THE REPRODUCIBILITY CRISIS

“If I have seen further, it is by standing on the shoulders of giants.” - Sir Isaac Newton

To build on previous findings, it helps to know exactly what was done. This not only facilitates understanding the previous study’s results and the methodology, but also it is important for confirmation of its findings.

***Computational***reproducibility is the ability to take the raw data from a study and re-analyze it to reproduce the final results, including the statistics.

***Empirical***reproducibility is demonstrated when, if the study is done again by another team, the critical results reported by the original are found again.

Poor **Computational** reproducibility

The COVID-19 crisis highlights that

- A more rapid research dissemination system is developing.

\* Preprints. Open access. Open evaluation.

- Data and code essential. Dynamic documents, with executable simulations, allowing change to parameters.

\* Ability to do robust analyses that others can check becoming more important.

\* Pooling data with others around the world becoming more important.

- Are research students prepared for this?

\* Coursework approval process too slow to provide the appropriate training in this changing landscape.

Economics Reinhart and Rogoff, two respected Harvard economists, reported in a 2010 paper that growth slows when a country’s debt rises to more than 90% of GDP. Austerity backers in the UK and elsewhere invoked this many times. A postgrad failed to replicate the result, and Reinhart and Rogoff sent him their Excel file. They had unwittingly failed to select the entire list of countries as input to one of their formulas. Fixing this diminished the reported effect, and using a variant of the original method yielded the opposite result than that used to justify billions of dollars’ worth of national budget decisions.

A systematic study of economics found that only about **55%** of studies could be reproduced, and that’s only counting studies for which the raw data were available (Vilhuber, 2018).

Cancer biology The Reproducibility Project: Cancer Biology found that for **0%** of 51 papers could a full replication protocol be designed with no input from the authors (Errington, 2019).

Not sharing data or analysis code is common. Ioannidis and colleagues (2009) could only reproduce about 2 out of 18 microarray-based gene-expression studies, mostly due to lack of complete data sharing.

Artificial intelligence (machine learning) A survey of reinforcement learning papers found only about **50%** included code, and in a study of publications associated with neural net recommender systems, only **40%** were found to be reproducible (Barber, 2019).

Poor **empirical** reproducibility

Wet-lab biology. Researchers at Amgen reported shock when they were only able to replicate **11%** of 53 landmark studies in oncology and hematology (Begley and Ellis, 2012). A Bayer team reported that ~**25%** of published preclinical studies could be validated to the point at which projects could continue (Prinz et al., 2011). Due to poor computational reproducibility and methods sharing, the most careful effort so far (Errington, 2013), of 50 high-impact cancer biology studies, decided only 18 could be fully attempted, and has finished only 14, of which 9 are partial or full successes.

Social sciences

**62%** of 21 social science experiments published in *Science* and *Nature* between 2010 and 2015 replicated, using samples on average five times bigger than the original studies to increase statistical power (Camerer et al., 2018).

**61%** of 18 laboratory economics experiments successfully replicated (Camerer et al., 2016).

**39%** of 100 experimental and correlational psychology studies replicated (Nosek et al.,, 2015).

**53%** of 51 other psychology studies (Klein et al., 2018; Ebersole et al., 2016; Klein et al. 2014)

Medicine

Trials: Data for **>50%** never made available, **~50%** of outcomes not reported, author-held data lost at ~7%/year (Devito et al., 2020)

CONSEQUENCES

-Research quality is low, with attendant effects on reputation.

-Science is slower than it should be.

-Politicians cite this as a reason to ignore scientific results (e.g. Bush, Holcombe, et al. 2019).

OPPORTUNITIES

-Action will improve the quality and influence of our research.

-Visible action will improve our reputation.

-Get one step ahead of expectations before they are requirements in funder policies.

-The automation needed for reproducibility will speed our science.

CAUSES (-) and potential university-based FIXES (°)

* Inadequate methods details.
  + eNotebooks – build sharing into lab workflow
  + Reproducibility checklists
* Inadequate and difficult-to-reproduce analyses
  + Literate programming OLEs (RMarkdown, Jupyter, git); builds code sharing into workflow
  + Software Carpentry
  + Data Carpentry
* Poor training in methodology and statistics, including reproducibility
  + Statistics and Methodology OLEs
  + Reproducibility OLE
  + More collaboration with methodologists and statisticians
  + Registered Reports
  + ResBaz (Research Bazaar) - festival of research skills and tool sharing
  + Hacky hours; R Ladies
  + Journal clubs such as ReproducibliTEA
  + Support for reproducibility- or methodology-related seminars in school colloquia
* Bias when analyzing data. E.g., analysis decisions contingent on the data (noise mining).
  + Preregistration of method and analysis plans.
  + Blind data analysts to group assignment.
* Demands to produce novel results breeds p-hacking and questionable research practices
  + Preregistration of method and analysis plans
  + Change of promotion policies (see Univ of Glasgow’s “Re-imagining Research Culture”)
* Low statistical power favours false positives.
  + Collaboration to collect bigger datasets – facilitated by contributorship policy
* Poor authentication of biological materials.
* Publication bias
  + Registered reports
  + Online lab notebooks allow immediate sharing of results
  + Preprints

REFERENCES

Barber, G. (n.d.). Artificial Intelligence Confronts a “Reproducibility” Crisis. Wired. Retrieved January 23, 2020, from https://www.wired.com/story/artificial-intelligence-confronts-reproducibility-crisis/

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533.

Bush, M., Holcombe, A. O., Wintle, B. C., Fidler, F., & Vazire, S. (2019). Real problem, wrong solution: Why the Nationals shouldn’t politicise the science replication crisis. The Conversation. <http://theconversation.com/real-problem-wrong-solution-why-the-nationals-shouldnt-politicise-the-science-replication-crisis-124076>

Camerer, C. F., et al., (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. <https://doi.org/10.1038/s41562-018-0399-z>

Camerer, C. F. et al. Evaluating replicability of laboratory experiments in economics. Science 351, 1433–1436 (2016). DOI: 10.1126/science.aaf0918

DeVito, N. J., Bacon, S., & Goldacre, B. (2020). Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: A cohort study. *The Lancet*, *0*(0). <https://doi.org/10.1016/S0140-6736(19)33220-9>

Ferrari Dacrema, Maurizio; Cremonesi, Paolo; Jannach, Dietmar (2019). "Are We Really Making Much Progress? A Worrying Analysis of Recent Neural Recommendation Approaches". Proceedings of the 13th ACM Conference on Recommender Systems. ACM: 101–109. doi:10.1145/3298689.3347058. hdl:11311/1108996.

Ebersole, C. R. Et al. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology, 67, 68–82. <https://doi.org/10.1016/j.jesp.2015.10.012>

Errington, T. (2019) https://twitter.com/fidlerfm/status/1169723956665806848

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *ELife*, *3*, e04333. <https://doi.org/10.7554/eLife.04333>

Errington, T. (2013). https://osf.io/e81xl/wiki/home/

Glasziou, P., et al. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet*, *383*(9913), 267–276. <https://doi.org/10.1016/S0140-6736(13)62228-X>

Ioannidis, J. P. A., Allison, D. B., et al. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*(2), 149–155. <https://doi.org/10.1038/ng.295>

Klein, R. A., et al. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. <https://doi.org/10.1177/2515245918810225>

Klein, R. A., et al. (2014). Investigating Variation in Replicability. Social Psychology, 45(3), 142–152. <https://doi.org/10.1027/1864-9335/a000178>

Nosek, B. A., Aarts, A. A., Anderson, C. J., Anderson, J. E., Kappes, H. B., & Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716–aac4716.

Prinz, F., Schlange, T. & Asadullah, K. Nature Rev. Drug Discov. 10, 712 (2011).

Vilhuber, L. (2018). Reproducibility and Replicability in