JAMA Network Open

Invited Commentary | Public Health

Limitations of Meta-analyses of Studies With High Heterogeneity

Peter B. Imrey, PhD

+ Related article

Author affiliations and article information are listed at the end of this article.

Sabitova et al[1] have performed an important service in compiling and summarizing 2 decades of studies on job burnout and satisfaction among physicians and dentists in middle-income countries and a few low-income countries. The authors followed a standard approach to performing a systematic review and meta-analysis to analyze studies that assessed job-related morale among physicians and dentists working in these countries, using levels of job burnout, job satisfaction, and job motivation as indicators of job morale. Data were extracted independently by several investigators following the Meta-analysis of Observational Studies in Epidemiology (MOOSE) reporting guidelines. The authors assessed the quality of the included studies for bias and conducted random-effects meta-analyses, planned subgroup analyses, and metaregression analyses. The study included results from 79 studies with 45 714 participants. The authors reported that, in their analysis of data from 21 studies including 9092 physicians and dentists, 32% of participants, who worked mainly in middle-income countries, exceeded the high threshold for job burnout, and in their analysis of 20 studies including 14 113 participants, 60% were satisfied with their jobs overall. Available data were insufficient for the authors to conduct an analysis of job motivation. The authors' meticulously documented structured literature review will be useful to health services researchers and policy makers. However, as the authors acknowledge, the meta-analytic portion of their study was limited by significant heterogeneity observed across studies that could not be explained by subgroup analyses or metaregressions. Their results illustrate why rigorously conducted meta-analyses of highly heterogeneous studies may be less interpretable and useful than initially anticipated.

The technical concept of study heterogeneity is central to understanding meta-analytic results. Results of multiple studies always differ to some degree, but studies are said to be heterogeneous when their underlying target parameters differ. Evidence for heterogeneity may be based on data or design, including differences in study target populations or targeted effects, survey recruitment and administration methods, measurement instruments, doses of interventions, timing of outcome measurements, and/or analytical methods, including covariate adjustments. Meta-analyses addressing broadly framed questions or the incidence or prevalence of a phenomenon in diverse environments may assemble highly heterogeneous studies, as did the study by Sabitova et al.[1] The most popular data-based measure of study heterogeneity, the I^2 statistic, approaches its maximum of 1 when heterogeneity substantially exceeds within-study sampling and measurement variability.[2] In the study by Sabitova et al,[1] I^2 exceeds 0.95 for most analyses. This degree of heterogeneity poses interpretive challenges.[2-4]

Heterogeneity requires, and Sabitova et al[1] have used, random-effects meta-analysis.[5-7] This estimates not a single correct overall answer to the research question but a distribution of particularized, situationally correct answers from an imagined universe of individual studies that might have been performed. Studies included in the meta-analysis are treated as randomly sampled from this universe. The hypothetical underlying distribution is generally assumed to be symmetric—usually Gaussian—in which case the meta-analysis estimates this distribution's mean and SD and, based on them, provides a confidence interval for the mean. Because the confidence interval accounts for between-study as well as within-study variation, it tends to be wider than that from a fixed-effects meta-analysis, which presumes homogeneity. Meta-analytic means are weighted averages of individual study means. For fixed-effects meta-analyses, they are weighted inversely to within-study variances; for random-effects meta-analyses, they are weighted inversely to total study variation, including between-study variation. When heterogeneity is very high and between-study

variation dominates, random-effects meta-analyses weight studies nearly equally, regardless of sample sizes, yielding a meta-analytic summary close to the more easily calculated arithmetic mean of the individual study results.[2-7]

In this context, consider the first meta-analytic summary reported by Sabitova et al.[1] "In 21 studies with 9092 participants working mainly in middle-income countries, 32% (95% CI, 27%-38%; I^2 = 95.32%; P < .001) reported job burnout."[1] Although the structured literature review includes 79 studies with responses from 37 countries, this meta-analysis is confined to the 21 studies that used the dichotomized results of the Maslach Burnout Inventory, which come from Brazil (7 studies), China, Mexico, and Serbia (3 studies each), and Argentina, Bosnia and Herzegovina, Cameroon, Lebanon, and Turkey (1 study each).[1] The 32% burnout prevalence estimate is a weighted average of these studies' results, wherein the largest, a study from China with 5590 participants,[8] receives only 30% more weight than the smallest, a study from Brazil with 43 participants.[9] The 32% summary differs by only 3 percentage points from the 35% arithmetic mean prevalence of the 21 studies. Because countries were weighted virtually proportionally to the number of included studies, the estimate of burnout is approximately 76% determined by reports from only 4 countries with multiple studies (ie, by 16 studies from Brazil, China, Mexico, and Serbia) and only slightly and roughly equally influenced by single reports from 5 other countries.[1]

This 32% estimate or the 35% arithmetic mean—both approximately one-third—can be properly interpreted only as the rough center of a widely dispersed range of burnout prevalences in different countries, work environments, subpopulations, and other particulars of the studies within them. The Supplement to the main article reports study-to-study SDs of approximately 15% overall and 9% within large geographic regions.[1] If the meta-analytic assumptions are correct and these were the respective true SDs, then true target burnout prevalences for one-third of studies would be outside the range of 17% to 47% and, for 10%, outside the range of 7% to 57%. Thus, the 32% average provides only very modest guidance on what to expect for any given situation because the studies are so heterogeneous. By metaregression, Sabitova et al[1] found that differences between broad geographic regions explain some of the heterogeneity. But even within a region, true burnout may analogously be expected to depart by more than 9% from mean regional burnout for about one-third of studies and by at least 15% for 1 in 10 studies.

Moreover, examination of the metaregressions on geographic region may also limit the interpretations of this study. In all but 1 region, studies were clustered in countries and may not be regionally representative. For example, all Central American studies were from Mexico (which is usually classified as a North American country), all Eastern Asian studies were from China, 7 of 8 South American studies were from Brazil, and all Southern European studies were from 2 bordering countries, both once part of Yugoslavia. Thus, the 9% within-region SD derives primarily from within-country variation among studies in countries such as Brazil, China, Mexico, and Serbia and likely understates possible broader regional variation. While it is reasonable to group studies by region, the regional burnout comparisons are dominated by specific countries in each region and may not be broadly representative geographic comparisons.

Sabitova et al[1] acknowledge that subgroup analyses indicated that the prevalence of burnout dimensions varied depending on the country's geographical region and other factors. They report that "high levels of within-group heterogeneity and uneven covariate distribution among groups were present, demonstrating that these subgroups could not account for the variance between studies" and that the "results of subgroup analyses might be uncertain because of uneven covariate distribution among groups and an insufficient number of studies per group."[1] Moreover, although geography was found to make some contribution to study heterogeneity, the phrasing of the geographic metaregression comparisons is technically somewhat ambiguous, and the interested reader may wish to refer to the supplementary forest plots, analytic tables, and computer code for a fuller understanding.[1]

These concerns do not diminish the authors' significant accomplishment in what in many respects is an exemplary and useful systematic review. But the formal meta-analysis illustrates how

high, unexplained heterogeneity limits meta-analytic results. Contributors to this unexplained heterogeneity may include variability in social environment and conditions of survey administration, which may not be discernible from published results and which may contribute to differences in response rates. In the end, despite exhaustive analyses, precise answers to broad meta-analytic questions about subjective issues may be difficult to achieve when there are important limitations in the studies and data included in the analyses.

ARTICLE INFORMATION

Published: January 10, 2020. doi:10.1001/jamanetworkopen.2019.19325

Corresponding Author: Peter B. Imrey, PhD, Department of Quantitative Health Sciences JJN3/291, Lerner Research Institute, Cleveland Clinic, 9500 Euclid Ave, Cleveland, OH 44195 (imreyp@ccf.org).

Author Affiliation: Department of Quantitative Health Sciences, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio.

Conflict of Interest Disclosures: None reported.

REFERENCES

1. Sabitova A, McGranahan R, Altamore F, Jovanovic N, Windle E, Priebe S. Indicators associated with job morale among physicians and dentists in low-income and middle-income countries: a systematic review and meta-analysis. *JAMA Netw Open*. 2020;3(1):e1913202. doi:10.1001/jamanetworkopen.2019.13202

2. Higgins J, Thompson S, Deeks J, Altman D. Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *J Health Serv Res Policy*. 2002;7(1):51-61. doi:10.1258/1355819021927674

3. Schroll JB, Moustgaard R, Gøtzsche PC. Dealing with substantial heterogeneity in Cochrane reviews: cross-sectional study. *BMC Med Res Methodol*. 2011;11:22. doi:10.1186/1471-2288-11-22

4. Alba AC, Alexander PE, Chang J, MacIsaac J, DeFry S, Guyatt GH. High statistical heterogeneity is more frequent in meta-analysis of continuous than binary outcomes. *J Clin Epidemiol*. 2016;70:129-135.

5. Israel H, Richter RR. A guide to understanding meta-analysis. *J Orthop Sports Phys Ther*. 2011;41(7):496-504. doi:10.2519/jospt.2011.3333

6. Murad MH, Montori VM, Ioannidis JPA, et al. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. *JAMA*. 2014;312(2):171-179. doi:10.1001/jama.2014.5559

7. Serghiou S, Goodman SN. Random-effects meta-analysis: summarizing evidence with caveats. *JAMA*. 2019;321(3):301-302. doi:10.1001/jama.2018.19684

8. Zhou X, Pu J, Zhong X, et al; China Neurologist Association. Burnout, psychological morbidity, job stress, and job satisfaction in Chinese neurologists. *Neurology*. 2017;88(18):1727-1735. doi:10.1212/WNL.0000000000003883

9. Barbosa FT, Eloi RJ, Santos LM, Leão BA, Lima FJCD, Sousa-Rodrigues CF. Correlation between weekly working time and burnout syndrome among anesthesiologists of Maceió-AL. *Braz J Anesthesiol*. 2017;67(2):115-121. doi:10.1016/j.bjan.2015.06.001