AML: Reproducible Research

Alan Chalk

May 7, 2020

Elements required for reproducibility

- The original data.
- The computer code used are made available (or fully specified)
- In all but the most trivial cases, it will be necessary to include full documentation (e.g. description of each data variable, an audit trail describing the decisions made when cleaning and processing the data, and full documented code).
- Full documented code can be achieved through literate statistical programming (as defined by Knuth) where the program includes an explanation of the program in plain language, interspersed with code snippets. With the R environment, a tool which allows this is R-markdown, and within Python - Jupyter notebooks.
- Although not strictly required to meet the definition of reproducibility a good version control process can ensure evolving drafts of code, documentation and reports are kept in alignment between the various stages of development and review, and changes are reversible if necessary. There are many tools that are used for version control. A popular command-line tool used for version control is git.

- In addition to version control, documenting the software environment, the computing architecture, the operating system, the software toolchain, external dependencies and version numbers can all be important in ensuring reproducility. As an example, in the R programming language, the command “sessionInfo()” provides information about the operating system, version of R and version of all R packages being used. Providing a Docker image achieves this aim in a particularly convenient manner.
- Where there is randomness in the statistical or machine learning techniques being used (for example random forests or neural networks) or where simulation is used, replication will require the random seed to be set.

Doing things “by-hand” is very likely to create problems in reproducing the work. Examples of doing things by hand are:

- Manually editing spreadsheets (rather than reading the raw data into a programming environment and making the changes there).
- Editing tables and figures (rather than ensuring that the programming environment creates them exactly as needed).
- Downloading data manually using a website (rather than doing it programmatically).
- Pointing and clicking (unless the software used creates an audit trail of what has been clicked).

The value of reproducibility

Many analyses are undertaken for commercial, not scientific, reasons and are not published, but reproducibility is still valuable:

- Reproducibility is necessary for a complete technical work review (which in many cases will be a professional requirement) to ensure the analysis has been correctly carried out and the conclusions are justified by the data and analysis;

- Reproducibility may be required by external regulators and auditors.
- Reproducible research is more easily extended to investigate the effect of changes to the analysis, or to incorporate new data;
- It is often desirable to compare the results of an investigation with a similar one carried out in the past; if the earlier investigation was reported reproducibly an analysis of the differences between the two can be carried out with confidence.
- The discipline of reproducible research, with its emphasis on good documentation of processes and data storage, can lead to fewer errors that need correcting in the original work and, hence, greater efficiency.

There are some issues that reproducibility does not address:

- Reproducibility does not mean that the analysis is correct. For example if an incorrect distribution is assumed, the results may be wrong – even though they can be reproduced by making the same incorrect assumption about the distribution. However, by making clear how the results are achieved, it does allow transparency so that incorrect analysis can be appropriately challenged.
- If activities involved in reproducibility happen only at the end of an analysis, this may be too late for resulting challenges to be dealt with. For example, resource may have been moved on to other projects.