Journal of Clinical Epidemiology 130 (2021) 149−151

Journal of Clinical Epidemiology

KEY CONCEPTS SERIES

Key concepts in clinical epidemiology: detecting and dealing with heterogeneity in meta-analyses

Cynthia P. Cordero[a,*], Antonio L. Dans[a,b]

[a]Department of Clinical Epidemiology, College of Medicine University of the Philippines, Manila, Philippines
[b]Department of Medicine College of Medicine University of the Philippines Manila-Philippine General Hospital, Manila, Philippines

Accepted 29 September 2020; Published online xxxx

Abstract

In a meta-analysis, a question always arises. Is it worthwhile to combine estimates from studies of different populations using various formulations of an intervention, evaluating outcomes measured differently? Sometimes even study designs differ. Differences are expected in a meta-analysis. These may be negligible, and a pooled estimate of effect can guide the clinical decision. However, when the differences are large, this estimate may mislead. Effect estimates from study to study differ because of real differences (between-study variability) and because of chance (within-study variability). To combine estimates when there is heterogeneity (between-study differences are large) may not be sensible. Two complementary methods may be used to detect heterogeneity: visual inspection of the forest plot and calculating numerical measures of heterogeneity (I^2 and Q). Visual inspection can show effects that are different from the rest. A large I^2 (proportion of overall variability attributed to between-study variation) or a small *P*-value associated with Q may suggest heterogeneity. Large *P*-values, however, do not mean the absence of heterogeneity. It is more informative to report the confidence interval of the I^2. If there is no heterogeneity, a pooled estimate of the true effect may be generated using only within-study variation (fixed-effect model). If there is substantial heterogeneity, reasons should be sought. Subgroup analysis or meta-regression using study-level characteristics may be done. Although more involved and potentially challenging, individual-level data (Individual Participant Data, IPD) may also be used. In the case of unexplained heterogeneity, both within- and between-study variation should be used to generate a pooled estimate (random-effects model). This estimate does not estimate a single true effect but estimates the average of a range of effects of the intervention on populations represented by the studies. If precise enough (narrow confidence interval), this estimate, together with the prediction interval (a measure of uncertainty in the effect one might see in a particular context), can guide clinical and policy decisions.

Keywords: Meta-analyses; Heterogeneity; I^2; Q-statistic; Subgroup analyses; Meta-regression; Individual participant data; Prediction interval; Fixed-effect model; Random-effects model

1. Introduction

A meta-analysis combines estimates from several studies to generate pooled estimates of the effect of a treatment or exposure on an outcome. The danger of reporting pooled estimates is that readers may overlook the overall picture—some studies having bigger effects than the other studies, some effects with different directions (harm) from the benefit shown by most studies. The variation may be due to subtle but important differences in the distributions of effect modifiers in the populations, the interventions compared, the observed outcome, or study design. For example, a meta-analysis comparing zinc supplementation with placebo to treat diarrhea in children showed a benefit of reducing the duration of diarrhea by 13 hours. Subgroup analysis showed a benefit for children older than 6 months but no benefit to possible harm for younger children [1].

Declaration of No conflict of interest: Authors Cynthia Cordero and Antonio Dans do not have any conflict of interest related to this manuscript.

Authors' statement: As authors of this article, we both conceived, developed, and finalized the manuscript. We thank our students and colleagues—they are the inspiration of this article, which is meant to guide readers on what to look for in a meta-analysis as a user and what to report as a doer. We thank our reviewers for their valuable comments, which we think made the article more informative and relevant. The process from conceptualization to the current version of the article required hours of literature review and drawing from our collective research and teaching experience. No funding or monetary support was needed. The article is the authors' original work, has not been previously published, and is not under consideration for publication elsewhere.

* Corresponding author. Department of Clinical Epidemiology, College of Medicine, University of the Philippines Manila, Room 103, Paz Mendoza Building 1000, Manila, Philippines. Tel.: 63-2-85254098.

E-mail address: cpcordero@up.edu.ph (C.P. Cordero).

https://doi.org/10.1016/j.jclinepi.2020.09.045

These findings are among the basis for the WHO recommendation for diarrhea management in children: ''Mothers, other caregivers, and health workers should provide children with 20 mg per day of zinc supplementation for 10–14 days (10 mg per day for infants under the age of 6 months).'' [2]. A local clinical practice guidelines recommend zinc supplementation (20 mg/day for 10–14 days) as adjunctive therapy for acute infectious diarrhea in children older than 6 months but not for children less than 6 months [3]. To avoid missing out on such important distinctions, researchers (and readers) must be alert to these differences.

2. Detecting heterogeneity

Differences in results are inevitable by chance alone and may be too small to consider. The key decision is when to worry. Several strategies have been used to assess heterogeneity. The simplest is an inspection of the forest plot for bigger or smaller estimates compared to the rest of the estimates that vary in direction (harm to benefit). It is also useful to check if the interval estimates have poor overlap (e.g., two or more estimates do not overlap), suggesting heterogeneity [4].

Sometimes differences may be hard to detect by visual inspection. Hence it is advisable to complement visual inspection with numerical measures of heterogeneity. One such measure is the Q statistic. Q is not interpreted as a measure of heterogeneity by itself. It is used to test the hypothesis that there is no heterogeneity, and *P*-values are reported. While this may be done, it should be noted that large *P*-values should not be interpreted as the absence of heterogeneity [4]. This test suffers from low power, especially when there are few studies, a common situation in a meta-analysis.

By a simple formula, another measure of heterogeneity can be derived based on the Q statistic—I^2. I^2 is easy to understand. A 95% confidence interval (CI) of the pooled estimate can become wide for two reasons: (1) the studies may have wide 95% CIs (within-study variability, Fig. 1A) and/or (2) the studies may vary greatly from one another (between-study variability, Fig. 1B). When both within-study and between-study variability are small, the 95% CI of the pooled estimate will be very narrow (Fig. 1C). I^2 describes the magnitude of heterogeneity as the proportion of the total variability (between + within) that is attributable to between-study variability. Cut-offs have been proposed but noted to be not applicable in all circumstances [4,5]. However, there is an agreement in the importance of reporting confidence intervals of I^2 for proper interpretation [5.6].

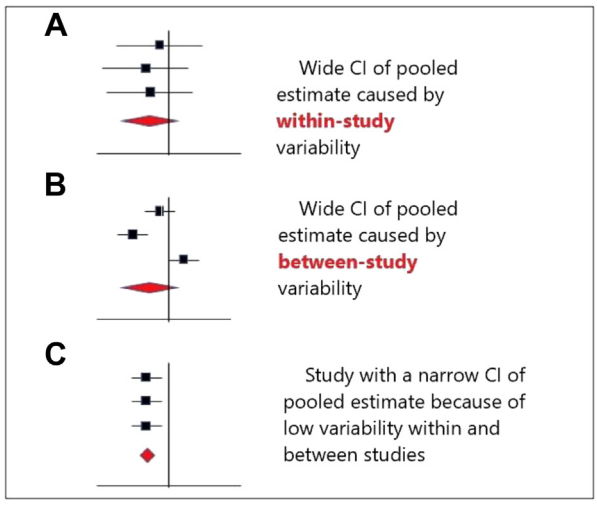

Fig. 1. Effect of within-study and between-study variability on the Confidence Interval of pooled estimates.

Whatever numerical measure of heterogeneity is used, this should be interpreted together with a visual inspection. Numerical measures should not replace visual inspection but should complement it.

3. Dealing with heterogeneity

When heterogeneity is low, the pooled estimate can consider only within-study variation (fixed-effect model). If heterogeneity is detected, an explanation should be sought. The populations studied may be different (varying ages and nutritional status in the zinc example), the interventions may not be exactly the same (zinc varies in dose and formulation), the outcome measures may have been defined differently (different standardization of pre-weighted disposable diapers to measure stool output), or the study design may have been dissimilar (randomized by individuals or by clusters) [1,4]. Subgroup analysis or meta-regression using study-level characteristics may be done. In the zinc study, for example, the studies were grouped according to their target population: age < 6 months, age >6 months, and age both < and >6 months [1]. This grouping highlights the limitation of using study-level characteristics. Children aged 6 months do not belong in any of the groups. The third group was needed to classify studies, which included both age groups. Individual-level characteristics (Individual Participant Data) may be used instead. Participants are more accurately classified into groups. Quantitative characteristic such as age is maximized when used in its original measurement. However, this method is intensive and potentially more challenging and requires a great deal of coordination with authors and their institutions [7].

If unexplained heterogeneity remains, the pooled estimate should include between- and within-study variability (random-effects model) [4]. The fixed-effect model assumes that there is a single unknown effect. The studies in a meta-analysis provide estimates of this single effect, which are pooled to yield a more precise estimate. The random-effects model postulates that there is variation in the effects of the intervention and what it estimates is the average of these effects among the populations represented by the

studies [4]. Given that we are dealing with a range of effects, it is intuitive to report the average. But this is an incomplete picture. Users should also know the extent of the spread of the effects. Prediction Intervals, although not commonly used, when reported together with the estimate of the average effect, can better guide clinical and policy decisions [4,8].

The Prediction Interval (PI) is different from the Confidence Interval (CI). CI is a measure of the precision of the estimate of the average effect in a range of contexts. PI is a measure of the uncertainty in the effect one might see in one particular context, taking into account between-study variation.

4. Conclusion

Forest plot showing estimates that are too far apart with a poor overlap of CIs suggests heterogeneity. This finding should be complemented with numerical measures of heterogeneity—I^2 and Q. I^2 should always be reported with its confidence interval. Large *P*-values associated with Q do not necessarily mean the absence of heterogeneity. When heterogeneity is detected, subgroup analysis or meta-regression may be done using study-level characteristics. Although intensive, Individual-level analysis (IPD) can be a better alternative. When unexplained heterogeneity remains, the random-effects model may be used to estimate the average effect. If precise enough (narrow confidence interval), this estimate, together with the prediction interval, can guide clinical and policy decisions. When heterogeneity is very large, average estimates can mislead instead of inform.

References

[1] Lazzerini M, Wanzira H. Oral zinc for treating diarrhea in children. Cochrane Database Syst Rev 2016;12:CD005436.
[2] World Health Organization. Zinc supplementation in the management of diarrhea. e-Library of Evidence for Nutrition Actions (eLENA). Available https://www.who.int/elena/titles/zinc_diarrhoea/en/. Accessed May 15, 2020.
[3] Department of Health Republic of the Philippines, San Lazaro Hospital and the Philippine Society of Microbiology and Infectious Diseases. Philippine Clinical Practice Guidelines on the Management of Acute Infectious Diarrhea in Children and Adults-A Pocket Guide. Manila, Philippines. 2019. Available https://www.doh.gov.ph/sites/default/files/publications/CPG%20AID_pocket%20guide.v7.pdf. Accessed September 23, 2020.
[4] Analyzing data and undertaking meta-analyses: heterogeneity. In: Deeks JJ, Higgins JPT, Altman DG, Green S, editors. Cochrane handbook for systematic reviews of interventions. Version 5.1.0. Chichester, West Sussex; Hoboken NJ: John Wiley & Sons; 2008.
[5] Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ 2003;327:557−60.
[6] Loannidis J, Patsopaulos N, Evangelou E. Uncertainty in heterogeneity estimate in meta-analyses. BMJ 2007;335:914−6.
[7] Riley RD, Lambert PC, Abo-Said G. Meta-analysis of individual-participant data: rationale, conduct and reporting. BMJ 2010;340:c221.
[8] Borenstein M, Higgins JPT, Hedges LV, Rothstein HR. Basics of meta-analysis: I^2 is not an absolute measure of heterogeneity. Res Syn Meth 2017;8(1):5−18.