Innovation and forward-thinking are needed to improve traditional synthesis methods: a response to Pescott & Stewart

Alec P. Christie[1,4,7*], Tatsuya Amano[1,2,3], Philip A. Martin[1,4,8], Gorm E. Shackelford[1,4], Benno I. Simmons[1,5,6], William J. Sutherland[1,4]

[1]Conservation Science Group, Department of Zoology, University of Cambridge, The David Attenborough Building, Downing Street, Cambridge, UK.

[2]Centre for the Study of Existential Risk, University of Cambridge, 16 Mill Lane, Cambridge, UK.

[3]School of Biological Sciences, University of Queensland, Brisbane, 4072 Queensland, Australia

[4]BioRISC, St Catharine's College, Cambridge, UK.

[5]Department of Animal and Plant Sciences, University of Sheffield, Sheffield, UK.

[6]Centre for Ecology and Conservation, College of Life and Environmental Sciences, University of Exeter, Penryn, UK.

[7]Downing College, Regent Street, Cambridge, UK.

[8]Basque Centre for Climate Change (BC3), Edificio sede no 1, planta 1, Parque científico UPV/EHU, Barrio Sarriena s/n, 48940, Leioa, Bizkaia, Spain

*Corresponding author, apc58@cam.ac.uk

Abstract

1. In Christie et al. (2019), we used simulations to quantitatively compare the bias of commonly used study designs in ecology and conservation. Based on these simulations, we proposed ‘accuracy weights’ as a potential way to account for study design validity in meta-analytic weighting methods. Pescott & Stewart (2021) raised concerns that these weights may not be generalisable and still lead to biased meta-estimates. Here we respond to their concerns and demonstrate why developing alternative weighting methods is key to the future of evidence synthesis.
2. We acknowledge that our simple simulation unfairly penalised Randomised Controlled Trial (RCT) relative to Before-After Control-Impact (BACI) designs as we assumed that the parallel trends assumption held for BACI designs. We point to an empirical follow-up study in which we more fairly quantify differences in biases between different study designs. However, we stand by our main findings that Before-After (BA), Control-Impact (CI), and After designs are quantifiably more biased than BACI and RCT designs. We also emphasise that our 'accuracy weighting’ method was preliminary and welcome future research to incorporate more dimensions of study quality.
3. We further show that over a decade of advances in quality effect modelling, which Pescott & Stewart (2021) omit, highlights the importance of research such as ours in better understanding how to quantitatively integrate data on study quality directly into meta-analyses. We further argue that the traditional methods advocated for by Pescott & Stewart (2021) (e.g., manual risk-of-bias assessments and inverse-variance weighting) are subjective, wasteful, and potentially biased themselves. They also lack scalability for use in large syntheses that keep up-to-date with the rapidly growing scientific literature.
4. *Synthesis and applications.* We suggest, contrary to Pescott & Stewart’s narrative, that moving towards alternative weighting methods is key to future-proofing evidence synthesis through greater automation, flexibility, and updating to respond to decision-

makers needs – particularly in crisis disciplines in conservation science where problematic biases and variability exist in study designs, contexts, and metrics used. Whilst we must be cautious to avoid misinforming decision-makers, this should not stop us investigating alternative weighting methods that integrate study quality data directly into meta-analyses. To reliably and pragmatically inform decision-makers with science, we need efficient, scalable, readily automated, and feasible methods to appraise and weight studies to produce large-scale living syntheses of the future.

Keywords: *evidence synthesis, meta-analysis, dynamic meta-analysis, living reviews, automation, quality effects modelling, meta-analyses, risk-of-bias, critical appraisal, bias adjustment.*

Introduction

Pescott & Stewart (2021) outlined their concerns over an alternative method of weighting in meta-analysis we proposed called "accuracy weights" in Christie et al. (2019). These weights were derived from our simulation study that aimed to quantitatively compare the performance of different experimental and observational study designs (Christie et al., 2019). Their two major concerns were that our accuracy weights were not generalisable and that quality score weightings, such as ours, may still lead to biased estimates in meta-analyses. Here we respond to their concerns and discuss why we believe alternative methods of weighting are central to the future of evidence synthesis.

1. Accuracy weights need improving and combining with other quality measures

As Pescott & Stewart suggest, we acknowledge that our simulation may have unfairly penalised Randomised Controlled Trial (RCT) designs, depending on whether researchers in ecology and conservation do take into account pre-impact sampling. However, in our experience, few Randomised Controlled Trials in conservation take account of pre-impact baseline data; this is supported by a recent study quantifying the use of different study designs in the environmental and social sciences (Christie et al., 2020a). We acknowledge that we did not discuss more of the shortcomings of Before-After Control-Impact (BACI) designs in terms of the bias that can be introduced by violating the ‘parallel trends’ assumption (Dimick and Ryan, 2014; Underwood, 1991; Wauchope et al., 2020). Therefore, with respect to comparing BACI and RCT designs, we acknowledge our simulation has limitations.

Nevertheless, our major motivation was to demonstrate the difference in study design performance between simpler designs (e.g., Before-After (BA), Control-Impact (CI), and After designs) and more rigorous designs (RCT and BACI). Thus, we intentionally made our simulation relatively simple to engage a wide audience of researchers. We have since built on our simulations in Christie et al. (2020a), which uses an empirical, model-based methodology to quantify the differences in bias affecting different study designs using raw (rather than simulated) data from a large number of within-study comparisons. This more fairly quantifies the bias associated with RCT versus BACI designs by making fewer, more statistically defensible assumptions about the ‘true effect’ (to estimate bias) and inherently accounts for the parallel trends assumption that can bias BACI designs (Christie et al., 2020a).

Pescott & Stewart also suggest our simulation weights do not capture the full range of potential sources of bias affecting study designs and advise that assessments of study quality should closely scrutinise the details of specific studies being summarised (e.g., using manual risk-of-bias assessments). In our study, we specifically acknowledged that our weights were relatively simple and need to be built upon to incorporate a wider range of study quality indicators; we outlined possible approaches in the future that could integrate scores from critical appraisal tools that exist for ecology and conservation (Mupepele et al., 2016). We are happy to see that others are building on our work and investigating the use of a broader set of quality or validity measures to weight studies in meta-analyses (e.g., Schafft et al. 2021, Mupepele et al. 2021). In the next sections, we address Pescott & Stewart's criticisms of weighting by quality scores and discuss statistical advances in applying quality score weightings to meta-analyses. We also discuss the problems associated with the traditional methods advocated for by Pescott & Stewart (such as inverse-variance weighting and manual risk-of-bias assessments).

2. Recent advances in directly integrating data on study quality into meta-analyses

In Pescott & Stewart's discussion on why they advocate against weighting by quality scores in meta-analyses, they omit over a decade of research in epidemiology on alternative quality score weighting methods that have overcome many of the problems they discuss (Doi, Barendregt and Mozurkewich, 2011; Doi et al., 2015a, 2015b; Doi and Thalib, 2008; Rhodes et al., 2020; Stone et al., 2020). In particular, 'bias adjustment' methods, such as quality effects models, represent an active and promising area of research in evidence synthesis in epidemiology (Doi, Barendregt and Mozurkewich, 2011; Doi and Thalib, 2008; Rhodes et al., 2020; Stone et al., 2020).

Critical appraisal is traditionally used to descriptively report the risk of bias for different studies, rather than trying to quantitatively incorporate those assessments within the analyses themselves (Johnson, Low and MacDonald, 2015). Instead, our accuracy weights are related to the field of ‘bias-adjustment’ methods which seek to directly integrate risk-of-bias assessments into meta-analytic results (Stone et al., 2020). Criticisms of quality score weightings have centered around four major issues: 1.) the choice of quality scale influences the weight of individual studies; 2.) the meta-estimate and its confidence interval depends on the scale; 3.) there is no reason why study quality should modify the precision of estimates; and 4.) poor studies are not excluded (Stone et al., 2020). Therefore, as Pescott & Stewart also appear to argue, any bias associated with poor quality studies can only be reduced at best, and not removed (Stone et al., 2020).

Whilst proponents of quality score approaches accepted these criticisms and ceased their development, an alternative, improved methodology called ‘quality effects models’ have subsequently been developed and refined in recent years. This approach uses a relative scale and ‘synthetic weights’ (yielding relative credibility ranks for different studies) that overcame the major issues that affected quality score approaches, and has been shown to yield an estimator with superior error and coverage to conventional estimators (Doi et al., 2015b, 2017). There are a range of possible ways, each with advantages or disadvantages, to derive the relative credibility weights for studies using numerical data generated by expert opinion (Turner et al., 2009), data-based distributions, or statistically combining expert opinion and data-based distributions (Rhodes et al., 2020). Therefore, results from further refining and improving our simulations and empirical analyses (Christie et al., 2019; Christie et al., 2020a) could provide valuable contributions to the active development of these methods to integrate data on study quality directly into meta-analyses.

Pescott & Stewart focus on the possibility of incorporating study quality scores into meta-regression approaches. Their criticism of our weights in their current form is that they are too unidimensional and not study-specific; this is a criticism that we partially accept. Indeed, we specifically discussed the need to expand and improve our weights to integrate other aspects of study quality (e.g., using expert opinion, data-based distributions, or critical appraisal tools to adjust relative credibility ranks; Rhodes et al., 2020). In hindsight, we should have dedicated more attention to how we would further develop and more robustly apply our accuracy weights alongside discussing advances in quality effects models.

Pescott & Stewart also suggest that we ignore issues relating to external validity. Given that traditional weightings, such as sample size or inverse variance, also fail to consider external validity, we find this an odd criticism, particularly given our simulation was clearly focused on addressing issues of study design quality and internal validity. We are in fact developing an alternative meta-analytic method, dynamic meta-analysis (Shackelford et al., 2021), based on the Metadataset platform (www.metadataset.com), which we plan to use to test different weighting methods, including ‘recalibration’ from the medical sciences (Kneale et al., 2019) which aims to adjust studies’ influence in meta-analyses based on their external validity (or relevance to decision-makers). Again this work is in the early stages of development and there are many methodological challenges to overcome, particularly in how to integrate ‘recalibration’ methods into random effects models and how to ensure such interactive meta-analytic tools are used robustly (Shackelford et al. 2021). Therefore, as Pescott & Stewart suggest, we believe it should be possible to integrate internal validity or quality items, and external validity items, into a hierarchical meta-regression framework, or to directly weight studies using new advances in quality effects models as discussed previously (see Stone et al. 2020 for a comparison and discussion of different approaches).

3. Integrating data on study quality into meta-analyses is essential to the future of evidence synthesis

We also believe Pescott & Stewart's discussion presents a narrow vision of the challenges faced by traditional critical appraisal and weighting methods. We believe that the traditional 'medical-style' approaches (e.g., manual risk-of-bias assessments combined with inverse-variance weighting) that Pescott & Stewart believe should be adhered to are ultimately inefficient and wasteful. The field of evidence synthesis is advancing at pace to respond to the challenges of rapidly growing evidence bases and fast-moving crises, which requires new methodologies that help to keep evidence bases 'up-to-date' or 'living', cost-efficient by working at massive discipline-wide scales, and dynamically adjustable to be relevant to different decision-makers' needs. Here we elaborate on why this is problematic to Pescott & Stewart's assertion that we should continue to rely on traditional methods, rather than alternative weighting methods such as the one we proposed in Christie et al. (2019).

3.a. Alternative weighting methods facilitate more efficient, automated, living, large-scale syntheses

First, there is growing recognition that decision makers need constantly updated evidence syntheses (Elliot et al., 2021) and that traditional synthesis methods (e.g., traditional systematic reviews) are often too time-consuming, quickly go out-of-date, and can miss important opportunities to influence practice and policy (Boutron et al., 2020; Grainger et al., 2019; Haddaway and Westgate, 2019; Koricheva and Kulinskaya, 2019; Nakagawa et al., 2020; Pattanittum et al., 2012; Shojania et al., 2007). Given that the scientific literature in most disciplines is growing rapidly (Bornmann and Mutz, 2015; Larsen and von Ins, 2010) and that publication delays already hamper evidence-based decision-making in crisis disciplines, such as conservation (Christie et al. 2021), evidence synthesis needs to be as time- and cost-efficient as possible (Grainger et al., 2019; Nakagawa et al., 2020). Generating large-scale, easily updateable, 'living' evidence databases containing results and metadata of scientific studies across subjects is therefore central to future-proofing evidence synthesis (Elliot et al., 2021; Shackelford et al. 2019). It is these forward-thinking approaches to synthesis that can be facilitated by alternative methods of weighting, which can use the metadata on studies to automatically and rapidly critically appraise studies and weight meta-analyses, without the need for cumbersome, inefficent, and ultimately subjective manual risk-of-bias assessments that Pescott & Stewart promote. Inverse-variance or sample size weighting could equally draw on these evidence databases, but are far poorer at capturing information on study validity (since our simulation study showed that certain study designs may have lower variance but higher bias (e.g., BA or CI) than other designs (e.g., RCT and BACI)).

Therefore, whilst we acknowledge Pescott & Stewart's belief that traditional methods to critical appraisal (e.g., manual risk-of-bias assessments) are a key part of the rigour of current evidence synthesis, we argue that such methods are not efficient, scaleable, or feasible enough for use in living synthesis projects that we need to deliver at scale to more comprehensively bridge the research-practice and policy gaps (e.g., Conservation Evidence produced using the subject-wide evidence synthesis methodology; Sutherland et al. 2019). We instead suggest that more efficient alternative weighting methods to rapidly critically appraise and weight studies by their validity and quality (e.g., through automating risk-of-bias assessments; Marshall et al. 2015) are urgently needed to ensure evidence synthesis can keep pace with the rapidly growing scientific literature (Marshall and Wallace, 2019; O'Connor et al., 2018; Thomas et al., 2017; Wallace et al., 2014). Increasingly using automated and data-based methods for estimating different dimensions of study quality will become a more important and necessary approach to speed up evidence synthesis (Marshall et al., 2020; Marshall, Kuiper and Wallace, 2015; Marshall and Wallace, 2019; O'Connor et al., 2018; Tsafnat et al., 2013; Wallace et al., 2014). Of course, it is important that automated methods, particularly for critical appraisal and weighting, balance increased efficiency against high standards and rigour to ensure they give a reliable reflection of the quality of studies – which we believe should act as a major motivation for follow-up research to improve and expand upon our weights.

We also argue that the traditional method of descriptively reporting the risk of bias for different studies in critical appraisal that Pescott & Stewart advocate for (rather than trying to quantitatively incorporate this information within the analyses themselves) represents an inefficient and wasteful use of this detailed information given the major investment of resources they require (Johnson, Low and MacDonald, 2015; Haddaway and Westgate, 2019). Guidance by the Cochrane Collaboration for medical syntheses currently makes risk-of-bias assessments mandatory, but there is still no consensus on how to use them to 'adjust' the results of evidence syntheses (Rhodes et al., 2020). The GRADE approach used in medicine, which Pescott & Stewart appear to support, is to use risk-of-bias assessments to define a threshold beyond which a recommendation can be supported, and then use subjective risk-of-bias assessment to determine which side the true effect lies of a particular threshold (or within a certain range) (Stone et al., 2020). This stratification of results (or regression models using risk of bias) has been criticised based on empirical evidence suggesting that conditioning on risk of bias may induce collider-stratification bias (Stone et al., 2019).

3.b. Alternative weighting methods facilitate considering study quality as a spectrum rather than a cut-off

Another concern we have with the manual risk-of-bias assessments that Pescott & Stewart advocate for is that these can be used to exclude studies from meta-analyses, which can have a major impact in disciplines where evidence bases often lack more rigorous study designs, such as conservation science (Christie et al., 2020b, 2020c, 2020a; Junker et al., 2020). Whilst this may be justifiable in cases where studies are clearly extremely flawed or unreliable, we believe the ‘rubbish in, rubbish out’ concept and idealised ‘best evidence’ approach (Slavin, 1986, 1995; Tugwell and Haynes, 2006) is dangerous and ignores the fact that studies of lower quality can add useful information to evidence syntheses (Davies and Gray, 2015; Gough and White, 2018; Lortie et al., 2015). Rather than excluding studies judged to be of lower quality, we believe that they should instead be included but treated with appropriate caution and uncertainty (Christie et al., 2020a). We need new, alternative methods of weighting in meta-analyses to do this because traditional inverse-variance weighting does not account for potential differences in bias introduced by different studies (hence the previous reliance on excluding studies below an arbitrary quality threshold; Doi and Thalib, 2008; Rhodes et al., 2018, 2020; Stone et al., 2020). Pescott & Stewart’s assertion that we should not stray from weighting by inverse-variance also ignores the fact that this traditional method is prone to bias when analysing both Hedges’ *d* (Hamman et al., 2018) and log-response ratios (Doncaster and Spake, 2018; Bakbergenuly, Hoaglin and Kulinskaya, 2020), which are two of the most commonly used effect size measures in ecology and conservation.

The alternative weighting methods that we are advocating for do not enforce a cut-off threshold in the evidence being used, but instead focus on weighting evidence on a spectrum of quality or validity, which we believe is a more defensible philosophical approach – i..e, placing greater weight behind studies that are more trustworthy and reliable. Furthermore, the approach of excluding studies via risk-of-bias assessments has been criticised because resulting recommendations from syntheses may rely on a small subset of studies (subjectively judged to be of sufficient quality), which may be less robust than analysing a much larger set of studies with variable quality (Davies and Gray, 2015; Gough and White, 2018; Lortie et al., 2015).

Indeed, during the Covid-19 pandemic, the overreliance of evidence-based recommendations on Randomised Controlled Trials has been criticised for delays in promoting wearing of face masks and coverings in Western countries, particularly when high quality non-RCT evidence was available from community settings (The Royal Society, 2020). Pescott & Stewart's comment: "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" surely justifies why excluding studies using manual, subjective risk-of-bias assessments can be dangerous – and why alternative weighting methods that consider a wide range of more objective data and statistics on study quality are needed. Excluding studies based on manual risk-of-bias assessments is also likely to strongly limit the scope, relevance, and external validity of any recommendations made by evidence syntheses in crisis disciplines with patchy evidence bases, such as conservation (Christie et al., 2020b; Gutzat and Dormann, 2020). Conversely, alternative weighting methods would maximise the number of studies (and study contexdts) considered (whilst directly accounting for study quality), which is likely to be more useful and efficient for informing rapid evidence-based decision-making in multiple different contexts.

Alternative weighting methods also facilitate the use of weighting by external validity (i.e., relevance), in addition to internal validity (i.e., reliability), as different components of the overall validity of a study can be considered. As discussed earlier, the alternative weighting method of 'recalibration' has been proposed in the medical sciences (Kneale et al., 2019) to adjust studies weights in meta-analyses by their relevance to a decision-maker's question and context of interest (as trialled in dynamic meta-analyses on the Metadataset platform for interventions on invasive species and agricultural management; Shackelford et al. 2019). The issue of external validity and relevance is a crucial issue for decision-makers and the usefulness of evidence syntheses, particularly in disciplines like ecology and conservation science where variation between studies based on their local context (e.g., biophysical environment, species studied, details of intervention carried out) is perceived to be large and important for management. Traditional methods of synthesis advocated for by Pescott & Stewart typically fail to account for or consider the relevance of different studies to decision-makers in any detail, and certainly do not directly integrate this meta-analytic weightings of studies. As different studies will have different levels of relevance to different decision-makers, these alternative weighting methods also support the movement towards more interactive, living meta-analytic platforms for evidence synthesis (rather than static publiations of meta-analyses) that can be dynamically adjusted to give bespoke evidence-based recommendations (Shackelford et al. 2019; Kneale et al. 2019). Developing alternative weighting methods, rather than avoiding them, is therefore central to increasing the usefulness and relevance of evidence syntheses to decision-makers.

Conclusion

We agree with Pescott & Stewart that synthesising the results of studies using different designs is a fundamental issue for evidence synthesis – hence the urgent need for further scientific investigation into how to directly integrate measures of study quality into meta-analyses (Boutron et al., 2020; Hamman et al., 2018; Jenicek, 1989; Nakagawa et al., 2020; Sutherland and Wordley, 2018). Pescott & Stewart’s concerns about straying from traditional methods of critical appraisal and weighting (e.g., manual risk-of-bias assessments and inverse-variance weightin) may be understandable, but we strongly believe that embracing alternative weighting methods present massive opportunities to advance and future-proof evidence synthesis against the mounting challenges we face in synthesising evidence. We should always strive to maintain high standards of rigour in evidence synthesis, but we must not let the pefect be the enemy of the good when it comes to integrating study quality data into meta-analyses. By embracing alternative weighting methods, we can ultimately increase the efficiency, usefulness, and scale of evidence synthesis through automating critical appraisal, weighting, the regular updating of syntheses.

Acknowledgements

Author funding sources: TA was supported by the Grantham Foundation for the Protection of the Environment, Kenneth Miller Trust and Australian Research Council Future Fellowship (FT180100354); WJS, PAM and GES were supported by Arcadia and The David and Claudia Harding Foundation; BIS and APC were supported by the Natural Environment Research Council via Cambridge Earth System Science NERC DTP (NE/L002507/1, NE/S001395/1), and BIS was supported by the Royal Commission for the Exhibition of 1851 Research Fellowship.

Conflict of Interest

The authors declare no conflicts of interest.

Authors contributions

APC led the writing of this manuscript and all other authors contributed critically to drafts, as well as giving final approval for publication.

Data Availability Statement

There is no data associated with this article.

Supporting Information

There is no Supporting Information associated with this article.

References

Bakbergenuly, I., Hoaglin, D.C. and Kulinskaya, E. (2020) “Estimation in meta-analyses of response ratios,” *BMC Medical Research Methodology*, 20(1), p. 263.

Borah, R., Brown, A.W., Capers, P.L. and Kaiser, K.A. (2017) “Analysis of the timeand workers needed to conduct systematic reviews of medicalinterventions using data from the PROSPERO registry,” *BMJ,* Open7:e012545.

Bornmann, L. and Mutz, R. (2015) “Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references,” *Journal of the Association for Information Science and Technology*, 66(11) John Wiley & Sons, Ltd, pp. 2215–2222.

Boutron, I., Créquit, P., Williams, H., Meerpohl, J., Craig, J.C. and Ravaud, P. (2020) “Future of evidence ecosystem series: 1. Introduction — Evidence synthesis ecosystem needs dramatic change,” *Journal of Clinical Epidemiology*, Elsevier Inc.

Christie, A.P., Abecasis, D., Adjeroud, M., Alonso, J.C., Amano, T., Anton, A., Baldigo, B.P., Barrientos, R., Bicknell, J.E., Buhl, D.A., Cebrian, J., Ceia, R.S., Cibils-Martina, L., Clarke, S., Claudet, J., Craig, M.D., Davoult, D., de Backer, A., Donovan, M.K., Eddy, T.D., França, F.M., Gardner, J.P.A., Harris, B.P., Huusko, A., Jones, I.L., Kelaher, B.P., Kotiaho, J.S., López-Baucells, A., Major, H.L., Mäki-Petäys, A., Martín, B., Martín, C.A., Martin, P.A., Mateos-Molina, D., McConnaughey, R.A., Meroni, M., Meyer, C.F.J., Mills, K., Montefalcone, M., Noreika, N., Palacín, C., Pande, A., Pitcher, C.R., Ponce, C., Rinella, M., Rocha, R., Ruiz-Delgado, M.C., Schmitter-Soto, J.J., Shaffer, J.A., Sharma, S., Sher, A.A., Stagnol, D., Stanley, T.R., Stokesbury, K.D.E., Torres, A., Tully, O., Vehanen, T., Watts, C., Zhao, Q. and Sutherland, W.J. (2020a) “Quantifying and addressing the prevalence and bias of study designs in the environmental and social sciences,” *Nature Communications*, 11(1), p. 6377.

Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R. and Sutherland, W.J. (2020b) “Poor availability of context-specific evidence hampers decision-making in conservation,” *Biological Conservation*, 248 Cold Spring Harbor Laboratory, p. 108666.

Christie, A.P., Amano, T., Martin, P.A., Petrovan, S.O., Shackelford, G.E., Simmons, B.I., Smith, R.K., Williams, D.R., Wordley, C.F.R. and Sutherland, W.J. (2020c) “The challenge of biased evidence in conservation," *Conservation Biology*, Wiley for Society for Conservation Biology, p. cobi.13577.

Christie, A.P., Amano, T., Martin, P.A., Shackelford, G.E., Simmons, B.I. and Sutherland, W.J. (2019) "Simple study designs in ecology produce inaccurate estimates of biodiversity responses," *Journal of Applied Ecology*, 56(12) John Wiley & Sons, Ltd (10.1111), pp. 2742–2754.

Christie, A.P., White, T.B., Martin, P.A., Petrovan, S.O., Bladon, A.J., Bowkett, A.E., Littlewood, N.A., Mupepele, A.C., Rocha, R., Sainsbury, K.A., Smith, R.K., Taylor, N.G., Sutherland, W.J., 2021. Reducing publication delay to improve the efficiency and impact of conservation science. PeerJ 9, e12245. https://doi.org/10.7717/PEERJ.12245/SUPP-17

Cornford, R., Deinet, S., de Palma, A., Hill, S.L.L., McRae, L., Pettit, B., Marconi, V., Purvis, A. and Freeman, R. (2021) "Fast, scalable, and automated identification of articles for biodiversity and macroecological datasets," *Global Ecology and Biogeography*, 30(1) John Wiley & Sons, Ltd, pp. 339–347.

Davies, G.M. and Gray, A. (2015) "Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring)," *Ecology and Evolution*, 5(22) John Wiley & Sons, Ltd, pp. 5295–5304.

Dimick, J.B. and Ryan, A.M. (2014) "Methods for Evaluating Changes in Health Care Policy," *JAMA*, 312(22), p. 2401.

Doi, S.A.R., Barendregt, J.J., Khan, S., Thalib, L. and Williams, G.M. (2015a) "Advances in the meta-analysis of heterogeneous clinical trials II: The quality effects model," *Contemporary Clinical Trials*, 45 Elsevier Inc., pp. 123–129.

Doi, S.A.R., Barendregt, J.J., Khan, S., Thalib, L. and Williams, G.M. (2015b) "Simulation Comparison of the Quality Effects and Random Effects Methods of Meta-analysis," *Epidemiology*, 26(4) Lippincott Williams & Wilkins, pp. e42–e44.

Doi, S.A.R., Barendregt, J.J. and Mozurkewich, E.L. (2011) “Meta-analysis of heterogeneous clinical trials: An empirical example,” *Contemporary Clinical Trials*, 32(2) Elsevier Inc., pp. 288–298.

Doi, S.A.R., Furuya-Kanamori, L., Thalib, L. and Barendregt, J.J. (2017) “Meta-analysis in evidence-based healthcare: A paradigm shift away from random effects is overdue,” *International Journal of Evidence-Based Healthcare*, 15(4) Lippincott Williams and Wilkins, pp. 152–160. Available at: 10.1097/XEB.0000000000000125 (Accessed: January 25, 2021).

Doi, S.A.R. and Thalib, L. (2008) “A quality-effects model for meta-analysis,” *Epidemiology*, 19(1), pp. 94–100.

Doncaster, C.P., Spake, R., 2018. Correction for bias in meta-analysis of little-replicated studies. Methods Ecol. Evol. 9, 634–644. https://doi.org/https://doi.org/10.1111/2041-210X.12927

Elliott, J., Lawrence, R., Minx, J.C., Oladapo, O.T., Ravaud, P., Tendal Jeppesen, B., Thomas, J., Turner, T., Vandvik, P.O., Grimshaw, J.M., 2021. Decision makers need constantly updated evidence synthesis. Nat. 2021 6007889 600, 383–385. https://doi.org/10.1038/d41586-021-03690-1

Gough, D. and White, H. (2018) *Evidence standards and evidence claims in web based research portals*. London.

Grainger, M.J., Bolam, F.C., Stewart, G.B. and Nilsen, E.B. (2019) “Evidence synthesis for tackling research waste,” *Nature Ecology & Evolution*, Springer US

Gutzat, F. and Dormann, C.F. (2020) “Exploration of Concerns about the Evidence-Based Guideline Approach in Conservation Management: Hints from Medical Practice,” *Environmental Management*, 66(3), pp. 435–449.

Haddaway, N.R. and Westgate, M.J. (2019) “Predicting the time needed for environmental systematic reviews and systematic maps,” *Conservation Biology*, 33(2) John Wiley & Sons, Ltd, pp. 434–443.

Hamman, E.A., Pappalardo, P., Bence, J.R., Peacor, S.D. and Osenberg, C.W. (2018) “Bias in meta-analyses using Hedges’ d,” *Ecosphere*, 9(9) John Wiley & Sons, Ltd, p. e02419.

Jenicek, M. (1989) “Meta-analysis in medicine Where we are and where we want to go,” *Journal of Clinical Epidemiology*, 42(1), pp. 35–44.

Johnson, B.T., Low, R.E. and MacDonald, H. v. (2015) “Panning for the gold in health research: Incorporating studies’ methodological quality in meta-analysis,” *Psychology and Health*, 30(1) Routledge, pp. 135–152. Available at: 10.1080/08870446.2014.953533 (Accessed: January 25, 2021).

Junker, J., Petrovan, S.O., Arroyo-RodrÍguez, V., Boonratana, R., Byler, D., Chapman, C.A., Chetry, D., Cheyne, S.M., Cornejo, F.M., CortÉs-Ortiz, L., Cowlishaw, G., Christie, A.P., Crockford, C., Torre, S.D. la, de Melo, F.R., Fan, P., Grueter, C.C., GuzmÁn-Caro, D.C., Heymann, E.W., Herbinger, I., Hoang, M.D., Horwich, R.H., Humle, T., Ikemeh, R.A., Imong, I.S., Jerusalinsky, L., Johnson, S.E., Kappeler, P.M., Kierulff, M.C.M., KonÉ, I., Kormos, R., Le, K.Q., Li, B., Marshall, A.J., Meijaard, E., Mittermeier, R.A., Muroyama, Y., Neugebauer, E., Orth, L., Palacios, E., Papworth, S.K., Plumptre, A.J., Rawson, B.M., Refisch, J., Ratsimbazafy, J., Roos, C., Setchell, J.M., Smith, R.K., Sop, T., Schwitzer, C., Slater, K., Strum, S.C., Sutherland, W.J., Talebi, M., Wallis, J., Wich, S., Williamson, E.A., Wittig, R.M. and KÜhl, H.S. (2020) “A Severe Lack of Evidence Limits Effective Conservation of the World’s Primates,” *BioScience*, 70(9), pp. 794–803.

Kneale, D., Thomas, J., O’Mara-Eves, A. and Wiggins, R. (2019) “How can additional secondary data analysis of observational data enhance the generalisability of meta-analytic evidence for local public health decision making?,” *Research Synthesis Methods*, 10(1), pp. 44–56.

Koricheva, J. and Kulinskaya, E. (2019) “Temporal Instability of Evidence Base: A Threat to Policy Making?,” *Trends in Ecology & Evolution*, 34(10), pp. 895–902.

Larsen, P.O. and von Ins, M. (2010) “The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index,” *Scientometrics*, 84(3), pp. 575–603.

Lortie, C.J., Stewart, G., Rothstein, H. and Lau, J. (2015) “How to critically read ecological meta-analyses,” *Research Synthesis Methods*, 6(2) John Wiley & Sons, Ltd, pp. 124–133.

Marshall, I.J., Johnson, B.T., Wang, Z., Rajasekaran, S. and Wallace, B.C. (2020) “Semi-Automated evidence synthesis in health psychology: current methods and future prospects,” *Health Psychology Review*, 14(1) Routledge, pp. 145–158.

Marshall, I.J., Kuiper, J. and Wallace, B.C. (2015) “Automating Risk of Bias Assessment for Clinical Trials,” *IEEE Journal of Biomedical and Health Informatics*, 19(4), pp. 1406–1412.

Marshall, I.J. and Wallace, B.C. (2019) “Toward systematic review automation: a practical guide to using machine learning tools in research synthesis,” *Systematic Reviews*, 8(1), p. 163.

Mupepele, A.-C., Walsh, J.C., Sutherland, W.J. and Dormann, C.F. (2016) “An evidence assessment tool for ecosystem services and conservation studies,” *Ecological Applications*, 26(5), pp. 1295–1301.

Mupepele, A.-C., Keller, M., Dormann, C.F., (2021). European agroforestry has no unequivocal effect on biodiversity: a time-cumulative meta-analysis. BMC Ecol. Evol. 21, 193. https://doi.org/10.1186/s12862-021-01911-9

Nakagawa, S., Dunn, A.G., Lagisz, M., Bannach-Brown, A., Grames, E.M., Sánchez-Tójar, A., O’Dea, R.E., Noble, D.W.A., Westgate, M.J., Arnold, P.A., Barrow, S., Bethel, A., Cooper, E., Foo, Y.Z., Geange, S.R., Hennessy, E., Mapanga, W., Mengersen, K., Munera, C., Page, M.J., Welch, V., Carter, M., Forbes, O., Furuya-Kanamori, L., Gray, C.T., Hamilton, W.K., Kar, F., Kothe, E., Kwong, J., McGuinness, L.A., Martin, P., Ngwenya, M., Penkin, C., Perez, D., Schermann, M., Senior, A.M., Vásquez, J., Viechtbauer, W., White, T.E., Whitelaw, M., Haddaway, N.R. and Participants, E.S.H. 2019 (2020) “A new ecosystem for evidence synthesis,” *Nature Ecology & Evolution*, 4(4), pp. 498–501.

O’Connor, A.M., Tsafnat, G., Gilbert, S.B., Thayer, K.A. and Wolfe, M.S. (2018) “Moving toward the automation of the systematic review process: a summary of discussions at the second meeting of International Collaboration for the Automation of Systematic Reviews (ICASR)," *Systematic Reviews*, 7(1), p. 3.

Pattanittum, P., Laopaiboon, M., Moher, D., Lumbiganon, P. and Ngamjarus, C. (2012) "A Comparison of Statistical Methods for Identifying Out-of-Date Systematic Reviews," *PLOS ONE*, 7(11) Public Library of Science, p. e48894.

Rhodes, K.M., Savović, J., Elbers, R., Jones, H.E., Higgins, J.P.T., Sterne, J.A.C., Welton, N.J. and Turner, R.M. (2020) "Adjusting trial results for biases in meta-analysis: combining data-based evidence on bias with detailed trial assessment," *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 183(1) John Wiley & Sons, Ltd, pp. 193–209.

Rhodes, K.M., Turner, R.M., Savović, J., Jones, H.E., Mawdsley, D. and Higgins, J.P.T. (2018) "Between-trial heterogeneity in meta-analyses may be partially explained by reported design characteristics," *Journal of Clinical Epidemiology*, 95, pp. 45–54.

Schafft, M., Wegner, B., Meyer, N., Wolter, C., Arlinghaus, R., (2021). Ecological impacts of water-based recreational activities on freshwater ecosystems: a global meta-analysis. Proc. R. Soc. B Biol. Sci. 288, 20211623. https://doi.org/10.1098/rspb.2021.1623

Shackelford, G.E., Kemp, L., Rhodes, C., Sundaram, L., HÉigeartaigh, S.S.Ó., Beard, S., Belfield, H., Weitzdörfer, J., Avin, S., Sørebø, D., Jones, E.M., Hume, J.B., Price, D., Pyle, D., Hurt, D., Stone, T., Watkins, H., Collas, L., Cade, B.C., Johnson, T.F., Freitas-Groff, Z., Denkenberger, D., Levot, M. and Sutherland, W.J. (2019) "Accumulating evidence using crowdsourcing and machine learning: a living bibliography about existential risk and global catastrophic risk," *Futures*, , p. 102508.

Shackelford, G.E., Martin, P.A., Hood, A.S.C., Christie, A.P., Kulinskaya, E. and Sutherland, W.J. (2021). Dynamic meta-analysis: a method of using global evidence for local decision making. BMC Biol 19(33)f. https://doi.org/10.1186/s12915-021-00974-w

Shojania, K.G., Sampson, M., Ansari, M.T., Ji, J., Doucette, S. and Moher, D. (2007) “How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis,” *Annals of Internal Medicine*, 147(4) American College of Physicians, pp. 224–233.

Slavin, R.E. (1995) “Best evidence synthesis: An intelligent alternative to meta-analysis,” *Journal of Clinical Epidemiology*, 48(1), pp. 9–18.

Slavin, R.E. (1986) “Best-Evidence Synthesis: An Alternative to Meta-Analytic and Traditional Reviews,” *Educational Researcher*, 15(9), pp. 5–11.

Stone, J., Gurunathan, U., Glass, K., Munn, Z., Tugwell, P. and Doi, S.A.R. (2019) “Stratification by quality-induced selection bias in a meta-analysis of clinical trials,” *Journal of Clinical Epidemiology*, 107 Elsevier USA, pp. 51–59. Available at: 10.1016/j.jclinepi.2018.11.015 (Accessed: January 25, 2021).

Stone, J.C., Glass, K., Munn, Z., Tugwell, P. and Doi, S.A.R. (2020) “Comparison of bias adjustment methods in meta-analysis suggests that quality effects modeling may have less limitations than other approaches,” *Journal of Clinical Epidemiology*, 117 Elsevier Inc, pp. 36–45.

Sutherland, W.J. and Wordley, C.F.R. (2017) “Evidence complacency hampers conservation,” *Nature Ecology & Evolution*, 1(9), pp. 1215–1216.

Sutherland, W.J. and Wordley, C.F.R. (2018) “A fresh approach to evidence synthesis comment,” *Nature*, 558(7710), pp. 364–366.

The Royal Society (2020) *Face masks and coverings for the general public: Behavioural knowledge, effectiveness of cloth coverings and public messaging*.

Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., Elliott, J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., Hodder, R., Pestridge, C., Albrecht, L., Horsley, T., Platt, J., Armstrong, R., Nguyen, P.H., Plovnick, R., Arno, A., Ivers, N., Quinn, G., Au, A., Johnston, R., Rada, G., Bagg, M., Jones, A., Ravaud, P., Boden, C., Kahale, L., Richter, B., Boisvert, I., Keshavarz, H., Ryan, R., Brandt, L., Kolakowsky-Hayner, S.A., Salama, D., Brazinova, A., Nagraj, S.K., Salanti, G., Buchbinder, R., Lasserson, T., Santaguida, L., Champion, C., Lawrence, R., Santesso, N., Chandler, J., Les, Z., Schünemann, H.J., Charidimou, A., Leucht, S., Shemilt, I., Chou, R., Low, N., Sherifali, D., Churchill, R., Maas, A., Siemieniuk, R., Cnossen, M.C., MacLehose, H., Simmonds, M., Cossi, M.-J., Macleod, M., Skoetz, N., Counotte, M., Marshall, I., Soares-Weiser, K., Craigie, S., Marshall, R., Srikanth, V., Dahm, P., Martin, N., Sullivan, K., Danilkewich, A., Martínez García, L., Synnot, A., Danko, K., Mavergames, C., Taylor, M., Donoghue, E., Maxwell, L.J., Thayer, K., Dressler, C., McAuley, J., Thomas, J., Egan, C., McDonald, S., Tritton, R., Elliott, J., McKenzie, J., Tsafnat, G., Elliott, S.A., Meerpohl, J., Tugwell, P., Etxeandia, I., Merner, B., Turgeon, A., Featherstone, R., Mondello, S., Turner, T., Foxlee, R., Morley, R., van Valkenhoef, G., Garner, P., Munafo, M., Vandvik, P., Gerrity, M., Munn, Z., Wallace, B., Glasziou, P., Murano, M., Wallace, S.A., Green, S., Newman, K., Watts, C., Grimshaw, J., Nieuwlaat, R., Weeks, L., Gurusamy, K., Nikolakopoulou, A., Weigl, A., Haddaway, N., Noel-Storr, A., Wells, G., Hartling, L., O'Connor, A., Wiercioch, W., Hayden, J., Page, M., Wolfenden, L., Helfand, M., Pahwa, M., Yepes Nuñez, J.J., Higgins, J., Pardo, J.P., Yost, J., Hill, S. and Pearson, L. (2017) “Living systematic reviews: 2. Combining human and machine effort,” *Journal of Clinical Epidemiology*, 91 Elsevier, pp. 31–37.

Tsafnat, G., Dunn, A., Glasziou, P. and Coiera, E. (2013) “The automation of systematic reviews,” *BMJ*, 346(jan10 1), pp. f139–f139.

Tugwell, P. and Haynes, R.B. (2006) “Assessing claims of causation,” *Clinical epidemiology: how to do clinical practice research*, Lippincott Williams & Wilkins Philadelphia, PA, pp. 356–387.

Turner, R.M., Spiegelhalter, D.J., Smith, G.C.S. and Thompson, S.G. (2009) “Bias modelling in evidence synthesis,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1) John Wiley & Sons, Ltd, pp. 21–47.

Underwood, A.J. (1991) “Beyond BACI: Experimental designs for detecting human environmental impacts on temporal variations in natural populations,” *Marine and Freshwater Research*, 42(5), pp. 569–587.

Wallace, B.C., Dahabreh, I.J., Schmid, C.H., Lau, J. and Trikalinos, T.A. (2014) “Modernizing Evidence Synthesis for Evidence-Based Medicine,” in Greenes, R. A. B. T.-C. D. S. (Second E. (ed.) *Clinical Decision Support*. Oxford: Elsevier, pp. 339–361.

Wauchope, H.S., Amano, T., Geldmann, J., Johnston, A., Simmons, B.I., Sutherland, W.J. and Jones, J.P.G. (2020) “Evaluating Impact Using Time-Series Data,” *Trends in Ecology & Evolution*