ILAR Journal, 2019, Vol. 60, No. 1, 17–23

doi: 10.1093/ilar/ilz015
Advance Access Publication Date: 5 November 2019
Review Article

Reproducibility and Rigor in Animal-Based Research

Malcolm Macleod[1,*] and Swapna Mohan[2]

[1]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom; and [2]National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland (previously in the Division of Policy and Education at the NIH Office of Laboratory Animal Welfare, NIH).

*__Corresponding Author:__ Malcolm R .Macleod, PhD FRCP, Centre for Clinical Brain Sciences, University of Edinburgh, 49 Little France Crescent, Edinburgh EH16 4TJ, United Kingdom. E-mail: Malcolm.macleod@ed.ac.uk.

Abstract

Increasing focus on issues of research reproducibility affords us the opportunity to review some of the key issues related in vivo research. First, we set out some key definitions, to guide the reader through the rest of the paper. Next we consider issues of epistemology, of how animal experiments lead to changes in our understanding of biomedicine and, potentially, to the development of new therapeutics. Here we consider the meaning of statistical significance; the importance of understanding whether findings have general truth; and the advances in knowledge which can result from 'failed' replication. Then, we consider weaknesses in the design, conduct and reporting of experiments, and review evidence for this from systematic reviews and from experimental studies addressing these issues. We consider the impact that these weaknesses have on the development of new treatments for human disease, and reflect on the response to these issues from the biomedical research community. Finally, we consider strategies for improvement including increased use of brief, pre-registered study protocols; pre-registration, open publication and open data; and the central importance of education in improving research performance.

Key words: rigor; reproducibility; research improvement activity; risks of bias; research ethics

Introduction

Definitions of Reproducibility and Rigor

It is important for research users to know how likely it is that reported research findings are true. Early definitions of 'reproducibility' related to the reanalysis of existing data following the same analytical procedures. 'Replication' was held to require the collection of new data, following the same methods, and applying the same analytical procedures. However, the interchangeable use of these terms (and others) is such that 'reproducible research' has come to have broader meaning than perhaps initially intended [1].

Articulating this broader definition, and borrowing from Goodman [1], one might consider a hierarchy of characteristics that might give confidence in the 'truth' of a research finding: first, 'reproducibility' as originally described, based on reanalysis of an existing dataset ('reproducibility of analysis'); secondly, the collection of new data in experiments as identical as possible to the first ('reproducibility of experimental findings'); and thirdly, the deliberate variation of experimental conditions or analytical approaches to establish whether the same conclusions can be drawn ('robustness'). Goodman considers 2 more levels: inferential reproducibility (making the same evidentiary claims for the same analytical findings) and generalizability (the extent to which predictions made by experiments are true outside of a research or laboratory setting; Figure 1).

The main focus of recent concern relates to reproducibility of experimental findings. This has been studied in retrospective observational [2,3] and prospective [4] studies. In these projects, it was not possible to confirm many findings previously considered to be 'true'. For instance, the recent Many Labs 2 replication project successfully replicated only 54% of findings [5]. The Cancer Biology Replication project found several instances of failed replication, including no evidence that *Fusobacterium nucleatum*

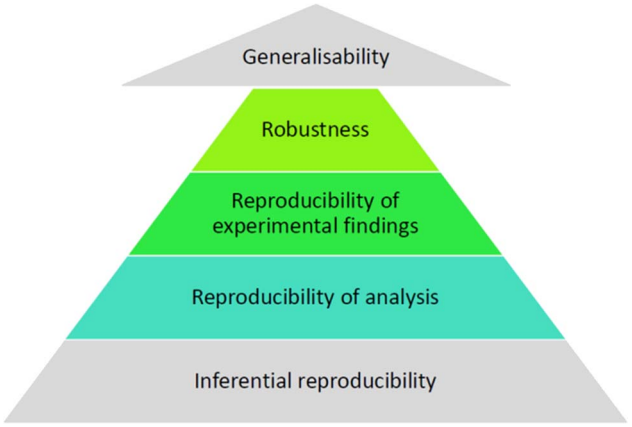

Figure 1: Definitions in reproducibility and replication.

was prevalent in human colonic carcinoma [6], in contrast to a highly cited previous report that it might be causally related to the development of cancer [7].

Failed replication ("reproducibility of experimental findings") in biomedical research may occur if the originator study was falsely positive (by chance or because the internal validity of the study was reduced because the experimental design placed the study at risk of bias); if our understanding of the literature is polluted by publication bias; or in the presence of some unknown (latent) independent variable that influences the phenomenon under study. In this last case, what was intended as a test of replication of experimental findings was in fact an inadvertent test of robustness. Identifying the nature of such previously latent factors that influence the biological phenomena observed might lead to better understanding of the mechanisms being studied. Finally, it may be that our understanding, drawn from the literature, is confounded by publication bias. That is, the reports in the literature represent experiments where an effect was observed by chance but similar experiments where no effect was observed were never published (publication bias [8,9]).

Sensible responses to failed replication might firstly seek to increase the probability that published research is true (through the development of organized research improvement strategies [10]). Next, we might establish a framework to select efficiently which research findings we should attempt to replicate (by establishing if there are study characteristics that predict whether a research finding can be replicated). Next, we could develop strategies to evaluate the robustness of key research findings (based on pre-registered, probably multicenter studies, with deliberate heterogeneity). This would provide an opportunity to substantially increase the value of existing and future research. Finally, adoption of animal study registries (such as preclinicaltrials.eu and animalstudyregistry.org) and the Registered Reports format (see, eg, https://openscience.bmj.com/pages/registered-reports-guidelines/) would do much to address the problem of publication bias.

Epistemology: How Do We Know What We Know?

What Does Statistical Significance Mean? Interpretation of experimental findings often gives cause for concern; many newspaper journalists (and some scientists) believe that a finding at $P < .05$ can be considered 'proven'. However, given low statistical power and an observation that might transform our biological understanding, the positive predictive value of such a finding may be very low indeed [11]. A recent supplement to *The American Statistician* explores these issues in greater depth than is possible here [12]. Where journals and individuals place a premium on novelty, the prior is by definition low, and so for a given statistical power and P value the chances of such findings being true are lower than they are for more ordinary research findings. Various prediction markets employed in the context of formal replication projects give approximate 'priors', derived from the findings of the originator study and the expert assessment of the credibility of those findings. From these, and the statistical power of the replication efforts, one can derive an estimate of the number of studies that will not replicate. For 41 studies in the Psychology Replication Project, the expected number of failed replications was 19 against an observed number of 25 [13]. From this, we can surmise that much of the replication crisis (at least in psychology) arises because of a tendency to accept the findings of such originator studies without acknowledging that the strength of evidence provided may actually be rather low. It seems that scientists, like everyone, love a good story.

From the Specific to the General In the context of in vivo models of human disease, the important considerations are firstly the reproducibility of experimental findings, and secondly their robustness. That is, firstly, can others observe the same effect under the same conditions? Secondly, is the effect observed under a suitably wide range of biologically relevant circumstances?

The exploitation of biomedical research for health gain passes through different domains, usually (but not always) starting with in vitro research and culminating in human clinical trials. At each transition, a judgement is made about whether the accumulated evidence is enough to justify that transition or whether further research in the current domain is required. The etymology of the term 'translation' refers to 'across from' and 'to carry', hence, the carrying across of research findings from 1 domain to another. We might also consider the process of conducting further research in the same domain, seeking more persuasive evidence before attempting translation. Using the same etymology, this might be described as "cis-lation," derived from 'on the side of' and 'to carry,' to mean the re-use of research findings in the same domain.

At each juncture, there is a choice to embark on translation or cis-lation; translate too early, and research in the next domain is based on an insufficiently complete understanding of in vivo biology, and so is at greater risk of failure. The ethical conduct

of all research requires consideration of the potential harms to research participants and the potential benefits from the knowledge that might be gained; the ethical status of research that is not well founded is diminished. Equally, translate too late, and further research conducted in the current domain is unnecessary, and the harms to laboratory animals and the costs of conducting research are not balanced by maximization of benefits from that research. The key consideration is the marginal utility of the next research investment—will we learn more from translation or cis-lation?

Is 'Failed' Replication Always a Bad Thing? Does the problem of failed replication mean that science is in crisis? If the failure to reproduce experimental findings were due to widely prevalent problems with originator experiments (risks of bias, or selective publication of positive findings, or flexibility in data analysis), this would indeed be a problem; this is discussed in detail below. However, the many biological variables that might impact observed outcomes and the limited reporting of even the most obvious of these (sex of the animal, housing conditions, circumstances of testing) make attempted reproducibility of many experimental findings difficult, if not impossible. Further, given the opportunities for important variables (eg, source of chow, ventilation, ambient noise) to differ unpredictably between laboratories, it is likely that many attempts at replication of experimental findings are in fact tests of robustness because of variation in latent independent variables differently represented in different research environments. This has been described by Voekl and Würbel [14]. Importantly, it follows that these latent independent variables are inducing important but unexplained biological effects, and attempts to better understand how this happens are likely to lead to increased understanding of basic biology.

For instance, the failure to replicate drug effects on extending lifespan in worms led initially to despair but, many years later, inspired the multicenter US National Institute on Aging (NIA)-funded *Caenorhabditis* Intervention Testing Program. [15] After much effort, the authors were able to show bimodal patterns of worm survival (some long lifespan, some short), with different strains falling into different survival categories in different labs. They could not explain these differences through any known biology. [16] This has provided a starting point for research that should deepen our understanding of the biology of ageing.

Weaknesses in the Design, Conduct, and Reporting of Experiments

There are some features of study design—such as the study population, interventions, and outcomes measured—that are so fundamental to understanding what was done, and to interpreting the findings, that they should be adequately described. Researchers should report all of their findings, not 'cherry-pick' the ones that are most interesting or favorable to their hypothesis. They should choose their statistical tests carefully at the time of conception of experiments and not subject the same data to multiple transformations and tests until they yield the desired answer (ie, p-hacking). They should articulate a hypothesis and then design an experiment to test it rather than developing a hypothesis to explain a series of observations and then claiming that this was the hypothesis they tested. This is also true—perhaps especially true—for observational studies seeking possible associations between certain characteristics and observed outcomes, where the risks of identifying spurious associations are high.

Risks of Bias Identified Through Systematic Reviews Systematic reviews in clinical medicine (as championed by the Cochrane Collaboration [17]) conventionally include an assessment of risks of bias in contributing studies. This practice has been adopted by systematic reviews of in vivo studies (ie, basic science), with reviewers recording whether a given study reports important aspects of experimental design. These may be general aspects of experimental design (such as randomization or blinding) or specific to the research field (for instance, a statement of control of body temperature in focal ischemia experiments). Importantly, these are (at least initially) aspects of study design that those conducting systematic reviews claim to be important, where they believe that studies not done to that standard are at risk of giving erroneous results. Further, it may be that although a publication does not describe blinding, this was in fact performed by the investigators but not reported (due for instance to limitations of space in manuscripts). However, where authors have been required to report whether they have blinded, there is only a small increase in prevalence. [18] This suggests that the problem is not with nonreporting but with blinding not having been performed.

In a meta-analysis, it is possible to group studies according to whether they did or did not report aspects of study design that might reduce the risk of bias (such as randomization or blinding) and to observe whether these groups of studies give different estimates of biological effect. This approach has somewhat limited statistical power [19], but across a range of models of various diseases we and others have shown higher reported efficacy in studies that do not report the blinded assessment of outcome, random allocation to group, blinded conduct of the experiment, or a sample size calculation [20–23].

Coupled with high prior beliefs that these issues are important, along with evidence from other research domains, there is now a consensus that, where possible, scientists should adopt these practices to reduce the risks of bias in their work (in some situations—such as observing behavior in transgenic animals with obvious differences in body habitus—this should be discussed as a limitation). This consensus has been reflected in several key recommendations including those from an NIH consensus panel [24], the ARRIVE guidelines [25], the US National Research Council, [26] and from the Nature Publication Group (NPG) [27]. For instance, the Landis guidelines recommend that investigators should report on randomization, blinding, sample size calculations, and animals excluded from analysis. Generally, the broader the group that developed the guidelines, the greater their range of applicability and the greater their impact.

Impact of Risks of Bias in Studies Using Laboratory Animals Identified Through Systematic Reviews To really understand the impact of bias in animal studies, we would need to do a series of studies, some at risk of bias and others not, and compare the findings. For instance, a series of studies could be randomized to be conducted either blinded or not—would the results be any different? However, the design of such research on research is challenging. The researchers conducting the individual studies would need to be unaware of the true purpose of the project (lest this impacted their behavior), and this makes recruitment, and issues of consent, difficult. It might be possible to test the impact of training programs for researchers, but this raises ethical concerns about potential harms done to experimental subjects where the investigator had been randomized to the control group.

In the early 1960s Rosenthal and Fode explored the impact of investigator expectations on tests of maze performance in rats [28]. They had graduate students conduct experiments where the students were led to believe that 2 cohorts of rats would have very different performance. The students did indeed observe such differences, but in fact the rats were identical and had been allocated to these different groups at random. The observed differences between the groups were consequently an effect of the observers' expectations and unintentional bias.

These findings are supported by a systematic review of 10 studies where outcome was assessed by both blinded and unblinded observers [29]; the observed effects were 59% greater in the unblinded studies.

Despite evidence from both observational and experimental studies, the quality of published scientific works as measured against relevant guidelines for reporting (for instance randomization and blinding) remains low. In a random sample of publications from PubMed published between 2008 and 2012, reporting of randomization was 33% and reporting of whether assessors were blinded to experimental group was only 7% [30]. In work from leading UK institutions (as determined by a national research assessment exercise) published in 2009 and 2010, only 1 of 1172 manuscripts reported whether experiments were randomized or blinded were designed on the basis of a sample size calculation and described criteria for excluding animals from analysis; 68% reported none of these things [31].

What Are the Costs and Consequences of Poor Practice? Scientific findings may never be used, be used to inform further research, or be used to inform changes in policy or practice. Where a scientific literature informs further research, it is important that that information is reliable; it is important for research users to know how likely it is that reported research findings are true. Where further research is planned based on flawed findings, the costs are propagated to that new research, which has a much lower chance of succeeding. The NXY059 clinical trial program in stroke was based on data from animal studies that, with the benefit of hindsight, were unreliable—most studies did not report randomization or blinding of investigators during the experiment or when assessing outcome. The NXY059 clinical trial program involved some 5500 patients, cost millions of dollars, and the publication of the statistically neutral results of the SAINT II trial [32] was associated with a substantial decrease in the market capitalization of the pharmaceutical company concerned. More recently, the development of a new vaccine for tuberculosis has been criticized on the grounds that the animal studies on which it was based were at substantial risks of bias [33]. The costs of failure were again high. The severe adverse effects experienced by participants in the phase I clinical study of the humanized monoclonal CD-28 agonist have been attributed to insufficient research done on the translatability of results between species (ie, from nonhuman primates to humans), specifically in the difference in mechanism of action of the study drug at the cellular level [34]. In each case, trial participants were exposed to potential harms in a context where the suggested benefits were less than predicted, because the premise for the trials was flawed.

The Community Response Interestingly, efforts to improve the conduct and reporting of animal research have met with some resistance, often from unexpected quarters. For instance, in 2015 the president of FASEB argued that guidelines introduced by the US National Institutes of Health were

> . . . premature, ..(lacked) the appropriate flexibility to be applicable across disciplines, and (were) likely (to) produce significant and in some cases unjustified burden associated with the preparation and review of scientific manuscripts. [35]

In a context where it has been argued that as much as 85% of the $300bn annual investment in biomedical research may be wasted [36] and where only 1 of over 1000 studies from leading UK institutions reported key aspects of study design [37], this seems at face value to have been an unusual although not a unique response. In fact, and in contrast to the reservations expressed above, FASEB has played an important role in addressing these issues, for instance in their Transparency and Rigor Initiative and their recommendations Enhancing Research Reproducibility. [38]

Institutional efforts have largely focused on the response to deliberate and unacceptable attempts by researchers to subvert the integrity of the scientific process. Such actions include falsification, fabrication, and plagiarism. [39]

One might consider the quality and rigor of research to exist in a spectrum, from these most egregious practices at one end, then through for instance the practice of hypothesizing after results are known (HARKing) and conducting research at risks of bias, to research of higher quality defined for instance by adoption of open science practices and the use of Registered Reports (Figure 2a). An institutional focus on worst practice (Figure 2b) might mean that efforts in research improvement are negatively perceived by researchers, being associated with a small group of malfeasant individuals rather than being something of relevance to all researchers.

Rather, the emphasis might more usefully be on improvement (Figure 2c). No matter how effective and rigorous a scientist is, there will always be room for some improvement, that they may become even better. If community efforts to address these issues were to improve everyone's performance even by a small amount, the impact on the value that accrues from biomedical research would be substantial.

Strategies to Improve Performance

Much effort has been expended on the development and implementation of guidelines for the design, conduct, and reporting of in vivo research. These now include the PREPARE guidelines [40], the ARRIVE guidelines [25], field-specific guidelines such as those for in vivo stroke research [41,42], journal-specific guidelines such as those introduced by Nature Publishing Group [27], and guidelines articulated by community groups. Given improvements in the conduct and reporting of clinical research, it was hoped that they would have a similar effect for in vivo research. The reality is more nuanced. It appears that the development, articulation, and endorsement of such guidelines, in and of itself, may do very little to improve performance; more sophisticated implementation strategies may be required. This may involve prioritizing those factors that are considered essential from those that are highly desirable.

Following publication of good laboratory practice guidelines for stroke research, a major journal in the field changed their peer review web platform to require reviewers to assess reporting of key measures to reduce the risks of bias. Over the following 4 years, reporting of design features such as randomization and blinding in manuscripts in that journal improved substantially, [43] and, critically, no improvement was seen in the reporting of animal research in 4 other journals from the same publisher even when adjusted for the disease under study and the species of animal used [44]. Some research domains seem to experience

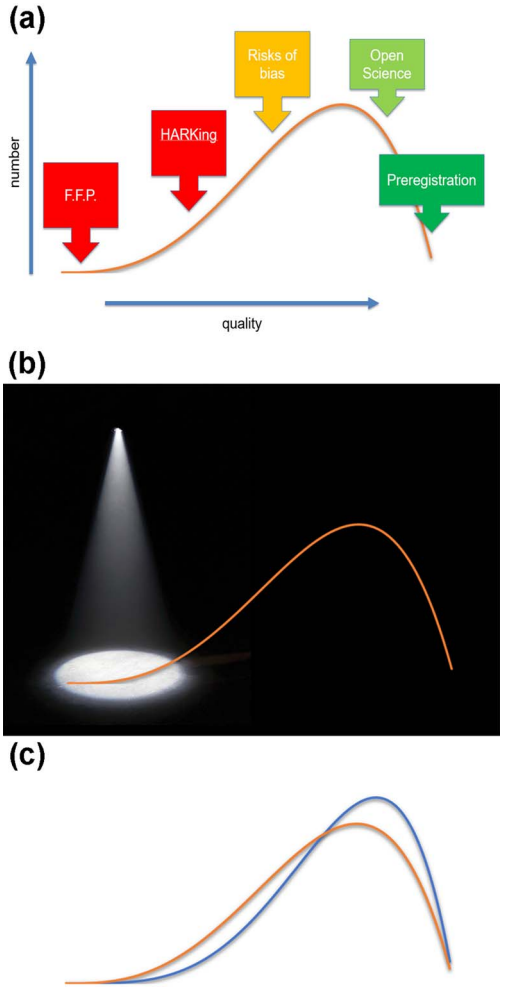

Figure 2: Illustrative histogram of the distribution of research behaviors. (a) Research practices defining points on the distribution. FFP = Falsification, Fabrication, and Plagiarism; HARKing = hypothesizing after results are known. (b) Spotlight on the most egregious practices. (c) Alternative strategy of shifting distribution of performance to higher level (move from gold to blue line).

greater challenges in translation, [45] perhaps because potential treatment effects are smaller or take longer to become manifest, but rigorous approaches to research are in our view important in increasing value across all research domains.

In 2013 the NPG introduced a new checklist for reporting in vivo research, based in part on the Landis criteria [27]. We have measured compliance with that checklist in NPG publications before and after the change in policy in the Nature Publications Quality In Publication (NPQIP) study and found a substantial increase in the proportion of studies reporting each of randomization, blinding, sample size calculations, and details of animal excluded from analysis. The proportion of studies reporting all of these increased from zero to 16%. In matched articles from non-NPG sources no improvement was seen. NPG editorial staff report making substantial efforts for each manuscript to maximize compliance, a process facilitated by the iterative and detailed review process occurring at those journals. In contrast, Baker et al reported that, 2 years after adoption of the ARRIVE checklist at PLoS, there had been little if any improvement in performance [46]. It was argued that journals need to do more to encourage guideline compliance, but in a randomized controlled study (IICARUS) we showed that even a requirement for submission of a completed ARRIVE checklist had no effect on the completeness of reporting of the ARRIVE items, being 0% in both the intervention and control groups [47,48].

Importantly, in both NPQIP and IICARUS, outcome assessments were performed in duplicate by at least 2 trained assessors, with differences reconciled by a third. This gives the opportunity to study the inter-observer agreement in the assessment of each of the checklist items, and we found that for some items agreement was no greater than would be expected by chance alone. This implies either that the wording of the checklist item was not sufficiently clear or that the underlying concept was not well understood even by trained assessors, or both. Although these checklists were developed by eminent and well-informed groups, this draws attention to the importance of testing guidelines under development for their in-use characteristics—that is, are the concepts articulated in such a way that informed scientists are able to agree on whether a manuscript does or does not meet the requirements of a checklist item?

Rigor in Animal Research in the Context of Rigor in Other Research Domains Animal research is by no means the only research domain that has been challenged by a replication crisis. Indeed, difficulties in replication have been observed in every field in which it has been studied. It is therefore reasonable to assume that it this a general feature of research of all kinds and animal researchers are not in any way unusual in experiencing difficulties in replication. Indeed, the NPQIP study suggested that in the in vitro literature, risks of bias were addressed much less frequently than was the case in the in vivo literature. As well as raising concerns for the development of non-animal alternatives, this sets the current performance of animal researchers in a more nuanced context, while accepting that research stakeholders in every domain should be seeking to improve the reliability of the work with which they are associated.

Ex Ante Statements of Experimental Intent One simple approach to improving the provenance of animal research is for investigators to assert, in advance, key aspects of their study. This might be accomplished using the Registered Report framework discussed above, or made available online, with the possibility (for instance on the web pages of the Open Science Framework) of it remaining private until a time of the author's choosing. The level of detail to include is a matter for the individual investigator, but we have suggested that at a minimum it include the study population, hypothesis, intervention, statistical analysis plan, primary outcome measure, and sample size calculation [49]. Research plans often change—appropriately—due to unforeseen circumstances. The existence of an a priori study protocol allows the research user to understand what those changes have been and allows the researcher to explain why the change was made.

Of course, the more detailed the protocol, the greater the protection afforded. An analysis plan that clearly describes the criteria for removing an animal from analysis is better than one that vaguely states they will remove animals where induction of the lesion has been unsuccessful. Wicherts et al have helpfully articulated 34 researcher degrees of freedom that should be constrained in a good study such as establishment of inclusion/exclusion criteria, defining outliers, and describing limitations and caveats of the study.

Preregistration, Open Publication, and Open Data In the context of in vivo research, previous publication models were limited by word count limits and page charges, which meant that for most publications the information provided was at best a precis of what had been done, condensed into the smallest possible

space. Demonstration of being the first to study and report a phenomenon was based on the date of submission or publication, and authors were understandably unwilling to share their ideas until they could establish primacy. As a result, they did not share research plans in advance, and the course of their study could only be established by inspection of signed and dated laboratory notebooks. Work was published months or years after the analyses were complete, following several rounds of peer review, a process often repeated at several journals until a home for the work was found. If the intention of (at least some) research is to contribute to treatments for human disease, these delays are measured in human death and disability. Often when work was published, only summary data were available, often only in graphical form, precluding for instance fruitful reuse and reanalysis. We are now in the fortunate position of having online tools that remove all of these barriers, but uptake of these tools is, to date, very slow.

Specifically, the Open Science Framework provides a medium for the confidential deposition of study-related materials, with date of accession recorded, so that these can at a later time point be made public, with a permanent digital object identifier (DOI) locator, at the time of the investigator's choosing. The development of preprint servers where work can be shared publically before peer review (eg, BioRxiv, www.biorxiv.org) and online journals offering post publication peer review allows research findings to be placed in the public domain—with a DOI and visible to indexing services such as Google Scholar—within days of submission. Finally, data archiving tools such as Figshare (www.figshare.org) allow scientists to make available entire complex datasets on which they have based their findings, again with a DOI. That these tools fill an important unmet need is not in doubt.

Challenge of Education There are now many, largely local initiatives seeking to improve research performance, and this is very much to be welcomed. The focus of most begins with efforts to provide education and training to scientists that they might improve their performance. However, scientific research is a complex ecosystem, and behavioral change will require more than education. Michie [50] has identified common themes in the behavioral change literature, articulated as a requirement that individuals have the capability, opportunity, and motivation to do things differently. It is likely, therefore, that research improvement will require complex interventions, and developing and validating these may well be beyond the capability of many research institutions. It would be desirable, therefore, if there were established some larger coalition of institutions, funders, and journals who could work together to establish which interventions work best in which circumstances and to provide support for research stakeholders embarking on improvement activity.

Conclusions

Increasing Ethics by Increasing Benefits

To be ethical, researchers need to demonstrate a positive benefit-harm calculus for their proposed work. For in vivo research, this has classically been approached through efforts to reduce harms experienced by experimental subjects, largely through the application of the three Rs, Russel and Burch's principles that scientific procedures should be carried out with replacement of animals where this is possible, reduction in the numbers of animals used, and refinement of techniques to minimize harms to the animals. In the context of the 3Rs, the benefits accruing from research have largely been taken as self-evident or evidenced by the fact that a funding agency has considered the work to have value. An alternative approach (complementary to reducing harm), with widespread potential application, is to improve the ethical position of research by taking reasonable efforts to increase benefit, as described above, by increasing the value and reliability of research findings.

References

1. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? *Sci Transl Med*. 2016;8(341):341ps312.
2. Begley CG, Ellis LM. Drug development: raise standards for preclinical cancer research. *Nature*. 2012;483(7391):531–533.
3. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*. 2011;10(9):712–717c711.
4. Macleod MR. The reproducibility opportunity. *Nature Human Behaviour*. 2018;2(9):616.
5. Klein RA, Vianello M, Hasselman F et al. Many labs 2: investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*. 2018;1(4):443–490. doi:10.1177/2515245918810225.
6. Repass J. Replication study: fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Elife*. 2018;7:e25801.
7. Castellarin M, Warren RL, Freeman JD et al. Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma. *Genome research*. 2012;22(2):299–306.
8. Sena ES, van der Worp HB, Bath PM et al. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol*. 2010;8(3):e1000344.
9. Tsilidis KK, Panagiotou OA, Sena ES et al. Evaluation of excess significance bias in animal studies of neurological diseases. *PLoS Biol*. 2013;11(7):e1001609.
10. No authors listed. Research integrity is much more than misconduct. *Nature*. 2019;570(7759):5.
11. Mogil JS, Macleod MR. No publication without confirmation. *Nature*. 2017;542(7642):409–411.
12. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$". *The American Statistician*. 2019;73(sup1):1–19.
13. Dreber A, Pfeiffer T, Almenberg J et al. Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*. 2015;112(50):15343–15347.
14. Voelkl B, Wurbel H. Reproducibility crisis: are we ignoring reaction norms? *Trends Pharmacol Sci*. 2016;37(7):509–10.
15. Lithgow GJ, Driscoll M, Phillips P. A long journey to reproducible results. *Nature*. 2017;548(7668):387–388.
16. Lucanic M, Plummer WT, Chen E et al. Impact of genetic background and experimental reproducibility on identifying chemical compounds with robust longevity effects. *Nature communications*. 2017;8:14256.
17. Grimshaw J, Craig J, Tovey D et al. The Cochrane collaboration 20 years in. *CMAJ*. 2013;185(13):1117–1118.
18. The NPQIP Collaborative Group. Did a change in nature journals' editorial policy for life sciences research improve reporting? *BMJ Open Science*. 2019;3(1):e000035.
19. Wang Q, Liao J, Hair K et al. Estimating the statistical performance of different approaches to meta-analysis of data from animal studies in identifying the impact of aspects of study design. *bioRxiv*. 2018;256776.

20. Rooke ED, Vesterinen HM, Sena ES et al. Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis. *Parkinsonism Relat Disord*. 2011; 17(5):313–320.
21. Hirst TC, Vesterinen HM, Sena ES et al. Systematic review and meta-analysis of temozolomide in animal models of glioma: was clinical efficacy predicted? *Br J Cancer*. 2013;108(1):64–71.
22. Zwetsloot PP, Végh AMD, Jansen of Lorkeers SJ et al. Cardiac stem cell treatment in myocardial infarction: a systematic review and meta-analysis of preclinical studies. *Circulation Research*. 2016;118(8):1223–1232.
23. Macleod MR, van der Worp HB, Sena ES et al. Evidence for the efficacy of NXY-059 in experimental focal cerebral ischaemia is confounded by study quality. *Stroke*. 2008;39(10):2824–2829.
24. Landis SC, Amara SG, Asadullah K et al. A call for transparent reporting to optimize the predictive value of preclinical research. *Nature*. 2012;490(7419):187–191.
25. Kilkenny C, Browne WJ, Cuthill IC et al. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol*. 2010;8(6):e1000412.
26. National Research Council. *Guidance for the Description of Animal Research in Scientific Publications*. National Academies Press. Washington, DC. 2011.
27. No authors listed. Announcement: Reducing our irreproducibility. *Nature* 2013;496–398.
28. Rosenthal R, Fode KL. The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*. 1963;8(3):183–189.
29. Bello S, Krogsboll LT, Gruber J et al. Lack of blinding of outcome assessors in animal model experiments implies risk of observer bias. *Journal of clinical epidemiology*. 2014;67(9):973–983.
30. Ioannidis JP, Greenland S, Hlatky MA et al. Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*. 2014;383(9912):166–175.
31. Macleod MR, Lawson MA, Kyriakopoulou A et al. Risk of bias in reports of in vivo research: a focus for improvement. *PLoS Biol*. 2015;13(10):e1002273.
32. Shuaib A, Lees KR, Lyden P et al. NXY-059 for the treatment of acute ischemic stroke. *New England Journal of Medicine*. 2007;357(6):562–571.
33. Macleod M. Learning lessons from MVA85A, a failed booster vaccine for BCG. *British Medical Journal Publishing Group* 2018;360:k66.
34. Attarwala H. TGN1412: From discovery to disaster. *J Young Pharm*. 2010;2(3):332–336.
35. Haywood JR. *Letter re: Principles and Guidelines for Reporting Preclinical Research* 2015. Available at: http://www.faseb.org/Portals/2/PDFs/opa/2015/4.8.15%20FASEB%20Response%20to%20NIH%20Preclin%20Guidelines.pdf. Accessed January 7, 2019.
36. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. *Lancet*. 2009;374(9683):86–89.
37. Macleod MR, Lawson MA, Kyriakopoulou A et al. Risk of bias in reports of in vivo research: A focus for improvement. *PLoS Biol*. 2015;13(10):e1002273.
38. *Enhancing Research Reproducibility: Recommendations from the Federation of American Societies for Experimental Biology* 2016. https://www.faseb.org/Portals/2/PDFs/opa/2016/FASEB_Enhancing%20Research%20Reproducibility.pdf. Accessed January 7, 2019.
39. Bulow W, Helgesson G. Criminalization of scientific misconduct. *Med Health Care Philos*. 2019;22(2):245–252.
40. Smith AJ, Clutton RE, Lilley E et al. PREPARE: Guidelines for planning animal research and testing. *Lab Anim*. 2018;52(2):135–141.
41. Macleod MR, Fisher M, O'Collins V et al. Good laboratory practice: preventing introduction of bias at the bench. *Stroke*. 2009;40(3):e50–2.
42. Percie du Sert N, Alfieri A, Allan SM et al. The IMPROVE guidelines (Ischaemia models: Procedural refinements of in vivo experiments). *J Cereb Blood Flow Metab*. 2017;37(11): 3488–3517.
43. Minnerup J, Zentsch V, Schmidt A et al. Methodological quality of experimental stroke studies published in the stroke journal: time trends and effect of the basic science checklist. *Stroke*. 2016;47(1):267–272.
44. Ramirez FD, Motazedian P, Jung RG et al. Methodological rigor in preclinical cardiovascular studies: targets to enhance reproducibility and promote research translation. *Circ Res*. 2017;120(12):1916–1926.
45. Kola I, Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat Rev Drug Discov*. 2004;3(8):711–715.
46. Baker D, Lidster K, Sottomayor A et al. Two years later: journals are not yet enforcing the ARRIVE guidelines on reporting standards for pre-clinical animal studies. *PLoS Biol*. 2014;12(1):e1001756.
47. Hair K, Macleod MR, Sena ES et al. A randomised controlled trial of an intervention to improve compliance with the ARRIVE guidelines (IICARus). *bioRxiv*. 2018;370874.
48. Hair K, Macleod MR, Sena ES et al. A randomised controlled trial of an intervention to improve compliance with the ARRIVE guidelines (IICARus). *Res Integr Peer Rev*. 2019;4:12.
49. Macleod M, Howells D. Protocols for laboratory research. *Evidence-based Preclinical Medicine*. 2016;3(2):e00021.
50. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement Sci*. 2011;6:42. doi: 10.1186/1748-5908-6-42.:42-46.