Journal of Clinical Epidemiology 76 (2016) 108−117

Change score or follow-up score? Choice of mean difference estimates could impact meta-analysis conclusions

Rongwei Fu[a,b,c,*], Haley K. Holmer[a,b]

[a]*Oregon Evidence-based Practice Center, Department of Medical Informatics and Clinical Epidemiology, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239, USA*

[b]*School of Public Health, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239, USA*

[c]*Department of Emergency Medicine, Oregon Health & Science University, 3181 SW Sam Jackson Park Rd, Portland, OR 97239, USA*

Accepted 3 January 2016; Published online 27 February 2016

Abstract

Objectives: In randomized controlled clinical trials, continuous outcomes are typically measured at both baseline and follow-up, and mean difference could be estimated using the change scores from baseline or the follow-up scores. This study assesses the impact of using change score vs. follow-up score on the conclusions of meta-analyses.

Study Design and Setting: A total of 63 meta-analyses from six comparative effectiveness reviews were included. The combined mean difference was estimated using a random-effects model, and we also evaluated whether the impact qualitatively varied by alternative random-effects estimates.

Results: Based on the Dersimonian−Laird (DL) method, using the change vs. the follow-up score led to five meta-analyses (7.9%) showing discrepancy in conclusions. Based on the profile likelihood (PL) method, nine (14.3%) showed discrepancy in conclusions. Using change score was more likely to show a significant difference in effects between interventions (DL method: 4 of 5; PL method: 7 of 9). A significant difference in baseline scores did not necessarily lead to discrepancies in conclusions.

Conclusions: Using the change vs. the follow-up score could lead to important discrepancies in conclusions. Sensitivity analyses should be conducted to check the robustness of results to the choice of mean difference estimates.

Keywords: Meta-analysis; Mean difference; Change score; Follow-up score; Baseline difference; Random-effects estimates

1. Introduction

In randomized controlled clinical trials (RCTs), continuous outcomes are typically measured at both baseline and follow-up time points, and mean difference is analyzed as the effect measure. Mean difference could be estimated using the change score from the baseline, the follow-up scores, or the analysis of covariance (ANCOVA) model. All these estimates provide unbiased estimates of mean difference when the clinical trials are adequately randomized, and the distribution of the baseline outcome scores is similar.

The distribution of the baseline outcome scores could become imbalanced in inadequately randomized trials, for example, due to chance, especially in small trials [1], or due to selection bias, often caused by inadequate randomization concealment [2]. In addition, systematic attrition that is linked to outcome may cause baseline imbalance (among patients with follow-up scores) [3]. Baseline imbalance occurs quite commonly in clinical trials. Hartling et al. [4] reported that 35% of RCTs at high/unclear risk of bias in child health had imbalanced baseline distribution.

When the baseline scores are imbalanced, using either the change scores from the baseline or the follow-up scores would produce biased effect estimates of mean difference and may lead to different conclusions in individual studies. Using the follow-up scores simply ignores baseline imbalance, and using the change score, contrary to common belief, does not address the issue of the baseline imbalance. 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

Funding: This project was funded under contract no. 290-2007-10057-I from the Agency for Healthcare Research and Quality, U.S. Department of Health and Human Services.

The authors of this report are responsible for its content. Statements in the report should not be construed as endorsement by the Agency for Healthcare Research and Quality or the U.S. Department of Health and Human Services.

* Corresponding author. Tel.: 503-494-6069; fax: 503-494-4981.
E-mail address: fur@ohsu.edu (R. Fu).

http://dx.doi.org/10.1016/j.jclinepi.2016.01.034

What is new?

Key findings

- Using the change score is more likely to produce significant results when there are discrepancies in conclusions; using the follow-up score is more likely to produce more conservative results.
- A significant difference in baseline scores did not necessarily lead to discrepancy in conclusions.
- Discrepancies in conclusions due to using the change score vs. the follow-up score could vary based on the choice of random-effects estimates.

What this adds to what was known?

- Using the change score versus the followup score to estimate mean difference could lead to important discrepancies in conclusions.

What is the implication and what should change now?

- Sensitivity analyses should be conducted to check the robustness of results to the choice of mean difference estimates.

the mean). For example, in an RCT [5] looking at the short-term effects of metformin (as compared with glibenclamide) in patients with type 2 diabetes mellitus, one outcome was total cholesterol. Patients randomized to the metformin group had a mean baseline level of total cholesterol of 5.1 mmol/L, and the mean baseline level for the patients randomized to the glibenclamide group was 4.8 mmol/L. After 12 weeks, the mean level of total cholesterol was 4.6 mmol/L in the metformin group and 4.9 mmol/L in the glibenclamide group. The metformin group had a higher mean baseline cholesterol level, and patients in that group did experience a greater decrease than those in the glibenclamide group. Using change scores to calculate the mean difference between the two groups produces a statistically significant estimate of 0.6 mmol/L, and the estimate is 0.3 mmol/L when using follow-up scores, only half of the above difference and not significant. In this particular example, the two estimates result in a different conclusion on the effectiveness of metformin. In addition, although the study concluded that metformin may have a favorable effect on the lipid profiles [5], the direction of mean difference estimates for high-density lipoprotein and triglycerides when using follow-up scores suggests favorable effects for glibenclamide. When baseline imbalance occurs by chance, the ANCOVA method removes conditional bias in treatment group comparisons and improves efficiency over unadjusted comparisons [6–8].

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 using study-level data, although important to consider [9–11], has not been well studied [12,13]. The difference between estimates based on the change score and the follow-up score would be reflected in the combined estimates and may be accumulated in a meta-analysis to create a potentially larger impact with possible implication in practice or health policy making. Of the 263 systematic reviews that the Agency of Healthcare Research and Quality (AHRQ) published between 1999 and June 2012, 247 reviews evaluated at least one continuous outcome, and 74 included a meta-analysis of continuous outcomes. However, only four reviews explicitly mentioned the issue of baseline imbalance [12]. ANCOVA estimates should be used in a meta-analysis whenever possible [9,11]; nonetheless, the choice of mean difference estimates has to depend on the reported data, and ANCOVA estimates may not always be available. Estimates using the change scores or the follow-up scores for mean difference often become the practical choice.

Therefore, in this meta-epidemiological study, we empirically evaluated how the choice of using the change scores or the follow-up scores to estimate the mean difference impacted the meta-analyses and whether the impact qualitatively varied by the comparator (whether the intervention was compared to a control group, or multiple interventions were compared to each other) or linked with differences in baseline scores. For this article, we specifically looked at mean difference for continuous outcomes measured in the same scale. In addition, we evaluated how different random-effects model estimates 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 (DL) random-effects model [14].

2. Methods

2.1. Selection and abstraction of data

Within the AHRQ Evidence-based Practice Center (EPC) Program, EPCs conduct comparative effectiveness reviews (CERs) of treatment options for the Effective Health Care Program [15]. From the 63 CERs conducted from 2005 to June 2012, 19 included a meta-analysis using mean difference and we selected six CERs to evaluate the impact of using the change score or the follow-up score to estimate the mean difference. To be included, the CER had to include at least one meta-analysis for continuous outcomes using mean difference. Only meta-analyses of at least three RCTs were included in this study, and we

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 [16]	2011	Acute pain	1
Diagnosis and treatment of obstructive sleep apnea in adults [17]	2011	Apnea-hypopnea index (AHI), Epworth Sleepiness Scale (ESS)	6
Nonpharmacologic interventions for treatment-resistant depression in adults [18]	2011	Depressive severity	3
Second-generation antidepressants in the pharmacologic treatment of adult depression: an update of the 2007 comparative effectiveness review [19]	2011	Montgomery–Åsberg Depression Rating Scale (MADRS)	1
Oral diabetes medications for adults with type 2 diabetes: an update of the 2007 report [20]	2011	Hemoglobin A1C, weight, LDL, HDL, triglycerides	50
Screening, behavioral counseling, and referral in primary care to reduce alcohol misuse [21]	2012	Weekly alcohol use	2

Abbreviations: HDL, high-density lipoprotein; LDL, low-density lipoprotein.

specifically evaluated mean difference 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 (e.g., if two CERs examined lipid variables or HbA1C, we picked only one of them). We only considered the updated review if there was an earlier review and an updated version. Finally, additional CERs were excluded if they only included variables that were not measured at both baseline and follow-up (e.g., birth weight, length of stay) or lacked forest plots to identify which studies and data were included in the meta-analyses.

A total of 63 meta-analyses were evaluated in this study. For each meta-analysis, we identified all original publications. One investigator abstracted data on outcomes, comparison groups, the analysis method, all data at baseline and follow-up, and any change data, including ANCOVA estimates; 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 in this study.

2.2. Statistical analysis

For each meta-analysis, we calculated two estimates of mean difference and the associated standard errors based on change score and follow-up score. When the standard deviation for baseline or follow-up score was missing, it was imputed using the mean standard deviation from studies with reported standard deviations in that meta-analysis. The standard deviation of the change score, when not reported, was calculated from baseline and follow-up standard deviations by assuming that the correlation between baseline and follow-up scores was 0.5. When mean difference could not be calculated based on change scores or follow-up scores due to inadequately reported data, we used the ANCOVA estimate or other estimate of mean difference 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 [22].

The mean difference estimates based on the change score or the follow-up score were combined using random-effects models. Given that different random-effects estimates might affect the results of using different mean difference estimates, we evaluated alternative random-effects estimators and each meta-analysis was conducted using six random-effects estimates: the 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 (KH) modification of random-effect estimate [23]. The primary analyses focused on results from the DL and PL methods as they are the methods with better performance [24,25]. We also assessed whether the baseline scores imbalance was due to chance by meta-analyzing the baseline score differences between treatment groups [10]. 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 [8,26,27].

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 [28]. In addition, we conducted sensitivity analyses by using ANCOVA estimates whenever available and by assuming different values of correlation between baseline and follow-up scores. Sensitivity analyses produced similar results and were not further reported. We did not use the ANCOVA estimates in the primary analyses as we aimed to focus on the comparison between using the change vs. the follow-up score. For each random-effects estimate, we qualitatively compared the combined estimates

using the change score and follow-up 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 (e.g., the combined estimate using mean difference based on the change score shows a significant difference, while using the follow-up score does not; or vice versa). Furthermore, qualitative difference means that two estimates show difference in the magnitude of effect, but no discrepancy in conclusion.

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

3. Results

A total of 63 meta-analyses from six CERs [16−21] were included in the following evaluation.

3.1. Impact of the mean difference estimates

These meta-analyses included 156 trials, among which 58 trials (37.2%) reported using the ANCOVA model, although only 16 trials (27.5% of 58) reported at least one ANCOVA estimate. Comparisons of results using the change score vs. the follow-up score based on DL and PL methods are shown in Tables 2 and 3. Table 3 presents only meta-analyses with discrepancy in conclusion or significant baseline differences. The complete results for all analyses based on all six random-effects estimates are shown in Appendices A and B at www.jclinepi.com.

Based on the DL method, using the change score vs. the follow-up score led to 5 of the 63 meta-analyses (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). 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 the change score, and one showed significant result when using the follow-up score. For the nine meta-analyses showing discrepancies using the PL method, the numbers were seven when using the change score and two when using the follow-up score. Therefore, using the follow-up score is more likely to produce conservative results in this analysis.

Compared to the DL method, estimates based on the ML and REML methods tended to have narrower 95% confidence intervals (CIs), although the REML method sometimes provided more conservative 95% CIs. The PE method worked only when there were at least six studies in a meta-analysis and provided 95% CIs that were much wider and often unrealistically too wide. In addition, the PE 95% CIs are often practically invalid with lower bounds like −4.8e+15 when there were only six studies. Using the ML method led to discrepancy in conclusions in six meta-analyses, and the impact of the change score vs. the follow-up 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 Appendices A and B at www.jclinepi.com). Estimates from the PE method are not available in many cases, and the 95% CIs of the PE method, when present, are too wide to make meaningful comparisons. Finally, discrepancy in conclusions was shown in 10 meta-analyses when using the KH modification, the highest among the different methods. Similarly, using the change score is more likely to show a significant difference in effects (seven meta-analyses) than using the follow-up score (three meta-analyses).

3.2. Meta-analysis of intervention vs. control

Ten meta-analyses from four CERs [16−18,21] compared an active intervention vs. a control or usual care; outcomes analyzed were pain, apnea-hypopnea, sleepiness, depression, and alcohol use. These meta-analyses included from 5 to 13 studies and from 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, bolded in Table 2), and the combined mean differences using change score vs. follow-up score also led to discrepancy in conclusions in another meta-analysis (depressive severity, bolded in Table 2). Therefore, the significant baseline imbalance does not necessarily coincide with the discrepancy in conclusions. For this particular case of baseline imbalance (apnea-hypopnea index), the magnitude of the mean difference is large 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 the PL and DL methods (as well as the REML method and the KH modification method, Appendix A at www.jclinepi.com). More interestingly, except for one meta-analysis (acute pain, skin traction vs. no traction), the combined mean difference using the change score consistently showed a larger intervention effect than the combined mean difference using the follow-up score, and the combined mean difference using the follow-up score produced an intervention effect about 20% smaller on average, ranging from 5% to more than 40%.

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

Table 2. Comparison of combined mean differences using change score and follow-up score (treatment vs. 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 scores (95% CI)	Difference in follow-up scores (95% CI)	Percentage difference in combined estimates
Pain management interventions for hip fracture [16]	Acute pain: skin traction vs. no traction	13 studies; #1: 568 (30–166) #2: 662 (34–151)	PL: −0.30 (−0.81, 0.26); DL: −0.28 (−0.84, 0.28); I^2 = 77.2%; P < 0.001	PL: 0.43 (−0.14, 1.00); DL: 0.43 (−0.13, 0.99); I^2 = 76.5%; P < 0.001	PL: 0.12 (−0.33, 0.62); DL: 0.12 (−0.32, 0.56); I^2 = 63.2%; P = 0.008	PL: 73%; DL: 72%
Diagnosis and treatment of obstructive sleep apnea in adults [17]	Apnea-hypopnea index: CPAP vs. control	6 studies; #1: 177 (12–66) #2: 159 (12–59)	PL: 2.21 (−3.44, 9.01); DL: 2.53 (−2.80, 7.86); I^2 = 64.7%; P = 0.015	PL: −27.02 (−41.19, −13.63); DL: −27.05 (−38.91, −15.19); I^2 = 93.3%; P < 0.001	PL: −24.18 (−36.05, −13.15); DL: −24.13 (−33.60, −14.67); 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: −0.02 (−0.52, 0.49); DL: −0.02 (−0.44, 0.41); I^2 = 0%; P = 0.805	PL: −2.68 (−4.31, −1.17); DL: −2.71 (−4.27, −1.16); I^2 = 84.3%; P < 0.001	PL: −2.33 (−3.65, −1.22); DL: −2.37 (−3.42, −1.33); I^2 = 54.2%; P = 0.042	PL: −14%; DL: −12%
	Apnea-hypopnea index: CPAP vs. sham CPAP	8 studies; #1: 163 (15–27) #2: 149 (10–29)	**PL: 7.16 (1.24, 13.09); DL: 7.16 (1.26, 13.06);** I^2 = 0%; P = 0.920	PL: −45.63 (−58.29, −34.18); DL: −45.89 (−57.53, −34.25); I^2 = 70.9%; P = 0.001	PL: −40.50 (−52.29, −29.40); DL: −40.69 (−52.07, −29.31); 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: 0.31 (−0.26, 0.88); DL: 0.31 (−0.26, 0.88); I^2 = 0%; P = 0.920	PL: −2.68 (−4.35, −1.03); DL: −2.69 (−4.40, −0.98); I^2 = 83.4%; P < 0.001	PL: −2.56 (−4.20, −0.92); DL: −2.56 (−4.21, −0.90); I^2 = 82.2%; P < 0.001	DL: −5%; PL: −5%
Nonpharmacologic interventions for treatment-resistant depression in adults [18]	Depressive severity: rTMS vs. sham (condition = Tier 1, MDD)	8 studies; #1: 116 (7–32) #2: 102 (5–31)	PL: 0.64 (−0.53, 2.67); DL: 0.64 (−0.46, 1.74); I^2 = 0%; P = 0.520	**PL: −5.44 (−7.79, −2.67); DL: −5.39 (−7.76, −3.03);** I^2 = 43.9%; P = 0.086	**PL: −3.07 (−6.17, 0.37); DL: −2.90 (−6.51, 0.71);** I^2 = 79.2%; P < 0.001	PL: −42%; DL: −45%
	Depressive severity: rTMS vs. sham (condition = Tier 1)	11 studies; #1: 197 (7–36) #2: 149 (5–31)	PL: 0.68 (−0.37, 2.28); DL: 0.68 (−0.28, 1.64); I^2 = 0%; P = 0.452	PL: −5.69 (−7.81, −3.48); DL: −5.67 (−7.78, −3.56); I^2 = 54.4%; P = 0.015	PL: −3.91 (−6.28, −1.40); DL: −3.84 (−6.48, −1.20); I^2 = 73.6%; P < 0.001	PL: −36%; DL: −37%
	Depressive severity: rTMS vs. sham (condition = Tier 1&2, MDD)	12 studies; #1: 374 (7–155) #2: 364 (5–146)	PL: −0.01 (−0.60, 0.76); DL: −0.01 (−0.60, 0.54); I^2 = 0%; P = 0.688	PL: −4.72 (−7.10, −2.32); DL: −4.71 (−7.13, −2.30); I^2 = 80.4%; P < 0.001	PL: −3.44 (−5.94, −0.80); DL: −3.44 (−5.83, −1.04); I^2 = 79.3%; P < 0.001	PL: −26%; DL: −26%

(*Continued*)

Table 2. 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 scores (95% CI)	Difference in follow-up scores (95% CI)	Percentage difference in combined estimates
Screening, behavioral counseling, and referral in primary care to reduce alcohol misuse [21]	Drinks/week: BCI vs. control (adults, 6 months)	11 studies; #1: 1,547 (39–353) #2: 1,556 (32–376)	PL: 0.61 (−0.25 1.61); DL: 0.61 (−0.23, 1.46); I^2 = 0%; P = 0.490	PL: −3.12 (−4.26, −2.31); DL: −3.23 (−4.21, −2.24); I^2 = 13.9%; P = 0.311	PL: −2.50 (−4.04, −1.21); DL: −2.59 (−4.06, −1.12); I^2 = 46.3%; P = 0.046	PL: −20%; DL: −20%
	Drinks/week: BCI vs. control (adults, 12 months)	13 studies; #1: 2,088 (33–371) #2: 2,012 (39–381)	PL: 0.46 (−0.40, 1.28); DL: 0.46 (−0.36, 1.27); I^2 = 0%; P = 0.649	PL: −3.70 (−4.64, −2.81); DL: −3.72 (−4.76, −2.68); I^2 = 16.9%; P = 0.274	PL: −2.92 (−4.91, −0.80); DL: −2.94 (−4.60, −1.27); I^2 = 57.2%; P = 0.005	PL: −21%; DL: −21%

The bold values indicate significant baseline differences or discrepancy in conclusion.

Abbreviations: BCI, behavioral counseling intervention; CI, confidence interval; CPAP, continuous positive airway pressure; DL, DerSimonian–Laird random-effects method; MDD, major depressive disorder; PL, profile likelihood method; rTMS, repetitive transcranial magnetic stimulation.

3.3. Meta-analysis of comparison between interventions

Fifty-three meta-analyses from three CERs [17,19,20] compared outcomes between active interventions. Most (50) of these are from one CER [20] 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 [17] and depression [19] 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 vs. the follow-up score led to 4 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 (bolded 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 using the follow-up score. Only two meta-analyses showed significant baseline differences using the PL method, and one is associated with a discrepancy in conclusions. We did not calculate the percent difference between the two combined mean differences because 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 estimates. When heterogeneity shows difference between the two estimates, there is no clear pattern of one estimate having more heterogeneity than the other. In the presence of a discrepancy in conclusions, the estimate with significant result does 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, scores on the Montgomery-Asberg Depression Rating Scale, and measures of HbA1C, weight, and lipids). These meta-analyses revealed no clear common characteristics.

4. Discussion

The issue of baseline imbalance was inadequately addressed in systematic reviews and CERs sponsored by AHRQ [12]. To our knowledge, this is the first empirical evaluation of how using the change vs. the follow-up score would affect the combined mean differences. Discrepancy in conclusions was shown in 5 of the 63 meta-analyses (7.9%) using the DL method and 9 (14.3%) using the PL method. Although the conclusions were consistent in most cases, it was concerning that up to 14.3% 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 [14]. The common misconception was that using change scores accounts for baseline difference, and five of the six CERs actually used the change score to estimate the mean difference [16–20], although it does not address the issue. In general, the mean difference based on the change score was more likely to show significant results, and using the follow-up score more likely produced more conservative results, although neither was necessarily more accurate. These results support the current AHRQ guidance that advises review authors to

Table 3. Discrepancy in comparison of combined mean differences using change score and follow-up score (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 scores (95% CI)	Difference in follow-up scores (95% CI)
Second-generation antidepressants in the pharmacologic treatment of adult depression: an update of the 2007 comparative effectiveness review [19]	MADRS: citalopram vs. escitalopram	6 studies; #1: 939 (125–214) #2: 932 (108–241)	PL: 0.18 (−0.34, 0.65); DL: 0.17 (−0.28, 0.63); I^2 = 12.7%; P = 0.333	**PL: 2.11 (0.08, 4.01);** DL: 2.08 (0.17, 3.98); I^2 = 67.8%; P = 0.008	**PL: 2.26 (−0.07, 4.43);** DL: 2.22 (0.06, 4.38); I^2 = 68.7%; P = 0.007
Oral diabetes medications for adults with type 2 diabetes: an update of the 2007 report [20]	HbA1c: metformin vs. thiazolidinediones	14 studies; #1: 1,132 (13–501) #2: 1,127 (14–499)	PL: −0.02 (−0.10, 0.06); DL: −0.02 (−0.11, 0.07); I^2 = 9.7%; P = 0.347	**PL: −0.11 (−0.19, −0.03);** DL: −0.11 (−0.21, 0.003); I^2 = 24.9%; P = 0.186	**PL: −0.05 (−0.19, 0.10);** DL: −0.05 (−0.19, 0.08); I^2 = 58.0%; P = 0.003
	HbA1c: metformin vs. metformin and thiazolidinediones	11 studies; #1: 1,428 (34–277) #2: 1,688 (60–296)	**PL: −0.16 (−0.28, −0.03); DL: −0.16 (−0.29, −0.03);** I^2 = 69.6%; P = 0.001	PL: 0.63 (0.43, 0.85); DL: 0.64 (0.44, 0.83); I^2 = 85.5%; P < 0.001	PL: 0.50 (0.28, 0.73); DL: 0.51 (0.27, 0.74); 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: −0.12 (−0.24, 0.01); **DL: −0.13 (−0.22, −0.03);** I^2 = 8.3%; P = 0.363	**PL: −0.05 (−0.15, 0.07); DL: −0.05 (−0.15, 0.05);** I^2 = 0%; P = 0.628	**PL: −0.17 (−0.26, −0.06); DL: −0.17 (−0.26, −0.07);** 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: 1.67 (0.12, 3.21); DL: 1.67 (0.16, 3.18);** I^2 = 0%; P = 0.939	**PL: −2.24 (−2.75, −1.79); DL: −2.23 (−2.64, −1.83);** I^2 = 2.8%; P = 0.408	**PL: −0.63 (−2.14, 0.85); DL: −0.63 (−2.12, 0.85);** 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: 1.75 (−2.86, 6.12); DL: 1.67 (−2.82, 6.16); I^2 = 87.5%; P < 0.001	**PL: −2.69** (**−3.84**, **−1.55**); **DL: −2.69** (**−3.75**, **−1.63**); I^2 = 78.1%; P = 0.003	**PL: −0.94** (**−6.11**, **3.97**); **DL: −1.04** (**−6.51**, **4.43**); I^2 = 90.8%; P < 0.001
	LDL: metformin vs. rosiglitazone	6 studies; #1: 198 (9–117) #2: 213 (14–128)	PL: 2.10 (−4.19, 8.89); DL: 2.10 (−3.96, 8.16); I^2 = 0.0%; P = 0.682	**PL: −12.54 (−22.24, −5.88);** DL: −13.26 (−20.55, −5.97); I^2 = 58.0%; P = 0.036	**PL: −14.01 (−29.49, 2.27);** DL: −13.94 (−27.56, −0.33); 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: −2.41 (−4.40, −0.42);** I^2 = 0.0%; P = 0.673	PL: −0.60 (−1.66, 2.32); DL: −0.05 (−1.88, 1.77); I^2 = 42.6%; P = 0.121	PL: 0.45 (−2.98, 3.83); DL: 0.45 (−2.94, 3.83); I^2 = 0%; P = 0.922

(*Continued*)

Table 3. 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 scores (95% CI)	Difference in follow-up scores (95% CI)
	HDL: metformin vs. DPP-4 inhibitors	3 studies; #1: 913 (241–427) #2: 791 (95–440)	PL: 0.71 (−0.27, 1.67); DL: 0.71 (−0.22, 1.64); I^2 = 0.0%; P = 0.787	**PL: 1.43 (−0.004, 3.25);** DL: 1.51 (0.13, 2.90); I^2 = 34.1%; P = 0.219	**PL: 2.11 (0.71, 3.90);** DL: 2.20 (0.81, 3.58); I^2 = 33.9%; P = 0.220
	HDL: pioglitazone vs. sulfonylurea	6 studies; #1: 304 (17–91) #2: 311 (18–109)	PL: 1.38 (−0.02, 2.70); **DL: 1.38 (0.07, 2.67);** I^2 = 0.0%; P = 0.706	PL: 5.30 (3.48, 6.96); DL: 5.28 (3.47, 7.08); I^2 = 43.3%; P = 0.117	PL: 6.37 (3.53, 8.85); DL: 6.33 (4.02, 8.63); I^2 = 60.5%; P = 0.027
	Triglycerides: metformin vs. sulfonylureas	11 studies; #1: 812 (19–210) #2: 858 (18–209)	PL: 17.19 (−2.50, 23.41); **DL: 17.19 (10.97, 23.41);** I^2 = 0.0%; P = 0.500	**PL: −17.22 (−28.68, −4.88); DL: −17.22 (−28.68, −5.75);** I^2 = 0.0%; P = 0.669	**PL: −5.85 (−23.61, 10.61); DL: −6.44 (−24.54, 11.65);** I^2 = 67.9%; P = 0.001

The bold values indicate significant baseline difference or discrepancy.

Abbreviations: 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.

conduct sensitivity analyses using both scores to assess the robustness of results to the choice of mean difference estimates [11].

It was interesting that in the meta-analyses of intervention vs. 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 over 40% higher. Because 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 [29]. This was also roughly shown by the combined baseline differences and suggested potential bias in randomization that may not be due to chance (patients with a condition, i.e., more severe 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 vs. the follow-up 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 follow-up is large, the standard error of the mean difference based on the change score will be smaller than that of the mean difference based on the follow-up 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 conclusions only reflects small shifts in numerical values. However, such shifts could be critical to consider and may have potential health policy implications because 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. Although 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 significant baseline difference is directly linked with discrepancy in conclusions, it highlights the importance to evaluate baseline imbalance for each meta-analysis.

Results on discrepancy in conclusions based on the ML method were similar to the DL method, and those based on

the REML method were similar to the PL method. Furthermore, 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 the ML and REML methods, although the PL method did not converge in a small proportion of meta-analyses, and the 95% CIs based on the PE method are too wide and conservative for general use [24,25]. Interestingly, the KH modification produced most discrepancies, and it generally produced a wider CI than the DL and PL methods. Although the KH modification has been shown to result in more adequate error rates than the DL method [30], its overall performance has been shown not to be better than the DL and PL methods [24].

We did not use the Trowman's method [10] or the modified Trowman's method [9] to adjust the summary baseline score using meta-regression because it has been shown that such methods could provide misleading results [9]. Meta-regression adjusts the summary baseline score only in the study level and ecological fallacy could play a role here. However, the evidence was still limited, and 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 [9,11]. Unfortunately, in this study, we found that the reporting of ANCOVA estimates has been poor, although 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 are conducted so they can be properly used in evidence synthesis and potential health policy decision making. On the other hand, this is one situation where availability of individual participant data (IPD) would provide the best solution, as baseline imbalance could be consistently adjusted using the ANCOVA model across studies. IPD could also help address the concern that systematic attrition may lead to baseline imbalance (among patients with follow-up scores) [3].

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 and by our focus on evaluating outcomes measured on the same scale using mean difference. Most outcomes were related to diabetes and came from one review [20] with multiple studies included in more than one meta-analysis. This would affect the generalizability of the results. However, these meta-analyses did address multiple research 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, likely using standardized mean difference (SMD). Baseline imbalance and the choice of the scores have further implications on estimating SMD. The standard deviation is also involved in calculating SMD, and the standard deviations are calculated differently when using the change score vs. the follow-up score. Although da Costa et al. [13] found no evidence for systematic differences between SMDs derived from the follow-up and the change pain scores without evaluating the impact of baseline scores, as with mean difference, the overall evidence is 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, such as cohort or other observational studies, were not considered. The issue of baseline imbalance in observational studies is inherently different from RCTs because baseline balance is not expected and Lord's paradox could further complicate the analysis [31].

In summary, using the change score vs. the follow-up 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, and using the follow-up score is more likely to produce more conservative results. Sensitivity analyses should be conducted to check the robustness of results to the choice of mean difference estimates.

Acknowledgments

The authors extend deep appreciation for the contributions of the reviewers who agreed to review this article, and the authors recognize Elaine Graham, M.L.S., and Amber L. Laurie, M.S., affiliated with Oregon Evidence-based Practice Center, Oregon Health & Science University, for their valuable contributions to this article. The funding source had no role in the study design, the collection, analysis, and interpretation of data, or the writing of the report.

Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.jclinepi.2016.01.034.

References

[1] Rosenberger W, Lachin J, editors. Randomization in clinical trials: theory and practice. New York: Wiley; 2002.
[2] Schulz KF, Grimes DA. Allocation concealment in randomised trials: defending against deciphering. Lancet 2002;359:614−8.
[3] Hewitt CE, Kumaravel B, Dumville JC, Torgerson DJ. Assessing the impact of attrition in randomized controlled trials. J Clin Epidemiol 2010;63:1264−70.

[4] Hartling L, Hamm MP, Fernandes RM, Dryden DM, Vandermeer B. Quantifying bias in randomized controlled trials in child health: a meta-epidemiological study. PLoS One 2014;9:e88008.
[5] Amador-Licona N, Guizar-Mendoza J, Vargas E, Sanchez-Camargo G, Zamora-Mata L. The short-term effect of a switch from glibenclamide to metformin on blood pressure and microalbuminuria in patients with type 2 diabetes mellitus. Arch Med Res 2000;31:571−5.
[6] Crager MR. Analysis of covariance in parallel-group clinical trials with pretreatment baselines. Biometrics 1987;43:895−901.
[7] Senn SJ. Covariate imbalance and random allocation in clinical trials. Stat Med 1989;8:467−75.
[8] Senn S. Testing for baseline balance in clinical trials. Stat Med 1994;13:1715−26.
[9] Riley RD, Kauser I, Bland M, Thijs L, Staessen JA, Wang J, et al. Meta-analysis of randomised trials with a continuous outcome according to baseline imbalance and availability of individual participant data. Stat Med 2013;32:2747−66.
[10] Trowman R, Dumville JC, Torgerson DJ, Cranny G. The impact of trial baseline imbalances should be considered in systematic reviews: a methodological case study. J Clin Epidemiol 2007;60:1229−33.
[11] Fu R, Vandermeer BW, Shamliyan TA, O'Neil ME, Yazdi F, Fox SH, 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; 2013. Available at www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed April 17, 2016.
[12] 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; 2015. Available at www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed April 17, 2016.
[13] da Costa BR, Nuesch E, Rutjes AW, Johnston BC, Reichenbach S, Trelle S, et al. Combining follow-up and change data is valid in meta-analyses of continuous outcomes: a meta-epidemiological study. J Clin Epidemiol 2013;66:847−55.
[14] Cornell JE, Mulrow CD, Localio R, Stack CB, Meibohm AR, Guallar E, et al. Random-effects meta-analysis of inconsistent effects: a time for change. Ann Intern Med 2014;160:267−70.
[15] U.S. Department of Health & Human Services. Agency for Healthcare Research and Quality. Effective Health Care Program. Available at http://effectivehealthcare.ahrq.gov/. Accessed November 19, 2014.
[16] Abou-Setta AM, Beaupre LA, Jones CA, Rashiq S, Hamm MP, Sadowski 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; 2011. Available at www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed April 17, 2016.
[17] Balk EM, Moorthy D, Obadan NO, Patel K, Ip S, Chung M, 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; 2011. Available at: www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed April 17, 2016.
[18] Gaynes BN, Lux L, Lloyd S, Hansen RA, Gartlehner G, Thieda P, 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) Evidencebased Practice Center under Contract No. 290-02-0016I.) AHRQ Publication No. 11-EHC056-EF. Rockville, MD: Agency for Healthcare Research and Quality; 2011. Available at www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed April 17, 2016.
[19] Gartlehner G, Hansen RA, Morgan LC, Thaler K, Lux LJ, Van Noord M, 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; 2011. Available at www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed April 17, 2016.
[20] Bennett WL, Wilson LM, Bolen S, Maruthur N, Singh S, Chatterjee R, 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; 2011. Available at www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed April 17, 2016.
[21] Jonas DE, Garbutt JC, Brown JM, Amick HR, Brownley KA, Council CL, 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; 2012. Available at www.effectivehealthcare.ahrq.gov/reports/final.cfm. Accessed April 17, 2016.
[22] Higgins JP, White IR, Anzures-Cabrera J. Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. Stat Med 2008;27:6072−92.
[23] Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. Stat Med 2003;22:2693−710.
[24] 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;21(4):409−26.
[25] 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;21(6):657−9.
[26] Begg CB. Suspended judgment. Significance tests of covariate imbalance in clinical trials. Control Clin Trials 1990;11:223−5.
[27] Roberts C, Torgerson DJ. Understanding controlled trials: baseline imbalance in randomised controlled trials. BMJ 1999;319:185.
[28] Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557−60.
[29] Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. BMJ 2001;323:1123−4.
[30] 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.
[31] Wright DB. Comparing groups in a before-after design: when t test and ANCOVA produce different results. Br J Educ Psychol 2006;76(Pt 3):663−75.