Reproducible Research

Gabriel Mateus Bernardo Harrington[*] Jade Perry[†] Professor Paul Cool[‡]

2021-02-20

Maybe it would interesting to provide an example of reproducible workflow? Perhaps include some basic data analysis (e.g. lm with iris), maybe write a very brief mock paper with RMarkdown and/or Jupyter. Could host on a git repo, and/or make a Docker image?

Points to raise:

1. containerisation
   - problem: dependencies for code can become difficult/impossible to source
   - solution: package all code and dependencies to run uniformly on any infrastructure
   - basic guide here - template
2. Importance of software being free and open-source
   - unclear what propriety software does - can't properly review work
   - Company may go bust/sell software rights - licence issues - can't reproduce work
   - Some may not have access to software - platform exclusivity, can't afford licence (often very expensive) - can't reproduce
3. Benefits of code for data analysis and writing
   - scripts by definition outlines all steps taken - written method section will never match this
   - more shareable - easier collaboration
   - For writing - if journal accepts markup, formatting much simpler - scientist spend more time doing science and less appeasing asinine journals
   - if raw data uploaded to repository it can get a doi - extra citations
4. References
   - Need for references to be in plain text - .bib files
   - Can use Zotero - free and open-source
   - Could also use more minimalist shell scripts to get .bib info

Introduction

With the increasing amount of data available, better methods are required to analyse large data files. Although spreadsheets can be used to collate and review data with up to around 100 patients. This becomes more difficult with larger data sets.

When performing the analysis, it is important to document the methods so that results can be verified. It is now common practice for journals to request data to be uploaded as part of the publication. This allows for independent verification and scrutiny. Although uncommon, there have been high profile retractions of published literature because results could not be verified [1–3].

Often, specialist advice is required for analysis. However, all studies should have a thorough open and transparent analysis that is robust to scrutiny, repeatable and reproducible [4].

[*] Keele University, g.m.bernardo.harrington@keele.ac.uk

[†] Keele University, email@here.com

[‡] The Robert Jones and Agnes Hunt Orthopaedic Hospital, john@example.org

In this short paper, the advantages of reproducible research and how to achieve it are discussed.

Data

Smaller data sets are often created on a spreadsheet where the variables are defined. When constructing a spreadsheet, it is important to adhere to the principles of tidy data [5] with observations in rows and variables in columns. Adhering to these principles will avoid a lot of subsequent data wrangling. In clinical practice, observations are often patients. However, if there are more than one observation on a patient, it is important to have an unique identifier for each observation. Frequently, a lot of time is spent wrangling and cleaning data. Any data manipulation technique could potentially introduce error. By obtaining the appropriate advice beforehand, it can be assured that data is recorded properly and consistently. Factorial variables frequently have to be redefined due to inconsistent format (e.g.: "f," "F," "female," "Female").

Dates require particular mention as parsing of dates in different types of software can be troublesome. It is strongly recommended to use the ISO standard when recording dates [7] to avoid unnecessary problems.

Variable names are often inconsistent with unnecessary information. Particularly spaces in a variable name can be troublesome in scripts of statistical software. It is recommended to avoid spaces in variable names and use popular styles like camelCase or snake_case. All styles are frequently used but it is important to be consistent [8].

It is recommended to collect primary data such as 'date_of_birth' and 'date_operation' and subsequently calculate the age with software as this approach is less prone to error. However, a dataset that is to be shared publicly should not contain the date of birth or date of procedure as this could make patients potentially identifiable. Although, these calculations can be made within a spreadsheet, it may be preferable to use statistical software for this. In a spreadsheet, calculated fields often become NA when the underlying data is removed.

It is important not to 'squirrel away' data and have multiple spread sheets nobody knows anything about. It is better to have the data in a format that has been agreed and can be accessed by all with proper version control. Version control can be very simple using a naming convention or software. Software such as GIT [9] needs some practice but allows easy review of changes that have been made and is more robust. A Shared, collaborative and open approach is more likely to result in reproducible research of high-quality worthy of publication.

Larger data sets are often downloaded from databases using queries dependent on the inclusion and exclusion criteria. Data is subsequently filtered to produce a raw data. Particularly with larger data sets, data integrity, reliability and accuracy can be a problem. Data validation test should be performed before analysis. Frequently there are NA values that need review. It can be difficult to know what to do with NA values as excluding them introduces bias. Often imputation techniques are used.

Data are nowadays more openly available and online scrutiny is increasingly vigorous. [10]. Consequently, it is important to be open about your methods. With this approach you can learn from your errors and improve.

Software

For simple analysis, spreadsheet software can be used. However, as mentioned it has its drawback particularly in relation to dates and calculated fields. However, for more complicated analysis often specialist statistical software is required. Using a non-standardised propriety format makes it necessary to have the software to access the data. Although some software is widely available, other software is prohibitively expensive. It is preferable to use a common format that is accessible to everybody.

Although not a defined standard, for data sharing comma separated values [11] files are in common use and can be read by freely available open-source software such as text editors, Python and R [12,13].

Propriety software often has a Graphical User Interface (GUI). When performing complicated analysis using a GUI, it is difficult to document all options that were chosen in multiple different menus. Consequently, it may be difficult to reproduce the result. It is preferable to use software that uses a script for analysis. Apart from performing the analysis, the script is also documentation of the method that has been used. Furthermore, it is possible to use the same script on future data, allowing a comparable analysis.

Contrary to open-source software, propriety software doesn't reveal what is going on 'under the hood.' Furthermore, open-source software is often freely available, though care should be taken to make sure the software licence is sufficiently permissive for copying, sharing and modification, such as the GPL, MIT or Creative Commons licences. This allows external scrutiny and development of better software. Software that is not up to scrutiny can be improved or it simply doesn't survive. Propriety software is also at greater risk of suddenly becoming unavailable, such as if the licence holder enters bankruptcy and removes access, or a competitor buys the licence and similarly limits access.

It should be noted that whilst free and open-source software is much longer lasting relative to propriety software, it is not immune to deprecation. As there are many software packages available to perform analysis, it is important to document the packages, version numbers and dates in the methods. However, if a parciular package or function used is script is not included in later versions and older versions of said package are not widely available, a piece of work may cease to be reproducible eventually. To counter this so-called "Containerisation" can be used.

Propriety menu driven software perhaps makes it easier to 'fish' for significant p-values, rather than selecting robust statistical methods. It is recommended to read the statement of the American Statistical Society regarding p-values [14]. Interestingly, some journals have now banned the use of p-values [15]. Particularly when performing multiple tests, it is important to have an assessment of the false positive rate and use the appropriate correction.

Summary

Robust data collection in a tidy format that is analysed with open-source software using scripts with full documentation of the packages used should be reproducible. This open and transparent analysis will improve your research and make you a better scientist.

This paper has been written with open-source software, including Zotero [16] reference manager.

References

[1] Retraction challenges. Nature News 2014;514:5. https://doi.org/10.1038/514005a.

[2] Retracted coronavirus (COVID-19) papers. Retraction Watch 2020.

[3] Davido B, Lansaman T, Lawrence C, Alvarez J-C, Bouchand F, Moine P, et al. Hydroxychloroquine plus azithromycin: A potential interest in reducing in-hospital morbidity due to COVID-19 pneumonia (HI-ZY-COVID)? medRxiv 2020:2020.05.05.20088757. https://doi.org/10.1101/2020.05.05.20088757.

[4] ASTM: Standard practice for use of the terms precision and bias in ASTM test methods. E177 ed: Subcommittee E11.20 on test method evaluation and quality control, ASTM International; 2013.

[5] 12 Tidy data | R for Data Science n.d.

[6] ISO 8601 Effectively Communicate Dates and Times Internationally. IONOS Digitalguide n.d.

[7] ISO - ISO 8601 Date and time format. ISO n.d.

[8] Naming convention (programming). Wikipedia 2020.

[9] Git n.d.

[10] Challenges in irreproducible research n.d.

[11] CSV, Comma Separated Values (RFC 4180) 2012.

[12] Welcome to Python.org. Pythonorg n.d.

[13] R: The R Project for Statistical Computing n.d.

[14] Wasserstein RL, Lazar NA. The ASA's Statement on p-values: Context, Process, and Purpose. The American Statistician 2016;70:129–33. https://doi.org/10.1080/00031305.2016.1154108.

[15] Trafimow D, Marks M. Editorial. Basic and Applied Social Psychology 2015;37:1–2. https://doi.org/10.1080/01973533.2015.1012991.

[16] Zotero | Your personal research assistant n.d.

Session Information

Packages Used

package	version	date
base	4.0.4	2021-02-15
rmarkdown	2.7	2021-02-19
knitr	1.31	2021-01-27