EDITORIALS

Why researchers should share their analytic code

Retraction of a trial shows the importance of transparency

Ben Goldacre *director*, Caroline E Morton *researcher*, Nicholas J DeVito *researcher*

DataLab, Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, UK

JAMA recently retracted and replaced an important clinical trial report from 2018 after a serious programming error was discovered.[1] Quantitative medical research relies on analytic scripts: a sequence of commands issued to extract, reshape, manage, and then analyse data. In this case, there was a catastrophe. The “randomisation assignment” variable coded the control group “1” and the intervention group “2”; this had to be converted to “0” and “1” for the statistical analysis to run, but an incorrect conversion command resulted in the intervention and control groups being mislabelled. The results of the trial were almost completely reversed.

It is laudable that this single error was acknowledged and corrected with a retraction. However, neither the retraction notice nor the accompanying editorial acknowledged the systemic problems and opportunities exemplified by this case.[1 2] Sharing analytic code is increasingly the norm across many fields.[3-5] It provides an unambiguous record of the analytical methods used, aiding reproducibility.[6 7] It also allows expert peer reviewers and the wider research community to audit the code, which increases the likelihood of errors being found and corrected.[8 9]

That benefit is exemplified by this retracted trial, and not only for the catastrophic central error leading to the retraction. While reviewing their code to correct their major error, the research team discovered at least two other areas of erroneous code (in the commands to impute missing values and to aggregate data into summary variables).[1] However, error checking is only one of the benefits that come from sharing code; more broadly, sharing code under open licence for reuse by others generates an archive of clinically relevant code that can help avoid duplicated effort and accelerate innovation.

Unwarranted secrecy

Some researchers object to this form of transparency. In our view these objections are either misplaced or fail to proportionately reflect the needs of patients and the scientific community. Sharing code, unlike sharing individual patient data, will typically present no privacy issues. We have been told that sharing code is difficult because the scripts are long, covering “many pages of information.”[10] But there are numerous free, open platforms to share version controlled code,[11 12] and the most commonly used, GitHub,[13] has a limit of 100 GB for each repository. For context, our group’s OpenPrescribing.net service is a substantial software project with 130 000 users a year: the whole project is over 30 000 lines of code, which is at least one order of magnitude bigger than any single epidemiological analysis script, but this equates to only 1.5 MB of storage.

Another objection is the time needed to create perfectly curated code, but there is no need for code to be converted into generalisable “libraries”; simply sharing practical working code is a good start.[14] Emerging best practice is to share full analyses using tools such as R Markdown and Jupyter Notebooks. These are easy to use and embed narrative text, analytic code, and the outputs of that code all in a single interactive notebook. Using these tools, our team aims to share analyses and code alongside every published quantitative study: we have shared over 100 notebooks to date (https://github.com/ebmdatalab).

Some researchers may feel they have earned a competitive advantage from software developed in-house to make data management and analysis more efficient. In our view such concerns do not legitimise any attempt to withhold code in a way that undermines transparency for reproducibility, but these resource concerns would be better addressed by recognising and supporting good open software contributions. For example, it is already common to cite code that is reused, but these norms could be expanded and reinforced, with compliance audited. Moreover, a strategic approach to fund shared open analytic resources would be likely to produce better software than the current code produced ad hoc by individual teams, often with duplicated effort.

Best practice

Overall there is much to be done. Firstly, journals should ask all submitting authors to share adequately documented code as supplementary material on publication and audit compliance. Secondly, institutions should ensure researchers can access tools and training to support sharing and other important practices such as code review and version control. Thirdly, as well as sharing their code, researchers should give credit when reusing others’ work and endeavour to critically review code as they do

Correspondence to: B Goldacre ben.goldacre@phc.ox.ac.uk

other aspects of a study's methods. Finally, funders have an important role: they should require all grant recipients to share code, in the same way that many already mandate sharing of data and results[15]; they should audit compliance and review applicants' previous sharing when assessing new applications; and they should explicitly support collaborative development of open analytic tools.

This is not an exhaustive list, and we are keen to hear further suggestions as well as objections. However, the prize is substantial. It is baffling that we are expected to rely on brief narrative text descriptions for complex technical data analysis. Medical research cannot progress at pace with its most foundational text—the code that analyses the data—withheld from view.

Competing interests: We have read and understood BMJ policy on declaration of interests and declare no relevant interests.

Provenance and peer review: Commissioned; not externally peer reviewed.

1 Aboumatar H, Wise RA. Notice of retraction. Aboumatar et al. Effect of a program combining transitional care and long-term self-management support on outcomes of hospitalized patients with chronic obstructive pulmonary disease: a randomized clinical trial. JAMA 2018;320(22):2335-2343. *JAMA* 2019;322:1417-8. 10.1001/jama.2019.11954 31593277

2 Rinne ST, Lindenauer PK, Au DH. Unexpected harm from an intensive COPD intervention. *JAMA* 2019;322:1357-9. 10.1001/jama.2019.12976 31593255

3 Masum H, Rao A, Good BM, etal . Ten simple rules for cultivating open science and collaborative R&D. *PLoS Comput Biol* 2013;9:e1003244. 10.1371/journal.pcbi.1003244 24086123

4 Perkel J. Democratic databases: science on GitHub. *Nature* 2016;538:127-8. 10.1038/538127a 27708327

5 Van den Eynden V, Knight G, Vlad A, et al. Towards open research: practices, experiences, barriers and opportunities. London School of Hygiene and Tropical Medicine, 2016.10. 6084/m9.figshare.4055448.v1

6 Reality check on reproducibility. *Nature* 2016;533:437. 10.1038/533437a 27225078

7 Munafò MR, Nosek BA, Bishop DVM, etal. A manifesto for reproducible science. *Nat Hum Behav* 2017;1:0021. 10.1038/s41562-016-0021

8 Wood BDK, Müller R, Brown AN. Push button replication: is impact evaluation evidence for international development verifiable?*PLoS One* 2018;13:e0209416. 10.1371/journal.pone.0209416 30576348

9 Naudet F, Sakarovitch C, Janiaud P, etal . Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in *The BMJ* and *PLOS Medicine*. *BMJ* 2018;360:k400. 10.1136/bmj.k400 29440066

10 Freemantle N. Author's reply to Goldacre. *BMJ* 2016;352:i889. 10.1136/bmj.i889 26887936

11 FigShare. https://figshare.com/

12 Open Science Framework. https://osf.io/

13 GitHub. https://github.com/

14 Barnes N. Publish your computer code: it is good enough. *Nature* 2010;467:753. 10.1038/467753a 20944687

15 DeVito NJ, French L, Goldacre B. Noncommercial funders' policies on trial registration, access to summary results, and individual patient data availability. *JAMA* 2018;319:1721-3. 10.1001/jama.2018.2841 29710154