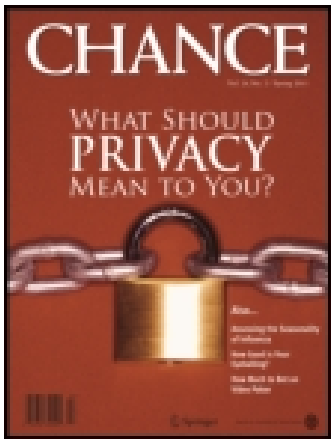


CHANCE
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucha20

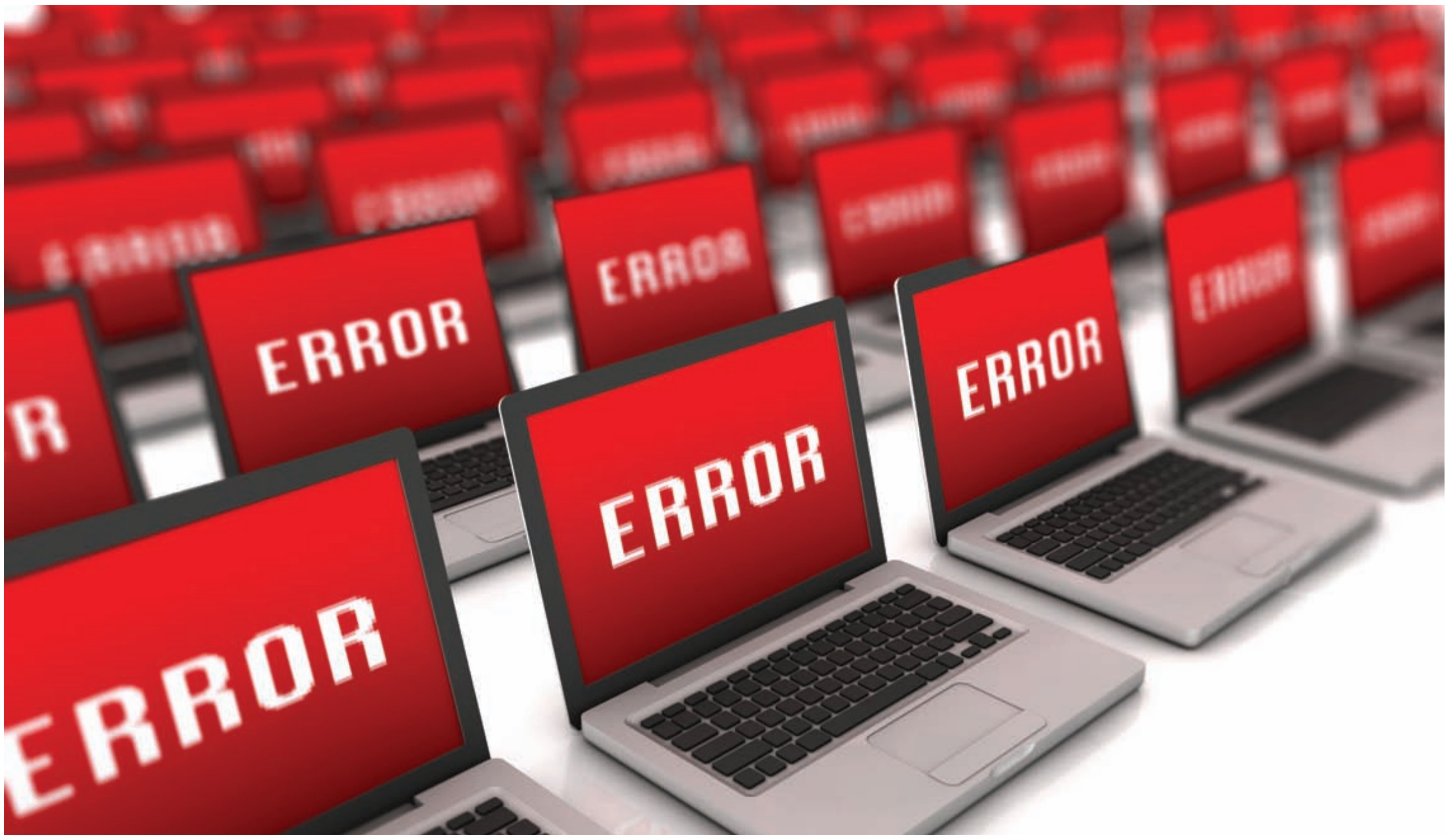
The Problem of Reproducibility

Darrel Ince [a]
[a] Open University

Published online: 10 Sep 2012.

To cite this article: Darrel Ince (2012) The Problem of Reproducibility, CHANCE, 25:3, 4-7

To link to this article: http://dx.doi.org/10.1080/09332480.2012.726554



The Problem of Reproducibility

Darrel Ince

A central tenet of computing is Moore's law. As expounded by Gordon Moore, co-founder of Intel, it implies that the power of a computer doubles every two years. When many talk about the advances computers have made possible, they often quote this law. What they forget, though, is that other factors also have been important, including the increasing miniaturization of components (RFID computers have even been attached to ants), major reductions in cost, and an increase in sturdiness (computers are attached to the hulls of ships to measure ocean temperatures). This has given rise to a major data explosion.

The computer has provided huge opportunities for researchers but, at the same time, made their job much more difficult in terms of validation. The aim of this article is to discuss some of these difficulties and to point the way forward.

Problems with Data

The intrusion of the computer within emerging areas of research is particularly worrying in areas such as genomics. In 2009, John Ioannidis and a number of colleagues published a landmark paper in *Nature Genetics* that examined research using a technology known as a microarray. Microarrays are a vital technology in bioinformatics. They are used by scientists to measure the expression levels of a large number of genes simultaneously. They generate large volumes of data.

Ioannidis and his colleagues examined 18 articles published in *Nature Genetics* that employed microarrays. They evaluated and tried to reproduce a table or figure from each paper. They reproduced two analyses in principle and six partially or with some discrepancies; 10 could not be reproduced. Ioannidis and his colleagues encountered problems with data reproducibility, the documentation of data, and the documentation of the processing and analysis that took place.

As well as the problems detailed by Ioannidis and his colleagues, there are also problems with statistical analyses.

Statistical Problems

The statistical literature contains a corpus of papers that examine the problems associated with statistical methods applied in medical research. In 2007, Alexander Strasak and colleagues from the Medical University in Innsbruck published an important paper in the *Swiss Medical Weekly*. Their aim in writing the article was to document common errors that occur in medical research by examining past publications that looked at statistical problems in medical domains. Their article brings together and reviews works appearing in major medical and statistical journals that critically examine the quality of medical statistics.

This is an important paper for two reasons. First, it is a major review of statistical problems in published medical research over the last 25 years—articles that examine papers in a specific medical journal, for example. Second, and equally important, it contains an implicit checklist that any researcher who employs statistics can use to evaluate their work.

Strasak and his colleagues looked at five areas of statistics in medicine—statistical design, data analysis, documentation, presentation of statistical data, and interpretation—covering each in terms of advice to the medical statistician. Some of their criticisms, taken from the 47 statistical errors and shortcomings the authors documented, include the following:

- Use of an inappropriate test for a hypothesis
- Inappropriate use of parametric methods
- Failure to use randomization
- Using an inappropriate control group
- Failure to state the number of tails
- Use of the mean to describe non-normal data
- Drawing conclusions not supported by the data

These are just a subset of the problems Strasak and his colleagues documented. They constitute an alarming review of problems with statistics in medicine. The whole paper is worth a careful read by anyone involved in medical statistics.

If this article stopped here, it should raise many alarm bells. However, there are problems researchers face over and above those of data curation, data validation, and specification of processing and poor statistical practice. There are problems with computation. These are both hardware and software related.

Problems with Computation

An area capable of generating error is floating point computation. A computer stores integers exactly; however, it stores floating point numbers in an approximate way. Following is a statement from a major article on floating point problems and solutions by David Monniaux that was published in *ACM Transactions on Programming Languages and Systems* in 2008:

> More subtly, on some platforms, the exact same expression, with the same values in the same variables, and the same compiler, can be evaluated to different results, depending on seemingly irrelevant statements (printing debugging information or other constructs that do not openly change the values of variables).

What this says is staggering: If you have a programming statement that involves floating point variables and constants, the result of that statement may be different depending on program code that came before and did not change any of the values. This is known as an order of evaluation problem, and many programming languages are subject to its ways. It arises from the way a programming language compiler is provided with more flexibility in its optimization strategy. Another problem is that floating point results can differ depending on the computer hardware and software used for implementation.

There are other problems. A major one is that the way floating point numbers are stored on a computer can give rise to errors if those numbers are subject to large amounts of repetitive processing.

Many of the problems with floating point computation have, at best, been solved only partially. What is particularly worrying is that the increasing power of the computer is enabling scientists to carry out more intensive processing with, for example, smaller grids, and leaving researchers of hardware and software issues of numerical error (a relatively small community, publishing in journals not routinely accessed by scientific and medical researchers) working hard to catch up.

Problems with Software Development

There are further problems with computer-based science. Many scientists carry out software development as a peripheral activity to their main work of advancing research hypotheses, gathering data, and validating the hypotheses. Consequently, their software skills are less than those who carry out systems development for a living. There have been many studies of software developed by commercial companies that show errors—many of them latent—being discovered only after many thousands of executions. The general figure given is between 1 and 10 errors per thousand lines of code in industrial software. Such software is produced by industrial staff working in teams for companies that have extensive quality assurance procedures, not by scientists working by themselves within a research environment.

Some progress is being made here. For example, the journal *Biostatistics* appointed an editor responsible for judging the reproducibility of work described in an article. The criteria used in evaluating reproducibility are whether the data used are made available on the journal's website, whether software is provided; and whether the reproducibility editor succeeds in executing the code on the data provided and produces the published results.

Another sign of progress is that the journal *Science* now asks authors for their program code in addition to data, and there is an increasing number of journals that either mandate or strongly recommend code release, a good example of the latter is the journal *Geoscientific Model Development*.

Solutions

Science is subject to statistical errors, computational errors, programming errors, and defects in data curation. Surprisingly, perhaps, I do not find this particularly depressing. Research is a low-yield process and errors have always been committed, even in the days when the computer was something of a curiosity in scientific research. What is troubling is that detecting these errors is getting much more difficult. In the past, researchers could spot poor research without a huge investment of effort. Now that the computer has interspersed itself between the data and results in a nontransparent way, we seem to have real problems in validation. We are, in effect, seeing a new era in which scientific intuition is much less useful. So, what should the research community do to improve current practices and enable reproducibility?

It is clear that we should all champion reproducibility. A glance at the major philosophers of science shows none deviate from the view of eminent philosopher Karl Popper: The result of a scientific experiment should stand until it is falsified and the longer it is unchallenged, the more confident we should be of it. Once a theory has been established, it stays at the party until it is either falsified or modified.

How do you make your work reproducible? There are some bright spots. For example, there is a system called Sweave that packages data files, programs written in the programming language R, and article text expressed in the document mark-up language LaTeX. It enables a researcher who receives the package to reproduce the results and even carry out further experiments such as modifying the data or making changes to the R programs.

If you are not an R programmer, you would be unable to use Sweave. However, this does not mean you should give up on reproducibility. There are a number of valuable aspects of packaging that can be implemented in a barefoot way (I am assuming the work involves some programming, but you can modify my suggestions if you are using a statistical package such as SPSS.):

- Comment your program code or script with cross-references to the parts of the research article to which it is relevant. Also, provide comments in each code module about what that module does: what data it reads, what it does with that data, and what output it produces.
- Make sure each module carries out just one action.
- Provide a document that describes the data you are manipulating. The data in this document—known as metadata—would describe what each element of your data was, what its format was, what its accuracy is. It also would assign a name to it that could be cross referenced in your stats code.
- Provide all the scripts you used to generate figures with a description of the data the scripts processed. For example, provide all your Gnuplot files. Cross reference these scripts to the output of your programs and to figures in your research articles.
- Provide a list of all the files' names and what data they contain (you will need to reference the metadata).
- If you have developed versions of your code and data—for example, when you discovered an error or developed a new version that might carry out extra functionality compared with the version reported in a research article—always package previous versions with a description of the difference between it and the previous version. Don't delete previous versions. Researchers are prone to improve their software after publication; hence, it will get out of synchronisation with the contents of an article.
- Provide all the tests you have programmed that generate results in a publication with a script that executes the tests.

These are processes that have been used in commercial software development over the last 30 years and have resulted in major improvements in quality. They are often used by developers of open-source software. The computer has impinged on scientific research so much that they should be regarded as standard.

I admire Ioannidis for his stance on the responsibilities of science and scientists. At the beginning of this article, I described a key experiment he carried out with a number of colleagues that uncovered problems with the reproducibility of work associated with a heavy-duty technology used in -omics work. In late 2011, he also published an article in *Science* with Muin Khoury that looked at the issue of validation in -omics-based research (e.g., genomics, transcriptomics, proteomics, and metabolomics). This is research that, for example, promises therapies and interventions based on the genetic make-up of a patient. However, it is an area that has huge collections of items of potentially error-prone data generated under different lab conditions by different technologies. The article is one of the best descriptions of problems that beset researchers who use the computer as a central tool in their work. Although it is oriented toward -omics research, the message is relevant to other areas in which the computer is a vital supporting technology. It also contains a number of suggestions to overcome the sort of problems that occur. Some of these, augmented by my own, include the following:

- Data, protocol, and software deposition should be made mandatory for publication and this policy should be applied consistently.

- Data, protocol, and software deposition should be made mandatory for receiving grant funding.
- A grant-funding application for a project whose results depend on the computer should, as a matter of course, be accompanied by a reproducibility plan.
- Funding or other agencies should outsource targeted reproducibility checks to bioinformatics experts, either for studies that have potential for high clinical impact and/or for the small proportion of studies that are to be translated into clinical trials.
- More units should be established in higher education that have reproducibility at their core. Most universities have computer units that provide advice on software issues, and many statistics department offer advice to other departments. However, more integration is required. This is vital. To expect scientists to carry out science and to act as top-quality statisticians and software developers is, in the age of the Internet, grossly unrealistic.
- There should be an increase in funding of the development of more tools that enable the packaging of research materials in a way that repeating an experiment is made much easier.

Clearly, detail needs to be fleshed out and there will be some potential exceptions—studies in which confidentiality issues are a factor, for example. There are also issues about when deposition should occur—on publication date or a relatively short time after—to give researchers who may have committed much financial and intellectual capital in developing software and gathering data to use the products for further research.

Although these suggestions are oriented toward -omics-based research, I would contend they are relevant to all research in which the computer plays an important part. To implement these suggestions, we need not only agreement, but also a lot of hard work from researchers. The hard work has a number of benefits.

David Donoho and colleagues have been at the forefront of reproducible research for many years. He and others who have pioneered the idea have clearly stated a number of reasons for reproducibility. First, if you want to return to your research some time later, you will find it properly documented and the start-up time will be much less than if you had to wade through a large amount of unstructured and undocumented data and software. Second, other researchers will use and cite your work, elevating your reputation within the research community. Third, if you are an academic, you will have generated excellent documentation to show the core of your work to your post-docs and doctoral students and teach them. Fourth, it enables research teams to work together efficiently. Fifth, you can feel a warm glow by not only delivering research results, but also data and software that can be built on by other researchers.

A Problem

I believe that at the heart of reproducibility there is a problem that transcends science. This problem is that, over the last 30 years, universities have been transformed from gentle, liberal institutions to businesses. It is inevitable, and I believe it has brought a number of benefits (e.g., a greater care for students). However, it has promoted a culture in which publication of new results is almost everything and labor-intensive activities such as carrying out confirmatory studies are rated less highly. It also has promoted a culture in which there is pressure on researchers to produce research quickly and neglect the packaging aspects. Change these and we should see major improvements. It is, however, a tough task indeed.

Further Reading

Carey, V.J., and V. Stodden. 2010. Reproducible research concepts and tools for cancer bioinformatics. In *Biomedical informatics for cancer research*, ed. Ochs, M.F., J.T. Casagrande, and R.V. Davuluri. New York: Springer.

Donoho, D.L., A. Maleki, I.U. Rahman, M. Shahram, and V. Stodden. 2009. Reproducible research in computational harmonic analysis. *Computing in Science and Engineering* 11(1): 8–18.

Donoho, D.L. 2010 An invitation to reproducible computational research. *Biostatistics* 11(3): 385–388.

Hey, T., S. Tansley, and K. Tolle, ed. 2009. *The fourth paradigm: data-intensive scientific discovery*. Microsoft Research. *http://research.microsoft.com/en-us/collaboration/fourthparadigm*.

Hothorn, T., and F. Leisch. 2011. Case studies in reproducibility. *Briefings in Bioinformatics* 12(3):288–300.

Ince, D.C., L. Hatton, and J. Graham-Cumming. 2012. The case for open computer programs. *Nature* 482(7386):485–488.

Ioannidis, J.P.A., D.B. Allison, C.A. Ball, et al. 2009. Repeatability of published microarray gene expression analyses. *Nature Genetics* 41(2):149–155.

Ioannidis, J.P.A., and M.J. Khoury. 2011. Improving validation practices in "omics" research. *Science* 334(6060):1230–1232.

Monniaux, D. 2008. The pitfalls of verifying floating point computations. *ACM Trans Programming Languages and Systems* 30(3):1–41.

Peng, R.D. 2009. Reproducible research and biostatistics. *Biostatistics* 10(3):405–408.

Strasak, A.M., Q. Zaman, K.P. Pfeiffer , G. Göbel, and H. Ulmer. 2007. Statistical errors in medical research—a review of common pitfalls. *Swiss Med. Wkly* 137(3-4):44–49.

About the Author

Darrel Ince is professor of computing at the Open University. His research centers on computation and software engineering. He is the author of more than 20 books, including the computation papers of Alan Turing.