Arbitrary, Inappropriate Grading Criteria Distort the Evaluation of Evidence: A Comment on Kim et al.'s (2022) Umbrella Review.

Marco Del Giudice

University of New Mexico

In press (2023): ***Neuroscience and Biobehavioral Reviews***

Marco Del Giudice, Department of Psychology, University of New Mexico. Logan Hall, 2001 Redondo Dr. NE, Albuquerque, NM 87131, USA; email: marcodg@unm.edu

In their recent umbrella review, Kim and colleagues (2022) examined the meta-analytic evidence on risk factors, protective factors, and biomarkers for postpartum depression. Using a classification scheme first codified by Bellou et al. (2016), they rated two candidate factors as showing "convincing" evidence, and seven as "highly suggestive". The evidence was deemed "suggestive" in 12 cases, "weak" in 22, and "not significant" in two. None of the biomarkers reached the level of "highly suggestive". The authors also graded the methodological quality of meta-analyses with AMSTAR 2 and the "certainty of evidence" with GRADE.

Umbrella reviews can be extremely useful—but of course, everything hinges on the soundness of the criteria used to evaluate the evidence. The criteria employed by Kim et al. (or slight variations thereof) have been applied in many high-profile reviews and recommended in tutorials (e.g., Fusar-Poli & Radua, 2018); readers may easily assume that they have a rigorous justification, and a demonstrated capacity to separate credible findings from unreliable ones. Unfortunately, this is not the case. As others have noted (e.g., Schlesinger et al., 2019) and I discuss below, most of these criteria rely on arbitrary thresholds without a clear statistical rationale; some are potentially valid but become problematic when applied mechanically; and others are simply invalid and/or based on questionable logic.

In addition, GRADE was specifically designed to assess the effectiveness of *treatments* in view of making clinical recommendations. Accordingly, it strongly privileges randomized controlled trials (RCTs) and meta-analyses showing large effects, with little heterogeneity across studies. But while a preference for large, homogeneous effects is sensible when recommending a medical treatment, it becomes misleading when trying to understand the etiology, epidemiology, and correlates of a disorder (e.g., the existence of certain risk factors can be theoretically important, even if effect sizes are comparatively small or variable across contexts/measures). The automatic downgrading of observational evidence is also a problem when RCTs are impossible/unethical, as is the case with many risk factors. When (mis)applied outside the domain of treatments, GRADE conflates logically distinct questions about the existence, strength, variability, and causal nature of an association (RCTs are the gold standard of causality in clinical research).

I now turn to the specific credibility criteria employed by Kim and colleagues:

a. Significance of the meta-analytic effect: $p < 10^{-6}$ ("highly suggestive" or "convincing"), $p < 10^{-3}$ ("suggestive"), $p < 0.05$ ("weak"). Significance thresholds are intrinsically arbitrary and should be used with caution. The 10^{-6} criterion was introduced by Bellou et al. (2016) without justification (the references cited in support recommended 10^{-3}), and its costs/benefits have not been tested (e.g., in simulations of plausible scenarios). It is unclear why an otherwise "convincing" meta-analysis with (say) $p < 10^{-5}$ should be downgraded to merely "suggestive" based on this criterion alone.

b. $P < 0.05$ for the largest study ("highly suggestive" or "convincing"). This criterion is easily met in practice, but irrelevant: a meta-analysis can provide robust, credible evidence for an effect even if *none* of the individual studies reach significance.

c. More than 1,000 cases of the disorder ("suggestive" and up). This criterion is entirely arbitrary, overly rigid (e.g., it ignores the size of the target effect), and untested. This is especially worrisome, because studies that fail to meet it get automatically rated as "weak".

d. $I^2 < 50\%$ ("convincing"). This and the next criterion focus on between-study heterogeneity; however, heterogeneity need not be a reason for concern outside of clinical practice (see above). This particular criterion is both arbitrary and invalid; it rests on an incorrect application of I^2, which is a *relative* index and says nothing about the absolute amount of heterogeneity (see Borenstein, 2019, pp. 103-120). Ironically, requiring $I^2 < 50\%$ penalizes meta-analyses of large studies: all else being equal, the *proportion* of heterogeneity due to true effect variability must increase as larger *N*s reduce the contribution of sampling error. Hence, meta-analyses that include many large, high-precision studies will tend to show larger I^2 values. Also note that I^2 cannot be estimated reliably if a meta-analysis includes too few studies ("less than 10" is a typical rule of thumb; Borenstein [2019], pp. 90 and 131-132).

e. 95% prediction interval (PI) excluding zero ("convincing"). This criterion uses PIs to perform a sort of improper significance test (Schlesinger et al., 2019). Moreover, if a meta-analysis includes too few studies the PI becomes unreliable, because the variance of the effects cannot be estimated with sufficient precision (see above). Mechanically computing PIs for meta-analyses that include only a handful of studies (see Table 2 in Kim et al.) is inappropriate and misleading.

f. Absence of small-study effects, which may indicate publication bias ("convincing"). This criterion is potentially informative, but should not be applied mechanically. For example, publication bias may still have a negligible impact if a meta-analysis includes enough large-*N* studies.

g. Evidentiary value in a *p*-curve analysis ("convincing"). This criterion is potentially informative; however, *p*-curve results are not dispositive and must be interpreted with caution (see e.g., Erdfelder & Heck, 2019), particularly when the study set is small. Importantly, *p*-curve analysis can detect evidentiary value even in presence of publication bias; thus, the logic of this criterion is inconsistent with the absence of small-study effects required by criterion (f).

h. $P < 0.05$ under a 10% credibility ceiling ("convincing"). As stressed by its inventors (Salanti & Ioannidis, 2009), this technique is highly subjective and context-dependent. It can be valuable as a "skeptical" tool for sensitivity analysis, but not as a mechanical criterion for evidence grading.

Together, these problems can seriously distort the evaluation of evidence. To give just one example, Kim et al. rated the evidence for poor marital relationships (summary OR = 3.38) as "weak" because the meta-analysis included 948 cases of postpartum depression instead of 1,001 or more. But even with more cases, the evidence would still have been deemed "highly suggestive" rather than "convincing" because the estimated $I^2 = 100\%$ fails the (invalid) criterion

of $I^2 < 50\%$. Leaving aside the fact that estimating I^2 from only six effects is ill-advised, supplementary Figure S28 (reproduced here as Figure 1) clearly shows that the large *relative* heterogeneity depends on the high precision of individual studies. (Of course, a sizable amount of *absolute* heterogeneity would not automatically invalidate the finding or make it unconvincing). Besides these purely statistical issues, it would have been useful to consider that all the studies in this meta-analysis came from a single country (Ethiopia); but this kind of information was not reported or discussed in the review, and can only be found by looking up the original papers.

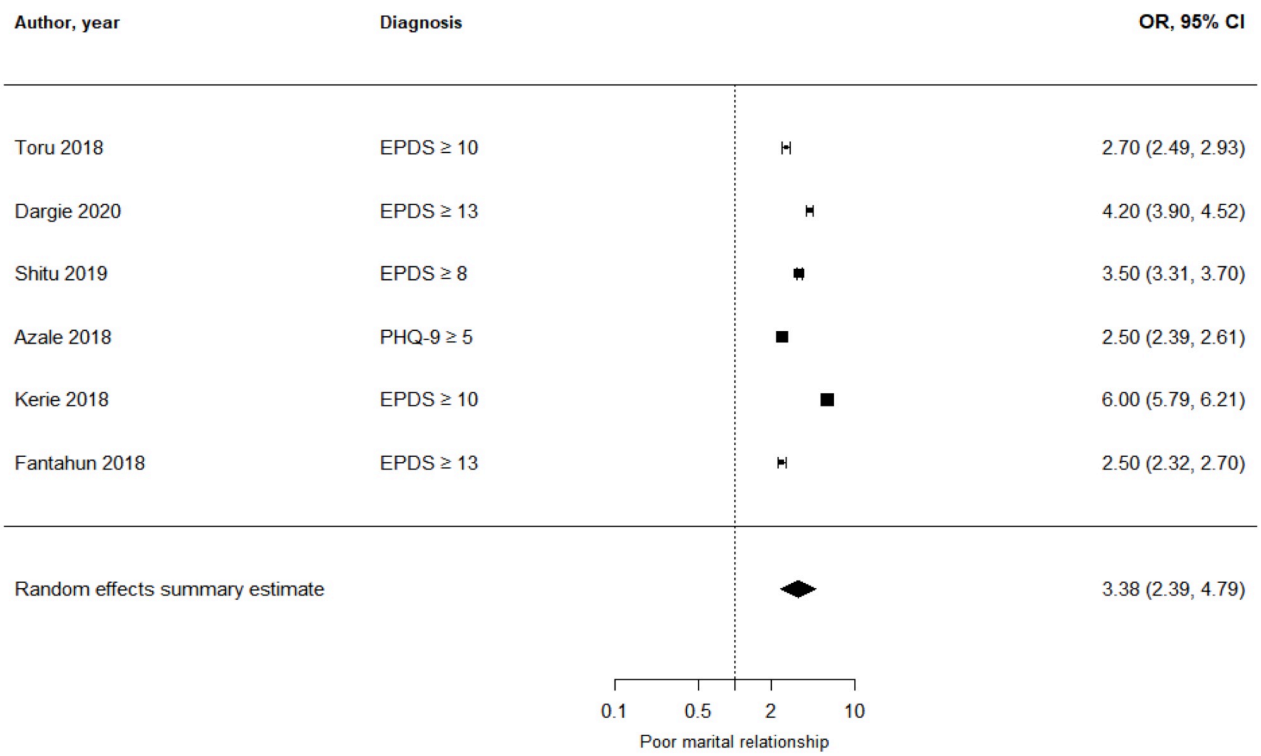

Figure 1. Forest plot of associations between poor marital relationships and postpartum depression, from a meta-analysis of studies performed in Ethiopia (reproduced from Figure S28 in Kim et al., 2022).

While these considerations cast doubts on the results of Kim et al.'s review, I do not mean to single out these authors or this particular paper for criticism. The same or very similar criteria have been widely used in the literature and crystallized into "how-to" guidelines. I hope that this note will alert researchers to the limitations of current approaches, prompting them to explore potential alternatives and reexamine published results. More broadly, I wish to stimulate reflection on the downsides of mechanical "evidence grading", and the risk of introducing biases and distortions under the appearance of rigor.

References

Bellou, V., Belbasis, L., Tzoulaki, I., Evangelou, E., & Ioannidis, J. P. (2016). Environmental risk factors and Parkinson's disease: An umbrella review of meta-analyses. *Parkinsonism & Related Disorders, 23,* 1-9. https://doi.org/10.1016/j.parkreldis.2015.12.008

Borenstein, M. (2019). *Common mistakes in meta-analysis and how to avoid them*. Biostat, Inc.

Erdfelder, E., & Heck, D. W. (2019). Detecting evidential value and *p*-hacking with the *p*-curve tool: A word of caution. *Zeitschrift für Psychologie*, *227,* 249–260. https://doi.org/10.1027/2151-2604/a000383

Fusar-Poli, P., & Radua, J. (2018). Ten simple rules for conducting umbrella reviews. *Evidence-Based Mental Health.* https://doi.org/10.1136/ebmental-2018-300014

Kim, J. H., Kim, J. Y., Lee, S., Lee, S., Stubbs, B., Koyanagi, A., ... & Fusar-Poli, P. (2022). Environmental risk factors, protective factors, and biomarkers for postpartum depressive symptoms: An umbrella review. *Neuroscience & Biobehavioral Reviews, 140,* 104761. https://doi.org/10.1016/j.neubiorev.2022.104761

Salanti, G., & Ioannidis, J. P. (2009). Synthesis of observational studies should consider credibility ceilings. *Journal of Clinical Epidemiology, 62,* 115-122. https://doi.org/10.1016/j.jclinepi.2008.05.014

Schlesinger, S., Schwingshackl, L., Neuenschwander, M., & Barbaresko, J. (2019). A critical reflection on the grading of the certainty of evidence in umbrella reviews. *European Journal of Epidemiology, 34,* 889-890. https://doi.org/10.1007/s10654-019-00531-4