Journal Club

Apisara Keesukphan

Change score or follow-up score?

Choice of mean difference estimates could impact meta-analysis conclusions

Rongwei Fu, Haley K. Holmer

Journal of Clinical Epidemiology, 2016,108 - 117

Introduction

Meta-analysis of continuous outcome

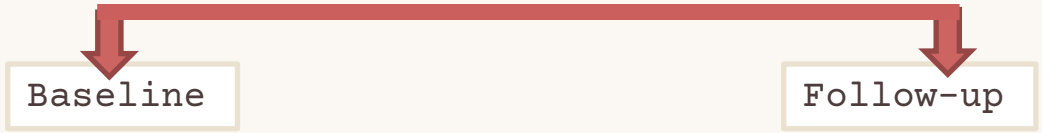
- Meta-analysis is a statistical tool of a systematic review
- Pooling data is generally based on type of outcome and number of treatment groups
- This paper focus on direct comparison of outcome between 2 groups, which the outcome is continuous data

Meta-analysis of continuous outcome

- Aim of pooling is to **compare mean values between two groups**
- **Mean difference** will be calculated and pooled across studies

Introduction

- In Randomized controlled trial (RCTS), continuous outcome are typically measured at both **baseline** and **follow-up** time points



Mean difference estimated

- **Mean difference** is analyzed as the effect measure
- Mean difference could be estimated using
  - The change score from baseline
  - The follow-up score
  - The analysis of covariance (ANCOVA) model

All these estimates provide **unbiased estimates** of mean difference when
- The clinical trials are adequately randomization
- The distribution of the baseline outcome scores is similar

Cause of baseline imbalance

- The distribution of baseline outcome scores could become **imbalanced** in **inadequately randomized trials** for example,
  - Due to chance, especially in small trials
  - Due to selection bias, inadequate randomization concealment

Result of baseline imbalance

Imbalanced baseline score → Biased effect estimates of mean difference → Different conclusion in individual studies

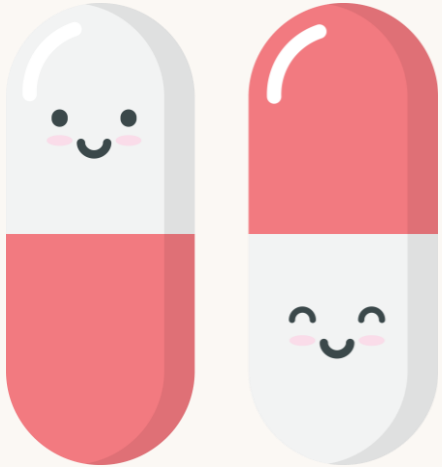
- Change score from base-line
- Follow-up score

Baseline imbalance

- **Follow-up score**
  - Simply ignores baseline imbalance
- **Change score from baseline**
  - Dose not address the issue of the baseline imbalance
  - Negatively associated with the baseline score
  - Patients with a worse baseline score are more likely to experience a high change score (regression to the mean)

Example

- RCT looking for short term effects of **Metformin** as compared with **Glibenclamide** in patients with type 2 diabetes mellitus
- Outcome: total cholesterol



Metformin vs Glibenclamide

	Metformin	Glibenclamide
Baseline total cholesterol (mmol/L)	5.1	4.8
Follow up (after 12 weeks) total cholesterol (mmol/L)	4.6	4.9
Change from baseline	4.6 - 5.1= -0.5	4.9 - 4.8 = 0.1

Mean difference between 2 group = 0.3 mmol/L

Mean difference between 2 group = 0.6 mmol/L

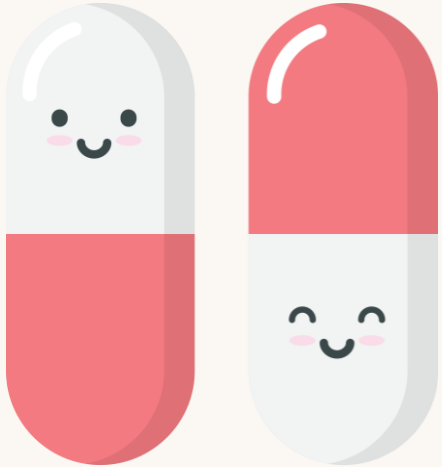
Example

- The two estimates result in a **different conclusion** on the effectiveness of metformin.

Mean difference form **Change score**

VS

Mean difference from **Follow up score**



Baseline imbalance

- **Baseline imbalance**
  - Could be attenuated by using ANCOVA model to adjust for baseline imbalance in individual studies
  - Its impact on meta-analysis using study-level data, although important to consider has not been well studied

ANCOVA

- ANCOVA estimates should be used in a meta-analysis whenever possible
- However, 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** to estimate the mean difference become the **practical choice**

Objective

- Therefore, in this meta-epidemiological study, empirically evaluated **how the choice of using the change scores or the follow-up scores to estimate the mean difference impacted the meta-analyses**

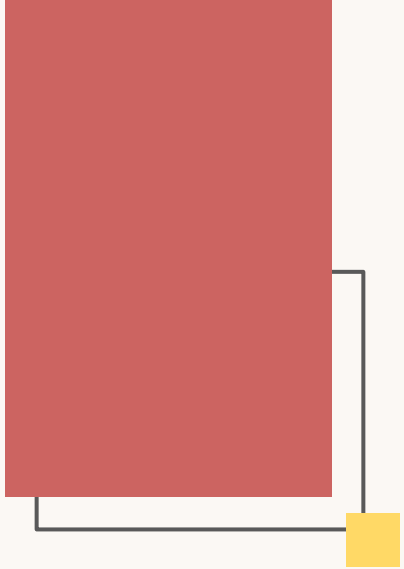
- And whether the impact qualitatively varied by the comparator or linked with difference baseline scores

Objective

- For this article, they specifically looked at **mean difference in the same scale.**

- In addition, they evaluated how different **random-effects model estimates** might affect the impact of the choice of mean difference estimates.

Methods



Selection and abstraction of data

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.
- From 63 CERs conducted from 2005 to 2012, 19 CERs included a meta-analysis using mean difference and they selected **6 CERs** to evaluate the impacted of using the change score or the follow-up score to estimate the mean difference.

AHRQ: Agency of Healthcare Research and Quality

Inclusion & Exclusion criteria

- **Inclusion**
  - 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
- **Exclusion**
  - Only included variables that were not measured at both baseline and follow-up
  - Lacked forest plots to identify which studies and data were included in the meta-analysis

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

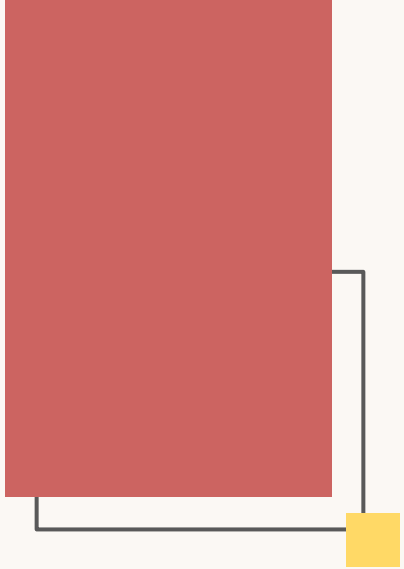
Abbreviations: HDL, high-density lipoprotein; LDL, low-density lipoprotein.

- A total of 63 meta-analyses were included in this study
- For each meta-analysis, they identified all original publications

Selection and abstraction of data

- 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

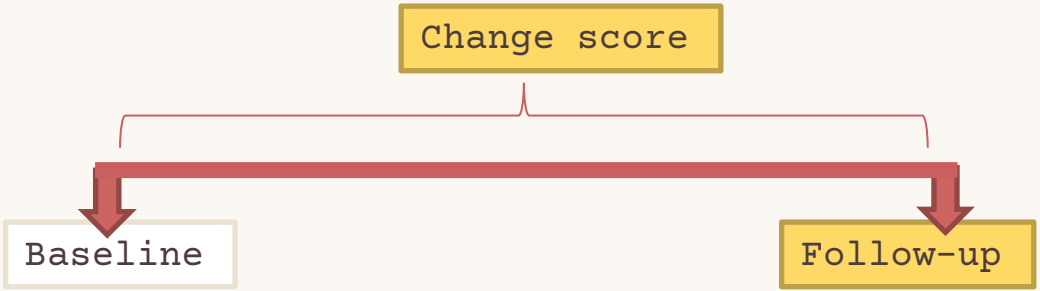
- A second investigator reviewed data abstraction for accuracy



Statistical analysis

Calculate mean difference

- For each meta-analysis, they calculated two estimates of **mean difference** and the associated **standard errors** based on **change score** and **follow-up score**



Calculate SD

- When standard deviation (SD) for **baseline** or **follow-up** score was missing
  - It was imputed using the **mean SD** from studies with reported SD in that meta-analysis

- When SD of the change score was not reported
  - It was calculated **from baseline and follow-up SD** by assuming that the correlation between baseline and follow score was 0.5

Statistical analysis

- When mean difference could not be calculated based on change scores or follow-up scores due to **inadequately reported data**, they used the ANCOVA estimate or other estimate of mean difference reported in the publication

Random-effects models

- The mean difference estimates based on the change score or the follow-up score were combined using **random-effects models.**
  - 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
  - The Knapp-Hartung (KH) method

- Focus
- Better performance

Baseline score imbalance assessment

- Assessed whether **the baseline score imbalance** was due to chance by meta-analyzing the baseline score differences between treatment groups

- If the baseline scores imbalance occurred by chance, the combined overall baseline score difference between treatment groups should be close to 0

Heterogeneity assessment

- Assessed the presence of statistical **heterogeneity** among the studies
  - The standard Cochran’s chi-square test
  - I^2 statistic

Statistical analysis

- Sensitivity analysis by using ANCOVA estimates whenever available
- Sensitivity analysis produced similar results and were not further reported

Statistical analysis

- For each random-effects estimate, they qualitatively compared the combined estimates using the **change score** and **follow-up score** to see whether there was **discrepancy in conclusion**

Discrepancy in conclusion

Discrepancy in conclusion

- One estimate show statistically significant difference and the other estimate does not

Change score → Significant difference

Follow up score → **Not** Significant difference

Qualitative difference

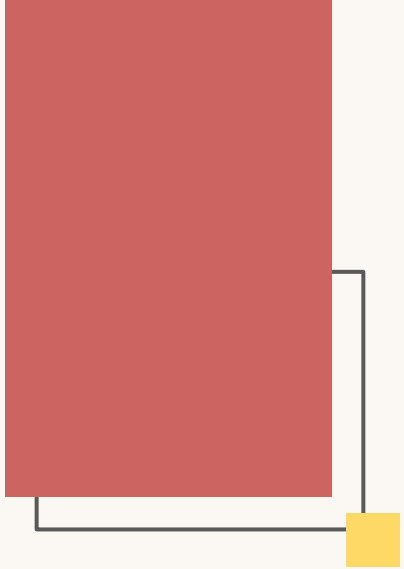
Qualitative difference

- two estimates show difference in the magnitude of effect, but no discrepancy in conclusion

Results

Results

- A total of 63 meta-analyses from six CERs were included in the following evaluation
  - **Impact of means difference estimates**
  - **Meta-analysis of intervention vs control**
  - **Meta-analysis of comparison between interventions**



Impact of the mean difference estimates

Impact of means difference estimates

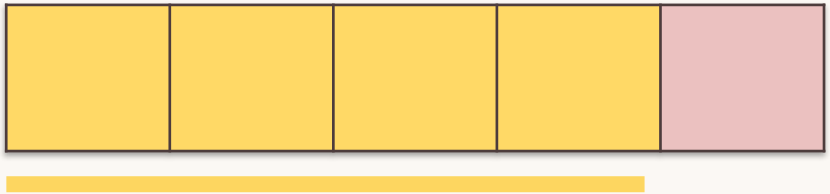
- These meta-analysis included 156 trials
- Base on **DL method**, using change score vs the follow-up score
  - 5 of 63 meta-analyses (7.9%) showing discrepancy in conclusion
- Base on **PL method**, using change score vs the follow-up score
  - 9 of 63 meta-analyses (14.3%) showing discrepancy in conclusion

Change score

- In general, using the **change score** is more likely to show **a significant difference** in the effects between interventions

DL method

- **DL** 5 of 63 meta-analyses (7.9%) showing discrepancy in conclusion

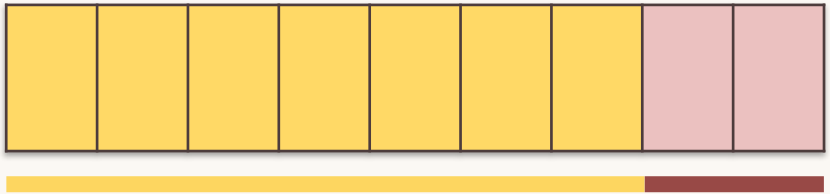


4 showed significant difference when using **change score**

1 showed significant difference when using the **follow-up score**

PL method

- **PL** 9 of 63 meta-analyses (14.3%) showing discrepancy in conclusion



7 showed significant difference when using **change score**

2 showed significant difference when using the **follow-up score**

Follow-up score

- Therefore, using the **follow-up score** is more likely to produce **conservative results** in this analysis

Meta-analysis of intervention vs control

Meta-analysis of intervention vs control

- 10 meta-analysis from 4 CERs compared an **active intervention** vs **a control or usual care**
- **Outcomes analyzed**
  - Pain, apnea-hypopnea, sleepiness, depression, and alcohol use
- These meta-analyses included
  - 5 to 13 studies
  - 218 to 4,100 patients

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
Diagnosis and treatment of obstructive sleep apnea in adults [17]	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%

- 1 meta-analysis showed baseline imbalance
  - The combined baseline difference differences from zero

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
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%

- 1 meta-analysis showed **discrepancy in conclusion** between using change score and follow-up score

Baseline imbalance vs Discrepancy in conclusion

- The significant **baseline imbalance** does not necessarily coincide with the **discrepancy in conclusion**

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
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%

- Using change score showed significant results based on both the PL and DL method.
  (as well as REML method and the KH modification method)

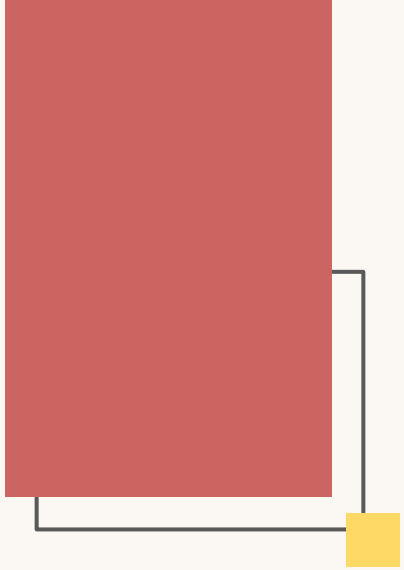
Table 2. Comparison of combined mean differences using change score and follow-up score (treatment vs. control) based on PL and DL methods

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%

- The combined mean difference using the **change score** consistently showed a larger intervention effect
- The combined mean difference using the **follow-up** score produced an intervention effect about **20% smaller** on average. (5 to >40%)

Meta-analysis of intervention vs control

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



Meta-analysis of comparison between interventions

Meta-analysis of comparison between interventions

- 53 meta-analysis from 3 CERs compared outcomes between **active interventions**
  - Comparing the various oral diabetes medications for adults with type 2 diabetes on weight, A1C and lipid variables
  - Sleepiness and depression variables
- These meta-analyses included
  - 3 to 17 studies
  - 215 to 3,252 patients

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

DL method

- **DL method:** 4 of 53 meta-analyses (7.5%) showing **discrepancy in conclusion**

3 showed significant results when using **change score**

- Significant **baseline difference** in 6 meta-analyses
  - 3 meta-analyses showed both significant baseline difference and discrepancy in conclusion

PL method

- **PL method**: 8 of 53 meta-analyses (15.1%) showing **discrepancy in conclusion**

6 showed significant results when using **change score**

- Significant **baseline difference** in 2 meta-analyses
  - 1 meta-analyses showed both significant baseline difference and discrepancy in conclusion

Heterogeneity

- For most meta-analyses, the magnitude of heterogeneity and 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

Discussion

Discrepancy in conclusion

- This is the first empirical evaluation of how using the **change** vs the **follow-up score** would affect the combined mean differences.

- Discrepancy in conclusion was shown
  - 7.9% using DL method
  - 14.3% using PL method

Mean difference estimate method

- Although the conclusion were consistent in most cases, it was concerning that up to 14.3% of results were not.

- Such discrepancy emphasize the need for **careful selection of mean difference estimates.**

Change score does not account for baseline difference

- The common misconception was that the using change score accounts for baseline difference.
- 5 of 6 CERs actually used the change score to estimate the mean difference, although it does not address the issue.

Results on discrepancy in conclusion

- Results on **discrepancy in conclusion** based on the
  - **ML method** were similar to the **DL method**,
  - **REML method** were similar to the **PL method**.

ANCOVA

- In the presence of baseline imbalance, the ANCOVA estimates should be used in meta-analysis whenever reported.

- Unfortunately, in this study, they found that the reporting of ANCOVA estimates has been poor, although many trials used an ANCOVA model in their analysis.

Limitation

- The choice of outcomes was limited by the outcomes studies in CERs, and by their focus on evaluating outcomes measure on the same scale using mean difference

- This would affect the generalizability of the results

Limitation

- Only RCTs were included in this study
- Other designs, such as cohort or other observational studies were not considered

Conclusion

Conclusion

- 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 **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.

Thanks!

Do you have any questions?