A framework for improving the reproducibility of data extraction for 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, shiny, metaDigitise

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 throughout the many facets of meta-analysis, the transparency and reproducibility of the data extraction phase are still lagging behind. This particular meta-analytic facet is critical because it facilitates error-checking and enables users to update older meta-analyses. Unfortunately, there is little guidance of how to make the process of data extraction more transparent and shareable, in part this is as a result of relatively few data extraction tools currently offering such functionality. Here, we suggest a simple framework that aims to help increase the reproducibility of data extraction for meta-analysis. We also provide suggestions of software that can further help users adopt open data policies. More specifically, we overview two GUI style software in the R environment, *shinyDigitise* and *juicr*, that both facilitate reproducible workflows while reducing the need for coding skills in R. Adopting the guiding principles listed here and using appropriate software will provide a more streamlined, transparent, and shareable form of data extraction for meta-analyses.

Introduction

In recent years, there has been a push to increase the reproducibility of meta-analyses, 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 now included alongside manuscripts. However, even as transparent reporting is improving over time, studies across a number of research fields have concluded that more still needs to be done to increase reproducibility and produce robust results (Lakens *et al.*, 2016, 2017; Maassen *et al.*, 2020; Polanin *et al.*, 2020; Chow *et al.*, 2021). One of the key methodological steps of meta-analyses is data extraction, but, unlike the study selection process, there is still no unified reporting framework. In addition, although effect sizes are typically associated with a particular study when data are available, the source of the data within that study (e.g. be it from text, table, or figure) used to compute these effect sizes is seldom reported, meaning this process lacks both transparency and, importantly, reproducibility. As a result, 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).

Typically, a large quantity of the data needed for meta-analysis are locked away within figures. Indeed, Simmonds *et al.* (2022) recently showed that in 3%-64% of studies (depending on the field of research), had measures of uncertainty needed for meta-analysis (e.g. standard error or variance) only contained within figures. Consequently, some form of figure-based data extraction software is required (reviewed in Pick *et al.*, 2019). However, often the data extraction process is mentioned simply as a sentence within a study's methodology. Which tool was used to extract data, which effect sizes originated from figures, which particular figures the effect sizes came from, and any record of the extraction process are rarely reported. Ultimately, this severely limits the potential for error-checking and replicability.

Here, to assess the extent of the data extraction problem, we review the current state of the literature. Firstly, by searching recent meta-analyses in Ecology and Evolution for evidence of adequate reporting of data extraction software. Secondly, by investigating papers that cited the R package *metaDigitise* as a case study. We then introduce a simple framework to help improve the reproducibility of data extraction, and in particular data extraction from figures (these should still readily apply to data extraction from other sources, such as tables and text). Finally, we introduce two GUI-based software packages in the R environment (one of the most commonly used statistical languages), *shinyDigitise* and *juicR*, which have both been designed to aid transparency and reproducibility. The use and sharing of resulting files from these programs will facilitate error-checking and updating of meta-analyses into the future.

State of the field

To start, we wanted to quantify what percentage of meta-analyses adequately report software packages used to extract data from figures. To do this, YY, ML, and JR re-examined the 102 meta-analyses reviewed in the 2021 PRISMA-EcoEvo guidelines (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).

Next, to assess transparency of the data extraction process itself, EIC reviewed all studies listed as citing the *metaDigitise* R package (Pick *et al.*, 2019) on Google Scholar in August 2022. Papers citing *metaDigitise* provide a good insight into the transparency of the data-extraction process. The recently published *metaDigitise* package (R package on CRAN in 2018, accompanying 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. Of these 70 published papers, 55 were meta-analyses (the other 15 papers cited were either general meta-analysis guidelines or systematic reviews). From these meta-analyses EIC recorded

a) If the data was open and accessible (interoperable; e.g. as a csv or txt file).
b) Whether the paper reported the origin of the effect size (text/table/figure).
c) Whether figures that had been used in the data extraction process were stored alongside the extracted data.
d) Whether a figure-specific data extraction record was stored alongside the appropriate figure.

The results of this survey are shown in Figure 1. 78% of the 55 meta-analyses (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 in those conducting meta-analysis. From these, only 24 (44% of the total) readily provided information about the origin of the effect sizes. These studies showed the large dependence of data from figures, with the number of effect sizes generated from figures ranging from 1-922 (median = 83) representing between 2-96% (median = 28%) of total effect sizes generated from figures. Finally, only four (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 (see Figure 1 and Table S1). The low reporting rates are even lower if one considers only 40% meta-analyses reported what software package was used to extract effect sizes from figures.

Framework for data extraction

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

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 does not seem many obvious reasons why data should not be made public with the paper. Furthermore, those conducting meta-analyses are in a prime position to understand

the importance of easily accessible data (given the shared experience and agony of extracting information from individual studies). Meta-analysts should therefore be expected to lead by example and provide their own data.

2. **Clearly state where each effect size was extracted from.** In addition to providing the title of the study and relevant additional information, a column in your meta-data file should clearly state whether your effect size was extracted from - text, table, figure or appendix/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 or page number where the data was extracted. This simple step greatly improves transparency, aids reproducibility and helps identify potential errors. Despite this, only 44% of meta-analyses in our sample gave this information, in line with the 44% in ecology and evolution meta-analyses (O'Dea *et al.*, 2021).

3. **Provide information on any transformation that may have been performed on the data.** Just 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 typical conversions include converting standard errors (SE) to standard deviations (SD), or calibrations of extracted data by back-transforming logarithms. These details are more challenging to report succinctly in tabular format, as they may require equations, but a textual description is better than nothing. Providing raw data and code also greatly aid transparency in how effect sizes have been generated, and allows them to be reproduced.

4. **Provide the figures in an open data repository with a record of the data extraction process.** Every figure that has undergone data extraction should be

provided in a digital data repository alongside the generated effect size. However, figures alone don't provide sufficient for reproducibility and error-checking of the data-extraction process. Data extraction files including calibration data are also needed for any researcher to be able to recreate the extraction process.

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, individuals looking to provide tools or software for data extraction should ensure this functionality. To further extend transparency and openness of these data, the file format of extractions should be 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 allow for reproducible figure-based data extraction, *metaDigitise* (Pick *et al.*, 2019) and *juicr* (Lajeunesse, 2021). We also introduce a new GUI-based package, *shinyDigitise*. We focus on these packages because R is one of the most widely used statistical environments for analysing meta-analytic data. Making use of packages already within the R environment will unify data extraction and analysis workflows which will promote transparency and reproducibility.

shinyDigitise is a streamlined and intuitive GUI interface which makes use of functions within the *metaDigitise* package. This includes the ability to extract data from a wide variety of plot-types (Pick *et al.*, 2019), and automatically save calibration data so users have a historical record of the figure-based data extraction process. *shinyDigitise* should reduce the barrier of entry by needing very little experience of writing code and knowledge of the R coding software. Users simply install the *shinyDigitise* package, which contains the functionality of *metaDigitise*, and direct *shinyDigitise* to the folder where images that need extracting are stored with a single line of R code. Using a simple shiny web application, which opens immediately, users progress through the various required steps of data extraction without having to engage with the R console directly. When the user has finished data extraction, a comma-delimited data file (.csv) containing extracted data, along with files containing a record of the data extraction process will automatically be created and ready to share. The raw extracted data can also easily be imported into R for further analysis if so required. The *shinyDigitise* enhancement to the popular *metaDigitise* package adds additional options such as image zooming, and will be more accessible to individuals with varying levels of coding experience in a clearer and cleaner data extraction environment.

Alongside *metaDigitise* and *shinyDigitise*, *juicr* offers savable and shareable records of retrieved data from images. Although being developed independently, this package has also converged on a similar philosophy of archiving the extraction process, but has approached the problem through the perspective of upholding the reproducibility of automated approaches. As in *metaDigitise* and *shinyDigitise*, *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. Although seemingly practical in terms of speed and accuracy, automated tools are neutral in what is extracted from images, and users will often need to parameterise models to target data (e.g., diamonds over circles in scatter plots), supplement missed data points, or cleanup false detections. Further, extraction success is also dependent on image identity—which is highly variable due to the diversity of ways published figures are available (e.g., high/low resolution images from publisher websites, inconsistent image sizes from user-extracted clippings or screenshots from PDFs). All of which are decisions and sources of variability that need to be transparent for reproducibility and validation, and *juicr* aims to embed this more detailed information in reports that are both human- and machine-readable. Moving forward, and given that technology is rapidly growing in this space, transparency and reporting of automated systems should be a core feature, like in *juicr*, in the design of model-assisted extraction tools.

Importantly, all these software packages provide a user with both an effect size and a record for each figure they choose to extract data from. The extraction process and data are archived and, after depositing on an appropriate data repository, can be subsequently viewed and error checked by the user or by anyone with access to both files. Whilst this is an important step for reproducibility, and directly adheres to point four on the framework above, very few people have adopted the use this archiving functionality. Figure 1 exemplifies this by highlighting the low percentage of studies that share source figures, their extracted data, and information as to when the extraction software tool was used, in addition to providing historical records of data extraction. Clearly there is an urgent need to increase transparency of data extraction, and the framework outlined above would go some way to addressing this. However, not only is this a data sharing issue, but it is also a functionality issue. As such, one of the chief priorities for future extraction software should be reproducibility, with the ability to save and reload the data generating process (i.e., the accuracy of extraction) and not just focus on the data itself.

Finally, it is our hope that the simple data extraction framework outlined in this article can serve as an important first step to help close the reproducibility gap that exists for this facet of meta-analysis. This general framework can be readily adopted by the many research areas that conduct meta-analysis and by doing so, will allow for greater reproducibility across disciplines.

Acknowledgement

We are grateful to Malgorzata Lagisz, Yefeng Yang and Joanna Rutkowska who surveyed the 102 meta-analyses as a part of a larger project.

References

Chow, J.C., Sjogren, A.L. & Zhao, H. (2021) Reporting and Reproducibility of Meta-Analysis in Speech, Language, and Hearing Research. *Journal of Speech, Language, and Hearing Research*, **64**, 2786–2793.

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.

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

Lakens, D., Page-Gould, E., Assen, M.A.L.M.v., Spellman, B., Schönbrodt, F., Hasselman, F., Corker, K.S., Grange, J.A., Sharples, A., Cavender, C., Augusteijn, H.E.M., Augusteijn, H., Gerger, H., Locher, C., Miller, I.D., Anvari, F. & Scheel, A.M. (2017) Examining the Reproducibility of Meta-Analyses in Psychology: A Preliminary Report.

Lakens, D., Hilgard, J. & Staaks, J. (2016) On the reproducibility of meta-analyses: six practical recommendations. *BMC Psychology*, **4**, 24. https://dx.doi.org/10.1186/s40359-016-0126-3.

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.

Maassen, E., Assen, M.A.L.M.v., Nuijten, M.B., Olsson-Collentine, A. & Wicherts, J.M. (2020) Reproducibility of individual effect sizes in meta-analyses in psychology. *PLOS ONE*, **15**, e0233107. https://dx.doi.org/10.1371/journal.pone.0233107.

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.

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.

Polanin, J.R., Hennessy, E.A. & Tsuji, S. (2020) Transparency and Reproducibility of Meta-Analyses in Psychology: A Meta-Review. *Perspectives on Psychological Science*, **15**, 1026–1041. https://dx.doi.org/10.1177/1745691620906416.

Simmonds, E.G., Adjei, K.P., Andersen, C.W., Aspheim, J.C.H., Battistin, C., Bulso, N., Christensen, H., Cretois, B., Cubero, R., Davidovich, I.A., Dickel, L., Dunn, B., Dunn-Sigouin, E., Dyrstad, K., Einum, S., Giglio, D., Gjerlow, H., Godefroidt, A., Gonzalez-Gil, R., Cogno, S.G., Grosse, F., Halloran, P., Jensen, M.F., Kennedy, J.J., Langsaether, P.E., Laverick, J.H., Lederberger, D., Li, C., Mandeville, E., Mandeville, C., Moe, E., Schroder, T.N., Nunan, D., Parada, J.S., Simpson, M.R., Skarstein, E.S., Spensberger, C., Stevens, R., Subramanian, A., Svendsen, L., Theisen, O.M., Watret, C. & OHara, R.B. (2022) How is model-related uncertainty quantified and reported in different disciplines? *arXiv preprint*. https://dx.doi.org/10.48550/arXiv.2206.12179.

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.

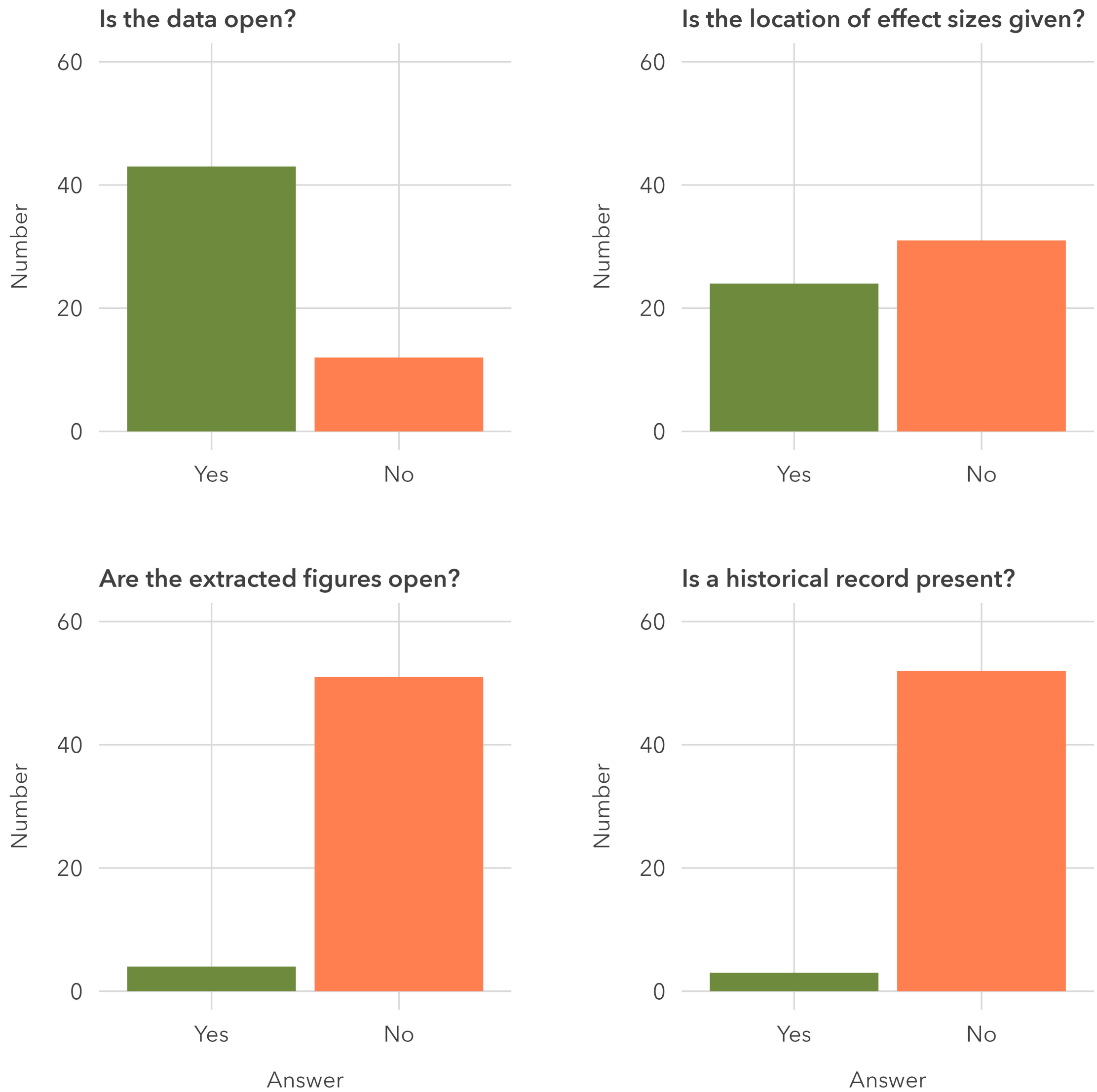

Figure 1: Barplots showing the number of meta-analysis papers that cite metaDigitise that provide open data, denote where effect sizes are located, provide the figures alongside open data, and provide a historical record of the figured-based data extraction process.