Reproducible Research: a Dissenting Opinion

Chris Drummond
National Research Council Canada
Ottawa, Ontario, Canada, K1A 0R6
Chris.Drummond@nrc-cnrc.gc.ca

Abstract

Reproducible Research, the de facto title of a growing movement within many scientific fields, would require the code, used to generate the experimental results, be published along with any paper. Probably the most compelling argument for this is that it is simply following good scientific practice, established over the years by the greats of science. It is further claimed that misconduct is causing a growing crisis of confidence in science. That, without this requirement being enforced, science would inevitably fall into disrepute. This viewpoint is becoming ubiquitous but here I offer a dissenting opinion. I contend that the consequences are somewhat overstated. Misconduct is far from solely a recent phenomenon; science has succeeded despite it. Further, I would argue that the problem of public trust is more to do with other factors. I would also contend that the effort necessary to meet the movement's aims, and the general attitude it engenders, would not serve any of the research disciplines well.

1 Introduction

There is a strong movement within many research communities, particularly those where computers are an essential part of experimentation, which would require the publication not only of a paper but also of all the computational tools and data used to generate the results reported therein. The code and data would be run together to prove that the results given in the paper could be reproduced. This movement has the de facto title of "Reproducible Research". Its growing influence can be seen by the increasing number of workshops and conferences on the topic in diverse areas of research (AAAS, 2011; AMP, 2011; NSF, 2010; ENAR, 2011; SIAM-CSE, 2011; SIAM-Geo, 2011).

One motivation for this movement, and its rapid expansion, is a series of well-known frauds that have occurred relatively recently. The number seems to be growing and has resulted in an increasing number of retractions from our journals, including *Science* and

Nature. Probably the most infamous episode being the Duke University cancer trials. When Anil Potti was found to have inflated his CV (Cancer Letter, 2010), it quickly became apparent that there were also major flaws with the data used to support his conclusions. This and other examples of scientific misconduct have been reported internationally in prestigious newspapers (New York Times, 2011, 2012; Economist, 2011; Guardian, 2011). This raises the concern that if not addressed the general public will inevitably lose its trust in science. This movement is progressively gaining more and more support in such diverse communities as Artificial Intelligence, Bio-statistics, Geoscience, to name but a few. What is surprising, perhaps, is the lack of any opposition to this movement. One argument for this would be that the problem is clear cut, the answer obvious. However, here I offer a dissenting opinion. In this paper, I raise some issues which should, at the very least, cause a pause for reflection.

I take the main arguments for Reproducible Research from a paper of the same title, representing the conclusions of the Yale Round-table on Data and Code Sharing in the Computational Sciences (Stodden, 2010). The round-table attracted 21 people from a good cross section of disciplines: Law, Medicine, Geoscience, Statistics, Physics, Bio-informatics and others. I believe it represents a broadly held viewpoint as evidenced by the number of researchers involved and the fields they represent. I also chose this paper because of the clarity of its position; clearly time was spent to work out the details and implications of their proposal.

I believe the main arguments for Reproducible Research are the following:

1. It is, and has always been, an essential part of science; not doing so is simply bad science.

2. It is an important step in the "Scientific Method" allowing science to progress by building on previous work; without it progress slows.

3. It requires the submission of the data and computational tools used to generate the results; without it results cannot be verified and built upon.

4. It is necessary to prevent scientific misconduct; the increasing number of cases is causing a crisis of confidence in science.

In the rest of the paper, I intend to show that each of these arguments is suspect. My main concern is to address the claim that "Reproducible Research", or some simple variant thereof, is the right and only way to do good science and that to achieve it, we must require the submission of code and data along with any paper. I am convinced that this would be a not inconsiderable burden on writer and reviewer alike. If the result was a considerable reduction in scientific misconduct that might represent a justifiable burden. However, I would suggest that rather than reducing problems it may have the opposite effect. Reviewers under time pressure will be less critical of a paper's content and more concerned about the correct form of the data and code received.

2 A Dissenting Opinion

In this section, I want to clarify what are, I claim, the problems for each of the arguments given for "Reproducible Research". Let me sketch my response here:

1. Reproducibility, at least in the form proposed, is not now, nor has it ever been, an essential part of science.

2. The idea of a single well defined scientific method resulting in an incremental, and cumulative, scientific process is highly debatable.

3. Requiring the submission of data and code will encourage a level of distrust among researchers and promote the acceptance of papers based on narrow technical criteria.

4. Misconduct has always been part of science with surprisingly little consequence. The public's distrust is likely more to with the apparent variability of scientific conclusions.

2.1 Not an Essential Part of Science

The first claim for "Reproducible Research" is that it has been a essential part of science of science for some substantial length of time. As the round-table puts it "Traditionally, ... papers contained... the information needed to effect reproducibility". Others have argued for the central role of something similar, termed "Replicability". It is important, therefore, to determine the similarities and differences of these concepts. To explore these issues let us look at a few papers in a special issue in the journal *Science* called *Data Replication and Reproducibility* (Jasny et al., 2011). The introduction suggests there is a single notion of replication and what is of interest is how it is achieved across multiple fields. Its importance, though, is not in dispute. As the first sentence claims "Replication ... is considered the scientific gold standard." and in the first paper, Peng (2011) argues "Replication is the ultimate standard by which scientific claims are judged."

I contend that the papers actually show that there are quite different views of what replicability means. It seems to cover two distinct ideas. Overall we need to consider three separate concepts: Reproducibility, Statistical Replicability and Scientific Replicability. There seems to be agreement that Reproducibility requires that the experiment originally carried out be duplicated as far as is reasonably possible. The aim is to minimize the difference from the first experiment including its flaws, to produce independent verification of the result as reported. Statistical Replicability addresses the problem of limited samples, that might have produced an apparently good result by chance. The aim here is also to minimize the difference from the first experiment. The single change is that the data, although drawn from the exactly the same source, should come from an independent sample. Scientific Replicability addresses the robustness and generalizability of the result. The point is to

deliberately increase the difference between the two experiments, to determine if the changes affect the outcome. To put it another way, it is the result that is being replicated not the experiment.

Ryan (2011) in his paper on "Frog eating bats" considers both forms. As far as Statistical Replication, he says "Was this observation replicated? Yes, we caught several bats feasting on these frogs". This seems to clearly reflect a statistical concern. Yet, he goes on "we replicated the same experiment in a flight cage". Here, deliberate changes were made to effectively support the validity of the claim. Although Ryan called this the "same experiment", I conjecture that is because he did not consider the changes should be critical to the outcome. In fact, the intent was to show they were not, so that the robustness of the result could be demonstrated. I would go further and say the more changes made the stronger the support, as I have argued elsewhere (Drummond, 2009).

It seems clear to me that reproducibility as proposed by the round-table has never been a central tenet of science. However, like Peng, some might still argue that it is useful as "a minimum standard" when Statistical Replication is not practical to do. It is true that that there is an increasing number of research fields where statistics takes on a fundamental role; Biostatistics and Geostatistics are two such examples. Yet, even in these areas, we primarily seek evidence for or against a scientific hypothesis not a statistical one. Statistical Replication, at least in a formal sense, is tied strongly to the idea of statistical hypothesis testing. This idea was introduced by Fisher in the 1920's and became an integral part of some, not all, sciences much later. This time line would not include many of the major events in science. Therefore, I would claim that to call Statistical Replication the "scientific gold standard" would overstate the importance of its role in science. Surely then, only Scientific Replicability has any real claim to be a gold standard. Reproducibility is far too weak to be considered even "a minimum standard".

2.2 No Single Scientific Method

Even if one concedes that reproducibility has little or nothing to do with the verification, or falsification, of scientific hypotheses, one might still claim that it is an important step in the scientific method. Without such a step, one could go on, it is impossible to build on previous work and scientific progress is slowed. The round-table puts it this way "Reproducibility will let each generation of scientists build on the previous generations' achievements." In a recent *Science* editorial Crocker and Cooper (2011) concur "... [it] is the cornerstone of a cumulative science."

The idea of a clear scientific method that has been, and should be, followed is pervasive and to many persuasive. It pervades education at all levels. It is taught to schoolchildren. It is taught to undergraduates. It is even taught to graduate students. One reason for its popularity is that it makes a clear distinction as to what falls under the rubric of science and what does not, separating science from pseudo-science. It also defines some clear and

simple steps that if followed should produce solid science, a clear pedagogical advantage. The round-table supports the idea of such a method with reproducibility as a critical step, "To adhere to the scientific method . . . we must be able to reproduce computational results." They go on to argue that following this method should be strongly encouraged through the policies of journals and funding agencies. Crocker and Cooper (2011) criticize journals for putting up barriers that discourage reproducibility, "Despite the need for reproducible results . . . findings are almost impossible to publish in top scientific journals."

This cumulative view of science is not universal by any means. Polanyi (1958) was an early critic of the idea of a single scientific method. His work strongly influenced Kuhn (1962) who, in his famous book *The Structure of Scientific Revolution*s, contended that science progressed through a series of paradigm shifts, rather than an incremental manner building on past successes. The philosopher Feyerabend (1970) went even further saying "The only principle that does not inhibit progress is: anything goes. " Some might feel that we should not worry too much about what philosophers say. As the science historian Holton (1986) puts it "the perception by the large majority of scientists, right or wrong, that the messages of more recent philosophers . . . may be safely neglected.". But it is not only the philosophers who feel that the idea of a single method is a considerable oversimplification. Polanyi was a scientist in his own right and Bridgman (1986), in his book *Reflections of a Physicist*, argued " . . . there are as many scientific methods as there are individual scientists".

It would seem that the claim for a single scientific method is at best debatable and therefore any argument for the necessity of particular steps is somewhat suspect. Although science progresses by ideas shared throughout a community, this is not to commit to a method where progress is achieved by taking the precise results of one paper and trying to incrementally improve on them. I have written elsewhere that we should be neither too orthodox or too anarchistic (Drummond, 2008). We need to find a balance between strongly enforced standards and a free for all that makes it difficult to assess another's research. I believe, we already have a generally shared sense of what it means to be scientific but to enforce much narrower standards would seem to be of dubious merit and without historical justification.

2.3 Submission of Data and Code Counterproductive

One might still argue that, as any author should have the code and data readily available, the additional costs of submission would be minimal. Peng (2011), for example, suggests that "Publishing code is something we can do now for almost no additional cost." Thus any tangible benefits should come cheaply. I have no wish to argue against the voluntary submission of code, although I would seriously question its value. Submitting code – in whatever language, for whatever system – will simply result in an accumulation of questionable software. There may be a some cases where people would be able to use it but I would doubt that they would be frequent. As Hamilton (1990) points out, many papers are

uncited and others have only a few citations. It is reasonable to infer, even setting aside any difficulties with execution, that the majority of code would not be used. The round-table is clearly concerned about problems arising from non-executable code. To address this it proposes "A system with a devoted scientific community that ... maintains the code and the reproducibility status." The round-table recognizes that this would be a not inconsiderable burden to the community, yet contend that the benefits would be large enough to significantly outweigh the costs.

I am less convinced that this trade-off is a good one. In fact, I am concerned that not all the costs have been identified and the apparent benefits might not be realized at all. Firstly, the process will seem to many to be little more than a policing exercise, checking that people have done exactly what they claim. This will undermine the level of trust between researchers that is important to any scientific community. I would question the old saw of "if you are not guilty, you have nothing to fear", the very policy would have negative consequences. Secondly, even if you feel that the effort needed is not large as the round-table suggests, it will clearly be another stage added to the reviewing process. Already, there is a large, and ever growing, workload for reviewers. The increase is due mainly to what I will term "paper inflation". It arises from the increasing pressures on scientists to publish and publish often. More papers can mean greater funding for an established scientist. More papers can mean a better position for a graduate student or post doc. I believe this increasing load is already reducing the time spent on each review. It also makes it more difficult to find an expert on any particular topic who has time for a review. When reviewers are under time pressure they will tend to make judgments that are easy to appraise and justify, e.g. checking that certain narrow technical criteria have been met. What the round-table proposes will increase this load substantially. We should be working to reduce, not increase, reviewer workload. I would claim that careful reviewing by experts is a much better defense against scientific misconduct than any execution of code.

2.4 Misconduct in Science is not New

One motivation for "Reproducible Research" is a putative recent increase in the number of cases of scientific misconduct. The round-table states "Relaxed attitudes about communicating computational experiments details is causing a large and growing credibility gap". Crocker and Cooper (2011) voice similar worries "The costs of the fraud ... for science, and for public trust in science are devastating." The round-table goes as far as to say there is an "... emerging credibility crisis.".

Misconduct is hardly new in science. Broad and Wade (1984) give a few examples that are some 2000 years old. They list many, more recent yet certainly not modern, cases, some by very eminent scientists. Some of the published results by Mendel, considered the originator of the study of genetics, are somewhat too good to be true. Even Newton is not above reproach. This is certainly not to condone such behavior but only to wonder how

science has been so successful in the past if its effect is so devastating. One case that is often on peoples minds when this issue is discussed is the announcement of "Cold Fusion" in March 1989 by Pons and Fleischmann. I concede this would seem to be a prototypical case showing the importance of reproducibility. However, this discovery would have had far greater impact than the vast majority of scientific results. The discovery of fusion that took place at low temperatures would have considerable consequences to the field of physics and enormous societal impact as a means of producing energy with little environmental consequence. Many scientists did attempt to reproduce the result and failed. In the end, the impact of this announcement was short lived. In different fields, there may be rare cases of very pivotal results that would also benefit from reproduction. I would doubt that these are anywhere close to the norm, even in our top journals. Again, researchers in those fields would identify what is important and challenge the appropriate result. There seems little need for additional safeguards.

I am also convinced that misconduct is not the main reason some of the general public have little trust in science. I would suggest that the public are mostly concerned about science when it affects them directly. An example is a health issue, such as the question "is coffee good or bad for me?" Science, as reported in the media, fails to give a clear answer. In the eighties, a link with established with pancreatic cancer (New York Times, 1981). In the nineties, coffee was found not to influence the frequency of bladder cancer (Chicago Sun-Times, 1993). Early in the new millennium carcinogens were found in coffee (Sunday Times, 2006). More recently coffee has been linked to reduction in brain cancer (USAToday, 2010). For scientists, this may seem unproblematic; these cases are not mutually contradictory. Even if they were, the overturning of early results by later studies is normal science. However, many people expect that science should have a single answer particularly about what they consider to be an important issue.

If I am right, it would seem that any crisis of confidence would best be addressed by better informing the general public of the way science works. That there is consensus on a number of broad theories. The idea of scientific consensus has been used to convince the public of the veracity of global warming (Oreskes, 2004). Nevertheless, it is quite common to have experimental results that conflict. That is why meta-analysis is so important in many fields. What worries me more is that some scientists also seem willing to put great faith in the results a single or small number of experiments. This is, at least partially, why they are so upset when misconduct is discovered. Even without misconduct, there are many potential sources of error that can readily creep into experiments. One such source, Meehl (1990) humorously called the crud factor. Perhaps another lesson we might take from the Duke University cancer trials fiasco is that we should be less willing to go to clinical trials based on such limited evidence. If we were somewhat more skeptical about the results of scientific experiments, cases of misconduct would probably have much less of an impact.

3 Conclusions

My main aim in this paper was to raise questions about the centrality of reproducibility in science. Some may generally support the broad idea of "Reproducible Research" without being committed to all the details. However, without those scientific credentials, the arguments for it must be considerably weakened. There may be practical reasons to do it, but such decisions are best left up to the individual researcher. I would contend that any move by editors of journals or funding agencies to enforce the ideas put forward would not serve any scientific community well.

References

AAAS (2011). AAAS annual meeting: Workshop on the digitization of science: Reproducibility and interdisciplinary knowledge transfer.

AMP (2011). Applied mathematics perspectives workshop on reproducible research: Tools and strategies for scientific computing applied mathematics perspectives.

Bridgman, P. (1986). *Reflections of a physicist*. Oxford science publications. Clarendon Press.

Broad, W. and Wade, N. (1984). *Betrayers of the truth*. Random House.

Cancer Letter (2010). Duke finds "issues of substantial concern" and sanctions potti.

Chicago Sun-Times (1993). Coffee 'not factor' in bladder cancer.

Crocker, J. and Cooper, M. L. (2011). Editorial: Addressing scientific fraud. *Science*, 334(6060):1182.

Drummond, C. (2008). Finding a balance between anarchy and orthodoxy. In *Proceedings of the Twenty-Fifth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning III (4 pages)*.

Drummond, C. (2009). Replicability is not reproducibility: Nor is it good science. In *Proceedings of the Twenty-Sixth International Conference on Machine Learning: Workshop on Evaluation Methods for Machine Learning IV (4 pages)*.

Economist (2011). An array of errors, investigations into a case of alleged scientific misconduct have revealed numerous holes in the oversight of science and scientific publishing.

ENAR (2011). Research ethics in biostatistics: Invited panel discussion at ENAR 2011 on the biostatistician's role in reproducible research.

Feyerabend, P. (1970). *Against Method: Outline of an Anarchistic Theory of Knowledge*. Humanities Press.

Guardian (2011). Scientific fraud in the UK: The time has come for regulation.

Hamilton, D. P. (1990). Publishing by – and for? – the numbers. *Science*, 250:1331–2.

Holton, G. (1986). *The advancement of science, and its burdens: the Jefferson lecture and other essays*. Cambridge University.

Jasny, B. R., Chin, G., Chong, L., and Vignieri, S. (2011). Introduction: Special issue on data replication and reproducibility. *Science*, 334(6060):1225.

Kuhn, T. (1962). *The Structure of Scientific Revolutions*. University of Chicago.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, 1(2):108–141.

New York Times (1981). Study links coffee use to pancreas cancer.

New York Times (2011). How bright promise in cancer testing fell apart.

New York Times (2012). University suspects fraud by a researcher who studied red wine.

NSF (2010). National science foundation workshop on changing the conduct of science in the information age summary.

Oreskes, N. (2004). Beyond the ivory tower: The scientific consensus on climate change. *Science*, 306(5702):1686.

Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334(6060):1226–1227.

Polanyi, M. (1958). *Personal knowledge: towards a post-critical philosophy*. Routledge.

Ryan, M. J. (2011). Replication in field biology: The case of the frog-eating bat. *Science*, 334(6060):1229–1230.

SIAM-CSE (2011). SIAM conference on computational science & engineering workshop on verifiable, reproducible computational science.

SIAM-Geo (2011). SIAM geosciences workshop on reproducible science and open-source software in the geosciences.

Stodden, V. C. (2010). Reproducible research: Addressing the need for data and code sharing in computational science yale law school roundtable on data and code sharing. *Computing in Science & Engineering*, 12(5):8–13.

Sunday Times (2006). Cancer chemical found in coffee.

USAToday (2010). Can coffee, tea lower brain cancer risk?