Published in final edited form as:

Circ Res. 2017 June 09; 120(12): 1852–1854. doi:10.1161/CIRCRESAHA.117.311114.

I'll have the rigor, but hold the mortis

Steven P. Jones

Department of Medicine—Cardiovascular, University of Louisville, Louisville, KY, USA

Keywords

reproducibility; study design; blinding; transparency; replication

There is no rigorous definition of rigor.

–Morris Kline, 20th century mathematician

Much has been written about a perceived lack of rigor in biomedical research[1], and this deficiency is implicated in the inability to replicate results from one laboratory in another. Such discussions may often be based more on anecdotal information rather than systematic studies; however, it is becoming accepted that clearer methodologies could benefit the scientific community. More importantly, this perceived lack of rigor supposedly underlies, in part, the failure to translate laboratory findings to the clinic. But what does such rigor look like? There are multiple elements of experimental design that can be affected by rigor, including how a study was designed, conducted, analyzed, and interpreted. Whether and how such elements are made evident in published studies is the subject of Ramirez et al[2], which stimulates a broader discussion on topics related to transparency, rigor, and reproducibility, all of which could ultimately relate to a common theme: bias.

There are several salient features of Ramirez et al[2]. Not surprisingly, the authors found that four out of five publications they evaluated did not report blinding. If the literature is replete with studies in which blinding was not used for the purpose of promoting experimental bias (which, I hope, it is not), it would erode one's confidence in the literature. Although it is safe to say that many of these 'unblinded' studies were indeed unblinded for unknown reasons, I am not convinced that 'not reporting blinding' equates to 'not blinding'. That is, many authors may not have reported blinding, though they incorporated blinding into their study execution. For some investigators, blinding has been a natural but unpublicized practice, like calibrating pH meters, pipets, and analytical balances—all things [some] experimentalists do, but do not tell people. So, what should change? Authors should explicitly report blinding or the lack thereof, and editors and reviewers should ask for and recognize when blinding may or may not make a difference. I think a generalized transition to blinding should happen sooner rather than later because it is a simple and valuable addition to the design of most studies.

Address for correspondence: Steven P. Jones, PhD, 580 South Preston Street, Delia Baxter Building, Room 321F, Louisville, KY 40202, Steven.P.Jones@Louisville.edu.

Disclosures: None.

There are other issues underlying the barriers to inter-laboratory experimental replicability. It should be acknowledged that discrepant results between laboratories should not necessarily mean that at least one of the laboratories engaged in substandard practices (e.g., not blinding). Although it is certainly possible that the results of two laboratories appear opposite from one another at the level of their conclusions, there is always an explanation. And while sometimes the explanation for the apparent contradiction may involve marginal practices, many times it does not. In fact, a transparent conversation between the two laboratories may often identify where one laboratory differed from the other. In a utopian scientific landscape, the discrepant parties could discuss it with one another, perhaps even prior to publishing their study. Such an approach would limit the inevitable divisions into factions of the devout and the heretical. It is, however, doubtful that collaboration across seemingly-religious scientific denominations will soon become common practice. Because many scientists are chasing acceptance by glamor journals, whose focus is on items that could potentially be picked up by the lay press, there is no reward for collaborating with other scientists who may have used a different experimental design. In addition to the absence of reward, it seems that some publishing outlets actually relish in conflict and discord. Thus, some journals (certainly not most) seem to be most interested in grabbing attention, and just like much of the lay media, thrive on controversy and fights. Such an ecosystem then encourages and enables studies that antagonize one another, and provoke scientists into making grand claims regarding the implications of their study. After all, haven't heart disease and cancer already been cured dozens of times, at least according to major media outlets? Ramirez and coworkers[2] ask another important question in their study: Is recent cardiovascular literature improving in reporting of experimental rigor? The answer, simply put, is 'not really'. The authors identify a general exception in the stroke literature, which appears to be driven by pro-active decisions at the journal level. The authors also indicate that during the time period of study, the NIH issued guidance on related topics. Specifically, in 2014, the NIH published its guiding principles regarding rigor and reproducibility (https://www.nih.gov/research-training/rigor-reproducibility). Although much can happen in three years, this is a short time to expect a cultural change in the way scientists conduct (or at least report) studies. Although it is true that there may not have been a significant difference in the way literature has been reported since the NIH's guidance, perhaps such an expectation of causality (i.e., between the NIH's 2014 statement and published studies) is premature. There is really little reason to think the 2014 publication of the NIH's principles regarding rigor and reproducibility would have yet changed what is occurring in the published literature—the time from grant preparation to study completion to publication is not short. I agree with the premise—that the NIH's guidance will eventually have a positive impact on methodological rigor—but we have not yet seen the impact of the NIH's engagement of vigilance regarding these principles. Moreover, journals' lack of coherence or action on such expectations has been slow, and even lip-service provided on the topic is not widespread. It is, therefore, not surprising that there has not been a significant change in methodological reporting, except when mandated by a journal.

Journals—not funding agencies—have the greatest potential for immediate and significant impact on methodological rigor and reporting, and by extension reproducibility. Journals must decide that such aspects of a study design/reporting figures prominently into their

decision-making algorithm. Also, it is not simply that journals lack initiative in encouraging transparency and rigor. Some journals indirectly contribute to problems with methodological reporting by restricting their page allowances, and Methods are the first section to be cut or relegated to the supplement. Perhaps the growing glut of journals should show restraint in their publication of an extraordinary number of positive studies at the expense of rejecting negative studies, particularly those performed rigorously. Likewise, scientists should be more careful and transparent in their reporting of methodologies. Unfortunately, many journals will likely continue to bias toward a flashy headline over a carefully conducted study. Thus, some level of reward for experimental rigor, or punition for lack thereof, must occur during the manuscript adjudication process.

Again, the issues triggered by Ramirez et al[2] are important and stimulating on many levels. For example, let us consider the topic of reporting ‘negative’ studies: is it ethical to continue a study that is, at an interim analysis, deemed futile? Clinical research usually involves an intermittent inspection of the data for safety, overwhelming benefit, or futility, and there are many examples of clinical studies being terminated for these reasons. Thus, should preclinical studies include ‘interim looks’ at the data to test for futility (which come with a statistical penalty)? And, if the studies are deemed [statistically] to be futile, how would this be reported? Or, should the negativity of the interim assessment be ignored and the study completed? If so, are there ethical implications? These and other questions should be considered thoroughly before anything approaching a blanket mandate is issued.

In this brief space, there are far too many issues to address, particularly related to rigorous conduct of preclinical research. There is, however, another area of general discussion stimulated by the work of Ramirez et al[2]—that is, the translation of preclinical studies to clinical practice. It is safe to say there is no single type of laboratory study. Some investigations may effectively be final tests of an intervention and these may be conducted in large animal models. In this condition, the most rigorous application of randomized, controlled trial (RCT)-like approaches should be applied to increase intrinsic value. In fact, this issue has been addressed in the cardiovascular field with the Consortium for preclinicAl assESsment of cARdioprotective therapies (aka, CAESAR), which had the goal of identifying promising infarct-sparing drugs by enhancing rigor and reproducibility in preclinical studies of infarct size reduction[3]. CAESAR was an example of conducting multi-site, RCT-like preclinical studies; however, this type of approach is expensive and cumbersome, and there are other types of studies that do not lend themselves easily to RCT-like approaches. Applying RCT-like standards to studies focused on discovery and elucidating fundamental biological processes may not add the same level of value as applying an RCT-like standard would add to an *in vivo* study. In fact, placing RCT-like expectations on a discovery study could limit scientific progress because “rigor pushed too far is sure to miss its aim, however good, as the bow snaps that is bent too stiffly” (Friedrich Schiller). Thus, expectations must be appropriately calibrated for the goals of the study. At another time, we can discuss the (ab)use of statistics in preclinical studies, which was also highlighted by Ramirez et al[2].

Rather than list all of the oft-cited preclinical issues that potentially limit translation to the clinic, consider the share of ownership of this failure from the clinical side. Too often the

failure to translate is blamed solely on various preclinical deficiencies *du jour*—to wit, the present 'reproducibility crisis'. It is potentially curious that preclinical research may be in a crisis because scientists are not reporting methodological details, implementing blinding, and adequately addressing statistical concerns. I am confident that multiple, specific examples of all combinations of such short-comings exist, but they are not likely the sole cause for clinical failure of all interventions. If we rewind the clock several decades—when, presumably, preclinical work was much more 'reliable'—we see the same deficiencies lamented by those describing a reproducibility crisis. That is, *a priori* statistical considerations were not made, blinding was not evident, etc. Yet, studies from such a seemingly lax (tongue in cheek) era translated to the clinic more frequently. So, perhaps more contemporary scientists are simply becoming less scrupulous because of temptations of fame, money, and other such corrupting pursuits.

I would like to mention one last element of the barrier of translation from the laboratory to the clinic. There is often ample evidence that interventions deemed ineffective in preclinical studies should also not have worked in the clinical environment. Many interventions taken to clinical trials had preclinical evidence to cast doubt on whether the intervention should have been expected to work. One example where this has been true is with cardioprotection (i.e., infarct size reduction). Preclinical studies have shown protection of various agents, some of which were known in preclinical studies NOT to be effective if, for example, the duration of ischemia was extended significantly. Not surprisingly, their subsequent clinical trials failed to translate, and the preclinical studies were blamed; however, better design of the clinical trials would have allowed a more direct comparison with the preclinical work. The inherent variation and imprecision of the clinical world plays a seemingly important, yet poorly defined, role as barriers to translation. A well-executed preclinical study will certainly involve a higher level of consistency, rigorous adherence to a protocol, and reporting of endpoints than a typical multi-center, Phase III clinical trial. And, yet, it could still fail when translated to the clinic. Does this imply that clinical trials are fatally flawed? Of course not. The point here is that treatment of human patients differs in heterogeneity and adherence to accepted clinical standards (think Phase IV, community-based experiences), whereas even poorly controlled preclinical studies would be more consistent (in terms of subject to subject variation). Germane to this topic, it is important to note that the majority of preclinical studies do not include risk factors (e.g., diabetes) or polypharmacy—imagine treating heart failure animals, like patients, i.e., with beta-blockers, ACE inhibitors, and aldosterone antagonists. Until the preclinical and clinical sides move closer to one another, the ownership of the failure to translate remains community property because we are part of a larger continuum in our pursuit of knowledge.

In closing, if there is a crisis in reproducibility, journals share with authors at least part of the responsibility (if you read this sentence, *Circulation Research* must be taking seriously their responsibility, too). It is abundantly clear that much more work and thought is needed to address these important topics discussed here and addressed by Ramirez et al[2]. Unfortunately, it is not clear who is sufficiently interested to pay, in terms of time and money, for such an effort. Unless and until funding bodies, and especially journals, establish clear sets of expectations, the perception of turmoil, uncertainty, and irreproducibility will persist. In the meantime, let's start being more rigorous—it won't kill us.

Acknowledgments

Funding: NIH.

References

1. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016; 533:452–4. [PubMed: 27225100]
2. Ramirez FD, Motazedian P, Jung RG, Di Santo P, MacDonald ZD, Moreland R, Simard T, Clancy AA, Russo J, Welch VA, Wells GA, Hibbert B. Methodological Rigor in Preclinical Cardiovascular Studies: Targets to Enhance Reproducibility and Promote Research Translation. Circ Res. 2017
3. Jones SP, Tang XL, Guo Y, Steenbergen C, Lefer DJ, Kukreja RC, Kong M, Li Q, Bhushan S, Zhu X, Du J, Nong Y, Stowers HL, Kondo K, Hunt GN, Goodchild TT, Orr A, Chang CC, Ockaili R, Salloum FN, Bolli R. The NHLBI-sponsored Consortium for preclinicAl assESsment of cARdioprotective therapies (CAESAR): a new paradigm for rigorous, accurate, and reproducible evaluation of putative infarct-sparing interventions in mice, rabbits, and pigs. Circ Res. 2015; 116:572–86. [PubMed: 25499773]