Published in final edited form as:

Epidemiology. 2018 November ; 29(6): 821–824. doi:10.1097/EDE.0000000000000857.

Performance of Between-Study Heterogeneity Measures in the Cochrane Library

Xiaoyue Ma^{a,^}, Lifeng Lin^{b,^}, Zhiyong Qu^c, Motao Zhu^d, and Haitao Chu^{e,*}

^aDepartment of Anesthesiology, Columbia University College of Physicians and Surgeons, New York, NY, 10032

^bDepartment of Statistics, Florida State University, Tallahassee, FL, 32306

^cSchool of Social Development and Public Policy, Beijing Normal University, Beijing, China, 100875

^dCenter for Injury Research and Policy, The Research Institute at Nationwide Children's Hospital, Department of Pediatrics and Division of Epidemiology, Ohio State University, Columbus, Ohio 43205

^eDivision of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, 55414

Summary

The growth in comparative effectiveness research and evidence-based medicine has increased attention to systematic reviews and meta-analyses. Meta-analysis synthesizes and contrasts evidence from multiple independent studies to improve statistical efficiency and reduce bias. Assessing heterogeneity is critical for performing a meta-analysis and interpreting results. As a widely used heterogeneity measure, the I^2 statistic quantifies the proportion of total variation across studies that is due to real differences in effect size. The presence of outlying studies can seriously exaggerate the I^2 statistic. Two alternative heterogeneity measures, the I_r^2 and I_m^2 , have been recently proposed to reduce the impact of outlying studies. To evaluate these measures' performance empirically, we applied them to 20,599 meta-analyses in the Cochrane Library. We found that the I_r^2 and I_m^2 have strong agreement with the I^2 , while they are more robust than the I^2 when outlying studies appear.

Keywords

Cochrane Library; heterogeneity; I^2 statistic; meta-analysis; outlying studies

*Corresponding author: chux0051@umn.edu.

[^]Ma and Lin have contributed equally to this paper

INTRODUCTION

Meta-analysis combines and contrasts multiple independent studies into a form of evidence for underpinning guidelines, patient decision aids, and other healthcare products.¹ The collected studies are heterogeneous if each study's underlying effect differs; otherwise, they are homogeneous. Heterogeneity may be caused by differences in study populations or methods used to conduct the studies. A random-effects model is customarily used for heterogeneous studies, while a fixed-effect model is commonly used for homogeneous studies.^{2,3} Because different models may lead to different conclusions, assessing heterogeneity is critical in meta-analyses.⁴ In addition, the perception of heterogeneity or homogeneity helps clinicians make important decisions such as whether a treatment is applicable to all patients.⁵

A classic statistical test for heterogeneity is the Q test, whose statistic is the sum of study-specific weighted squared deviations.⁶ However, its power is low in many situations.^{7,8} In addition to examining whether studies are heterogeneous, meta-analysts are more interested in interpreting the source of heterogeneity. Accordingly, several measures (e.g., $\hat{R}_I^2$, $\hat{P}^2$, and $\hat{R}_b^2$) have been proposed to quantify the proportion of total variation between studies that is due to heterogeneity rather than within-study sampling error.^{9–12} Because of their intuitive interpretations, these measures have been widely used in meta-analyses to assess heterogeneity.⁴ This article focuses on the $\hat{P}^2$ statistic because it is the most popular measure so far.

Outlying studies can have great impact on conventional methods for heterogeneity, including the $\hat{P}^2$, and on the conclusions of a meta-analysis.¹³ Outlying studies are included in a meta-analysis by inappropriate study selection, unit-of-analysis error, and typos in editorial process. Hedges and Olkin¹⁴ and Viechtbauer and Cheung¹⁵ introduced outlier detection methods under the fixed-effect and random-effects settings, respectively. Even if outlying studies are identified, excluding them is unwise as it may introduce bias, unless these studies are evidently erroneous.¹⁶ Outliers may yield overestimated heterogeneity, which in turn will affect procedures to detect them: if we use the above two methods under inappropriate setting, outlying studies may be masked or over-detected.

To reduce the impact of outliers, two alternative heterogeneity measures I_r^2 and I_m^2 were recently proposed using the weighted absolute deviations instead of the weighted squared deviations as in the $\hat{P}^2$ statistic.¹³ Specifically, the I_r^2 uses the overall meta-analytic mean estimate, while I_m^2 uses the overall median estimate; see eAppendix A for details. Lin et al.¹³ studies the large-sample properties of I_r^2 and I_m^2 , and investigated their false positive rates and statistical power using simulations. Specifically, the type I error rate of the $\hat{P}^2$ statistic is seriously inflated in the presence of outliers, while those of the I_r^2 and I_m^2 are lower than that of the $\hat{P}^2$. However, simulations cannot reproduce the exact mechanism that creates outliers in real meta-analyses.

This article compares the three heterogeneity measures (i.e., $\hat{P}^2$, I_r^2 , and I_m^2) using published meta-analyses in the Cochrane Library, which is a leading resource for systematic reviews in healthcare. We conduct pairwise comparisons among the three heterogeneity measures first; then compare the changes due to removing outliers. In addition, eAppendix F provides a case study for detailed illustration. These findings will guide researchers in properly assessing outliers and heterogeneity for systematic reviews.

METHODS

We used a data scraper tool¹⁷ to download the data in the Cochrane Library available up to January 2016, and collected 5,677 systematic reviews containing over 180,000 meta-analyses. We focused on 20,599 meta-analyses with continuous or dichotomous outcomes containing at least 5 studies. Specifically, 14,666 meta-analyses from 1,853 systematic reviews have 5–9 studies and 5,933 meta-analyses from 757 systematic reviews have at least 10 studies. The odds ratio is used as the measure of effect size for dichotomous outcomes. eFigure 1 presents the flow chart of data selection.

Pairwise agreement among $\hat{P}^2$, I_r^2 , and I_m^2 was assessed using the Bland–Altman plot¹⁸, which presents the difference between two heterogeneity measures for each meta-analysis against their average. Clustered points around the overall mean difference with most points within the limits of agreement (mean difference $\pm$ 1.96 times its standard deviation) suggest strong agreement between the two measures. A large proportion of points outside the limits indicates poor agreement.

Moreover, potential outliers were detected using the methods under both fixed-effect and random-effects settings.^{14,15} A study was considered outlying if its absolute standardized residual was greater than three. The heterogeneity measures' robustness was assessed by comparing the changes caused by removing outliers.

RESULTS

Assessing agreement

For the 14,666 meta-analyses with 5–9 studies, 42.1%, 37.6%, and 49.0% were truncated at zero for the $\hat{P}^2$, I_r^2 , and I_m^2 respectively. For the 5,933 meta-analyses with at least 10 studies, these proportions were 30.5%, 28.5%, and 33.5%. eFigure 2 plots pairwise comparisons among the $\hat{P}^2$, I_r^2 , and I_m^2 . Although the heterogeneity measures were generally close to each other, some meta-analyses yielded dramatically different measures. The Bland-Altman plots in Figure 1 further demonstrate the measures' agreement. Among the meta-analyses with 5–9 studies, the proportions of data points within the limits of agreement were 92.7% for I_r^2 vs. $\hat{P}^2$, 93.2% for I_m^2 vs. $\hat{P}^2$, and 94.9% for I_r^2 vs. I_m^2 . Among those with at least 10 studies, these proportions were 93.2%, 93.4%, and 94.3%. The nonparametric local regression fit curves in Figure 1 reflect the change of agreement between the paired measures as a function of their average. The agreement was relatively low when the average measure was between 0.2 and

0.4; otherwise, the agreement was generally excellent. Taking a closer look at the comparisons between the heterogeneity measures, $I_r^2 > I^2$, $I_m^2 > I^2$, and $I_m^2 > I_r^2$ in 46.5%, 13.9%, and 2.77% of the meta-analyses with 5–9 studies, while $I_r^2 > I^2$, $I_m^2 > I^2$, and $I_m^2 > I_r^2$ in 17.2%, 45.1%, 59.8% of those meta-analyses. For meta-analyses with at least 10 studies, 42.0%, 23.2%, and 8.04% of them yielded $I_r^2 > I^2$, $I_m^2 > I^2$, and $I_m^2 > I_r^2$, while 31.3%, 47.9%, and 63.7% of them yielded $I_r^2 > I^2$, $I_m^2 > I^2$, and $I_m^2 > I_r^2$. Overall, Figure 1 implies the rough pattern of $I_r^2 > I^2 > I_m^2$.

Moreover, Figure 1 indicates that meta-analyses with more studies had smaller variations in the Bland–Altman plots. The heterogeneity measures also had better agreement in larger meta-analyses, because the local regression curves in the lower panels were closer to zero than those in the upper panels.

The impact of outlying studies

We used meta-analyses with at least one outlying study to empirically compare the heterogeneity measures' robustness. Using the outlier detection method under the random-effects setting,¹⁵ we identified 1,273 meta-analyses with 5–9 studies and 1,054 meta-analyses with at least 10 studies that had at least one outlying study; 97% of them contained only one outlier. eFigure 3 compares the changes of $\hat{P}$, I_r^2 , and I_m^2 caused by removing the identified outlying studies. Most changes were negative because removing outliers likely reduced heterogeneity. The I_r^2 was less impacted by outliers than $\hat{P}$, because its absolute change was smaller than the $\hat{P}$ in 72.5% and 78.7% of the meta-analyses in eFigures 3(a) and 3(d), respectively. In eFigures 3(b) and 3(e), 66.2% and 74.8% of the meta-analyses had smaller absolute changes in I_m^2 than $\hat{P}$, so I_m^2 was also less impacted by outliers than $\hat{P}$. Moreover, I_m^2 had similar robustness as I_r^2 from eFigures 3(c) and 3(f): 49.3% and 43.5% of the meta-analyses had smaller absolute changes in I_m^2 than I_r^2 .

Figure 2 shows the relationship between the mean changes of the heterogeneity measures due to removing outliers and the number of studies. The mean changes of I_r^2 and I_m^2 were consistently smaller than the $\hat{P}$, further reflecting their better robustness. The mean change of I_m^2 was slightly smaller than that of I_r^2 for most sizes of meta-analyses. As the number of studies in a meta-analysis increase, removing outliers has smaller influence on the heterogeneity measures. eAppendix E presents the results using the outlier detection method under the fixed-effect setting¹⁴, which lead to similar conclusions.

DISCUSSION

Using the meta-analyses from the Cochrane Library, this article has empirically shown strong agreement among the $\hat{P}$, I_r^2 , and I_m^2 statistics in measuring heterogeneity and has demonstrated the robustness of I_r^2 and I_m^2 in the presence of outliers. The robustness of I_r^2

does not mean that it is more conservative than the $\hat{P}$ and underestimates heterogeneity. In fact, I_r^2 is smaller than $\hat{P}$ in only 17.2% and 31.3% of the meta-analyses with 5–9 studies and with at least 10 studies, respectively.

Recall that I_m^2 is based on the overall meta-analytic weighted median estimate, instead of the overall mean used in I_r^2 , so ideally I_m^2 may be even more robust than I_r^2 . Due to the intractable finite-sample properties of weighted median, it is infeasible to theoretically derive a relationship between I_r^2 and I_m^2 . This article has provided an empirical comparison between them and has shown that I_m^2 did not dramatically outperform I_r^2 ; also, I_m^2 likely had the smallest value among the three measures. Therefore, despite of its robustness to outliers, researchers may carefully examine if heterogeneity is underestimated when using I_m^2 .

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Professor James Hodges for helpful comments and discussion that led to a better presentation of this article. This research was supported in part by NIDCR R03 DE024750 (HC), NLM R21 LM012197 (HC), NIDDK U01 DK106786 (HC), and the Doctoral Dissertation Fellowship from the University of Minnesota Graduate School (LL).

References

1. Egger M, Davey Smith G, Altman DG. Systematic Reviews in Health Care: Meta-Analysis in Context. 2. London, UK: John Wiley & Sons; 2008.
2. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*. 2010; 1(2):97–111. [PubMed: 26061376]
3. Riley RD, Higgins JPT, Deeks JJ. Interpretation of random effects meta-analyses. *BMJ*. 2011; 342:d549. [PubMed: 21310794]
4. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *BMJ*. 2003; 327(7414):557–560. [PubMed: 12958120]
5. Ioannidis JPA, Patsopoulos NA, Evangelou E. Uncertainty in heterogeneity estimates in meta-analyses. *BMJ*. 2007; 335(7626):914–916. [PubMed: 17974687]
6. Cochran WG. The combination of estimates from different experiments. *Biometrics*. 1954; 10(1): 101–129.
7. Jackson D. The power of the standard test for the presence of heterogeneity in meta-analysis. *Statistics in Medicine*. 2006; 25(15):2688–2699. [PubMed: 16374903]
8. Hardy RJ, Thompson SG. Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*. 1998; 17(8):841–856. [PubMed: 9595615]
9. Takkouche B, Cadarso-Suárez C, Spiegelman D. Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology*. 1999; 150(2):206–215. [PubMed: 10412966]
10. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*. 2002; 21(11):1539–1558. [PubMed: 12111919]

11. Takkouche B, Khudyakov P, Costa-Bouzas J, Spiegelman D. Confidence intervals for heterogeneity measures in meta-analysis. *American Journal of Epidemiology*. 2013; 178(6):993–1004. [PubMed: 23921232]
12. Crippa A, Khudyakov P, Wang M, Orsini N, Spiegelman D. A new measure of between-studies heterogeneity in meta-analysis. *Statistics in Medicine*. 2016; 35(21):3661–3675. [PubMed: 27161124]
13. Lin L, Chu H, Hodges JS. Alternative measures of between-study heterogeneity in meta-analysis: reducing the impact of outlying studies. *Biometrics*. 2017; 73(1):156–166. [PubMed: 27167143]
14. Hedges LV, Olkin I. *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press; 1985.
15. Viechtbauer W, Cheung MWL. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*. 2010; 1(2):112–125. [PubMed: 26061377]
16. Barnett V, Lewis T. *Outliers in Statistical Data*. 3. Chichester, UK: John Wiley & Sons; 1994.
17. Springate DA, Kontopantelis E. *Cochrane_scraper: tools for downloading data from the Cochrane Library of Systematic Reviews*. Version 1.1.0. 2014
18. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*. 1986; 327(8476):307–310.

Figure 1.

Bland-Altman plots of the I^2 , I_r^2 , and I_m^2 for meta-analyses with 5–9 studies (upper panels) and for meta-analyses with at least 10 studies (lower panels). In each panel, the horizontal solid line is zero, representing no difference, the dashed line is the overall mean difference, and the two dotted lines represent the lower and upper limits of agreement. The red solid curve is the nonparametric local regression fit curve, which shows the change in agreement between two measures as a function of their average.

Figure 2.

Mean changes of I^2 , I_r^2 , and I_m^2 due to removing outliers identified using the random-effects model, categorized by the number of collected studies in a meta-analysis.