MedCrave
Step into the World of Research

Biometrics & Biostatistics International Journal

Editorial

Open Access

CrossMark

Statistical harmonization methods in individual participants data meta-analysis are highly needed

Volume 3 Issue 3 - 2016

E R van den Heuvel,[1] L E Griffith[2]

[1]Department of Stochastics Biostatistics, Eindhoven University of Technology, Netherlands

[2]Department of Clinical Epidemiology, McMaster University, Canada

Correspondence: E R van den Heuvel, Department of Stochastics Biostatistics, Faculty of Mathematics and Computer Science, Eindhoven University of Technology, Email e.r.v.d.heuvel@tue.nl

Received: January 28, 2016 | **Published:** February 01, 2016

Editorial

Meta-analysis has a long history within the medical sciences and epidemiology.[1-3] The main goal of a meta-analysis is to improve the precision of a specific effect size of a treatment or an exposure on a clinical or disease outcome by pooling or combining multiple studies. It is frequently conducted within a systematic review of the scientific literature to guarantee that studies with appropriate information are not ignored or overlooked and to make sure that the pooled estimate represents an unbiased and precise estimate of the true effect size. Moreover, the pooled effect estimate is evaluated in the context of study heterogeneity. In case substantial heterogeneity in the effect sizes across studies is present, the pooled estimate is considered less reliable or even questionable. Thus not only the pooled estimate must be unbiased, also the estimate of the measure of heterogeneity should be correct. Within the area of medical statistics, meta-analysis has evolved into an important and rich research field.[4,5]

Meta-analysis can roughly be divided into aggregate data (AD) meta-analysis and individual participant data (IPD) meta-analysis. AD meta-analysis is the traditional form of meta-analysis and fully focuses on pooling effect sizes that are formulated at the study level, i.e. this analysis represents an analysis of analyses.[6] AD meta-analysis typically combines reported summary statistics from published articles. IPD meta-analysis considers the full data sets from the included studies and analyzes these data sets either as one large data set (one-stage IPD) or in two steps (two-stage IPD). In this two-stage approach the same statistical model is typically fitted to each study separately before the estimated summary statistics or effect sizes from these models are combined using methods from the AD meta-analysis. IPD meta-analysis is considered more appropriate than AD meta-analysis,[7-10] in particular when observational studies are combined or pooled, since it provides the opportunity to correct for (the same) potential confounders at the individual level.

The IPD meta-analysis for pooling observational cohort studies, in particular when a one-stage analysis is applied, has triggered an important issue that was not recognized with traditional AD meta-analysis in the past or at least it was not considered relevant. The issue is that outcome and/or risk variables can be measured with different instruments across studies. For instance, memory can be measured with the Rey Auditory Verbal Learning Test[11] or the Buschke Cued Recall Procedure.[12] Both approaches record a discrete or integer score per individual that indicates how many items have been correctly remembered from a set of tested items, but this does not imply at all that these tests perfectly measure the same construct of memory, since one test is visually while the other is verbally. Other examples that frequently provide differences between studies are variables like education and income. School systems and teaching content are quite different across countries and are therefore difficult to compare when the IPD meta-analysis includes international studies. This is also true when just the number of years in education is used. Income can differ due to time differences among studies (i.e. due to inflation), in their monetary unit (e.g. dollars versus euros), and if salaries across occupations are not valued in the same way across countries. Even if variables are identically measured across studies, they may not be recorded in the same way, since variables can be observed either numerically or categorically.

Although substantial emphasize has been given to the way that a meta-analysis should be analyzed and reported,[13,14] issues regarding this lack of content equivalence of variables across studies have received little attention in the statistical literature. One reason is that AD meta-analysis and two-stage IPD meta-analysis may simply overlook the lack of content equivalence, since the statistical analysis pools only the study-level summary statistics. In an AD meta-analysis, the selected articles may not have reported the details on the measurement instruments for all included variables. Indeed, smoking (yes/no) does not tell us the time frame of smoking nor what is being smoked. When data from multiple studies is not allowed to be shared with others, a two-stage IPD meta-analysis is executed by the researchers at the location of the study using the same set of statistical codes (this is called a federated data meta-analysis). The collected variables, on for instance memory, income, and education, may just be put in the statistical analysis of the study without giving it much thought, whether they are identically measured or whether they represent something different across studies. Even if lack of content equivalence is recognized, researchers may argue that these non-identically measured variables may still be strongly correlated, indicating that it is okay to use them across studies as if they would have been measured with the same instruments. This can however never be the sole argument, since many variables may be (strongly) correlated, but this does not mean that we can interchange them for a statistical analysis or any other purpose for that matter.

In case the multi-study data is pooled at one location, differences in variables, that should have been content equivalent, would become more easily apparent. For instance, if memory is measured with a 15-item or 12-items test, this may easily be detected in an explorative data analysis. Then some kind of data manipulation is conducted to align the scales of the variables before they can be reported into one variable memory. Adjusting the scales though would make the range identical, but it does not change the difference in resolution. Additionally, if the instruments for memory are also different, adjustments of the scales alone might not be satisfactory enough whenever the tests do not capture the exact same construct. To make variables content equivalent, it would be necessary to converse each value on an

individual observed with one instrument to an equivalent value with the other instrument for the exact same individual. This implies a precise conversion or calibration model that could interchange values from different instruments per individual. This process or activity is what we call statistical harmonization of variables.

Calibration of test forms (e.g. IQ tests, candidate assessment tests) has been the subject domain of psychometricians for a long time and they have called this discipline "test equating and linking".[15] However, many of the suggested statistical approaches to create equivalent test forms have focused on methods that are independent of subject characteristics, typically referred to as "population invariance". The basic idea is that psychometric tests are measurement invariant, which means that they capture the true construct within a varying and diverse population.[16] Indeed, differences in scores among subgroups quantifies differences in performance and not differences in interpretation. Under population invariance, researchers will map the distribution of test scores from one test form onto the distribution of test scores from another test form using sampling data from the population. Typical statistical approaches are location-scale and quantile normalization, but in the case that invariance is violated, there is no real solution anymore to test equating and linking. Our view is that lack of invariance suggests the need for subject-specific calibration models, i.e. calibration models that can correct for subject-specific variables.[17,18]

IPD meta-analysis within the medical sciences, has used subject-specific normalization based methods to make variables content equivalent across studies,[19] but little is known about the performance of how well these statistical methods harmonize non-equivalent variables across studies for IPD meta-analysis.[20] Approaches that have been considered are algorithmic procedures, subgroup normalization, linear regression, and item response theory (IRT). Algorithmic procedures map the variable into a set of categories using specific thresholds that may depend on covariates like age, gender, and education.[21] For subgroup normalization a well-defined subgroup or control group that exists in each study is selected and the values of the variables are normalized per study with the mean and standard deviation of the selected subgroup.[22] For linear regression the variable is regressed on covariates to eliminate their influence on the variable and to generate covariate independent residuals per study. These residuals are then normalized using simple methods from equating and linking.[22] IRT models typically harmonize the latent ability of individuals by using bridging items, i.e. items that are measured in multiple studies and that can connect all the studies together, assuming that bridging items across studies are identically measured.[23,24]

To push the research agenda on harmonization of variables in meta-analysis forward, we believe that new and innovative statistical methods for harmonization, extending the existing methods, are needed that would incorporate participants characteristics. Moreover, these methods should be compared on real data and in simulation studies to judge which approach works best and under what conditions. For instance, methods may want to make a distinction between harmonization of outcomes, exposures of interest, or potential confounders. Additionally, the methods should not just investigate whether effect sizes of interest are unbiasedly estimated, but also how measures of study heterogeneity are affected by harmonization. The only form of heterogeneity that should be captured in a meta-analysis is the heterogeneity due to populations, since harmonization should eliminate the inconsistencies in measuring variables. We also believe that it would be highly beneficial if separate sampling studies are performed to quantify calibration models for complex measurements across subgroups of participants. This should lead to accepted and standardized conversion procedures, similar to the work that has been conducted in metrology on physics measurements in the past (1 inch = 2.54 cm and 1 kg = 2.2046 lbs). Finally, we strongly recommend that papers on meta-analysis should mention the lack of content equivalence for their selected set of variables in detail and report the approach that they have used to accommodate the issue in the analysis.

Acknowledgement

There are no acknowledgements and funding.

Conflict of interest

No conflict of interest

References

1. Jones DR. Meta-analysis: weighing the evidence. *Statistics in Medicine*. 1995;14(2):137–149.
2. Kulinskaya E, Morgenthaler S, Staudte RG. Combining statistical evidence. *International Statistical Review*. 2014;82(2):214–242.
3. Hedges LV. The early history of meta-analysis research synthesis methods. 2015;6(3):284–286.
4. Van Houwelingen HC, Arends LR, Stijnen T.Tutorials in biostatistics Advanced methods in meta-analysis: multivariate approach and meta-regression. *Statistics in Medicine*. 2002;21:589–624.
5. Brockwell SE, Gordon IR. A comparison of statistical methods for meta-analysis. *Stat Med*. 2001;20(6):825–840.
6. Glass GV. Primary, secondary, and meta-analysis of research. *Educational Researcher*. 1976;5(10):3–8.
7. Stewart LA, Parmar MK. Meta-analysis of the literature or of individual patient data: is there a difference? *Lancet*. 1993; 341(8842): 418–422.
8. Houwelingen HC. The future of biostatistics: Expecting the unexpected. *Statistics of Medicine*. 1997;16(24):2773–2784.
9. Stewart LA, Tierney JF. To IPD or not to IPD? advantages and disadvantages of systematic reviews using individual participant data. *Evaluation & Health Professions*. 2002;25(1):76–97.
10. Stewart GB, Altman DG, Askie LM, et al. Statistical Analysis of Individual Participant Data Meta-Analysis: A Comparison of Methods and Recommendations for Practice. *PLoS ONE*. 2012;**7**(10): e46042.
11. Taylor EM. The appraisal of children with cerebral deficits. Harvard university press, Cambridge, USA; 1959.
12. Buschke H. Cued recall in amnesia. *J Clin Neuropsychol*. 1984;6(4):433–440.
13. Moher D, Liberati A, Tetzlaff J, et al. The PRISMA Group. Preferred reporting Items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ*. 2009;339:b2535.
14. Stewart LA, Clarke M, Rovers M, Riley RD, Simmonds M, et al. (2015) Preferred Reporting Items for Systematic Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement. JAMA 313(16): 1657-1665.
15. Kolen MJ, Brennan RL. Test equating, scaling, and linking. *Statistics for Social and Behavioral Sciences*. New York, USA; 2004.
16. Streiner DL, Norman GR. Health measurement scales: A practical guide to their development and use 4th ed, Oxford University Press, USA; 2008.

17. Dorans NJ, Holland PW. Population invariance and the equitability of tests: basic theory and the linear case. *ETS Research Report Series*. 2000;2:1–35.

18. Kolen MJ. Population invariance in equating and linking: concept and history. *Journal of Educational Measurement*. 2004;41(1):3-14.

19. Griffith LE, van den Heuvel E, Fortier I, et al. Statistical approaches to harmonize data on cognitive measures in systematic reviews are rarely reported. *Journal of Clinical Epidemiology*. 2015;68(2):154–162.

20. Griffith LE, van den Heuvel E, Raina P, et al. Comparison of standardization methods for the harmonization of phenotype data: an application to cognitive measures. *American Journal of Epidemiology*. 2015.

21. Fortier I, Burton PR, Robson PJ, et al. Quality, quantity and harmony: The DataSHaPER approach to integrating data across bioclinical studies. *Int J Epidemiol*. 2010;39(5):1383–1393.

22. Tuokko H, Woodward TS. Development and validation of a demographic correction system for neuropsychological measures used in the canadian study of health and aging. *J Clin Exp Neuropsychol*. 1996;18(4):479–616.

23. Van Buuren S, Eyres S, Tennant A, et al. Improving comparability of existing data by response conversion. *Journal of Social Statistics*. 2005;21(1):53-72.

24. Bauer DJ, Hussong AM. Psychometric approaches for developing commensurate measures across independent studies: traditional and new models. *Psychological Methods*. 2009;14(2):101–125.

Citation: Heuvel ERVD, Griffith LE. Statistical harmonization methods in individual participants data meta-analysis are highly needed. *Biom Biostat Int J.* 2016;3(3):70–72. DOI: 10.15406/bbij.2016.03.00064