0.156) REML: 0.039 (-0.068, 0.146) ML: 0.043 (-0.054, 0.140) PE: 0.033 (-0.155, 0.110) KH: 0.039 (-0.101, 0.179) I^2 = 42.3%, P = 0.048	PL: 0.848 (0.627, 1.098) DL: 0.851 (0.629, 1.073) REML: 0.852 (0.627, 1.077) ML: 0.848 (0.632, 1.063) PE: 0.851 (0.571, 1.115) KH: 0.852 (0.598, 1.106) I^2 = 80.7%, P < 0.001	PL: 0.832 (0.564, 1.117) DL: 0.837 (0.558, 1.116) REML: 0.834 (0.568, 1.100) ML: 0.832 (0.576, 1.087) PE: 0.837 (0.511, 1.136) KH: 0.834 (0.535, 1.134) I^2 = 92.6%, P < 0.001

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
HbA1c: metformin vs. metformin and DPP-4 Inhibitors	6 studies #1: 1177 (88–355) #2: 2075 (91–552)	PL: -0.001 (-0.083, 0.067) DL: -0.001 (-0.069, 0.067) REML: -0.001 (-0.069, 0.067) ML: -0.001 (-0.069, 0.067) PE: -0.001 (-5.1e+14, 0.033) KH: -0.001 (-0.090, 0.088) I^2 = 0%, P = 0.478	PL: 0.596 (0.469, 0.764) DL: 0.603 (0.475, 0.731) REML: 0.605 (0.473, 0.736) ML: 0.596 (0.488, 0.704) PE: 0.603 (-1.000, 0.912) KH: 0.601 (0.420, 0.782) I^2 = 51.8%, P = 0.065	PL: 0.592 (0.479, 0.712) DL: 0.593 (0.485, 0.701) REML: 0.593 (0.485, 0.700) ML: 0.592 (0.494, 0.689) PE: 0.593 (0.426, 0.775) KH: 0.593 (0.452, 0.734) I^2 = 41.2%, P = 0.131
HbA1c: thiazolidinediones vs. sulfonylureas	13 studies #1: 1202 (15–384) #2: 1013 (15–251)	PL: -0.017 (-0.150, 0.084) DL: -0.025 (-0.141, 0.092) REML: -0.020 (-0.124, 0.083) ML: -0.017 (-0.108, 0.074) PE: -0.025 (-0.295, 0.050) KH: -0.020 (-0.147, 0.106) I^2 = 26.5%, P = 0.176	PL: 0.029 (-0.143, 0.207) DL: 0.029 (-0.128, 0.185) REML: 0.031 (-0.136, 0.198) ML: 0.029 (-0.128, 0.186) PE: 0.029 (-0.284, 0.131) KH: 0.031 (-0.155, 0.216) I^2 = 35.7%, P = 0.097	PL: -0.077 (-0.262, 0.103) DL: -0.078 (-0.257, 0.101) REML: -0.077 (-0.253, 0.098) ML: -0.077 (-0.244, 0.091) PE: -0.078 (-0.504, 0.064) KH: -0.077 (-0.272, 0.117) I^2 = 71.9%, P < 0.001
HbA1c: sulfonylureas vs. meglitinides	7 studies #1: 563 (15–171) #2: 924 (29–338)	PL: 0.034 (-0.136, 0.193) DL: 0.033 (-0.107, 0.173) REML: 0.032 (-0.120, 0.184) ML: 0.034 (-0.104, 0.171) PE: 0.033 (-0.846, 0.099) KH: 0.032 (-0.158, 0.222) I^2 = 8.0%, P = 0.367	PL: 0.072 (-0.146, 0.322) DL: 0.074 (-0.123, 0.271) REML: 0.078 (-0.137, 0.293) ML: 0.072 (-0.119, 0.263) PE: 0.074 (-2.076, 0.254) KH: 0.078 (-0.19, 0.347) I^2 = 44.7%, P = 0.093	PL: 0.109 (-0.111, 0.359) DL: 0.113 (-0.096, 0.321) REML: 0.114 (-0.101, 0.330) ML: 0.109 (-0.082, 0.300) PE: 0.113 (-0.204, 0.229) KH: 0.114 (-0.159, 0.388) I^2 = 54.2%, P = 0.042
Hba1c: metformin and thiazolidinediones vs. metformin and sulfonylureas	6 studies #1: 1055 (48–285) #2: 1039 (47–288)	PL: 0.038 (-0.059, 0.118) DL: 0.037 (-0.040, 0.114) REML: 0.035 (-0.045, 0.116) ML: 0.038 (-0.036, 0.113) PE: 0.037 (-0.218, 0.077) KH: 0.035 (-0.070, 0.141) I^2 = 3.3%, P = 0.395	PL: 0.090 (-0.091, 0.268) DL: -0.090 (-0.061, 0.241) REML: 0.090 (-0.075, 0.254) ML: 0.090 (-0.056, 0.236) PE: 0.090 (-0.168, 0.208) KH: 0.09 (-0.138, 0.317) I^2 = 68.5%, P = 0.007	PL: 0.136 (-0.049, 0.308) DL: 0.135 (-0.021, 0.291) REML: 0.134 (-0.031, 0.298) ML: 0.136 (-0.012, 0.284) PE: 0.135 (-0.236, 0.260) KH: 0.134 (-0.088, 0.356) I^2 = 71.5%, P = 0.004
Hba1c: metformin and Sulfonylureas vs. Thiazolidinediones and Sulfonylureas	6 studies #1: 847 (37–320) #2: 871 (34–319)	PL: -0.123 (-0.238, 0.008) **DL: -0.126 (-0.217, -0.034)**[b] *REML: -0.119 (-0.231, -0.006)*[c] *ML: -0.123 (-0.221, -0.025)*[c] PE: -0.126 (-3.493, 1.1e+15) KH: -0.119 (-0.266, 0.029) I^2 = 8.3%, P = 0.363	**PL: -0.050 (-0.151, 0.069)**[b] **DL: -0.050 (-0.148, 0.047)**[b] *REML: -0.050 (-0.148, 0.047)*[c] *ML: -0.050 (-0.148, 0.047)*[c] PE: -0.050 (-1.296, 0.016) *KH: -0.05 (-0.178, 0.078)*[c] I^2 = 0%, P = 0.628	**PL: -0.165 (-0.262, -0.061)**[b] **DL: -0.165 (-0.262, -0.068)**[b] *REML: -0.165 (-0.262, -0.068)*[c] *ML: -0.165 (-0.262, -0.068)*[c] PE: -0.165 (-3.106, 0.008) *KH: -0.165 (-0.292, -0.038)*[c] I^2 = 30.7%, P = 0.205

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
Weight: metformin vs. thiazolidinediones	8 studies #1: 802 (13–501) #2: 816 (14–499)	PL: 1.328 (-0.697, 2.836) DL: 1.328 (-0.125, 2.781) REML: 1.328 (-0125, 2.781) ML: 1.328 (-0.125, 2.781) PE: 1.328 (-6.058, 2.089) KH: 1.328 (-0.320, 2.976)[e] I^2 = 0%, P = 0.550	PL: -2.877 (-4.229, -1.488) DL: -2.873 (-4.118, -1.629) REML: -2.870 (-4.161, -1.579) ML: -2.877 (-4.079, -1.675) PE: -2.873 (-19.057, -0.271) KH: -2.87 (-4.433, -1.307) I^2 = 83.0%, P < 0.001	PL: -3.184 (-4.446, -1.025) DL: -3.104 (-4.428, -1.780) REML: -2.968 (-4.486, -1.450) ML: -3.184 (-4.405, -1.963) PE: -3.104 (-39.375, -0.038) KH: -2.968 (-4.800, -1.137) I^2 = 21.6%, P = 0.258
Weight: metformin vs. sulfonylureas	12 studies #1: 987 (16–210) #2: 1066 (17–209)	PL: 0.623 (-0.895, 1.889) DL: 0.515 (-1.005, 2.034) REML: 0.623 (-0.638, 1.883) ML: 0.623 (-0.638, 1.883) PE: 0.515 (-1.427, 1.439) KH: 0.623 (-1.006, 2.252) I^2 = 24.5%, P = 0.203	PL: -2.635 (-3.157, -2.127) DL: -2.637 (-3.155, -2.120) REML: -2.636 (-3.132, -2.140) ML: -2.635 (-3.112, -2.158) PE: -2.637 (-8.701, -0.653) KH: -2.636 (-3.192, -2.079) I^2 = 52.7%, P = 0.016	PL: -1.757 (-3.282, -0.423) DL: -1.847 (-3.434, -0.260) REML: -1.759 (-3.049, -0.468) ML: -1.757 (-3.033, -0.482) PE: -1.847 (-7.651, -0.161) KH: -1.759 (-3.454, -0.063) I^2 = 27.8%, P = 0.172
Weight: metformin vs. sulfonylureas (studies < 24 weeks in duration)	8 studies #1: 718 (21–164) #2: 797 (18–161)	**PL: 1.669 (0.115, 3.208)**[b] **DL: 1.669 (0.163, 3.175)**[b] *REML: 1.669 (0.163, 3.175)*[c] *ML: 1.669 (0.163, 3.175)*[c] PE: 1.669 (0.890, 2.533) KH: 1.669 (-0.148, 3.486) I^2 = 0%, P = 0.939	**PL: -2.240 (-2.747, -1.792)**[b] **DL: -2.234 (-2.640, -1.828)**[b] *REML: -2.250 (-2.700, -1.801)*[c] *ML: -2.240 (-2.659, -1.821)*[c] PE: -2.234 (-14.999, -0.310) *KH: -2.250 (-2.793, -1.708)*[c] I^2 = 2.8%, P = 0.408	**PL: -0.634 (-2.144, 0.852)**[b] **DL: -0.634 (-2.118, 0.850)**[b] *REML: -0.634 (-2.118, 0.850)*[c] *ML: -0.634 (-2.118, 0.850)*[c] PE: -0.634 (-6.145, 0.211) *KH: -0.634 (-2.424, 1.157)*[c] I^2 = 0%, P = 0.883
Weight: metformin vs. sulfonylureas (studies ≥ 24 weeks in duration)	4 studies #1: 269 (16–210) #2: 269 (17–209)	PL: -2.979 (-8.174, 0.976) DL: -3.259 (-7.386, 0.867) REML-3.256 (-7.375, 0.863) ML: -2.979 (-6.542, 0.583) PE: Not calculable[a] KH: -3.256 (-9.945, 3.432) I^2 = 50.5%, P = 0.109	PL: -3.531 (-4.232, -2.940) DL: -3.531 (-4.041, -3.022) REML: -3.531 (-4.041, -3.022) ML: -3.531 (-4.041, -3.022) PE: Not calculable[a] *KH: -3.531 (-4.359, -2.704)*[c] I^2 = 0%, P = 0.904	PL: -4.940 (-11.360, -2.359) DL: -5.837 (-9.383, -2.290) REML: -6.213 (-10.240, -2.186) ML: -4.940 (-7.521, -2.359) PE: Not calculable[a] *KH: -6.213 (-12.751, 0.325)*[c] I^2 = 21.4%, P = 0.282
Weight: metformin vs. DPP-4 Inhibitors	3 studies #1: 937 (243–446) #2: 783 (100–458)	PL: -0.300 (-2.920, 2.320) DL: -0.300 (-2.920, 2.320) REML: Not calculable ML: -0.300 (-2.920, 2.320) PE: Not calculable[a] KH: -0.300 (-0.496, -0.104)[e] I^2 = Not applicable[d] P = Not applicable[d]	PL: -1.280 (-1.793, -0.677) DL: -1.251 (-1.830, -0.672) REML: -1.254 (-1.796, -0.712) ML: -1.280 (-1.594, -0.966) PE: Not calculable[a] *KH: -1.254 (-2.628, 0.120)*[c] I^2 = 47.8%, P = 0.147	PL: -1.360 (-1.959, -0.944) DL: -1.360 (-1.689, -1.032) REML: -1.360 (-1.689, -1.032) ML: -1.360 (-1.689, -1.032) PE: Not calculable[a] *KH: -1.360 (-2.082, -0.639)*[c] I^2 = 0%, P = 0.522

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome **Comparison (Group #1 vs. Group #2)**	**Number of Studies (N)** **Group #1 and Group #2 Total Sample Size (Min-Max)**	**Baseline Difference (95% CI)**	**Difference in Change Score (95% CI)**	**Difference in Endpoint Score (95% CI)**
Weight: metformin vs. combination metformin and thiazolidinediones	5 studies #1: 665 (86–280) #2: 786 (83–288)	PL: -1.194 (-3.881, 1.884) DL: -1.194 (-3.420, 1.031) REML: -1.194 (-3.420, 1.031) ML: -1.194 (-3.420, 1.031) PE: Not calculable[a] KH: -1.194 (-15.621, 13.232) I^2 = 0%, P = 0.706	PL: -2.359 (-2.768, -1.982) DL: -2.359 (-2.737, -1.981) REML: -2.359 (-2.737, -1.981) ML: -2.359 (-2.737, -1.981) PE: Not calculable[a] KH: -2.359 (-2.895, -1.824) I^2 = 0%, P = 0.893	PL: -2.686 (-3.500, -1.993) DL: -2.686 (-3.361, -2.012) REML: -2.686 (-3.361, -2.012) ML: -2.686 (-3.361, -2.012) PE: Not calculable[a] KH: -2.686 (-3.642, -1.731) I^2 = 0%, P = 0.931
Weight: metformin vs. combination metformin and sulfonylureas	10 studies #1: 992 (16–210) #2: 1535 (23–323)	PL: 0.428 (-0.935, 1.903) DL: 0.428 (-0.933, 1.789) REML: 0.428 (-0.933, 1.789) ML: 0.428 (-0.933, 1.789) PE: 0.428 (-0.965, 0.709) KH: 0.428 (-1.174, 2.029) I^2 = 0%, P = 0.960	PL: -2.197 (-2.896, -1.440) DL: -2.168 (-2.960, -1.375) REML: -2.189 (-2.884, -1.493) ML: -2.197 (-2.861, -1.533) PE: -2.168 (-9.086, -0.468) KH: -2.189 (-2.991, -1.386) I^2 = 84.4%, P < 0.001	PL: -2.394 (-3.057, -1.481) DL: -2.394 (-3.020, -1.768) REML: -2.394, -3.020, -1.768) ML: -2.394 (-3.020, -1.768) PE: -2.394 (-27.001, -0.197) KH: -2.394 (-3.117, -1.671) I^2 = 0%, P = 0.868
Weight: metformin vs. combination metformin and sulfonylureas	3 studies #1: 579 (88–248) #2: 910 (91–523)	PL: 0.683 (-1.821, 3.545) DL: 0.683 (-1.384, 2.751) REML: 0.683 (-1.384, 2.751) ML: 0.683 (-1.384, 2.751) PE: Not calculable[a] KH: 0.683 (-12.719, 14.086) I^2 = 0%, P = 0.714	PL: -0.200 (-0.601, 0.247) DL: -0.200 (-0.584, 0.184) REML: -0.200 (-0.584, 0.184) ML: -0.200 (-0.584, 0.184) PE: Not calculable[a] KH: -0.200 (-1.043, 0.642) I^2 = 0%, P = 0.639	PL: 0.004 (-0.810, 1.124) DL: 0.004 (-0.644, 0.652) REML: 0.004 (-0.644, 0.652) ML: 0.004 (-0.644, 0.652) PE: Not calculable[a] KH: 0.004 (-0.595, 0.603)[e] I^2 = 0%, P = 0.838
Weight: thiazolidinediones vs. sulfonylureas	5 studies #1: 883 (23–384) #2: 694 (18–251)	PL: 0.298 (-1.479, 2.327) DL: 0.298 (-1.479, 2.076) REML: 0.298 (-1.479, 2.076) ML: 0.298 (-1.479, 2.076) PE: Not calculable[a] KH: 0.298 (-2.588, 3.184) I^2 = 0%, P = 0.676	PL: 0.407 (-1.509, 2.210) DL: 0.402 (-1.136, 1.940) REML: 0.385 (-1.310, 2.079) ML: 0.407 (-1.094, 1.909) PE: Not calculable[a] *KH: 0.385 (-2.015, 2.785)*[c] I^2 = 80.7%, P < 0.001	PL: 1.114 (-0.378, 2.592) DL: 1.114 (-0.363, 2.590) REML:1.114 (-0.363, 2.590) ML: 1.114 (-0.363, 2.590) PE: Not calculable[a] *KH: 1.114 (0.001, 2.226)*[c,e] I^2 = 0%, P = 0.889
Weight: sulfonylureas vs. meglitinides	6 studies #1: 518 (15–182) #2: 808 (29–362)	PL: 0.428 (-1.172, 2.318) DL: 0.507 (-1.509, 2.522) REML: 0.428 (-0.902, 1.758) ML: 0.428 (-0.902, 1.758) PE: Not calculable[a] KH: 0.428 (-2.109, 2.966) I^2 = 44.9%, P = 0.123	PL: -0.176 (-0.894, 0.649) DL: -0.157 (-0.853, 0.538) REML: -0.156 (-0.859, 0.547) ML: -0.176 (-0.791, 0.439) PE: -0.157 (-15.284, 0.270) KH: -0.156 (-1.078, 0.767) I^2 = 47.3%, P = 0.091	PL: 0.192 (-0.990, 1.585) DL: 0.254 (-1.150, 1.659) REML: 0.192 (-0.963, 1.346) ML: 0.192 (-0.963,1.346) PE: 0.254 (-4.8e+15, 1.371) KH: 0.192 (-1.550, 1.933) I^2 = 24.4%, P = 0.251

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
Weight: combination metformin and thiazolidinediones vs. combination metformin and sulfonylureas	5 studies #1: 845 (48–294) #2: 864 (47–301)	PL: 0.396 (-1.871, 2.863) DL: 0.457 (-1.748, 2.661) REML: 0.446 (-1.672, 2.565) ML: 0.396 (-1.454, 2.247) PE: Not calculable[a] KH: 0.446 (-2.994, 3.887) I^2 = 65.3%, P = 0.034	PL: 0.261 (-0.785, 1.339) DL: 0.273 (-0.860, 1.407) REML: 0.267 (-0.706, 1.239) ML: 0.261 (-0.619, 1.141) PE: Not calculable[a] KH: 0.267 (-1.111, 1.644) I^2 = 90.3%, P < 0.001	PL: 0.546 (-2.003, 3.253) DL: 0.568 (-1.757, 2.894) REML: 0.576 (-1.821, 2.973) ML: 0.546 (-1.601, 2.692) PE: Not calculable[a] KH: 0.576 (-2.820, 3.972) I^2 = 88.3%, P < 0.001
Weight: metformin and sulfonylureas vs. combination thiazolidinediones and sulfonylureas.	4 studies #1: 653 (37–320) #2: 646 (34–319)	PL: 1.746 (-2.856, 6.117) DL: 1.672 (-2.815, 6.160) REML: 1.702 (-2.320, 5.724) ML: 1.746 (-1.792, 5.284) PE: Not calculable [a] KH: 1.702 (-4.828, 8.233) I^2 = 87.5%, P < 0.001	**PL: -2.689 (-3.840, -1.550)**[b] **DL: -2.689 (-3.747, -1.631)**[b] *REML: -2.688 (-3.712, -1.665)*[c] *ML: -2.689 (-3.574, -1.804)*[c] PE: Not calculable[a] *KH: -2.688 (-4.350, -1.027)*[c] I^2 = 78.1%, P = 0.003	**PL: -0.935 (-6.105, 3.972)**[b] **DL: -1.038 (-6.510, 4.434)**[b] *REML: -0.981 (-5.511, 3.549)*[c] *ML: -0.935 (-4.944, 3.075)*[c] PE: Not calculable[a] *KH: -0.981 (-8.336, 6.374)*[c] I^2 = 90.8%, P < 0.001
LDL: metformin vs. rosiglitazone	6 studies #1: 198 (9–117) #2: 213 (14–128)	PL: 2.100 (-4.188, 8.886) DL: 2.100 (-3.960, 8.159) REML: 2.100 (-3.960, 8.159) ML: 2.100 (-3.960, 8.159) PE: 2.100 (-13.062, 8.122) KH: 2.100 (-5.848, 10.047) I^2 = 0.0%, P = 0.682	**PL: -12.535 (-22.237, -5.876)**[b] DL: -13.263 (-20.553, -5.974) *REML: -13.249 (-20.495, -6.003)*[c] ML: -12.535 (-17.993, -7.077) PE: -13.263 (-133.809, -0.000) *KH: -13.249 (-23.870, -2.628)*[c] I^2 = 58.0%, P = 0.036	**PL: -14.009 (-29.491, 2.268)**[b] DL: -13.944 (-27.561, -0.327) *REML: -13.858 (-28.448, 0.733)*[c] ML: -14.009 (-26.990, -1.028) PE: -13.944 (-126.845, 2.438) *KH: -13.858 (-33.699, 5.984)*[c] I^2 = 65.7%, P = 0.012
LDL: metformin vs. pioglitazone	6 studies #1: 676 (16–501) #2: 679 (14–499)	PL: 4.325 (-5.729, 14.373) DL: 4.325 (-5.644, 14.294) REML: 4.325 (-5.644, 14.294) ML: 4.325 (-5.644, 14.294) PE: Not calculable[a] KH: 4.325 (-11.861, 20.512) I^2 = 0.0%, P = 0.935	PL: -12.394 (-17.705, -6.587) DL: -12.394 (-17.684, -7.105) REML: -12.394 (-17.684, -7.105) ML: -12.394 (-17.684, -7.105) PE: -12.394 (-354.099, -0.000) KH: -12.394 (-19.332, -5.457) I^2 = 0%, P = 0.848	PL: -11.513 (-17.022, -2.810) DL: -11.141 (-17.152, -5.130) REML: -11.304 (-17.056, -5.552) ML: -11.513 (-16.946, -6.080) PE: -11.141 (-367.467, 0.797) KH: -11.304 (-19.063, -3.545) I^2 = 10.6%, P = 0.348
LDL: metformin vs. sulfonylureas	8 studies #1: 679 (16–210) #2: 671 (17–209)	PL: 1.146 (-4.930, 6.028) DL: 1.096 (-3.724, 5.916) REML: 0.950 (-4.130, 6.029) ML: 1.146 (-3.594, 5.886) PE: 1.096 (-52.943, 4.163) KH: 0.950 (-5.392, 7.291) I^2 = 32.0%, P = 0.184	PL: Does not converge DL: -7.717 (-13.175, -2.260) REML: -7.130 (-11.644, -2.616) ML: Does not converge PE: -7.717 (-51.181, -0.977) KH: -7.130 (-12.576, -1.684) I^2 = 75.6%, P < 0.001	PL: -7.229 (-13.707, -2.382) DL: -7.743 (-13.328, -2.158) REML: -7.494 (-12.651, -2.338) ML: -7.229 (-11.991, -2.467) PE: -7.743 (-49.429, -0.748) KH: -7.494 (-13.715, -1.274) I^2 = 80.6%, P < 0.001

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
LDL: metformin vs. DPP-4 Inhibitors	3 studies #1: 910 (239–426) #2: 791 (94–441)	PL: -1.733 (-9.626, 5.760) DL: -1.828 (-8.387, 4.730) REML: -1.831 (-8.447, 4.784) ML: -1.733 (-7.0874, 3.618) PE: Not calculable[a] KH: -1.831 (-16.414, 12.751) I^2 = 75.5%, P = 0.017	PL: -6.281 (-10.313, -0.622) DL: -6.094 (-10.015, -2.172) REML:-5.993 (-10.141, -1.845) ML: -6.281 (-9.848, -2.714) PE: Not calculable[a] KH: -5.993 (-15.099, 3.114) I^2 = 13.9%, P = 0.313	PL: -6.390 (-11.167, -2.276) DL: -6.444 (-10.253, -2.635) REML: -6.390 (-9.815, -2.966) ML: -6.390 (-9.815, -2.966) PE: Not calculable[a] KH: -6.390 (-14.597, 1.817) I^2 = 16.1%, P = 0.304
LDL: metformin vs. combination metformin and rosiglitazone	7 studies #1: 1006 (34–277) #2: 1149 (71–268)	PL: 0.905 (-1.887, 3.775) DL: 0.905 (-1.887, 3.697) REML: 0.095 (-1.887, 3.697) ML: 0.905 (-1.887, 3.697) PE: 0.905 (-4.920, 1.893) KH: 0.905 (-2.581, 4.391) I^2 = 0.0%, P = 0.612	PL: -13.651 (-15.891, -11.342) DL: -13.651 (-15.759, -11.542) REML: -13.651 (-15.759, -11.542) ML: -13.651 (-15.759, -11.542) PE: -13.651 (-88.012, -1.948) KH: -13.651 (-16.283, -11.019) I^2 = 0%, P = 0.668	PL: -13.029 (-16.242, -9.463) DL: -13.029 (-16.234, -9.825) REML: -13.029 (-16.234, -9.825) ML: -13.029 (-16.234, -9.825) PE:-13.029 (-73.250, -2.159) KH: -13.029 (-17.030, -9.029) I^2 = 0%, P = 0.663
LDL: metformin vs. combination metformin and sulfonylureas	6 studies #1: 580 (16–210) #2: 808 (60–213)	PL: -2.719 (-8.339, 2.904) DL: -2.727 (-7.494, 2.041) REML: -2.732 (-7.790, 2.325) ML: -2.719 (-7.213, 1.775) PE: Not calculable[a] KH: -2.732 (-9.897, 4.432) I^2 = 9.1%, P = 0.354	PL: -2.552 (-8.390, 1.539) DL: -2.564 (-6.499, 1.371) REML: -2.930 (-7.434, 1.574) ML: -2.552 (-6.471, 1.367) PE: -2.564 (-31.611, 1.202) KH: -2.930 (-8.837, 2.977) I^2 = 61.8%, P = 0.022	PL: -4.212 (-11.003, 0.517) DL: -4.617 (-9.796, 0.562) REML: -4.641 (-9.865, 0.583) ML: -4.212 (-8.764, 0.339) PE: -4.617 (-97.587, 0.248) KH: -4.641 (-11.492, 2.210) I^2 = 66.2%, P = 0.011
LDL: metformin vs. combination metformin and DPP-4 inhibitors	4 studies #1: 773 (83–245) #2: 1279 86–479)	PL: -0.795 (-6.649, 5.569) DL: -0.639 (-6.663, 5.386) REML: -0.698 (-6.192, 4.797) ML: -0.795 (-5.673, 4.084) PE: Not calculable[a] KH: -0.698 (-9.619, 8.224) I^2 = 76.8%, P = 0.005	PL: 1.739 (-1.432, 5.037) DL: 1.739 (-1.413, 4.890) REML: 1.739 (-1.413, 4.890) ML: 1.739 (-1.413, 4.890) PE: Not calculable[a] KH: 1.739 (-3.378, 6.855) I^2 = 0%, P = 0.857	PL: 1.556 (-3.203, 6.393) DL: 1.553 (-2.625, 5.731) REML: 1.554 (-2.728, 5.836) ML: 1.556 (-2.116, 5.227) PE: Not calculable[a] KH: 1.554 (-5.399, 8.507) I^2 = 40.3%, P = 0.170
LDL: combination metformin and rosiglitazone vs. combination metformin and sulfonylureas	4 studies #1: 532 (48– 202) #2: 592 (47– 240)	PL: -2.170 (-5.335, 0.840) DL: -2.170 (-5.163, 0.823) REML: -2.170 (-5.163, 0.823) ML: -2.170 (-5.163, 0.823) PE: Not calculable[a] KH: -2.170 (-7.029, 2.689) I^2 = 0.0%, P = 0.724	PL: 14.825 (9.650, 19.554) DL: 14.768 (10.447, 19.088) REML: 14.748 (10.287, 19.209) ML: 14.825 (10.861, 18.789) PE: Not calculable[a] KH: 14.748 (7.505, 21.991) I^2 = 36%, P = 0.196	PL: 12.736 (7.395, 17.171) DL: 12.577(8.279, 16.876) REML: 12.565 (8.214, 16.915) ML: 12.736 (9.000, 16.472) PE: Not calculable[a] KH: 12.565 (5.500, 19.629) I^2 = 39.0%, P = 0.178

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome **Comparison (Group #1 vs. Group #2)**	**Number of Studies (N)** **Group #1 and Group #2 Total Sample Size (Min-Max)**	**Baseline Difference (95% CI)**	**Difference in Change Score (95% CI)**	**Difference in Endpoint Score (95% CI)**
HDL: metformin vs. rosiglitazone	6 studies #1: 198 (9–117) #2: 213 (14–128)	PL: Does not converge **DL: -2.408 (-4.400, -0.416)**[b] *REML: -1.995 (-4.393, -0.403)*[c] ML: Does not converge PE: -2.408 (-4.4e+16, 0.016) KH: -1.995 (-5.140, 1.150) I^2 = 0.0%, P = 0.673	PL: -0.595 (-1.658, 2.319) DL: -0.052 (-1.877, 1.774) REML: -0.066 (-1.870, 1.738) ML: -0.595 (-1.658, 0.468) PE: -0.052 (-50.905, 0.798) KH: -0.066 (-2.773, 2.641) I^2 = 42.6%, P = 0.121	PL: 0.445 (-2.983, 3.832) DL: 0.445 (-2.941, 3.831) REML: 0.445 (-2.941, 3.831) ML: 0.445 (-2.941, 3.831) PE: 0.445 (-5.546, 1.723) KH: 0.445 (-3.996, 4.886) I^2 = 0%, P = 0.922
HDL: metformin vs. pioglitazone	8 studies #1: 736 (16–501) #2: 740 (14–499)	PL: -1.228 (-2.737, 0.170) DL: -1.228 (-2.736, 0.281) REML: -1.157 (-2.761, 0.447) ML: -1.228 (-2.736, 0.281) PE: -1.228 (-47.266, 0.214) KH: -1.157 (-3.261, 0.946) I^2 = 0.0%, P = 0.745	PL: -2.780 (-3.461, -0.818) DL: -2.989 (-5.460, -0.519) REML: -2.837 (-4.852, -0.823) ML: -2.780 (-4.683, -0.876) PE: -2.989 (-19.849, -0.191) KH: -2.837 (-5.268, -0.407) I^2 = 78.9%, P < 0.001	PL: -3.256 (-6.058, -0.633) DL: -3.387 (-6.749, -0.026) REML: -3.285 (-5.870, -0.700) ML: -3.256 (-5.708, -0.805) PE: -3.387 (-26.620, -0.321) KH: -3.285 (-6.404, -0.166) I^2 = 87.3%, P < 0.001
HDL: metformin vs. sulfonylureas	11 studies #1: 812 (19–210) #2: 858 (18–209)	PL: -0.004 (-1.111, 0.827) DL: 0.019 (-0.787, 0.824) REML: -0.054 (-0.942, 0.835) ML: -0.004 (-0.834, 0.825) PE: 0.019 (-3.215, 0.490) KH: -0.054 (-1.064, 0.957) I^2 = 0.0%, P = 0. 458	PL: 0.243 (-0.538, 1.002) DL: 0.243 (-0.505, 0.991) REML:0.243 (-0.505, 0.991) ML: 0.243 (-0.505, 0.991) PE: 0.243 (-0.342, 0.562) KH: 0.243 (-0.607, 1.094) I^2 = 0%, P = 0.961	PL: 0.242 (-0.727, 1.073) DL: 0.232 (-0.608, 1.071) REML: 0.242 (-0.574, 1.057) ML: 0.242 (-0.571, 1.056) PE: 0.232 (-1.730, 0.675) KH: 0.242 (-0.702, 1.185) I^2 = 3.8%, P = 0.407
HDL: metformin vs. DPP-4 inhibitors	3 studies #1: 913 (241–427) #2: 791 (95–440)	PL: 0.711 (-0.271, 1.673) DL: 0.711 (-0.217, 1.639) REML: 0.711 (-0.217, 1.639) ML: 0.711 (-0.217, 1.639) PE: Not calculable[a] KH: 0.711 (-1.326, 2.748) I^2 = 0.0%, P = 0. 787	**PL: 1.432 (-0.004, 3.251)**[b] DL: 1.514 (0.132, 2.897) REML: 1.520 (0.117, 2.923) ML: 1.432 (0.288, 2.575) PE: Not calculable[a] KH: 1.520 (-1.561, 4.600) I^2 = 34.1%, P = 0.219	**PL: 2.112 (0.705, 3.899)**[b] DL: 2.199 (0.814, 3.584) REML: 2.195 (0.827, 3.562) ML: 2.112 (1.001, 3.222) PE: Not calculable[a] KH: 2.195 (-0.876, 5.265) I^2 = 33.9%, P = 0.220
HDL: metformin vs. combination metformin and rosiglitazone	7 studies #1: 1069 (34–302) #2: 1248 (71–309)	PL: -0.074 (-1.293, 1.201) DL: -0.072 (-1.139, 0.995) REML: -0.063 (-1.224, 1.097) ML: -0.074 (-1.122, 0.975) PE: -0.072 (-11.071, 0.668) KH: -0.063 (-1.512, 1.386) I^2 = 10.5%, P = 0. 349	PL: -3.229 (-4.168, -2.313) DL: -3.230 (-4.134, -2.325) REML: -3.229 (-4.065, -2.394) ML: -3.229 (-4.065, -2.394) PE: -3.230 (-20.158, -0.387) KH: -3.229 (-4.338, -2.120) I^2 = 11.5%, P = 0.342	PL: -3.083 (-4.683, -1.541) DL: -3.087 (-4.515, -1.658) REML: -3.088 (-4.544, -1.632) ML: -3.083 (-4.405, -1.762) PE: -3.087(-19.775, -0.138) KH: -3.088 (-4.906, -1.270) I^2 = 25.5%, P = 0.234

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
HDL: metformin vs. combination metformin and sulfonylureas	5 studies #1: 415 (19–164) #2: 666 74–204)	PL: -0.441 (-2.091, 1.043) DL: -0.441 (-1.867, 0.984) REML: -0.441 (-1.867, 0.984) ML: -0.441 (-1.867, 0.984) PE: Not calculable[a] KH: -0.441 (-2.461, 1.578) I^2 = 0.0%, P = 0. 619	PL: 1.908 (-4.330, 8.218) DL: 1.888 (-2.546, 6.321) REML: 1.922 (-3.791, 7.634) ML: 1.908 (-3.182, 6.998) PE: Not calculable[a] KH: 1.922 (-6.229, 10.072) I^2 = 96.0%, P < 0.001	PL: 1.370 (-5.498, 8.106) DL: 1.361 (-4.395, 7.117) REML: 1.346 (-4.854, 7.545) ML: 1.370 (-4.170, 6.910) PE: Not calculable[a] KH: 1.346 (-7.444, 10.135) I^2 = 95.6%, P < 0.001
HDL: metformin vs. combination metformin and DPP-4 inhibitors	4 studies #1: 775 (83–245) #2: 1329 (86–528)	PL: -0.596 (-1.529, 0.354) DL: -0.596 (-1.527, 0.334) REML: -0.596 (-1.527, 0.334) ML: -0.596 (-1.527, 0.334) PE: Not calculable[a] KH: -0.596 (-2.107, 0.914) I^2 = 0.0%, P = 0. 967	PL: 0.210 (-0.954, 1.316) DL: 0.210 (-0.861, 1.281) REML: 0.210 (-0.861, 1.281) ML: 0.210 (-0.861, 1.281) PE: Not calculable[a] KH: 0.210 (-1.528, 1.948) I^2 = 0%, P = 0.644	PL: -0.324 (-1.626, 0.885) DL: -0.324 (-1.420, 0.773) REML: -0.324 (-1.420, 0.773) ML: -0.324 (-1.420, 0.773) PE: Not calculable[a] KH: -0.324 (-2.104, 1.457) I^2 = 0%, P = 0.494
HDL: pioglitazone vs. sulfonylurea	6 studies #1: 304 (17–91) #2: 311 (18–109)	PL: 1.375 (-0.020, 2.695) **DL: 1.375 (0.074, 2.675)**[b] *REML: 1.375 (0.074, 2.675)*[c] *ML: 1.375 (0.074, 2.675)*[c] PE: 1.375 (-0.783, 2.297) KH: 1.375 (-0.331, 3.081) I^2 = 0.0%, P = 0. 706	PL: 5.295 (3.480, 6.964) DL: 5.278 (3.472, 7.084) REML: 5.306 (3.786, 6.825) ML: 5.295 (4.068, 6.522) PE: 5.278 (-0.000, 7.4e+14) KH: 5.306 (2.896, 7.716) I^2 = 43.3%, P = 0.117	PL: 6.369 (3.528, 8.853) DL: 6.325 (4.021, 8.630) REML: 6.298 (3.864, 8.733) ML: 6.369 (4.252, 8.486) PE: Does not converge KH: 6.298 (3.059, 9.538) I^2 = 60.5%, P = 0.027
HDL: sulfonylureas vs. meglitinides	6 studies #1: 529 (18–182) #2: 885 (16–362)	PL: 0.699 (-1.148, 2.312) DL: 0.699 (-0.876, 2.274) REML: 0.699 (-0.876, 2.274) ML: 0.699 (-0.876, 2.274) PE: Not calculable[a] KH: 0.699 (-2.759, 4.156) I^2 = 0.0%, P = 0. 750	PL: 0.018 (-0.179, 0.133) DL: -0.457 (-1.566, 0.653) REML:-0.584 (-1.955, 0.787) ML: 0.018 (-0.047, 0.083) PE: Not calculable[a] KH: -0.584 (-2.526, 1.358) I^2 = 53.5%, P = 0.072	PL: 0.029 (-0.120, 0.152) DL: -0.011 (-0.453, 0.431) REML: 0.029 (-0.036, 0.094) ML: 0.029 (-0.036, 0.094) PE: Not calculable[a] KH: 0.029 (-0.067, 0.125) I^2 = 8.1%, P = 0.360
HDL: combination metformin and rosiglitazone vs. combination metformin and sulfonylureas	4 studies #1: 577 (48–210) #2: 613 (47–240)	PL: -0.539 (-1.836, 0.696) DL: -0.552 (-1.681, 0.577) REML: -0.552 (-1.679, 0.576) ML: -0.539 (-1.513, 0.436) PE: Not calculable[a] KH: -0.552 (-2.388, 1.284) I^2 = 35.0%, P = 0. 202	PL: 3.048 (1.714, 4.862) DL: 3.155 (1.776, 4.535) REML: 3.160 (1.767, 4.552) ML: 3.048 (1.909, 4.188) PE: Not calculable[a] KH: 3.160 (0.856, 5.464) I^2 = 25.8%, P = 0.257	PL: 2.490 (0.968, 4.439) DL: 2.576 (1.064, 4.087) REML: 2.582 (1.049, 4.116) ML: 2.490 (1.207, 3.774) PE: Not calculable[a] KH: 2.582 (0.025, 5.139) I^2 = 61.5%, P = 0.051

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome Comparison (Group #1 vs. Group #2)	Number of Studies (N) Group #1 and Group #2 Total Sample Size (Min-Max)	Baseline Difference (95% CI)	Difference in Change Score (95% CI)	Difference in Endpoint Score (95% CI)
Triglycerides: metformin vs. rosiglitazone	6 studies #1: 198 (9–117) #2: 213 (14–128)	PL: Does not converge DL: -9.516 (-37.750, 18.717) REML: -9.657 (-37.090, 17.777) ML: -10.011 (-34.080, 14.058) PE: -9.516 (-1.4e+17, 10.651) KH: -9.657 (-47.354, 28.041) I^2 = 65.9%, P = 0. 012	PL: does not converge DL: -10.005 (-43.660, 23.651) REML: -9.362 (-45.739, 27.015) ML: -10.051 (-43.533, 23.430) PE:-10.005 (-190.295, 12.301) KH: -9.362 (-57.072, 38.348) I^2 = 89.2%, P < 0.001	PL: -9.166 (-22.621, 5.619) DL: -9.166 (-21.899, 3.568) REML: -9.166 (-21.899, 3.568) ML: -9.166 (-21.899, 3.568) PE: -9.166 (-2.1e+17, 2.401) KH: -9.166 (-25.866, 7.535) I^2 = 0.0%, P = 0.542
Triglycerides: comparing metformin vs. pioglitazone	8 studies #1: 736 (16–501) #2: 740 (14–499)	PL: Does not converge DL: -17.691 (-44.411, 9.029) REML: -18.430 (-46.652, 9.792) ML: -5.445 (-17.344, 6.454) PE: -17.691 (-147.142, 0.280) KH: -18.430 (-56.346, 19.486) I^2 = 51.9%, P = 0. 052	PL: 4.561 (-7.327, 16.511) DL: 4.561 (-7.321, 16.442) REML: 4.561 (-7.321, 16.442) ML: 4.561 (-7.321, 16.442) PE: 4.561 (0.821, 32.654) KH: 4.561 (-9.774, 18.895) I^2 = 0.0%, P = 0.820	PL: 2.015 (-8.812, 15.063) DL: 2.015 (-8.523, 12.553) REML: 2.015 (-8.523, 12.553) ML: 2.015 (-8.523, 12.553) PE :2.015 (-4.551, 9.978) KH: 2.015 (-10.699, 14.728) I^2 = 0.0%, P = 0.981
Triglycerides: metformin vs. sulfonylureas	11 studies #1: 812 (19–210) #2: 858 (18–209)	PL: 17.186 (-2.503, 23.406) **DL: 17.186 (10.965, 23.406)**[b] REML: 11.553 (-0.041, 23.147) *ML: 17.186 (10.965, 23.406)*[c] PE: 17.186 (-18.623, 22.103) KH: 11.553 (-1.628, 24.733) I^2 = 0.0%, P = 0. 500	**PL: -17.217 (-28.683, -4.876)**[b] **DL: -17.217 (-28.684, -5.750)**[b] *REML: -17.217 (-28.684, -5.750)*[c] *ML: -17.217 (-28.684, -5.750)*[c] PE: -17.217 (-94.192, -0.855) *KH: -17.217 (-30.253, -4.181)*[c] I^2 = 0.0%, P = 0.669	**PL: -5.847 (-23.606, 10.612)**[b] **DL: -6.444 (-24.536, 11.648)**[b] *REML:-6.120 (-22.472, 10.232)*[c] *ML: -5.847 (-21.399, 9.706)*[c] PE: -6.444 (-84.395, 5.507) *KH: -6.120 (-24.709, 12.470)*[c] I^2 = 67.9%, P = 0.001
Triglycerides: metformin vs. DPP-4 Inhibitors	3 studies #1: 940 (241–427) #2: 820 (96–441)	PL: 7.335 (-5.673, 20.817) DL: 7.335 (-4.245, 18.916) REML: 7.335 (-4.245, 18.916) ML: 7.335 (-4.245, 18.916) PE: Not calculable[a] KH: 7.335 (-18.086, 32.757) I^2 = 0.0%, P = 0. 616	PL: 3.047 (-11.286, 18.285) DL: 3.047 (-10.969, 17.063) REML: 3.047 (-10.969, 17.063) ML: 3.047 (-10.969, 17.063) PE: Not calculable[a] KH: 3.047 (-27.722, 33.815) I^2 = 0.0%, P = 0.845	PL: 9.797 (-5.000, 28.534) DL: 9.797 (-4.105, 23.699) REML: 9.797 (-4.105, 23.699) ML: 9.797 (-4.105, 23.699) PE: Not calculable[a] KH: 9.797 (-20.722, 40.316) I^2 = 0%, P = 0.419
Triglycerides: metformin vs. combination metformin and rosiglitazone	7 studies #1: 1035 (34–266) #2: 1210 (71–271)	PL: -3.495 (-12.625, 6.168) DL: -3.400 (-13.238, 6.437) REML: -3.495 (-11.744, 4.754) ML: -3.495 (-11.744, 4.754) PE: -3.400 (-147.146, 2.542) KH: -3.495 (-15.169, 8.179) I^2 = 22.2%, P = 0. 260	PL: does not converge DL: -4.990 (-18.190, 8.210) REML: -4.966 (-17.227, 7.295) ML: -4.824 (-16.079, 6.430) PE: -4.990 (-160.681, 3.206) KH: -4.966 (-20.274, 10.342) I^2 = 67.8%, P = 0.005	PL: -8.487 (-22.309, 9.286) DL: -7.806 (-22.264, 6.651) REML: -7.840 (-22.219, 6.538) ML: -8.487 (-21.426, 4.452) PE: -7.806 (-131.261, 4.355) KH: -7.840 (-26.855, 11.175) I^2 = 60.3%, P = 0.019

Appendix B. Comparison of Combined Mean Differences Using Change Score and Endpoints (Comparison Between Different Treatments) (continued)

Outcome **Comparison (Group #1 vs. Group #2)**	**Number of Studies (N)** **Group #1 and Group #2 Total Sample Size (Min-Max)**	**Baseline Difference (95% CI)**	**Difference in Change Score (95% CI)**	**Difference in Endpoint Score (95% CI)**
Triglycerides: metformin vs. combination metformin and sulfonylureas	6 studies #1: 625 (19–210) #2: 879 (74–213)	PL: 0.700 (-18.082, 18.057) DL: 0.700 (-16.590, 17.990) REML: 0.700 (-16.590, 17.990) ML: 0.700 (-16,590, 17.990) PE: 0.700 (-82.625, 7.379) KH: 0.700 (-21.977, 23.376) I^2 = 0.0%, P = 0. 850	PL: 9.941 (-2.404, 22.002) DL: 9.941 (-2.084, 21.967) REML: 9.941 (-2.084, 21.967) ML: 9.941 (-2.084, 21.967) PE: 9.941 (-34.244, 11.815) KH: 9.941 (-5.831, 25.713) I^2 = 0.0%, P = 0.826	PL: does not converge DL: 6.214 (-12.041, 24.470) REML: 6.186 (-12.214, 24.587) ML: 6.637 (-9.710, 22.985) PE: 6.214 (-1.1e+17, 16.196) KH: 6.186 (-18.328, 30.700) I^2 = 47.1%, P = 0.092
Triglycerides: metformin vs. combination metformin and DPP-4 inhibitors	4 studies #1: 802 (83–272) #2: 1331 (86–528)	PL: 2.980 (-7.643, 13.705) DL: 2.980 (-6.932, 12.891) REML: 2.980 (-6.932, 12.891) ML: 2.980 (-6.932, 12.891) PE: Not calculable[a] KH: 2.980 (-13.114, 19.073) I^2 = 0.0%, P = 0. 683	PL: 27.447 (15.424, 39.419) DL: 27.447 (15.611, 39.283) REML: 27.447 (15.611, 39.283) ML: 27.447 (15.611, 39.283) PE: Not calculable[a] KH: 27.447 (8.228, 46.666) I^2 = 0.0%, P = 0.904	PL: 30.498 (17.956, 42.971) DL: 30.498 (18.169, 42.827) REML: 30.498 (18.169, 42.827) ML: 30.498 (18.169, 42.827) PE: Not calculable[a] KH: 30.498 (10.479, 50.517) I^2 = 0.0%, P = 0.905
Triglycerides: pioglitazone vs. sulfonylureas	6 studies #1: 304 (17–91) #2: 311 (18–109)	PL: -2.571 (-18.550, 16.510) DL: -2.349 (-18.616, 13.919) REML: -2.571 (-17.931, 12.789) ML: -2.571 (-17.931, 12.789) PE: -2.349 (-324.811, 7.121) KH: -2.571 (-23.511, 18.369) I^2 = 7.4%, P = 0. 369	PL: -28.553 (-45.047, -13.327) DL: -28.553 (-42.577, -14.529) REML: -28.553 (-42.577, -14.529) ML: -28.553 (-42.577, -14.529) PE: -28.553 (-5.1e+17, 0.000) KH: -28.553 (-46.946, -10.159) I^2 = 0.0%, P = 0.463	PL: does not converge DL: -28.042 (-46.283, -9.801) REML: -29.172 (-44.610, -13.734) ML: -30.302 (-43.159, -17.445) PE: -28.042 (-5.6e+17, 2.996) KH: -29.172 (-53.547, -4.797) I^2 = 42.4%, P = 0.122
Triglycerides: sulfonylureas vs. meglitinides	4 studies #1: 320 (18–182) #2: 580 (16–362)	PL: Does not converge DL: 11.258 (-5.843, 28.359) REML: 11.258 (-5.843, 28.359) ML: 13.733 (2.823, 24.643) PE: Not calculable[a] KH: 11.258 (-99.605, 122.121) I^2 = 39.7%, P = 0. 198	PL: 1.095 (-11.167, 11.568) DL: 1.095 (-9.197, 11.388) REML: 1.095 (-9.197, 11.388) ML: 1.095 (-9.197, 11.388) PE: Not calculable[a] KH: 1.095 (-15.617, 17.807) I^2 = 0.0%, P = 0.898	PL: does not converge DL: 7.724 (-7.293, 22.742) REML: 7.086 (-8.897, 23.069) ML: 9.169 (-4.029, 22.366) PE: Not calculable[a] KH: 7.086 (-18.866, 33.039) I^2 = 32.2%, P = 0.219
Triglycerides: combination metformin and rosiglitazone vs. combination metformin and sulfonylureas	4 studies #1: 576 (48–210), #2: 611 (47–240)	PL: 7.028 (-2.966, 16.345) DL: 6.851 (-2.177, 15.878) REML: 6.930 (-1.204, 15.065) ML: 7.028 (-0.271, 14.327) PE: Not calculable[a] KH: 6.930 (-7.690, 21.550) I^2 = 27.9%, P = 0. 245	PL: does not converge DL: 6.739 (-15.969, 29.448) REML: 6.303 (-13.188, 25.795) ML: 5.833 (-11.500, 23.165) PE: Not calculable[a] KH: 6.303 (-25.346, 37.953) I^2 = 83.9%, P < 0.001	PL: does not converge DL: 13.608 (-9.217, 36.434) REML: 13.499 (-7.695, 34.692) ML: 13.242 (-5.156, 31.639) PE: Not calculable[a] KH: 13.499 (-20.913, 47.911) I^2 = 91.8%, P < 0.001

[a] Not calculable as there are fewer than six studies.
[b] Significant baseline difference or discrepancy in conclusion using DL or PL methods.
[c] Significant baseline difference or discrepancy in conclusion using REML, ML, PE or KH methods.
[d] Only one study.
[e] Hartung-Knapp-Sidik-Jonkman method as implemented in IntHout et al. 2015[28] was used. Otherwise, the KH estimate was obtained using Stata command *metareg*.

CI = confidence interval; CPAP = continuous positive airway pressure; DL = DerSimonian-Laird random-effects method; DPP-4 = Dipeptidyl peptidase-4; HbA1c = hemoglobin A1c/glycated hemoglobin; HDL = high-density lipoprotein; LDL = low-density lipoprotein; MADRS = Montgomery-Asberg Depression Rating Scale; ML = maximum likelihood method; PE = permutations method; PL = profile likelihood method; REML = restricted maximum likelihood method.