Received: 26 July 2020 | Accepted: 20 October 2020

DOI: 10.1111/1365-2664.14153

Journal of Applied Ecology BRITISH ECOLOGICAL SOCIETY

FORUM

Simulation-based study design accuracy weights are not generalisable and can still lead to biased meta-analytic inference: Comments on Christie et al. (2019)

Oliver L. Pescott[1] | Gavin B. Stewart[2]

[1]UK Centre for Ecology & Hydrology, Oxfordshire, UK

[2]Evidence Synthesis Lab, School of Natural and Environmental Science, University of Newcastle, Newcastle-upon-Tyne, UK

Correspondence
Oliver L. Pescott
Email: oliver.pescott@ceh.ac.uk

Handling Editor: Júlio Louzada

Abstract

1. Variable study quality is a challenge for all the empirical sciences, but perhaps particularly for disciplines such as ecology where experimentation is frequently hampered by system complexity, scale and resourcing. The resulting heterogeneity, and the necessity of subsequently combining the results of different study designs, is a fundamental issue for evidence synthesis.
2. We welcome the recognition of this issue by Christie et al. (2019) and their attempt to provide a generic approach to study quality assessment and meta-analytic weighting through an extensive simulation study. However, we have reservations about the true generality and usefulness of their derived study 'accuracy weights'.
3. First, the simulations of Christie et al. rely on a single approach to effect size calculation, resulting in the odd conclusion that before-after control-impact (BACI) designs are superior to randomised controlled trials (RCTs), which are normally considered the gold standard for causal inference. Second, the so-called 'study quality' scores have long been criticised in the epidemiological literature for failing to accurately summarise individual, study-specific drivers of bias and have been shown to be likely to retain bias and increase variance relative to meta-regression approaches that explicitly model such drivers.
4. *Synthesis and applications*. We suggest that ecological meta-analysts spend more time critically, and transparently, appraising actual studies before synthesis, rather than relying on generic weights or weighting formulas to solve assumed issues; sensitivity analyses and hierarchical meta-regression are likely to be key tools in this work.

KEYWORDS
causal inference, epidemiology, evidence synthesis, meta-analysis, meta-regression, multilevel modelling, quality scoring, study design

1 | INTRODUCTION

Christie et al. (2019; CEA hereafter) outline a simulation-based approach to the generation of study 'accuracy' weights for use in ecological meta-analyses. An index of the error in the effect size estimates resulting from study designs of differing quality, that is differing levels of internal validity or risk of bias (Turner et al., 2009), was used by CEA to create a set of metric weights for the design types investigated. We discuss two issues with this approach: one relating to the specific simulations, effect size calculations and resulting study design accuracy weights proposed by CEA; and another relating to the fact that the use of unidimensional study quality scores as meta-analytic weights has long been criticised as a suboptimal approach to the generation of minimum error meta-analytic effect estimates, and as such has been regularly advised against in the epidemiological literature (Greenland, 1994a, 1994b; Greenland & O'Rourke, 2001; Higgins et al., 2011; Moreno et al., 2012; Shea et al., 2017; Turner et al., 2009).

1.1 | The Christie et al. accuracy weights are effect size calculation and simulation study specific

A perhaps surprising result of the simulations presented by CEA is the conclusion that randomised controlled trials (RCTs), normally considered the gold standard for causal inference (e.g. Rubin, 2007), can display more error than before-after control-impact (BACI) designs. This conclusion relates to the fact that CEA assume that meta-analyses of RCTs always calculate effect sizes using post-impact measures only (e.g. Cohen's *d*). In the situation, common in ecology, where RCTs are small, then metrics that do not take into account the potential lack of balance (which may or may not be adjustable using known covariates) will do worse at estimating a known true effect than those that do (e.g. Roberts & Torgerson, 1999). It seems to us unlikely that ecologists who go to the trouble of setting up a randomised controlled experiment would also not collect baseline data and subsequently use this to draw the most robust conclusions possible. For example, in RCTs investigating the impact of trampling on plant communities, the standard effect size reported takes baseline plant cover into account (Pescott & Stewart, 2014). The metric used by CEA for their BACI effect size estimates overcomes this issue (resulting in the superior performance of this approach reported by CEA), because here effect size estimates are calculated relative to baseline in each treatment arm of the simulated studies. In general, where there is some sort of imbalance between different arms of a study, methods that take the baseline state into account, for example using repeated measures-type analyses or a within-arm change score as the dependent variable, normally result in lower overall error (but see Glymour et al., 2005). Where imbalance can be adjusted for by known baseline covariates, then ANCOVAs, or similar regression-based methods, can also be used (Senn, 1989). (Note that the ANCOVA approach yields higher precision relative to change score approaches; Senn, 2005, 2020.) If, as with RCTs, error is expected to be only of random origin, then meta-analytic summaries of such trials will be unbiased, and weighting may be unnecessary. However, small RCTs may be associated with other issues causing systematic error (Turner et al., 2009), a fact that has been used to justify inverse-variance weighting of meta-analyses of RCTs, although this may be more profitably adjusted for using regression-based methods (Moreno et al., 2012; also see Section 1.2 below).

Taking the estimates of error, and the resulting weights, from a simulation exercise as of broader relevance for the weighting of studies in other meta-analyses therefore ignores a number of facts, including that RCTs should logically possess greater validity than a BACI design focused on the same research question. This is always an assumption, but is more likely to be true for RCTs, particularly as they increase in size; this also applies to unreported, unmeasured and/or unknown variables that may be of importance for the monitored response. All else being equal, the use of the CEA accuracy weights downweights RCTs relative to BACIs simply on the basis of a set of simulations not adjusting for baseline information resulting in an estimator with higher error, despite the fact that in general RCTs are logically superior for causal inference, that methods exist to adjust for such imbalances and that workers running RCTs in ecology would be very likely to collect such information. The overall error seen in simulations that do not take these points into account is therefore not a sensible or generalisable measure of study design relative quality for the purpose of weighting studies in meta-analysis. Given this, CEA-type simulations could perhaps be extended to investigate the impact of different effect size calculations on the resulting weights.

In our opinion, however, assessments of study quality for evidence synthesis should pay close attention to the details of the specific studies being summarised, rather than working from the position that a single set of assumptions used in a simulation exercise, whether based on estimated parameter values from the wider literature or not, accurately captures the only important features relating to the potential for bias, or overall error, in any particular study. As noted by CEA, frameworks for such assessments already exist in other disciplines (e.g. Deeks et al., 2003; Higgins et al., 2011; Shea et al., 2017; Turner et al., 2009), and could be further developed for ecology beyond standard study design classifications (Lortie et al., 2015). This issue of whether or not a generic set of simulation-based weights are likely to capture all important determinants of the bias or overall error that might be exhibited by a study also underlies our second criticism.

1.2 | Using quality scores as meta-analytic weights can still result in biased inference

Even if we were broadly happy that a set of simulation-based metrics represented the key features capturing bias in a set of studies of variable design quality, the use of these to create unidimensional weights for meta-analysis may still result in biased overall inference (Greenland & O'Rourke, 2001). Greenland and O'Rourke (2001), developing arguments put forward by Greenland (1994a, 1994b), noted that while such quality score weightings may reduce bias, they will often do this at the expense of increased variance in the weighted estimator. Overall inference may therefore not be improved. Greenland and O'Rourke (2001) pointed out that in this situation, methods that formally trade-off bias and variance to minimise overall error, such as hierarchical meta-regression, are likely to be more appropriate (Greenland, 2000; see also Moreno et al., 2012). Regression-based approaches to study quality adjustment also have the advantage of being able to adjust summary estimates for the directions in which different biases might act (Greenland & O'Rourke, 2001), another reason why they are recommended by modern epidemiological bias assessment frameworks (Higgins et al., 2011; Shea et al., 2017).

Greenland and O'Rourke (2001) outlined various meta-regression approaches that could potentially overcome the limitations of unidimensional study quality scores by regressing outcomes on individual study 'quality item' covariates. This approach is based on the suggestion of Rubin (1990, 1992) that using a regression framework to extrapolate to an 'ideal' study is theoretically justified in preference to estimating an average effect from the literature (Moreno et al., 2012). One could argue that the simulation-based accuracy weights of CEA could be used as a quality item in a hierarchical meta-regression framework; however, summary quality scores resulting from generic simulation approaches are unlikely to adequately represent the study-specific, multi-dimensional nature of bias. The accuracy weight approach of CEA necessarily combines, and therefore ignores, many different and separate features of study quality. For example, as discussed above, the weights derived by CEA are dependent on the particular effect size metrics calculated, on assumptions about balance across different treatment arms, and on whether or not the parameter values and distributions used in any simulation are truly appropriate for any set of studies to which the resulting weights might be applied. They also ignore issues relating to external validity (i.e. the representativeness of a study relative to its inferential target population; Turner et al., 2009) that will be important for some types of evidence synthesis. The individual simulation parameters of CEA, and the weights themselves, collapse the multiple dimensions of individual study bias into a single score of a type that is highly unlikely to be perfectly proportional to the actual overall biases exhibited across and within the study design types (Greenland & O'Rourke, 2001). The use of the CEA accuracy weights as a covariate in a meta-regression framework would therefore likely still result in bias. For this reason, Greenland and O'Rourke (2001, 2008) and others (Higgins et al., 2011; Shea et al., 2017) recommend keeping quality items separate and estimating them individually for each study included in a meta-analysis.

2 | CONCLUSION

Detailed assessments of individual study validity are most likely to provide appropriate and focused measures of quality that can be used in ways that are known to avoid, or minimise, the biases associated with the use of unidimensional quality scores that subsume or ignore the detail of primary studies under review (Greenland & O'Rourke, 2001; Turner et al., 2009). Even studies that appear to be high quality may still contain non-obvious biases, and apparently lower quality studies could in fact be unbiased for the effect of interest. Ultimately, corrections for study bias rely on arguments that may not be empirically provable in any given case: context and expert judgement must be used to assess whether or not claims for study-specific bias are either corroborated by available evidence, plausible but uncorroborated, or implausible (Greenland & Pearce, 2015). Subsequently adjusting for study-level drivers of bias is the current best practice in epidemiology (Greenland & O'Rourke, 2008; Thompson et al., 2011), a science that has already grappled with this issue for several decades. We suggest that ecologists undertaking evidence synthesis focus critical appraisal on the generation of transparent value judgements about causation, the potential for confounding (risk of bias) and the appropriateness of the outcome measure (directness), using sensitivity analyses and hierarchical meta-regression to quantify and model uncertainty arising from study quality, rather than automatically using generic weights. The considerable risk presented by such weighting processes to thoughtful ecological synthesis should be clear.

AUTHORS' CONTRIBUTIONS

O.L.P. and G.B.S. conceived the ideas; O.L.P. led the writing of the manuscript. Both authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

We will not be archiving data because this manuscript does not use data.

ORCID

Oliver L. Pescott https://orcid.org/0000-0002-0685-8046

REFERENCES

Christie, A. P., Amano, T., Martin, P. A., Shackelford, G. E., Simmons, B. I., & Sutherland, W. J. (2019). Simple study designs in ecology produce inaccurate estimates of biodiversity responses. *Journal of Applied Ecology*, *56*(12), 2742–2754. https://doi.org/10.1111/1365-2664.13499

Deeks, J. J., Dinnes, J., D'Amico, R., Sowden, A. J., Sakarovitch, C., Song, F., Petticrew, M., & Altman, D. G. (2003). Evaluating non-randomised intervention studies. *Health Technology Assessment*, *7*(27), 1–173.

Glymour, M. M., Weuve, J., Berkman, L. F., Kawachi, I., & Robins, J. M. (2005). When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *American Journal of Epidemiology*, *162*(3), 267–278.

Greenland, S. (1994a). A critical look at some popular meta-analytic methods. *American Journal of Epidemiology*, *140*(3), 290–296. https://doi.org/10.1093/oxfordjournals.aje.a117248

Greenland, S. (1994b). Quality scores are useless and potentially misleading. *American Journal of Epidemiology*, *140*(3), 300–301.

Greenland, S. (2000). When should epidemiologic regressions use random coefficients? *Biometrics*, *56*, 915–921.

Greenland, S., & O'Rourke, K. (2001). On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*, *2*(4), 463–471. https://doi.org/10.1093/biostatistics/2.4.463

Greenland, S., & O'Rourke, K. (2008). Meta-analysis. In K. J. Rothman, S. Greenland, & T. L. Lash (Eds.), *Modern epidemiology* (3rd ed., pp. 652–683). Lippincott Williams & Wilkins.

Greenland, S., & Pearce, N. (2015). Statistical foundations for model-based adjustments. *Annual Review of Public Health*, *36*, 89–108.

Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., Sterne, J. A. C., Cochrane Bias Methods Group, & Cochrane Statistical Methods Group. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, *343*, d5928. https://doi.org/10.1136/bmj.d5928

Lortie, C. J., Stewart, G., Rothstein, H., & Lau, J. (2015). How to critically read ecological meta-analyses. *Research Synthesis Methods*, *6*(2), 124–133. https://doi.org/10.1002/jrsm.1109

Moreno, S. G., Sutton, A. J., Thompson, J. R., Ades, A. E., Abrams, K. R., & Cooper, N. J. (2012). A generalized weighting regression-derived meta-analysis estimator robust to small-study effects and heterogeneity. *Statistics in Medicine*, *31*(14), 1407–1417.

Pescott, O. L., & Stewart, G. B. (2014). Assessing the impact of human trampling on vegetation: A systematic review and meta-analysis of experimental evidence. *PeerJ*, *2*, e360. https://doi.org/10.7717/peerj.360

Roberts, C., & Torgerson, D. J. (1999). Baseline imbalance in randomised controlled trials. *BMJ*, *319*(7203), 185. https://doi.org/10.1136/bmj.319.7203.185

Rubin, D. B. (1990). A new perspective. In K. W. Wachter & M. L. Straf (Eds.), *The future of meta-analysis* (pp. 155–166). Russell Sage.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics*, *17*, 363–374.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, *26*(1), 20–36. https://doi.org/10.1002/sim.2739

Senn, S. J. (1989). Covariate imbalance and random allocation in clinical trials. *Statistics in Medicine*, *8*(4), 467–475. https://doi.org/10.1002/sim.4780080410

Senn, S. J. (2005). An unreasonable prejudice against modelling? *Pharmaceutical Statistics*, *4*, 87–89.

Senn, S. J. (2020). *Being just about adjustment*. Retrieved from https://errorstatistics.com/2020/03/16/stephen-senn-being-just-about-adjustment-guest-post/

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., Moher, D., Tugwell, P., Welch, V., Kristjansson, E., & Henry, D. A. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ*, *358*, j4008. https://doi.org/10.1136/bmj.j4008

Thompson, S., Ekelund, U., Jebb, S., Lindroos, A. K., Mander, A., Sharp, S., Turner, R., & Wilks, D. (2011). A proposed method of bias adjustment for meta-analyses of published observational studies. *International Journal of Epidemiology*, *40*(3), 765–777. https://doi.org/10.1093/ije/dyq248

Turner, R. M., Spiegelhalter, D. J., Smith, G. C. S., & Thompson, S. G. (2009). Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 21–47. https://doi.org/10.1111/j.1467-985X.2008.00547.x