European Journal of Epidemiology (2019) 34:891–892
https://doi.org/10.1007/s10654-019-00535-0

CORRESPONDENCE

Author Reply: A critical reflection on the grading of the certainty of evidence in umbrella reviews

Stefania Papatheodorou[1,2]

Received: 6 June 2019 / Accepted: 13 June 2019 / Published online: 21 June 2019

I would like to thank Dr. Schlesinger and colleagues for their critical reflection on the grading of the certainty of evidence in umbrella reviews. Even though this field is relatively new and we will learn a lot as it develops, there are certain aspects of this approach that are not as arbitrary as Schlesinger and colleagues are presenting.

The goal of my commentary [1] is not to provide a short "manual" on how to perform an umbrella review. The commentary specifically highlights the complexity and the expertise needed to perform such a review by elaborating on the different aspects that need to be very carefully taken into consideration in the design, analysis and interpretation of the evidence. The classification for meta-analyses is described as Convincing (Class I), Highly suggestive (Class II), Suggestive (Class III) and Weak (Class VI). The wording aims to work against simple black-and-white dichotomization of significant and non-significant results. There may be no formal consensus yet on the use of this classification from recognized professional organizations but they have been repeatedly used from large teams with long lasting expertise in evidence synthesis. Moreover, I should clarify that the criteria for class I–IV categories are meant to be applied to meta-analyses of observational studies of risk factors. They are not aimed to be applied to meta-analyses of interventions (typically studied with randomized trials).

I am surprised to see the authors' claim that the overall approach and grading criteria are relying solely on statistical significance expressed as *p* values. Actually, the opposite is true. The proposed criteria should not be judged on a standalone basis but collectively, because they complement each other and do not substitute the careful appraisal of the evidence before any classification is performed. Passing a *p* value threshold of 0.05 provides very weak evidence, if any, to support an observational association, it is very easy to achieve but it means close to nothing nowadays [2]. Several methodologists proposed to shift the threshold from 0.05 to 0.005 for new discoveries [3] and many fields, e.g. in omics already use far more stringent *p* value thresholds. Therefore, the proposed threshold for umbrella reviews is 10^{-6} for classes I and II and 10^{-3} for class III. The strength of the proposed association is reinforced by applying the number of events or cases criterion which is more than 1000 and requires a large total sample size and decent power to detect any other than small effects.

The proposed criteria use two different ways to evaluate inconsistency and its impact in meta-analysis. One is the 50% cut-off for I^2 and the other is the 95% predictive intervals, which convey much more information than statistical significance of the meta-analysis they derive from. The I^2 is a relative measure of heterogeneity which can be used to compare the amount of inconsistency across different meta-analyses and even with different number of studies. There is a long lasting debate about the use and the interpretation of these cut-offs, but the Cochrane handbook still refers to them in finer categorization than dichotomy. In addition, the 95% prediction intervals, which unfortunately are not so commonly reported but much more insightful, are very well suited to evaluate the variability of the effect size in different settings [4], not to judge the statistical significance.

Several of the Bradford Hill criteria for causation (specificity, temporality, biological gradient, plausibility, coherence, and analogy) do not necessarily involve quantification and they can be addressed in the steps described before applying the criteria described in the commentary. Bradford Hill was very careful not to consider any of them as definitive and the performance of these criteria in the current research environment is often limited or problematic [5]. Schlesinger et al. also did not notice that there is a clear referral to the quality assessment of the included studies in the general appraisal of the evidence.

✉ Stefania Papatheodorou
spapathe@hsph.harvard.edu

[1] Harvard TH Chan School of Public Health, Boston, USA

[2] Cyprus International Institute for Environment and Public Health, Limassol, Cyprus

I agree that GRADE [6] may be a good tool in the classification of synthesis of randomized or non-randomized trials of interventions but it is not optimal in other fields (observational epidemiology of risk factors, genetic epidemiology, or mendelian randomization) which are occupying a large part of the available evidence or are often the only source of evidence for epidemiological questions. There is substantial overlap between the proposed criteria and the GRADE tool which also involves statistical tests and subjectivity. The focus on the *p* values of the statistical tests to make recommendations on the population level is clearly inappropriate (a short-sighted and biased approach) and was not proposed in my commentary. Inevitably, we have to rely on statistical tests to make inferences but we need to adopt a thoughtful and careful approach in the interpretation of the totality of evidence that takes into account all the points that were reported in this commentary—and more.

References

1. Papatheodorou S. Umbrella reviews: what they are and why we need them. Eur J Epidemiol. 2019;34(6):543–6. https://doi.org/10.1007/s10654-019-00505-6.
2. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting p values in the biomedical literature, 1990–2015. JAMA. 2016;315(11):1141–8. https://doi.org/10.1001/jama.2016.1952.
3. Benjamin DJ, Berger JO, Johannesson M, et al. Redefine statistical significance. Nat Hum Behav. 2018;2(1):6–10. https://doi.org/10.1038/s41562-017-0189-z.
4. IntHout J, Ioannidis JP, Rovers MM, Goeman JJ. Plea for routinely presenting prediction intervals in meta-analysis. BMJ Open. 2016;6(7):e010247. https://doi.org/10.1136/bmjopen-2015-010247.
5. Ioannidis JP. Exposure-wide epidemiology: revisiting Bradford Hill. Stat Med. 2016;35(11):1749–62. https://doi.org/10.1002/sim.6825.
6. Guyatt G, Oxman AD, Akl EA, et al. GRADE guidelines: 1. Introduction-GRADE evidence profiles and summary of findings tables. J Clin Epidemiol. 2011;64(4):383–94. https://doi.org/10.1016/j.jclinepi.2010.04.026.