4.4 The Replication Crisis in Research

Many articles have appeared over the last decade or so in eminent journals such as Nature and Science expressing great concern over observations that many scientific and clinical publications, when tested, are showing poor reproducibility. That is, many prominently published papers contain data that others skilled in the art are unable to repeat. Poor reproducibility is bad news because reproducible observations form the bedrock of science. It is important to understand the nature of the problem and how it may be changing attitudes about scientific study design and funding. Positive changes are already evident, such as the rise of the clinical systematic review. As well as this there are significant implications for poor clinician trying to avoid wasting time on a flawed project.

When an observation or experimental result is easily replicated by others it can be accepted as solid progress, a verified fact that can support scientific arguments and serve as a sound basis for further progress. However, various studies attempting to reproduce the central findings of prominent published work have indicated that an alarming proportion either contain flaws or simply contain observations that cannot be repeated[1,2,3]. This is be entirely surprising since most worthwhile scientific studies are technically hard to perform (and to repeat) and the publication peer review process is imperfect; however the problems clearly go deeper than that[2,3]. Lack of replication is likely go some way to explain why so many promising drug targets fail in the early stages of drug development. Indeed, pharmaceutical companies are always painfully aware of the need to verify carefully a new scientific findings before they invest too much in it.

What studies can be trusted to form a basis of new work?

There is no straight answer to this question, only general rules of thumb. A simple one is that older studies have more information that help judge trustworthiness. Thus, if a study has been cited by others and its outcome repeated successfully over a long period then more weight can be placed on it. A study on a clinically important topic published 10 years ago has had plenty of time to be reproduced and built upon, and if it has not (perhaps evidenced by absence of later related publications) then it may not be wrong, but it needs to be marked down. Unlike an old study, a recently performed study may be perfectly executed but it has not had time to earn as much trust.

Studies where the outcomes or observations are made by a simple method can be more robust as there is less chance of systematic errors,. Another good indicator is that a study reports secondary independent parameters that help to validate the main outcome. Thus, patients receiving treatment may show improvement in a targeted function (e.g., limb mobility, cardiac output, urine flow) but should show changes in expected correlates, such as lifestyle parameters. Other indicators that the study has been done well include blinding of researchers during the study, publication of all data (not just selected data), and use of appropriate statistical tests.

What about peer review? Are reviewers no good?

Peer review must be viewed as a filtering system of quality control that is imperfect, although the least imperfect (and most practical) system devised so far. Alternative assessment systems exist such as online feedback on medRxiv and bioRxiv preprint manuscripts but are still developing. There are two particular issues of note with peer reviews. The first is that reviewers can only judge what is put in front of them and ferret out flaws they can detect, and

most do a good job at this. A perfectly executed study may give outcomes that no one can repeat because of some obscure issue: an unstable drug, faulty computer code or a biased patient sampling not evident in the manuscript.

The second issue is that journals (and peer reviewers) can themselves be part of the problem. They expect novel exciting and important results. This creates huge perverse incentives for authors to provide that. If a fascinating outcome has not been seen by 99 researchers but is seen by 1, whose study will be published? If one clinical trial did not show any positive result but a second one did which will be published? This is the well recognised problem of publication bias, which is proving hard to fix although there has been progress.

Why is this happening?
A number of factors contribute to replication issues. Pressure to publish (for career enhancement and funding), pressure to report exciting new results (or risk job loss), poor usage of statistical techniques (very common), protocols that are difficult to follow and studies that have a high risk of systematic bias in collecting data - all add to the problems.

What can be done to avoid this type of trouble?
There is no panacea, but the classic scientific virtues will go a long way: good experimental design, dispassionate concern that truth is uncovered, robust critiques of the work from many sources, a high quality of statistical analysis and interpretation, good mentorship, extensive documentation and a painstaking approach to performing well even the smaller aspects of study protocols. At issue is that the pressures to publish noted above create an ever increasing conflict of interest that undermines those classic scientific virtues

It is also important to be aware of the impact of cognitive biases that affect researcher judgement - we fool no one better than we fool ourselves. Again, the scientific virtues are a large part of the answer. Indeed the scientific method developed in part as a way to overcome such biases. This can be seen in clinical studies, for example, in the need to take seemingly excessive trouble over blinding of both subjects and researchers.

Is there no hope?
There is certainly hope, as replication issues have gained mainstream recognition, leading gradually to improvements and debates over how to respond to new challenges. Should a funding agency spend precious resources to pay for replication studies? Publication bias is being addressed, and publication of purely replicative papers becoming respectable at least in clinical fields. Preregistration of clinical trials and systematic reviews are now a near-universal requirement for publication. A concern for statistical rigor is starting to emerge in journal peer review, and more powerful analysis methods gaining traction. One is the use of a Bayesian statistical framework, which this needs some trouble to use. This views a data analysis outcome less as hard proof than as a *modifier* of a prior belief, to a degree that depends on the strength of the evidence. This approach leads to a conclusion we can instinctively accept: that extraordinary claims need extraordinary evidence, but less so with less extraordinary claims. Thus, smudgy photos may be proof to a court of a speeding offence, but poor proof that Martians are among us.

Another hope is the emerging clinical practice of systematic review and meta-analysis of prior studies, noted above. Not only does this involve careful assessment of biases but it aggregates studies that have their own features (good and bad), and so a diversity of practices may average out their effects.

There are also many other emerging strategies that can help ensure better reproducibility. It is a good idea to read something on the subject and take it seriously. A bracing place to start may be the famous inflammatory essay from 2005 by John Ioannides “Why most published research findings are false” [2]. Alternatively, the interested researcher can enter the terms “scientific reproducibility” or “scientific replication crisis” into a search engine and stand well back.

[1]*Begley C G et al. (2012).* "Drug Development: Raise Standards for Preclinical Cancer Research". *Nature.* ***483****: 531–533.* PMID 22460880*.*

[2]Ioannides JPA "Why most published research findings are false" PLOS Medicine 2005 2(8):e124, PMID: 16060722

Author: Julian Quinn ***Version*** 2.1 (Nov 2020)

Thanks to Dr Venu Chalasani and Dr Richard Piper for reviewing and critiquing to this article.

Royal North Shore Hospital, DIVISION of SURGERY and ANAESTHESIA