Openness and Computational Reproducibility in Plant Pathology: Where do we Stand and a Way Forward

Adam H. Sparks1,2,✉, Emerson M. Del Ponte3, Kaique S. Alves3, Zachary S. L. Foster4, and Niklaus J. Grünwald5

1 Department of Primary Industries and Regional Development, Farming Systems Innovation, Perth WA 6000, Australia  
2 University of Southern Queensland, Centre for Crop Health, Toowoomba Qld 4350, Australia  
3 Departmento de Fitopatologia, Universidade Federal de Viçosa, Brazil  
4 Department of Botany and Plant Pathology, Oregon State University, Corvallis OR 97331, USA  
5 Horticultural Crops Research Unit, USDA Agricultural Research Service, Corvallis OR 97330, USA

✉ Corresponding author: [Adam H. Sparks <[Adam.Sparks@dpird.wa.gov.au](mailto:Adam.Sparks@dpird.wa.gov.au)>](mailto:Adam.Sparks@dpird.wa.gov.au)

**Keywords: Data Science, Computational Biology, Techniques**

# Abstract

Open research practices have been highlighted extensively during the last ten years in many fields of scientific study as essential standards needed to promote transparency and reproducibility of scientific results. Scientific claims can only be evaluated based on how protocols, materials, equipment and methods were described; data were collected and prepared; and, analyses were conducted. Openly sharing protocols, data and computational code is central for current scholarly dissemination and communication, but in many fields, including plant pathology, adoption of these practices has been slow. We randomly selected 450 articles published from 2012 to 2021 across 21 journals representative of the plant pathology discipline and assigned them scores reflecting their openness and computational reproducibility. We found that most of the articles were not following protocols for open science, and were failing to share data or code in a reproducible way. We also propose that use of open-source tools facilitates computationally reproducible work and analyses benefitting not just readers, but the authors as well. Finally, we also provide ideas and tools to promote open, reproducible computational research practices among plant pathologists.

# Introduction

Modern plant pathological research has many facets given the array of disciplines and sub-disciplines currently involved. Collectively, they contribute to increasing our basic and applied knowledge of several aspects of pathogen biology and disease development to ultimately improve plant disease management. Scientific research in the field varies from the purely observational or descriptive nature to inferential, based on experimental or simulation-derived data sets. Whatever the case, research findings are verifiable based on how much of the research materials, processes and outcomes are made available beyond what is reported in the scientific article and the ability of others to make use of the methods and results. These research findings include biological materials (host and pathogen genotypes), nucleic/protein sequences, experimental and simulated raw data annotations, drawings and photographs and statistical analysis code among other materials and data generated during the course of the research.

Recently, open science has been highlighted with many funding agencies expecting data to be available upon the conclusion of the research project (Government of Canada 2016; Noorden 2017; Australian Research Council 2018), journals in the field promoting the sharing of data (Del Ponte 2020), and other scientists interested in sharing and collaborating more widely (Wald 2010).

Reproducibility is one component under the umbrella of open science. By proactively practising open science, the work increases the chance to become more reproducible through the availability of data and code. That is, open science leads to reproducibility and replicability.

For us to easily discuss the topic, we first must define what we mean so that we may clearly communicate. Many of the terms used in this area have varying definitions that may or may not agree with each other. For instance, reproducible research was recently highlighted by many authors (Tiwari et al. 2021; Eckert et al. 2020; Dienlin et al. 2020; Preeyanon et al. 2018; Wallach et al. 2018; Baker 2016; Iqbal et al. 2016; Editors at Nature 2016; Patil et al. 2016; Weissgerber et al. 2016; Brunsdon 2015; Sweedler 2015; Fitzjohn et al. 2014; Ioannidis 2014; Fidler and Gordon 2013; Stodden et al. 2013) as an important issue.

In the biological sciences it is not always possible to use identical test material or perhaps the time or resources are not available for full reproducibility, *e.g.*, field trials that span years and locations or complex glasshouse experiments. Therefore, we will follow Peng’s (2009) definition that provides clear guidelines for a minimum standard of “reproducible research”:

“The replication of scientific findings using independent investigators, methods, data, equipment, and protocols has long been, and will continue to be, the standard by which scientific claims are evaluated. However, in many fields of study there are examples of scientific investigations that cannot be fully replicated because of a lack of time or resources. In such a situation, there is a need for a minimum standard that can fill the void between full replication and nothing. One candidate for this minimum standard is ‘reproducible research’, which requires that data sets and computer code be made available to others for verifying published results and conducting alternative analyses”. Peng (2009). *Reproducible research and Biostatistics*. Biostatistics, 10 (3): 405-408.

Therefore our definition of reproducibility will be that the computer code and data are made freely available to others for verification and conducting alternate analyses or for use in instructional purposes. And that the software used are also easily obtained and preferably open source to avoid licensing or other issues related to accessibility for end-users related to costs or non-standard file formats, etc.

Plant pathologists routinely provide information on protocols and chemicals allowing for reproducibility. However, frequently biological specimens such as strains, cultures or cultivars are not available after publication. These cases do constitute a lack of reproducibility but will not be covered here.

# Materials and Methods

## Article Selection and Evaluation Events

In an effort to understand where as a discipline we stand as plant pathologists regarding open science and reproducible research, we surveyed a broad selection of articles to represent a wide swathe of publications to evaluate our collective status. We hand-picked 21 journals which we felt represented research publications in the field of plant pathology (Table 1) that encompassed a range of subject matter foci, applied and fundamental work, country of origin and ranking metrics, *e.g.* quartile range or citation index, to represent the discipline. The aim was to gather as complete an overview of the status of computational reproducibility in plant pathology journals as possible and avoid bias in the findings by skewing towards high-impact journals that may have a greater influence. From those 21 journals we randomly selected 450 articles published from 2012 to 2021. Using R (R Core Team 2022), two lists were created, the first was a list of the 21 journals and the second was a list of the evaluators that were evaluating the articles for reproducibility. Initially there were four evaluators, later a fifth was added and the list was recreated. There were three evaluation events that took place to increase the yearly coverage of the evaluations. Each evaluator was assigned 50 articles to evaluate during an evaluation event. A list of randomly generated numbers representing page numbers from one to 150, sampled with replacement, was assigned to a randomized list of the 21 journals for each sampling event. This was done because some journals restart their numbering with each issue and also ensures that the journal is more likely to have a page number corresponding to the randomly generated value. This also assumes that there is no effect or bias on reproducibility based on the time of year that an article was published since most journals start with page number one at the beginning of the year.

Articles were then manually selected by visiting the journals’ websites and selecting the articles within which the randomly assigned page numbers fell, *i.e.*, if the page number was 32, the article that started on page 28 and ended on page 35 was selected as it contained page 32. In cases where an article was not suitable, *e.g.*, a review or otherwise not related to plant pathology or the randomly assigned numbers for that journal and year fell within the same article, the next article in that journal was selected until a suitable article was found.

Depending on the scoring event, three to five of this paper’s authors were each assigned to rate a randomised list of journal articles using scoring criteria devised for the purposes of this research in each of the three different scoring events as their time allowed for their participation. In the first event, three authors (AS, EDP, and ZF) evaluated 200 articles, in the second event two authors evaluated 100 articles (EDP, KA) and in the third round, five authors (AS, EDP, KA, ZF, NG) evaluated 150 articles for a total of 450 articles.

## Scoring Criteria

The five-year impact factor for 2022 for each journal was retrieved from InCites Journal Citation Reports, Clarivate Analytics and entered in a separate sheet in the Google sheets file. This was downloaded, saved as an Open Document Spreadsheet (ODS) file for analysis.

Each article was rated on a 0 – 3 scale for its data and code availability. ‘Code availability’ rated how easily and openly available the computational methods used in the article were. Scores were assigned as ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no code was created to conduct the work that was detectable. Second, the ‘Data availability’ evaluated how freely and openly available the data presented in the article were. This was evaluated as ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no data were generated, e.g., a methods paper.

Where possible, the software used in conducting the research behind the publication was recorded in the notes when it was cited or otherwise specified in the article text.

## Data Cleaning

A custom function, import\_notes(), was written to import the data and format the columns properly R (R Core Team 2022). Values for software packages were checked for spelling consistency and corrections were made manually where necessary. When working with these data in R, all strings of software character values were converted to fully upper-case to standardise the capitalisation and alleviate any issues with capitalisation used between evaluators.

## Statistical Analysis

The statistical analysis was performed using R version R version 4.2.0 (2022-04-22) (R Core Team 2022) on an Apple MacBook Pro (13-inch, M1, 2020).

Inter-rater differences were evaluated using percent agreement and Fliess’ Kappa (Del Ponte et al. 2019).

We fitted Bayesian logistic mixed models (estimated using MCMC sampling with 4 chains of 10000 iterations and a warmup of 5000 and thinning of 1) using cumulative family function with a logit link for ordinal data using the contributed package ‘brms’ (version 2.17.0) (Bürkner 2017, 2018, 2021). The cumulative family allows for the probability of a given value is the probability of that value or any smaller value. Priors were selected to be weakly informative and deemed suitable through using pp\_check() to examine the the models’ predictions based on priors only (using the parameter sample\_prior = only in brm()). Priors over all parameters were set as: normal (mean = 0.00, SD = 1.0) distributions for both fixed parameters and random effects. The best fitting models were selected by using Widely Applicable Information Criterion (WAIC) and the expected log pointwise predictive density via loo\_compare().

The models’ were structured as follows to predict two responses, the availability of computational methods and data that were created or collected for the manuscript. To predict the responses with publishing journal, as the journals’ abbreviation, the formula took the form: response ~ abbreviation and the models included assignee as a random effect (formula: ~1 | assignee). To predict the response with year were predicted as (formula: response ~ year) and the models included abbreviation and assignee as random effects (formula: list(~1 | abbreviation, ~1 | assignee)). To predict the responses with IF\_5year the formula took the form response ~ IF\_5year and included assignee as random effect (formula: ~1 | assignee) without journal as a random effect to mitigate correlation between impact factor and the journal of publication.

The adapt\_delta() and max\_treedepth() values were adjusted as necessary on a case-by-case basis for each of the models to ensure a good model fit to the data and that the chains mixed well.

Model fitness’s were evaluated using model summaries and diagnostic plots from ‘brms’ (version 2.17.0) and posterior fits using pp\_check() (Figure 1) from the contributed R package ‘bayesplot’ (version 1.9.0) (Gabry et al. 2019).

A test for practical equivalence, equivalence\_test(), was performed for each model using the contributed R package ‘bayestestR’ (version 0.12.1) (Makowski et al. 2019) with the Region of Practical Equivalence (ROPE) set to -0.1 to 0.1 and confidence interval of 0.95.

Using the Sequential Effect eXistence and sIgnificance Testing (SEXIT) framework, the median of the posterior distribution and its 95% CI (Highest Density Interval), along the probability of direction (pd), the probability of significance and the probability of being large are reported. The thresholds beyond which the effect is considered as significant (i.e., non-negligible) and large are 0.05 and 0.3 as suggested by Makowski et al. (2019). Convergence and stability of the Bayesian sampling was assessed using R-hat, which should be below 1.01 (Vehtari et al. 2021) and Effective Sample Size (ESS), which should be greater than 1000 (Bürkner 2017).

A full report of all parameters and model details are available in the Supplementary Materials. Additionally, all methods are described in greater detail with the code necessary to reproduce the work in Sparks et al. (2022).

# Results

## Inter-rater Agreement

All authors agreed on the five inter-rater article evaluations for the data availability, ‘0’, not available for an inter-rater score of 100%. However, one of the authors rated one article as ‘NA’ or not having any computational methods used rather than ‘0’ as with the other four authors, giving the inter-rater percent agreement a score of 100% with the ‘NA’ value dropped from the code availability.

## Code and Data Sharing Findings

Most articles did not make any computational methods available in any fashion with 4 (0.9%) classified as ‘3’, which was the highest score available (Figure 2A). Four hundred and thirty-five (96.7%) were classed as ‘0’ and ten articles (2%) appeared to not use any computational methods.

Additionally, data that supported the articles were mostly unavailable with 361 (80%) scoring ‘0’ where the data was not available or mentioned in the article. However, more articles, 45 (10%), scored ‘3’ than scored ‘1’ or ‘2’ combined, 42 (9%) with two articles not producing shareable data (Figure 2D).

## Software Used

There were two hundred and fifty-three unique software applications recorded being used in the articles that were evaluated. These included desktop programs, web-based software and databases. From the top ten most frequently cited software, the most frequently cited program was MEGA, where different versions were not distinguished for this work. The next three programs were statistical software, SAS (2nd), SPSS (3rd) and R (4th) with two other statistical programs also frequently cited, GenStat and Statistica (tied 8th). The remainder of the top ten programs we found were mostly related to sequence analysis, *i.e.*, Clustal (5th), BLAST (6th), and BioEdit (9th) or phylogenetic analysis MrBayes (8th), tied with the most general-purpose software cited in the top 10, Excel®. In tenth place was FigTree, used for phylogenetic trees. The Python programming language was cited only twice.

## Statistical Analysis

When predicting the availability of computational methods (code) and availability, all journals were compared with *Phytopathology* as the reference with latent scores having a mean of zero.

There were no clear differences that the analysis could detect between any of the journals as compared with *Phytopathology* for the sharing of computational methods (code); the effects of all parameters were undecided (Figure 3A).

When predicting the availability of data availability as compared with *Phytopathology*, publications in two journals, *Phytopathologia Mediterrania* and *Molecular Plant Pathology*, can be considered clearly more likely to share data than publications in *Phytopathology*, but no effects were detectable in any other journal title. The effect of *Molecular Plant Microbe Interactions* (Median = 0.67, 95% CI [-0.28, 1.57]) has a 92.15% probability of being positive (> 0), 90.61% of being significant (> 0.05), and 78.94% of being large (> 0.30). The estimation successfully converged (Rhat = 1.000) and the indices are reliable (ESS = 25478). The effect of *Phytopathologia Mediterranea* (Median = 1.25, 95% CI [0.32, 2.15]) has a 99.58% probability of being positive (> 0), 99.42% of being significant (> 0.05), and 97.73% of being large (> 0.30). The estimation successfully converged (Rhat = 1.000) and the indices are reliable (ESS = 24986).

There was no effect detected for year when predicting the availability of computational methods (code) or data availability.

There were no detectable effects when predicting the effects of the five-year impact factor on the availability of computational methods (code) or data.

# Discussion

Except for a few isolated cases, most papers were not fully computationally reproducible. Very few authors choose to share both data and code. More authors shared data due to journal requirements to share sequence data, but other types of data related to field experiments or other laboratory studies were not likely to be shared. Code sharing was extremely rare, but in cases where it was shared, most were included as a part of the journal article’s publication as extra materials rather than through data repositories. The reasons for not sharing code or data were not clearly available from the papers themselves and so this work was unable to determine possible reasons for this situation. We recognise that in some cases there may be commercial or intellectual property (IP) reasons for not sharing data or code, but these reasons should be clearly stated. However, in most cases, the data are collected with public funding and the code was developed using similar funding and there are no commercial or IP issues that preclude sharing and in fact the funding agency may have guidelines in place for sharing these materials.

Much of the software used are Free Open Source Software (FOSS) packages, which means that the workflow can easily be recreated by anyone with the proper skills in using the software. However, the top three most widely used software packages were not FOSS, limiting the ability of the authors to share workflows with others. MEGA was the most widely cited software in this evaluation reflecting the wide spread use of sequence analysis in the field of plant pathology. The second, third and fourth, and eighth most frequently cited software packages were all statistical programs. SAS remains firmly entrenched in the discipline as the choice of software for statistical analysis followed by SPSS but R, the most frequently occurring FOSS, is not far removed from SPSS in fourth place with Statistica and Genstat tied for eighth. It is interesting to note however, that even with the popularity of Python in scientific programming it only appears in the survey twice in 2020 for both articles, not enough times to be in the top 10 most-cited.

## Examples of Different Levels of Reproducibility

There are different ways to make research more open and reproducible, which can be thought of as levels of reproducibility. Assuming that a general work flow for producing academic research involves clearly defining a research question, obtaining data for testing the hypothesis, summarizing/analyzing and presenting data and results, writing the manuscript then the next logical steps are depositing the code used along with the raw data in proper repositories (Figure 6). Here we define three levels of reproducibility which are also related to the evolution of computational methods and reproducible practices and highlight papers that fit these definitions.

The first level is including tables of raw data or code with the paper as a supplemental file or even within the paper as tables where possible. This is suitable for studies that may have a small data set or simple analysis or for demonstrations purposes as Madden et al. (2015) demonstrate in their discussion regarding the use of values in statistical analyses where they supply an *e-****X****tra* with reproducible examples for readers to refer to. Hill et al. (2019) also shared code and example data to reproduce the results of the paper as supplementary materials, citing all packages used and the versions. The scripts provided used an R package, *checkpoint* (Ooi et al. 2022), to provide a mechanism to help ensure reproducibility by installing the package versions used by the authors that are necessary to reproduce the work rather than defaulting to the latest versions. However, the scripts did not run unhindered, as the supplementary materials suggested that they should, without changes to the data, but with some modifications to the data, and file name changes, the scripts ran allowing examples of the research to be reproduced using the definition of Peng (2009). One drawback to this approach is that journals are often not equipped to handle code, e.g., script files, that may be developed as a part of the research process and so they are often archived as PDF or Word documents which hinder the ability to easily ingest and start working with them or if they are provided in a native text format, they do not render via the web properly. Though this should be possible given some effort from the publishing journal to share simple text files rather than binary formats along with the proper instructions and handling of these file formats rather than a one-size-fits all approach that we commonly see for supplementary materials. For example, the R scripts for Hill et al. (2019) were provided as supplementary materials but do not actually appear in the browser window when requested, which may confuse readers and makes downloading them more difficult. While this allows the reader to quickly view the extra materials and a DOI is assigned as a part of the article itself, the data or code are not readily findable and accessible. Furthermore, in many cases this does not allow prompt access to the data and running the code because of a journal paywall.

The second level is providing machine readable text files of the raw data and code in a public code repository such as GitHub, GitLab, but without a DOI or other long-term repository as provided by a library, Zenodo, FigShare, etc. Fewer authors choose to follow this method, but in one instance that we found, which was not a part of this analysis, Vogel et al. (2021) deposited Fastq files in the National Center of Biotechnology Information Sequence Read Archive (BioProject accession number PRJNA616021) and provided the scripts for analysis via a GitHub repository (<https://github.com/gmv23/pcap-gwas>) but we were unable to find a DOI that refers to these materials or a proper citation to properly cite them here.

The third level is the use of proper code repositories such as GitHub and data repositories like FigShare, Zenodo or OSF.io among others, allowing for the deposition and updating of code, figures, data preprints or any other materials that support the article itself while providing a DOI and citation. As an example, Sparks et al. (2011, 2014) used FigShare to provide models, data and code (Sparks 2016) necessary to replicate model development and the subsequent study on the effects of climate change on potato late blight. Similarly, Carleson et al. (2019) hosted the code for reproducing a population genomic analysis of *Phytophthora plurivora* on GitHub, while providing all data on OSF.io (the Open Science Framework). Lehner et al. (2017) used GitHub to host a code repository of their research compendium website with data and a reproducible report that explains in detail all steps of the analysis and the R code for conducting a meta-analysis for assessing heterogeneity in the relationship between white mold incidence and soybean yield and between incidence and soybean yield (<https://emdelponte.github.io/paper-white-mold-meta-analysis/>). The website clearly demonstrates the analysis to readers and uses R so that anyone can easily replicate the work. Using a public code repository resource like GitHub allows readers to easily contact the authors by opening “Issues” and report bugs or ask questions in an open forum that are not as straightforward when the data are provided as supplementary material.

Taking this approach even further, packaging the full analysis in a containerized software application, *e.g.*, Docker is a way to help ensure computational reproducibility but at the expense of added complexity. Docker is an open source containerization platform that enables users to package several applications and an operating system into containers, thereby standardising the executable components by combining application source code (or analyses) with the operating system required to run that analysis on a user’s computer. However, there are drawbacks to using Docker. It can be difficult to understand for a new user and new platforms, like the Apple M1 chip, may not be fully supported, which hinder the ability to share the container. But, in most cases using an open source language like R, Julia or Python allows us to share our work in a fashion where we know that the analysis will run exactly the same on every computer. For more on using Docker for reproducible research, we would refer readers to Nüst et al. (2020). As an example of this approach, Khaliq et al. (2020a) provided a research compendium as a Docker container with a DOI and a full R package that enables readers to fully replicate their analysis of *Ascochyta rabiei* conidia dispersal in chickpea using the data collected and stepping through other points where weather data were investigated and various models were fit before deciding on the best fit and recreate any figures as published in the article (Khaliq et al. 2020b). When this level is employed, tools such as Binder, <https://mybinder.org>, can be used that allow readers and reviewers to launch an interactive session in their web browser and interact with the data and rerun the analysis in an RStudio instance or Jupyter notebooks as Mioroni et al. used (2018, 2019). Kamvar et al. (2015) took a slightly different approach by including all files necessary for the analysis and most output files in a repository (Kamvar et al. 2014) that also included an installable R package that was used for the original analysis. While there are many other methods, these two approaches illustrate some of the best-practices where the data and other files were deposited in repositories with DOIs and reproducibility issues were addressed by using R packages to handle dependencies and other versioning issues making the work more portable.

Efforts to annotate structured raw data (FAIR – Findable, Accessible, Interoperable and Re-usable) (Wilkinson et al. 2016) and fully document the analysis using open source code which are deposited in public repositories and can be run by anyone following the download of data and code. The first level, as reported, is an essential step that is not substituted by the other practices and eventually researchers fail to provide sufficient description or correct citations. When making your science more open and reproducible, methods, software used (this includes items such as R, Julia or Python packages that were directly used in the analysis or production of the paper, etc.) should be cited properly. Provide a full description of all equipment used, *e.g.*, “a Spectrum Technologies Watchdog 2700 weather station was used to record wind speed, direction, rainfall, temperature and relative humidity at one-hour intervals”. This allows end-users to identify what was used and identify the methods used more accurately. Just as importantly, this acknowledges the contributions of others whose works were instrumental in your research. This also helps ensure that you as a researcher can reconstruct what you have done since you will have good notes and documentation and be able to identify if something changes, *e.g.*, a package version, what effect it had on your research.

The use of programming or scripting languages such as R, Julia, SAS or Python enables you as a scientist to keep very detailed records of what was computationally performed. This is as opposed to using software such as spreadsheet programs like Excel, Google Sheets, Numbers, Calc or others which can be used for simple statistical analyses and visualization or other point-and-click software packages that do not enable you to keep an accurate record of the steps taken to import, format, visualise and analyse data.

Text files for saving small sets of data are preferable. Data that are saved in binary formats such as PDF files are difficult to reuse because they are not easily machine readable. In many cases, data sets are small enough and curated in spreadsheets, which should be saved as a plain text file, *e.g.*, comma separated (CSV) or tab separated (TSV) files. This also helps ensure that the data are reusable. Larger data sets may warrant the use of a proper database like lightweight personal databases, e.g. SQLite or DuckDB, larger more robust databases, e.g., MariaDB, PostgreSQL or a specialised database like GenBank which provide users with several benefits but two important benefits to mention here are (i) data redundancy, ensuring no records are duplicated and (ii) data consistency, ensuring that all records in a data set are recorded in the same format for every observation. Databases such a GenBank are preferable for molecular data ensuring data integrity and machine readability. While databases may offer many advantages for speed and data integrity, the trade-off is that they are more complex to set up and administer, especially for a small data set.

Ideally, once the data are complete, best practices for keeping your data as you perform your work include treating the raw data as read-only and using file permissions to prevent changes to the raw data files. It should be noted, that the use of a database management system also allows for both of these at the expense of added complexity. Saving files in proprietary formats such as .xls(x) can also lead to issues in the future when opening using newer (or older) software versions. Unexpected changes to values in the data (Ziemann et al. 2016) may also occur when using proprietary formats.

If steps are followed to make the data FAIR then it will be readable by humans and machines alike and this will help support discoveries and support further research. In turn, sharing data will lead to new citations for your work as others discover and use it. To make your data the most widely discoverable and usable, ensure that it has a persistent identifier. A digital object identifier (DOI) is the most common (<https://www.doi.org/>) but the Handle.Net Registry (HNR) (<https://handle.net/>) is also an option. There are different options for generating a DOI for your data and other materials. FigShare, Zenodo and OSF.io all offer persistent archives along with a services to generate a DOI for your materials. The use of a persistent identifier works to ensure that even if the data are moved, they can still be located using that unique identifier. For more on FAIR data, visit Go-Fair <https://www.go-fair.org/fair-principles/>.

Once you have determined how to best manage your source code and the data sets for analysis the next step is to consider how to share your data. Providers such as FigShare, Dataverse, OSF.io and Zenodo allow for you to deposit your data and generate a DOI for sharing your project once you are finished with it.

Other providers exist that allow for you to not only track changes but also to share the data openly, these include GitHub, Gitlab and Bitbucket. GitHub is arguably the most popular and widely used software development platform currently. Data that is encoded in CSV or other plain text formats can easily be deposited in a repository along with code for analysis to enable tracking of changes and other users to download and replicate the work.

We would advise against the practice of depositing data on a laboratory website or a site such as GitHub only as these are not an optimal way to preserve and share your work over the long term. Doing either of these leaves the work in an unstable state where future users may be unable to access the work as they are fraught with link-rot and other issues. It is a best practice to always ensure that you have deposited the data with a provider such as GenBank, Zenodo, FigShare or OSF.io and generated a DOI for the materials to help ensure continued accessibility. Many of these providers provide rather easy ways to link the project with a software development repository to help ensure that the data are available in perpetuity. If readers are uncertain, we suggest to also consult with local librarians about possible resources. Most universities and other research-focused workplaces provide a facility for staff to deposit papers and other academic materials, but this may extend to software development or data repositories as well. The important aspect is that data once deposited cannot be modified any more.

## Final Thoughts

In this work we have evaluated the state of computational reproducibility in the plant pathology literature and presented suggestions for areas of improvement. As we prepared this letter, we became more aware of the urgent need to spread and establish an open science attitude and culture among plant pathologists. To assist in fostering this sort of change in our discipline, Open Plant Pathology (OPP), <https://www.openplantpathology.org>, an institution-independent and non-funded initiative, was founded in January 2018 by two of this letter’s co-authors, Del Ponte and Sparks, in 2018 with the following vision: “foster a diverse community culture that values openness, transparency and reproducibility of scientific research data and methods in our field”. We started OPP with a minimal infrastructure and support from other enthusiastic colleagues that allows members to interact sharing and gathering ideas on how we can improve the openness and reproducibility in our discipline.

We believe that adopting an attitude of open and collaborative science and using the best reproducibility practices in our daily work, directly benefits us as researchers. For example, between complicated analyses, reviews and revisions and questions years later about the data that was collected or analysis that was conducted, it is extremely beneficial to be able to easily reproduce your work quickly and easily. This manuscript was drafted over the course of several years as the authors had time to devote to it. Having everything in a reproducible framework made it easy to resume work and set-aside as necessary without losing information and having everything well-documented made it more simple to do this. Second, it benefits the reviewer by aiding their understanding of the work done and gives them more materials to use to make suggestions for improvements when reviewing and the end-user or reader is better able to verify the validity of the methods used and recreate the analysis. Perhaps more importantly, sharing these details helps with knowledge transfer by showing other interested parties how something was done rather than simply describing it. Lastly, openly sharing your work and making it discoverable can lead to new collaborations and synergistic ideas. One of the most important messages that we would like to share is that there is more to the work than just the paper. Sharing materials detailing the analysis that was performed and documenting the data provide citable products and enhance the manuscript providing the reader with a richer set of information with which to understand the work that was performed. This open sharing of code and data leads to greater impact as work is cited if resources such as code or data are reused. But it is not just up to the authors to ensure that their work is reproducible. Journals can and should provide clear instruction for how to deposit the data and code in a repository and mint a DOI to accompany these resources to encourage authors to share the data and code that support the manuscript. With mandates for openly sharing data becoming more common with funding agencies, we can and should embrace this and move the discipline forward and have a greater impact with our work.

Without the code used to generate the paper and the data that were gathered for the paper, we as a community cannot adequately evaluate the work as it is presented. That is, while we may be standing on the shoulders of giants in doing our work, we cannot see where they have been to do our work, which hinders our efforts to be able to see farther ahead.

# Authors’ Contributions

A.H.S, E.DP., Z.F. and N.J.G. conceived of the presented idea and scoring system. E.DP. provided five-year impact factors. A.H.S., E.DP., Z.F. K.A. and N.G. evaluated articles for scoring. A.H.S. designed the computational framework for analysis and created the research compendium with feedback from E.DP. and Z.F. A.H.S. wrote the manuscript in consultation with E.DP. All authors provided critical feedback and helped contribute to the final version of the manuscript.

# Data and Code Availability

## Data

The raw data for this work are documented and available from DOI [https://doi.org/10.5281/zenodo.4941722](10.5281/zenodo.4941722).

## Code

All code used in the analyses and data visualization and associated materials have been made available as research compendium available from DOI [https://doi.org/10.5281/zenodo.1250664](10.5281/zenodo.1250664). A webpage version of the compendium is available from <https://openplantpathology.github.io/Reproducibility_in_Plant_Pathology/>.

# Acknowledgements

The authors are grateful for insightful comments from Dr. David Ferris and the USQ Centre for Crop Health Advisory Group on the final manuscript and from Ms. Anna Hepworth for statistical guidance on evaluating inter-rater scores.

# Literature Cited

Australian Research Council. 2018. ARC Open Access Policy. Available at: <https://www.arc.gov.au/policies-strategies/policy/arc-open-access-policy> [Accessed April 20, 2020].

Baker, M. 2016. [Muddled meanings hamper efforts to fix reproducibility crisis](https://doi.org/10.1038/nature.2016.20076). Nature News.

Brunsdon, C. 2015. [Quantitative methods I: Reproducible research and quantitative geography](https://doi.org/10.1177/0309132515599625). Prog. Hum. Geogr.

Bürkner, P.-C. 2018. [Advanced Bayesian multilevel modeling with the R package brms](https://doi.org/10.32614/RJ-2018-017). The R Journal. 10:395–411.

Bürkner, P.-C. 2021. [Bayesian item response modeling in R with brms and Stan](https://doi.org/10.18637/jss.v100.i05). Journal of Statistical Software. 100:1–54.

Bürkner, P.-C. 2017. [brms: An R package for Bayesian multilevel models using Stan](https://doi.org/10.18637/jss.v080.i01). Journal of Statistical Software. 80:1–28.

Carleson, N. C., Fieland, V. J., Scagel, C. F., Weiland, J. E., and Grünwald, N. J. 2019. Population structure of *Phytophthora plurivora* on rhododendron in Oregon nurseries. Plant disease. 103:1923–1930.

Del Ponte, E. M. 2020. [Towards a more open and transparent plant pathology research](https://doi.org/10.1007/s40858-020-00388-z). Tropical Plant Pathology. 45:361–362.

Del Ponte, E. M., Nelson, S. C., and Pethybridge, S. J. 2019. [Evaluation of app-embedded disease scales for aiding visual severity estimation of cercospora leaf spot of table beet](https://doi.org/10.1094/pdis-10-18-1718-re). Plant Disease. 103:1347–1356.

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., et al. 2020. [An agenda for open science in communication](https://doi.org/10.1093/joc/jqz052). Journal of Communication. 71:1–26.

Eckert, E. M., Cesare, A. D., Fontaneto, D., Berendonk, T. U., Bürgmann, H., Cytryn, E., et al. 2020. [Every fifth published metagenome is not available to science](https://doi.org/10.1371/journal.pbio.3000698). PLOS Biology. 18:e3000698.

Editors at Nature. 2016. [Reality check on reproducibility](https://doi.org/10.1038/533437a). Nature. 533:437.

Fidler, F., and Gordon, A. 2013. Science is in a reproducibility crisis: How do we resolve it? Phys.org. Available at: <https://phys.org/news/2013-09-science-crisis.html>.

Fitzjohn, R., Pennell, M., Zanne, A., and Cornwell, W. 2014. Reproducible research is still a challenge. ROpensci.org. 2014 Available at: <http://ropensci.org/blog/2014/06/09/reproducibility/>.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. 2019. [Visualization in Bayesian workflow](https://doi.org/10.1111/rssa.12378). J. R. Stat. Soc. A. 182:389–402.

Government of Canada. 2016. Tri-Agency Open Access Policy on Publications - Science.gc.ca. Tri-Agency Open Access Policy on Publications. Available at: <http://www.science.gc.ca/eic/site/063.nsf/eng/h_F6765465.html?OpenDocument> [Accessed April 12, 2019].

Hill, G. N., Beresford, R. M., and Evans, K. J. 2019. [Automated analysis of aggregated datasets to identify climatic predictors of botrytis bunch rot in wine grapes](https://doi.org/10.1094/phyto-10-17-0357-r). Phytopathology. 109:84–95.

Ioannidis, J. P. A. 2014. [How to make more published research true](https://doi.org/10.1371/journal.pmed.1001747). PLoS Medicine. 11.

Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D., and Ioannidis, J. P. A. 2016. [Reproducible research practices and transparency across the biomedical literature](https://doi.org/10.1371/journal.pbio.1002333). PLoS Biology. 14:1–13.

Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grünwald, N. J. 2015. [Spatial and temporal analysis of populations of the sudden oak death pathogen in Oregon forests](https://doi.org/10.1094/PHYTO-12-14-0350-FI). Phytopathology. 105:982–989.

Kamvar, Z. N., Larsen, M. M., Kanaskie, A. M., Hansen, E. M., and Grünwald, N. J. 2014. [Sudden Oak Death in Oregon Forests: Spatial and temporal population dynamics of the sudden oak death epidemic in Oregon Forests](https://doi.org/10.5281/zenodo.13007).

Khaliq, I., Fanning, J., Melloy, P., Galloway, J., Moore, K., Burrell, D., et al. 2020a. [The role of conidia in the dispersal of *Ascochyta rabiei*](https://doi.org/10.1007/s10658-020-02126-2). European Journal of Plant Pathology. 158:911–924.

Khaliq, I., Melloy, P., Burrell, D., Fanning, J., Galloway, J., Moore, K., et al. 2020b. *ChickpeaAscoDispersal: A research compendium to accompany “the role of conidia in the dispersal of Ascochyta rabiei”*. Available at: <https://doi.org/10.5281/zenodo.3810826>.

Lehner, M. S., Pethybridge, S. J., Meyer, M. C., and Del Ponte, E. M. 2017. [Meta-analytic modelling of the incidence–yield and incidence–sclerotial production relationships in soybean white mould epidemics](https://doi.org/10.1111/ppa.12590). Plant Pathology. 66:460–468.

Madden, L. V., Shah, D. A., and Esker, P. D. 2015. [Does the *p* value have a future in plant pathology?](https://doi.org/10.1094/PHYTO-07-15-0165-LE) Phytopathology. 105:1400–1407.

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. 2019. [bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework](https://doi.org/10.21105/joss.01541). Journal of Open Source Software. 4:1541.

Miorini, T. J. J., Kamvar, Z. N., Higgins, R. S., Raetano, C. G., Steadman, J. R., and Everhart, S. E. 2019. [Differential aggressiveness of *Sclerotinia sclerotiorum* isolates from North and South America and partial host resistance in Brazilian soybean and dry bean cultivars](https://doi.org/10.1007/s40858-018-00273-w). Tropical Plant Pathology. 44:73–81.

Miorini, T. J. J., Kamvar, Z. N., Higgins, R., Raetano, C. G., Steadman, J., and Everhart, S. E. 2018. Data and analysis for variation in pathogen aggression and cultivar performance against *Sclerotinia sclerotiorum* in soybean and dry bean from the U.S. And Brazil. Available at: <https://osf.io/2x7fc/> [Accessed July 29, 2020].

Noorden, R. van. 2017. Gates Foundation demands open access. Nature. 541:270–270.

Nüst, D., Sochat, V., Marwick, B., Eglen, S. J., Head, T., Hirst, T., et al. 2020. [Ten simple rules for writing dockerfiles for reproducible data science](https://doi.org/10.1371/journal.pcbi.1008316) ed. Scott Markel. PLOS Computational Biology. 16:e1008316.

Ooi, H., Vries, A. de, and Microsoft. 2022. *checkpoint: Install packages from snapshots on the Checkpoint server for reproducibility*. Available at: <https://CRAN.R-project.org/package=checkpoint>.

Patil, P., Peng, R. D., and Leek, J. 2016. A statistical definition for reproducibility and replicability. bioRxiv. Available at: <http://biorxiv.org/content/early/2016/07/29/066803>.

Peng, R. D. 2009. [Reproducible research and biostatistics](https://doi.org/10.1093/biostatistics/kxp014). Biostatistics. 10:405–408.

Preeyanon, L., Pyrkosz, A. B., and Titus Brown, C. 2018. [Reproducible bioinformatics research for biologists](https://doi.org/10.1201/9781315373461-7). In Chapman; Hall/CRC, p. 185–217.

R Core Team. 2022. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Available at: <https://www.R-project.org/>.

Sparks, A. 2016. Global-late-blight (meta)modelling. Available at: <https://figshare.com/articles/dataset/Global_Late_Blight_Modelling/1066124/6>.

Sparks, A. H., Del Ponte, E. M., Alves, K. S., Foster, Z., and Grünwald, N. J. 2022. Compendium of r code and data for ‘status and best practices for reproducible research in plant pathology. Available at: <https://doi.org/10.5281/zenodo.1250665>.

Sparks, A. H., Forbes, G. A., Hijmans, R. J., and Garrett, K. A. 2011. [A metamodeling framework for extending the application domain of process-based ecological models](https://doi.org/10.1890/ES11-00128.1). Ecosphere. 2:art90.

Sparks, A. H., Forbes, G. A., Hijmans, R. J., and Garrett, K. A. 2014. [Climate change may have limited effect on global risk of potato late blight](https://doi.org/10.1094/PDIS-04-11-031). Global Change Biology. :3621–3631.

Stodden, V., Guo, P., and Ma, Z. 2013. [Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals](https://doi.org/10.1371/journal.pone.0067111). PLoS One. 8:2–9.

Sweedler, J. V. 2015. [Striving for reproducible science](https://doi.org/10.1021/acs.analchem.5b04300). Analytical Chemistry. 87:11603–11604.

Tiwari, K., Kananathan, S., Roberts, M. G., Meyer, J. P., Shohan, M. U. S., Xavier, A., et al. 2021. [Reproducibility in systems biology modelling](https://doi.org/10.15252/msb.20209982). Molecular Systems Biology. 17.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. 2021. [Rank-normalization, folding, and localization: An improved Rˆ for assessing convergence of MCMC (with discussion)](https://doi.org/10.1214/20-ba1221). Bayesian Analysis. 16.

Vogel, G., Gore, M. A., and Smart, C. D. 2021. [Genome-wide association study in new york iphytophthora capsici/i isolates reveals loci involved in mating type and mefenoxam sensitivity](https://doi.org/10.1094/phyto-04-20-0112-fi). Phytopathology. 111:204–216.

Wald, C. 2010. Scientists embrace openness. Available at: <https://www.science.org/content/article/scientists-embrace-openness> [Accessed May 1, 2022].

Wallach, J. D., Boyack, K. W., and Ioannidis, J. P. A. 2018. [Reproducible research practices, transparency, and open access data in the biomedical literature, 20152017](https://doi.org/10.1371/journal.pbio.2006930) ed. Ulrich Dirnagl. PLOS Biology. 16:e2006930.

Weissgerber, T. L., Garovic, V. D., Winham, S. J., Milic, N. M., and Prager, E. M. 2016. [Transparent reporting for reproducible science](https://doi.org/10.1002/jnr.23785). Journal of Neuroscience Research. 00:1–6.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. 3:1–9 Available at: <https://www.nature.com/articles/sdata201618> [Accessed April 20, 2020].

Ziemann, M., Eren, Y., and El-Osta, A. 2016. [Gene name errors are widespread in the scientific literature](https://doi.org/10.1186/s13059-016-1044-7). Genome Biology. 17:177.

# Tables

Table 1: Journal titles selected for inclusion, their respective five year impact factors as of 2022 and the number (n) of articles from each journal that were evaluated.

| Journal | 5 Year IF | n |
| --- | --- | --- |
| Forest Pathology | 1.595 | 18 |
| Phytopathologia Mediterranea | 2.080 | 20 |
| Revista Mexicana de Fitopatología | 0.000 | 19 |
| Journal of Phytopathology | 1.574 | 19 |
| Journal of Plant Pathology | 1.681 | 23 |
| Phytoparasitica | 1.569 | 23 |
| Virology Journal (Plant Viruses Section) | 3.719 | 25 |
| Australasian Plant Pathology | 1.708 | 11 |
| Canadian Journal of Plant Pathology | 1.993 | 19 |
| Crop Protection | 3.110 | 23 |
| European Journal of Plant Pathology | 2.022 | 19 |
| Journal of General Plant Pathology | 1.416 | 24 |
| Molecular Plant Pathology | 5.626 | 29 |
| Nematology | 1.485 | 16 |
| Physiological and Molecular Plant Pathology | 2.388 | 24 |
| Molecular Plant-Microbe Interactions | 4.836 | 22 |
| Phytopathology | 4.394 | 28 |
| Plant Disease | 4.700 | 24 |
| Plant Health Progress | 0.000 | 16 |
| Plant Pathology | 2.924 | 30 |
| Tropical Plant Pathology | 1.675 | 18 |

Table 2: The top ten sofware programs that were found to be used in the papers that were surveyed.

| Software | Frequency |
| --- | --- |
| MEGA | 73 |
| SAS | 70 |
| SPSS | 48 |
| R | 41 |
| CLUSTAL | 28 |
| BLAST | 22 |
| EXCEL | 17 |
| MRBAYES | 17 |
| GENSTAT | 14 |
| STATISTICA | 14 |
| BIOEDIT | 10 |
| FIGTREE | 9 |

# Figures

![Figure 1: Posterior distribution visualisations for each of six models fitted to scoring data that were used to evaluate factors on reproducibility of 450 papers published in 21 plant pathology journals or plant pathology focused articles from other specialised journals. Six factors were evaluated, (A, B) Publication (Journal Title), (C, D) Year of Publication, (E, F) Five Year Impact Factor of Journal, and (G, H) Journal Open Access Policy where y is the data’s density and yrep are ten random posterior draws. Model fitnesses were found to be good for all models.](data:application/eps;base64,)

Figure 1: Posterior distribution visualisations for each of six models fitted to scoring data that were used to evaluate factors on reproducibility of 450 papers published in 21 plant pathology journals or plant pathology focused articles from other specialised journals. Six factors were evaluated, (A, B) Publication (Journal Title), (C, D) Year of Publication, (E, F) Five Year Impact Factor of Journal, and (G, H) Journal Open Access Policy where y is the data’s density and yrep are ten random posterior draws. Model fitnesses were found to be good for all models.

![Figure 2: Aggregated article scores for each of the two categories evaluated, (A) displays ‘Code availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no code was created to conduct the work that was detectable. (B) shows ‘Data availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no data were generated, e.g., a methods paper.](data:application/eps;base64,)

Figure 2: Aggregated article scores for each of the two categories evaluated, (A) displays ‘Code availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no code was created to conduct the work that was detectable. (B) shows ‘Data availability’, where ‘0’ was ‘Not available or not mentioned in the publication’; ‘1’ was ‘Available upon request to the author; ’2’ was ‘Online, but inconvenient or non-permanent (e.g., login needed, paywall, FTP server, personal lab website that may disappear, or may have already disappeared)’; and ‘3’ was ‘Freely available online to anonymous users for foreseeable future (e.g., archived using Zenodo, dataverse or university library or some other proper archiving system)’; ‘NA’ indicates that no data were generated, e.g., a methods paper.

![Figure 3: 95% credible interval for the probability associated with the journal title for (A) computational materials (code) and (B) data being made readily available to the public. A test for equivalence was unable to detect any clear differences between Phytopathology and all other journals sharing computational materials (code). However, a test for equivalence found that Molecular Plant Pathology and Phytopathologia Mediterranea were more likely have data available than articles found in Phytopathology. Intervals in grey have a median value less than that of Phytopathology, the base level used for the analysis and articles in these journals could be less likely to share computational materials and data than articles published in Phytopathology. Intervals in black have a median value that is greater than that of Phytopathlogy, the base level for the analysis and articles in these journals could be more likely to share data or code than those found in Phytopathology.](data:application/eps;base64,)

Figure 3: 95% credible interval for the probability associated with the journal title for (A) computational materials (code) and (B) data being made readily available to the public. A test for equivalence was unable to detect any clear differences between *Phytopathology* and all other journals sharing computational materials (code). However, a test for equivalence found that *Molecular Plant Pathology* and *Phytopathologia Mediterranea* were more likely have data available than articles found in *Phytopathology*. Intervals in grey have a median value less than that of *Phytopathology*, the base level used for the analysis and articles in these journals could be less likely to share computational materials and data than articles published in *Phytopathology*. Intervals in black have a median value that is greater than that of *Phytopathlogy*, the base level for the analysis and articles in these journals could be more likely to share data or code than those found in *Phytopathology*.

![Figure 4: 95% credible interval for the probability associated with the year of publication for (A) computational materials (code) and (B) data being made readily available to the public. A test for practical equivalence was undecided in any detectable effects of year of publication on computational materials or data being shared. Intervals in grey indicate that the median was less than zero or less likely that code or data were shared and intervals in black indicate that the median value is above zero or more likely to share data and code.](data:application/eps;base64,)

Figure 4: 95% credible interval for the probability associated with the year of publication for (A) computational materials (code) and (B) data being made readily available to the public. A test for practical equivalence was undecided in any detectable effects of year of publication on computational materials or data being shared. Intervals in grey indicate that the median was less than zero or less likely that code or data were shared and intervals in black indicate that the median value is above zero or more likely to share data and code.

![Figure 5: 95% credible interval for the probability associated with the publishing journal’s five-year impact factor for (A) computational materials (code) and (B) data being made readily available to the public. A test for practical equivalence was undecided on any detectable effects of impact factor. Intervals in grey indicate that the median was less than zero or less likely that code or data were shared and intervals in black indicate that the median value is above zero or more likely to share data and code.](data:application/eps;base64,)

Figure 5: 95% credible interval for the probability associated with the publishing journal’s five-year impact factor for (A) computational materials (code) and (B) data being made readily available to the public. A test for practical equivalence was undecided on any detectable effects of impact factor. Intervals in grey indicate that the median was less than zero or less likely that code or data were shared and intervals in black indicate that the median value is above zero or more likely to share data and code.

![Figure 6: An example of an open and reproducible research workflow. Starting with the question, determine the methodology, describe it, make it available and cite it. Data are used in analysis and any binary files or code are made available as supplements to the manuscript. Source code and raw data are made available in a public repository, preferably with version control for tracking changes through time, and a DOI for final released products.](data:application/eps;base64,)

Figure 6: An example of an open and reproducible research workflow. Starting with the question, determine the methodology, describe it, make it available and cite it. Data are used in analysis and any binary files or code are made available as supplements to the manuscript. Source code and raw data are made available in a public repository, preferably with version control for tracking changes through time, and a DOI for final released products.