Advice for improving the reproducibility of data extraction in meta-analysis

Edward R. Ivimey-Cook[1*], Daniel W. A. Noble[2], Shinichi Nakagawa[3], Marc J. Lajeunesse[4] & Joel L. Pick[5]

[1] School of Biodiversity, One Health and Veterinary Medicine, University of Glasgow, Glasgow, G12 8QQ, UK

[2] Division of Ecology and Evolution, Research School of Biology, The Australian National University, Canberra ACT 2600

[3] Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, The University of New South Wales, Sydney, NSW 2052

[4] Department of Integrative Biology, University of South Florida, Tampa, FL USA

[5] Institute of Ecology and Evolution, University of Edinburgh, Charlotte Auerbach Road, Edinburgh, EH9 3FL, UK

[*] Corresponding author, email address: e.ivimeycook@googlemail.com

Running title: Reproducibility in meta-analysis

Keywords: reproducibility, meta-analysis, data extraction, metaDigitise, juicr, shiny-Digitise

Abstract

Extracting data from studies is the norm in meta-analyses, enabling researchers to generate effect sizes when raw data are otherwise not available. While there has been a general push for increased reproducibility in meta-analysis, the transparency and reproducibility of the data extraction phase is still lagging behind. Unfortunately, there is little guidance of how to make this process more transparent and shareable. To address this, we provide several steps to help increase the reproducibility of data extraction in meta-analysis. We also provide suggestions of R software that can further help with reproducible data policies: the *shinyDigitise* and *juicr* packages. Adopting the guiding principles listed here and using the appropriate software will provide a more transparent form of data extraction in meta-analyses.

Introduction

In recent years, there has been a push to increase the reproducibility of meta-analyses (the ability to recreate the same findings if the same project was reconducted; see Ihle *et al.*, 2017), with the expectation that exact search strings, screening steps (e.g. the PRISMA flowchart: Moher *et al.*, 2011; O'Dea *et al.*, 2021), and metadata of accepted papers are included alongside manuscripts. However, unlike the study selection process, the reproducibility of steps taken during data extraction is typically overlooked, and no unified reporting guidelines currently exist. Indeed, several papers have highlighted the prevalence of errors in meta-analysis, particularly surrounding the data extraction process (Gøtzsche *et al.*, 2007; Mathes *et al.*, 2017; Wong & Bouchard, 2022). As a result, if studies provide neither the data needed to reproduce the analysis nor the source of the effect size within the screened study (e.g. in text, table or figure, reporting of which is typically low; see O'Dea *et al.*, 2021)), then there can be a lack of repeatability, where independent screeners are unable to locate and extract the same values (see Buscemi *et al.*, 2006). Altogether, this suggests that this vital stage of the meta-analysis workflow lacks both transparency and, importantly, reproducibility.

Here, to assess the extent of problems with data extraction reporting, we review the current state of the literature. Firstly, we review the evidence of reporting of data extraction software in recent meta-analyses in Ecology and Evolution. Secondly, we investigate the reporting practices of papers that have cited the R package *metaDigitise* as a case study. We then introduce a simple five-step guide to help improve the replicability and reproducibility of data extraction. We note that this will not reduce user-specific errors made during the data extraction process, but will enable a higher probability of detecting and correcting any errors made. Finally, we introduce two R-based GUI packages, *shinyDigitise* and *juicr*, which have both been designed to aid transparency and reproducibility.

State of the field

To start, we quantified the percentage of meta-analyses that reported any software packages used to extract data from figures. To do this, Y.Y., M.L., and J.R. re-examined the 102 meta-analyses reviewed in the 2021 PRISMA-EcoEvo guidelines paper (O'Dea *et al.*, 2021). From these 102 studies, only 39 cited the data extraction software that was used to extract data from figures (representing 38% of the total number). We note that whilst this survey and results focus on meta-analysis within the fields of Ecology and Evolution, no such survey has yet been conducted in other disciplines despite the common nature of figure-based data extraction.

Next, to assess transparency of the data extraction process itself, E.I-C reviewed all studies listed as citing the R package *metaDigitise* (Pick *et al.*, 2019) in August 2022 (for full methodology, see SM1). The *metaDigitise* package (on CRAN in 2018, associated paper published in 2019) was in part designed to help improve transparency and reproducibility of data extraction (Pick *et al.*, 2019). It provides a simple way of storing figures and associated extraction data which can easily be uploaded as part of the data archiving process. Papers citing *metaDigitise* therefore provide a good insight into the transparency of data-extraction reporting in recently published meta-analyses. In total, 55 published meta-analyses were obtained that covered several subject areas, ecology and evolution, medicine, environmental science, and psychology.

The results of this survey are shown in Figure S1. 78% of the 55 meta-analyses using metaDigitise (n = 43) had available data in an interoperable format, despite the open access policy of many journals and increased awareness of the importance of open-data. From these, only 24 (44% of the total) readily provided information about the origin of the effect sizes which is in line with the 39% reported from a recent survey in ecology and evolution meta-analyses (O'Dea *et al.*, 2021). Of these studies between 2-96% (median = 28%) of all effect sizes were generated from figures. Finally, only four studies (7%) provided the figures from which data was extracted and only two provided the calibration data needed to recreate the extraction (5%) in addition to the figure and metadata required to reproduce the analysis (Figure S1). The low reporting rates are even more extreme when one considers only 38% of meta-analyses reviewed by O'Dea *et al.* (2021) reported the software used to extract these effect sizes from figures.

Advice for data extraction

Based on this survey it is clear that we need to improve the transparency and reproducibility of data extraction in meta-analyses. To achieve this, we introduce a simple five-step guide.

1. **Provide data.** As discussed at length elsewhere (Miyakawa, 2020), providing data is a minimum requirement for reproducibility. We found that 78% of meta-analyses provide data, similar to the 77% in a recent review of ecology and evolution meta-

analyses (2010 - 2019; O'Dea *et al.*, 2021). Although this shows an improvement over the last decade (from 31% shared data in Ecology meta-analyses between the years 1996–2013 Koricheva & Gurevitch, 2014), and is substantially greater than in other fields (e.g. 3% of studies provided interoperable data in clinical psychological meta-analyses from 2000–2020; López-Nicolás *et al.*, 2022), data in meta-analysis typically come from open sources (i.e. published literature) and so there are few obvious reasons why data should not be made public. Meta-analysts should therefore be expected to lead by example and provide their own data.

2. **Clearly state where each effect size was extracted.** In addition to providing other relevant metadata, it should be clearly stated where effect sizes were extracted from (e.g. text, table, figure or supplementary material), including a reference to the exact location, e.g. "Figure 2a", "table 3", "main text p275". Curtis *et al.* (2013) suggested a shorthand for reporting this information in tabular form (e.g., F2a, T3), and we extend this formatting to T=table, M=main text, F=figure, A=appendix, S=supplementary material, R=raw data, followed by the figure and/or page number where the data was extracted. In addition to providing copies of the extracted figures, uploading a screenshot or section of PDF which clearly highlights the location of the extracted effect size would be useful, particularly when considering data in text or in table (although note the caveats listed below). Lastly, under some circumstances, data might be provided from unpublished studies through personal contact with authors. In this case, it is still important to provide a location of where or how the effect size was obtained (i.e. personal communication or unpublished data), in order to allow for others to similarly obtain the data.

3. **Provide transformation information.** Providing effect sizes alone does not give information on how they have been generated. For example, transformations have to be used to generate means and standard deviations from the quantiles in a boxplot (e.g. Wan *et al.*, 2014). Other transformations include converting stan-

dard errors (SE) to standard deviations (SD), or calibrations of extracted data by back-transforming logarithms. Generating effect sizes from figures always requires additional steps in order to make them usable in meta-analysis. These details are more challenging to report succinctly, as they may require equations, but a textual description alongside raw data and code is better than nothing. Indeed, O'Dea *et al.* (2021) showed that only 39% of papers provided the raw data used to generated effect sizes, compared with the 77% that provided processed effect sizes.

4. **Provide figures alongside a record of the data extraction process.** A considerable amount of data for meta-analysis comes from figures (e.g., in the above survey, 28% of effect sizes, on average, originated from figures). Therefore, every figure that has undergone data extraction should be provided in a digital data repository (e.g. Open Science Framework, Zenodo, or Dryad) alongside the generated effect size. Data extraction files including calibration data are also needed for any researcher to be able to recreate and check the extraction process. Importantly, it is also worth considering (and noting in the metadata) whether the source paper was open or non-open access. Whilst a breach of copyright may not be an issue with figures from open access papers, this could be a potential problem with non-open access papers. In this case, we suggest three actions: 1) note in the metadata which figures might be restricted due to copyright infringement; 2) seek permission from the journal and/or author of the paper; 3) store all of the figures on a private repository (such as those listed above) which can be made available upon request. It is also a requirement, regardless of whether the paper is open or non-open access, to appropriately cite the primary literature where the figure has been obtained.

5. **For software developers, enable the saving and reloading of the data extraction process.** Whilst there exists a multitude of data extraction tools, few allow users to easily save and reload the data extraction process. Therefore, to increase reproducibility, the development of new tools or software for data extraction

should ensure this functionality. The file format of extractions should be also tool agnostic with a format accessible to all (interoperable; e.g., a .csv file).

Tools for increasing reproducibility in figure-based extraction

Here we highlight two R-based packages that are being developed that allow for reproducible figure-based data extraction. Firstly, *shinyDigitise*, a GUI for the *metaDigitise* (Pick *et al.*, 2019) package, and secondly, *juicr* (Lajeunesse, 2021). We focus on these packages because R is one of the most widely used statistical environments for analysing meta-analytic data. We note that whilst these packages should be suitable for extraction of many of the commonly used figures across disciplines (scatterplots, mean-error plots, boxplots, and histograms), they may not be as well equipped to extract data from highly specialised domain-specific figures.

shinyDigitise (developed by E.I-C & J.L.P) is a streamlined and intuitive GUI interface which is built upon the functions of the *metaDigitise* package (Pick *et al.*, 2019). This includes the ability to extract data from a wide variety of plot-types (scatterplots, mean-error plots, boxplots, and histograms), and automatically saves calibration data so users have a historical record of the data extraction process. *shinyDigitise* should reduce the barrier of entry by requiring very little experience of writing code or the R coding software. To install this package, see the GitHub: https://github.com/EIvimeyCook/shinyDigitise.

Alongside *shinyDigitise*, *juicr* (developed by M.J.L.) offers savable and shareable records of retrieved data from images. *juicr* offers a point-and-click solution to extracting data from images; however for some tasks, decision-making of what to extract can be delegated to automated (full algorithmic) or semi-automated (algorithmic with user assistance) tools. The *juicr* package extends the automated extraction tools first developed in the *metagear* package for research synthesis (Lajeunesse, 2016); to install this package, see the GitHub: https://github.com/mjlajeunesse/juicr.

Importantly, these software packages provide the user with an effect size *in addition* to a record of the extraction process for each figure. After depositing into an appropriate data repository, these can be subsequently viewed and error checked by the user or by anyone with access to both the figure and record files. Whilst this is an important step for reproducibility, and directly adheres to step four above, very few people have adopted the use of this archiving functionality. Figure S1 highlights the low percentage of studies that share source figures, their extracted data, and information as to when and what extraction software tool was used, in addition to providing records of the data extraction process. Clearly there is an urgent need to increase transparency of data extraction, and the steps outlined above should go some way to addressing this.

Acknowledgements

We are grateful to Malgorzata Lagisz, Yefeng Yang and Joanna Rutkowska who surveyed the 102 meta-analyses as a part of a larger project. We also thank two anonymous reviewers for helpful comments on the manuscript. Lastly, we thank Stuart Taylor from the Royal Society for guidance on issues with copyright and non-open access papers. Note, we refer to authors in text using the MeRIT system (Method Reporting with Initials for Transparency) as per (Nakagawa *et al.*, 2023).

Data Availability Statement

The data that support the findings of this study are openly available on GitHub: https://github.com/EIvimeyCook/DataExtraction or Zenodo: 10.5281/zenodo.8187175

Conflict of Interest

The authors declare no conflict of interest.

Highlights

- In meta-analysis, large quantities of data need to be extracted from published literature. However, the transparency and reproducibility of the data extraction process is often limited, both in terms of its description in the methods section and also when data is later uploaded to an open data repository.
- In order to increase the reproducibility of data extraction in meta-analysis, we introduce a simple five-step guide which includes suggestions for future research. Furthermore, we highlight two packages in R that readily facilitate reproducible workflows and allow for shareable records of the data extraction process.
- Adopting the principles and suggestions provided here will help to make the entire meta-analysis process more transparent, open, and reproducible.

References

Buscemi, N., Hartling, L., Vandermeer, B., Tjosvold, L. & Klassen, T.P. (2006) Single data extraction generated more errors than double data extraction in systematic reviews. *Journal of Clinical Epidemiology*, **59**, 697–703. https://dx.doi.org/https://doi.org/10.1016/j.jclinepi.2005.11.010.

Curtis, P.S., Mengersen, K., Lajeunesse, M.J., Rothstein, H.R. & Stewart, G.B. (2013) Extraction and critical appraisal of data. J. Koricheva, J. Gurevitch & K. Mengersen, eds., *Handbook of meta-analysis in ecology and evolution*, pp. 52–60. Princeton University Press, Princeton, New Jersey, USA.

Gøtzsche, P.C., Hróbjartsson, A., Marić, K. & Tendal, B. (2007) Data Extraction Errors in Meta-analyses That Use Standardized Mean Differences. *JAMA*, **298**. https://dx.doi.org/10.1001/jama.298.4.430.

Ihle, M., Winney, I.S., Krystalli, A. & Croucher, M. (2017) Striving for transparent and credible research: practical guidelines for behavioral ecologists. *Behavioral Ecology*, **28**, 348–354.

Koricheva, J. & Gurevitch, J. (2014) Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, **102**, 828–844. ISBN: 1365-2745, https://dx.doi.org/10.1111/1365-2745.12224.

Lajeunesse, M.J. (2016) Facilitating systematic reviews, data extraction and meta-analysis with the metagear package for R. *Methods in Ecology and Evolution*, **7**, 323–330.

Lajeunesse, M.J. (2021) Squeezing data from scientific images using the juicr package for R. R package version 0.1.

López-Nicolás, R., López-López, J.A., Rubio-Aparicio, M. & Sánchez-Meca, J. (2022) A meta-review of transparency and reproducibility-related reporting practices in published meta-analyses on clinical psychological interventions (2000–2020). *Behavior Research Methods*, **54**, 334–349. https://dx.doi.org/10.3758/s13428-021-01644-z.

Mathes, T., Klaßen, P. & Pieper, D. (2017) Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. *BMC Medical Research Methodology*, **17**, 152. https://dx.doi.org/10.1186/s12874-017-0431-4.

Miyakawa, T. (2020) No raw data, no science: another possible source of the reproducibility crisis. *Molecular Brain*, **13**, 24. https://dx.doi.org/10.1186/s13041-020-0552-2.

Moher, D., Altman, D.G., Liberati, A. & Tetzlaff, J. (2011) PRISMA statement. *Epidemiology*, **22**, 128.

Nakagawa, S., Ivimey-Cook, E.R., Grainger, M.J., O'Dea, R.E., Burke, S., Drobniak, S.M., Gould, E., Macartney, E.L., Martnig, A.R., Paquet, M., Morrison, K., Pick, J.L., Pottier, P., Ricolfi, L., Wilkinson, D.P., Willcox, A., Williams, C., Wilson, L.A.B., Windecker, S.M., Yang, Y. & Lagisz., M. (2023) Method reporting with initials for transparency (merit) promotes more granularity and accountability for author contributions. *Nature Communications*, **14**, 1788. https://dx.doi.org/10.1038/s41467-023-37039-1.

O'Dea, R.E., Lagisz, M., Jennions, M.D., Koricheva, J., Noble, D.W., Parker, T.H., Gurevitch, J., Page, M.J., Stewart, G., Moher, D. & Nakagawa, S. (2021) Preferred reporting items for systematic reviews and meta-analyses in ecology and evolutionary biology: a PRISMA extension. *Biological Reviews*, **96**, 1695–1722. https://dx.doi.org/10.1111/brv.12721.

Pick, J.L., Nakagawa, S. & Noble, D.W.A. (2019) Reproducible, flexible and high-throughput data extraction from primary literature: The metaDigitise r package. *Methods in Ecology and Evolution*, **10**, 426–431. https://dx.doi.org/10.1111/2041-210X.13118.

Wan, X., Wang, W., Liu, J. & Tong, T. (2014) Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC medical research methodology*, **14**, 135. ISBN: 1471-2288 (Electronic)\r1471-2288 (Linking), https://dx.doi.org/10.1186/1471-2288-14-135.

Wong, J.S. & Bouchard, J. (2022) Do Meta-Analyses of Intervention/Prevention Programs in the Field of Criminology Meet the Tests of Transparency and Reproducibility? *Trauma, Violence, & Abuse*, p. 15248380211073839. https://dx.doi.org/10.1177/15248380211073839.