Carcinogenesis, 2018, Vol. 39, No. 1, 21–25

doi:10.1093/carcin/bgx085
Advance Access publication August 31, 2017
Commentary

COMMENTARY

Getting rigorous with scientific rigor

Lorne J. Hofseth*

College of Pharmacy, University of South Carolina, Columbia, SC 29208, USA

*To whom correspondence should be addressed. Tel: 803-403-5588; Fax: 803-777-2775; Email: hofseth@cop.sc.edu

Two of the cornerstones of science advancement are rigor in designing and performing scientific research and the ability to reproduce biomedical research findings. The application of rigor ensures robust and unbiased experimental design, methodology, analysis, interpretation, and reporting of results. When a result can be reproduced by multiple scientists, it validates the original results and readiness to progress to the next phase of research. This is especially important for clinical trials in humans, which are built on studies that have demonstrated a particular effect or outcome

—Francis S. Collins, MD, PhD, NIH Director (www.nih.gov/research-training/rigor-reproducibility).

If science isn't rigorous, it's reckless. With this, the NIH announced its implementation of rigor and transparency on 9 October 2015 (notice number: NOT-OD-16-011). Now, reviewers of NIH grants are instructed to consider scientific rigor (intertwined with the consideration of sex and other relevant biological variables) in their final evaluation. The aim is for robust and unbiased results to increase the likelihood: (i) for accuracy of results; (ii) that these accurate results can be independently repeated. Most strive for the former, and hope for the latter. We scientists are quite frankly, relieved when our results are repeated. And, for the most part, this is not because we do not believe our results are true. Scientists just inherently question things, including ourselves. Ultimately, the more both personal integrity and scientific rigor are addressed during the journey, the less we (or the public) will question ourselves. However, regardless of the extent of scientific rigor implemented, there is always a nagging question of accuracy of results.

Scientific rigor can be visualized as a ship that has to be built with expert and sturdy craftsmanship. This will not only ensure it is kept afloat, but with ethical scientists to captain the ship, it will keep sailing in the appropriate direction (Figure 1A). The endpoint would be solid, repeatable, enduring data that can be confidently built upon to move a field forward with the full confidence of our colleagues and the public. The front end (the bow) involves proper project design before grant submission and initial project review by funding agencies (e.g. NIH Notice: NOT-OD-16-011). The middle of this ship is in the laboratory, performing experiments during testing of the hypothesis. For the most part, this is not scrutinized by peers, and is investigator-driven in that it is up to the investigator to be rigorous both scientifically and ethically. At the back end are peer and editor reviews during the publication process. Finally, at the far end are the published results and further scrutiny by peers and the public (Figure 1A). All these parts need to be there for the ship to sail in the right direction.

At the investigator level, the individual is charged with 'getting it right'. There needs to be vigilance in our personal integrity and scientific rigor when carrying out experiments to test a hypothesis (Figure 1B). The NIH defines scientific rigor as '*the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results. This includes full transparency in reporting experimental details so that others may reproduce and extend the findings*'. The rules of scientific rigor can differ from study-to-study, depending on methodology (e.g. *in vitro* studies vs. epidemiological studies), but generally include (where appropriate) proper negative and positive controls (in every experiment), appropriate replicates within experiments (as well as repeat experiments—preferably at least in triplicate), randomization, blinding, measures to control bias, controlling for inter-operator variability, robust and accurate statistical methods, appropriate and accurate experimental design (e.g. inclusion/exclusion criteria), suitable and authenticated models, consideration of sex and other relevant biological variables, and others as appropriate to the field. 'Getting it right' is the goal (Figure 1B). To address the complexity and intertwining nature of scientific rigor, it might help to place it into six general categories: insidious, creative, careless, selective, careful, and enduring rigor (Table 1). Unfortunately, placement is often only revealed post-publication or post-award (far end, Figure 1A), which is the premise for this commentary. The official, ethical, and personal rules of scientific rigor on the entire ship (from front to back) must be followed. This increases the likelihood of accuracy and trueness of results to keep the ship sailing forward with personal confidence and the confidence of our peers and the public.

Received: April 12, 2017; **Revised:** July 14, 2017; **Accepted:** December 7, 2017

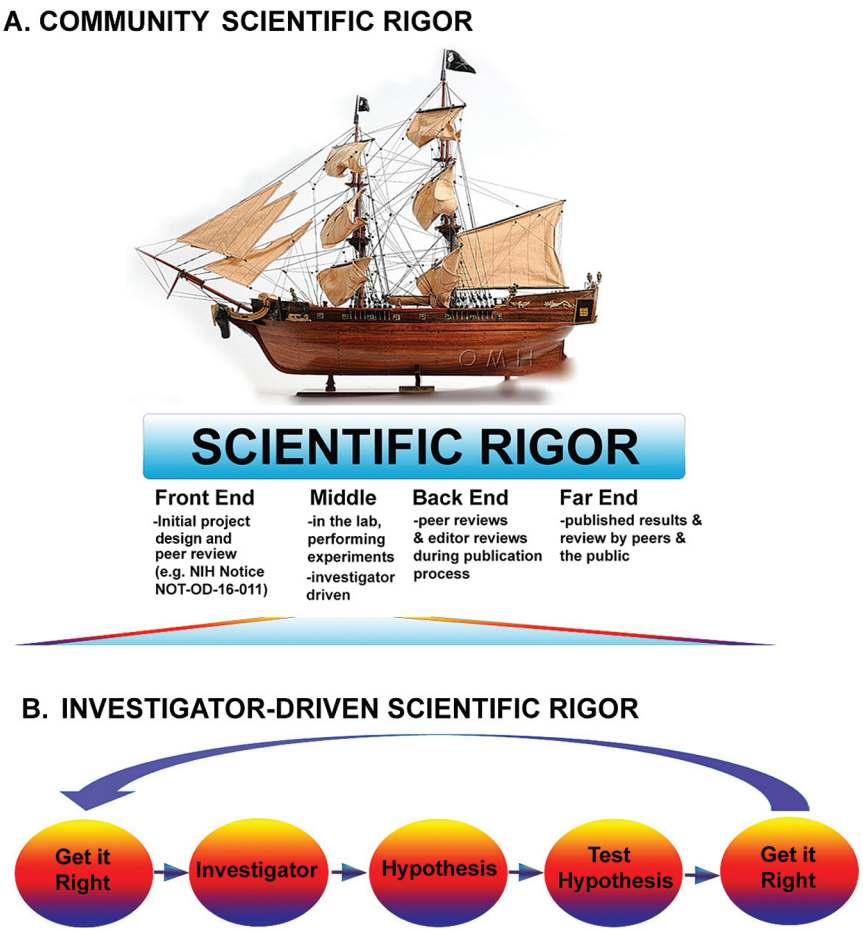

Figure 1. (**A**) Community scientific rigor: scientific rigor policed by the scientific community. (**B**) Investigator-driven scientific rigor. Scientific rigor policed by the individual investigator(s). We are all charged and most of us strive to get it right. In the end, the onus of scientific rigor is on the investigator. If rigor is carried out appropriately, the investigator and the scientific community can build upon those results, and the cycle continues towards (accurate) progress.

Table 1. Categories of scientific rigor

Rigor level	Name	Description	Outcome
Rigor L6	Insidious rigor	Scientist purposely engages in falsifying data from initial grant review to publication	Misleading Misconduct Possibly criminal
Rigor L5	Creative rigor	Scientist deliberately targets or avoids targets where rigor need to be applied; shows best results to support hypothesis; cherry-picking data	Misleading Low chance of reproducibility
Rigor L4	Careless rigor	Scientist randomly applies rigor only when necessary or if asked to (e.g. verify cell lines)	Modest chance of reproducibility
Rigor L3	Selective rigor	Scientist applies rigor where their experience dictates it necessary. Logic.	Good chance of reproducibility
Rigor L2	Careful rigor	Scientist carefully applies rigor	High chance of reproducibility
Rigor L1	Enduring rigor	Results are independently repeated	Reproducible

Insidious rigor (Level 6)

Here, the scientist intentionally and insidiously fabricates methods or data making it appear scientifically rigorous. It compromises scientific rigor due to deliberate falsification or fabrication of data (often due to the external forces described in the accompanying commentary).

The percentage of scientific articles retracted because of fraud has increased approximately 10-fold since 1975 (1). At Level 6, Insidious rigor will lead to, at a minimum, misinformation; and can ultimately lead to findings of research misconduct upon reporting and potentially jail-time. Research misconduct is defined by the US Department of Health and Human Services (DHHS), Office of Research Integrity, as '*a fabrication, falsification, or plagiarism in proposing, performing or reviewing research, or in reporting research results*' (https://ori.hhs.gov/definition-misconduct). Misconduct has been reported as the reason for paper retraction in, on average (depending on the definition of misconduct used), 40% of cases (1–5). Fang *et al.* (1), using the definition from the DHHS Office of Research Integrity, found the majority of retracted articles were due to some form of misconduct (fraud, duplicate publication, and plagiarism) (1).

Interwoven into this alarming data in the front and middle of the ship is Insidious rigor. The investigator will give the perception of rigor by deliberately fabricating methods/data/figures, making it look scientifically rigorous, but mostly lying about it. I liken it to pirates on this ship that must be tossed overboard. However, in an increasing number of cases, we are playing 'catch up'; identifying the fraud long after-the-fact, leading to retraction.

One of the most infamous and globally impactful examples of insidious rigor and scientific misconduct is the study published in The Lancet indicating a connection between autism and the measles-mumps-rubella vaccine (6); eventually retracted (7). Although many high-profile reports have refuted this retracted study (8–10), parents across the world did not vaccinate their children out of fear of the risk of autism. Measles outbreaks in the USA are at record highs, and attributed to the non-vaccination of children (11,12). The Wakefield fraud is one of the most serious frauds in medical history, and used here as an example of the minority of scientists that insidiously give the perception of accurate and rigorous results by purposely misleading colleagues and the public. Reinforcing the timely, dynamic and evolving nature of this topic, a new report from the US National Academies of Sciences, Engineering, and Medicine (NASEM), Fostering Integrity in Research, has recommended specific steps to secure a future based on research integrity and reliability (13). Key recommendations include a new Research Integrity Advisory Board (RIAB) and taking stronger steps to discourage and eliminate practices that harm research and the reputation of its' scientists. Shockingly, there is currently no US organization that promotes research integrity across sectors and disciplines on a continuing basis and as its core mission. This step will help keep the ship sailing in the right direction, with a goal of 'getting it right', and striving for enduring rigor, Level 1. Ultimately, this is a failure of personal integrity (could be described as 'ethical rigor'—regardless of the cause), spawning unrepeatable results, and sinking the ship.

Creative rigor (Level 5)

Although this is not as blatant as insidious rigor, Level 5 creative rigor is serious because it can, and probably will lead to erroneous conclusions. Cherry-picking (selectively choosing) data, only showing a result (different than showing 'representative results') that supports their hypothesis, not repeating a promising result, are all good examples. Whether with good or shady intentions, the investigator has a cognitive bias and deliberately targets data to omit or include in order to be consistent with their original hypothesis. Results are therefore, misleading, and have a low probability of independent reproducibility.

Careless rigor (Level 4)

In attempts to gather data quickly, many scientists are tempted—and may be guilty—at one or more occasions in their career of careless rigor. The scientist randomly applies rigor where it is easy and does not impinge on the urgency of results. Unfortunately, in today's environment, there is always a sense of urgency. So, temptation is high. This type of scientific rigor is a common result of the external pressures described in the accompanying commentary which outlines some of the external pressures (e.g. publish or perish model of academia that prizes production over scientific integrity). Culturally, this category might best represent many of the current ills of the reproducibility crisis. Instead of following the scientific method and setting out to disprove one's own hypothesis, there is a bias to believe the experimental results that agree with the hypothesis; and possibly not confirm such results. Such carelessness can lead to erroneous conclusions (e.g. identifying a band on a western blot without having the appropriate positive and negative controls), with a low to modest chance of independent reproducibility.

Selective rigor (Level 3)

Although the proper rules of scientific rigor are followed at this level, there are certain times that it may not be absolutely necessary to do this with each and every experiment. Selective rigor, Level 3 is more experience, intuition, and logic-driven. The likelihood of reproducibility will vary, and depends not only on adherence to the rules of scientific rigor, but can depend on investigator experience, assumptions and intuition. One simple example here, is running a western blot for the first few times with the appropriate positive and negative controls. Thereafter, experience dictates that the investigator will know where the specific band s/he is looking for. At this point, is it necessary to run positive and negative controls each time a western is run to detect this protein? In the end, if scientific rigor is carried out in a selective way, ultimately this has less of a chance of reproducibility than careful rigor, Level 2. However, the likelihood will depend on the experience of the investigator, and where selective rigor is applied. In the end, performing scientific rigor this way leaves open the possibility of the investigator selecting where to be rigorous inappropriately; therefore leading to ambiguous results, and a low chance of reproducibility.

Careful rigor (Level 2)

Careful rigor, Level 2 or above should be the primary goal, and becoming increasingly mandated. Here, by definition, the scientist is careful to avoid misleading results by following the rules of scientific rigor outlined by funding agencies, journals, publishers, and here. On the back end, journals, publishers and editors have put into place a set of standards ranging from minor to robust guidelines for authors. For example, most, if not all, journals now have supplementary information. Many encourage (but few mandate) authors to supply original data, such as large data sets from genomic studies. More detailed methodology is also encouraged to be supplied in supplementary information. Journals can also encourage (but again, do not mandate) authors to put detailed protocols on protocol exchanges (e.g. the one provided by Nature: Nature Protocols' Protocol Exchange). This allows not only full transparency, but also gives authors an opportunity to refine their published protocols. Many journals, including this one, are members of the Committee on Publication Ethics (COPE), and strive to adhere to its code of conduct and guidelines (http://publicationethics.org/). Additionally, Carcinogenesis and other journals mandate positive and negative controls where appropriate (e.g. on western blots). Journals, such as Nature Research Journals have implemented a checklist (e.g. www.nature.com/authors/policies/checklist.pdf), which includes addressing transparency, replicates, detailing sample sizes and adequate power, defining and detailing statistical methods, inclusion/exclusion criteria for samples or animals excluded from analysis (addressing 'cherry picking'), randomization procedures, approval protocols, antibody specifics, cell line identity, authentication, contamination (e.g. mycoplasma), data accession codes, deposition of data into a public repository, and other details. Although these kinds of checklists address transparency and scientific rigor, it is ultimately up to the individual investigator to carry out experiments in an ethically and scientifically rigorous manner. The development of a rigor

scale by editors, journals and publishers, could be implemented and connected to each published journal article. This would not only help the reader, but hold authors accountable in carrying out proper scientific rigor. No one wants to see their science attached to low scientific rigor, right? It must be cautioned, however, that this may—in itself—seed insidious behaviour for those who want to corrupt the system. Finally, it is important to note that addressing rigor with these checks and balances does not guarantee that we 'got it right'. This might be from an inappropriately used chemical that was simply copied by the next laboratory that confirmed the original results. Outlets such as Nature's Protocols' Protocol Exchange that allow protocols to be updated will help keep the rigor ship sailing in the right direction.

Casadevall and Fang eloquently addressed the issue of scientific rigor in terms of a Pentateuch with five pillars: recognition of error, intellectual honesty, sound statistical analysis, experimental redundancy, and logic (14). Importantly, their commentary is a complementary perspective of scientific rigor, and highlights a key issue that needs to be addressed: personal integrity and scientific rigor are not mutually exclusive. If there is not personal integrity and accountability, scientific rigor will not succeed. Indeed, one could argue that scientific rigor is an ethical imperative because of opportunity costs due to, for example, following up on fraudulent or poorly conducted research. Steps to guide ethical standards that underlies scientific rigor can be taken and include enhancing biomedical training on the subject. For example, required graduate courses on good experimental practice and statistics; journal clubs focusing on methods and approaches; and teaching aids to enhance the quality of peer review (14). Such training and tools are especially necessary to drive home the importance of rigor and scientific integrity when a hypothesis is disproven through stringent and rigorous science with the appropriate positive and negative controls. Although it is difficult to measure the individual scientist's reaction to such results, it is critical that the rigor behind the result is strong, that the scientist re-evaluates the hypothesis without personal bias, and follow the scientific results wherever they may lead. As long as this direction is ethically and scientifically sound, the ship will continue to be guided in the right direction. Addressing these issues will not only improve the likelihood of reproducing intra-lab experimental results, but will increase the probability of independent inter-lab reproducibility, which leads to enduring rigor, Level 1.

Enduring rigor (Level 1)

Enduring rigor, Level 1, is the gold standard, and corroborates previous findings independently and at multiple levels. This is largely dependent on following the rules involved in careful scientific rigor. By definition, those studies that fall into enduring rigor, Level 1 will play a large part in addressing the coined term, 'reproducibility crisis' (15). Although there are important external forces that are more difficult to control (e.g. increased levels of scrutiny, complexity of experiments and statistics, and resulting pressures on researchers) (see commentary by Wyatt and Pittman in this issue), paying careful attention to scientific rigor will play a key role in steering the ship in the right direction.

Recent initiatives aimed at enduring rigor spawning from a perceived 'reproducibility crisis' (16) will help steer the ship in the right direction. Although there are currently over 600 published 'replication studies' in the literature, only recently have there been coordinated efforts to tackle this issue. These include 'The Reproducibility Project', a collaborative effort by the Center for Open Science (https://cos.io/) and Center for Scientific Exchange (https://www.scienceexchange.com/); with results published in eLife (17–27). Another initiative with a similar goal of repeating landmark findings has recently been launched by The Dutch Organization for Scientific Research (NWO) (28). Such initiatives will have the dual benefit of confirming (or disproving) results, and 'keeping us on our toes' and encouraging adherence to the rules of scientific rigor. However, another predictable and cautionary note is that although replication of landmark studies will be embraced, there will be a negative perception towards the science (and scientists) that cannot be replicated in these 'replication studies'. We should not assume insidious rigor, Level 6 (or even Levels 3–5) has taken place. We should not assume there is a 'stowaway' on our proverbial ship. There are natural pitfalls of replication studies: Unknown (or difficult to govern) variables and conditions. Another issue that arises is the accuracy of the repeated results. In many cases, many pieces of these targeted high profile and impactful studies have already been replicated in multiple laboratories. So, for such studies, although the experiments cannot be repeated in a formal setting, the scientific premise is enduring. This is a grey area, and how this plays out is yet to be determined.

Concluding remarks

For the most part, when our own results are repeated, the natural tendency to question our data is dissipated. Data confirmed by at least two separate groups in animal models, *in vitro*, or in humans, not only gives the original investigator the confidence that s/he 'got it right', but gives peers and the public the confidence to build upon these data [albeit carefully, and with continued scientific rigor, both retrospectively (with an open mind) and prospectively]. The scientific process is fundamentally dependent on trust and intellectual honesty. In an attempt to cut corners and rush to publication, unfortunately there is an element of human temptation to disregard proper scientific rigor. However, retracted science (regardless of the reason) is costly and erodes public perception and confidence. In learning from our mistakes, we are now mandating the necessary tools for scientific rigor oversight. Indeed, addressing the rules and regulations of scientific rigor will solve only a portion of the problem. Merged with this is personal integrity. Importantly, it comes from within each investigator (*anyone* who performs a scientific experiment) to carry out proper rigor and transparency with both a strong Scientific and Ethical standard. With this, robust scientific rigor will translate into robust public perception, getting it right for public consumption, and keeping the ship sailing in the right direction.

References

1. Fang, F.C. *et al.* (2012) Misconduct accounts for the majority of retracted scientific publications. Proc. Natl. Acad. Sci. U. S. A., 109, 17028–17033.
2. Budd, J.M. *et al.* (1998) Phenomena of retraction: reasons for retraction and citations to the publications. JAMA, 280, 296–297.
3. Nath, S.B. *et al.* (2006) Retractions in the research literature: misconduct or mistakes? Med. J. Aust., 185, 152–154.
4. Wager, E. *et al.*; COPE Council. (2010) Retractions: guidance from the Committee on Publication Ethics (COPE). Obes. Rev., 11, 64–66.
5. Wager, E. *et al.* (2011) Why and how do journals retract articles? An analysis of Medline retractions 1988–2008. J. Med. Ethics, 37, 567–570.
6. Wakefield, A.J. *et al.* (1998) Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet, 351, 637–641.
7. (2010) Retraction--Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet, 375, 445.

8. DeStefano, F. *et al.* (1999) Negative association between MMR and autism. Lancet, 353, 1987–1988.
9. Taylor, B. *et al.* (1999) Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. Lancet, 353, 2026–2029.
10. Dales, L. *et al.* (2001) Time trends in autism and in MMR immunization coverage in California. JAMA, 285, 1183–1185.
11. Gastanaduy, P.A. *et al.* (2014) Measles—United States, January 1–May 23, 2014. MMWR Morb Mortal Wkly Rep, 63, 496–499.
12. Eggertson, L. (2010) Lancet retracts 12-year-old article linking autism to MMR vaccines. CMAJ, 182, E199–E200.
13. McNutt, M. *et al.* (2017) Research integrity revisited. Science, 356, 115.
14. Casadevall, A. *et al.* (2016) Rigorous science: a how-to guide. MBio, 7, e01902–e01916.
15. Anonymous (2017) The challenges of replication. Elife, 6, e23693.
16. Baker, M. (2015) Reproducibility crisis: blame it on the antibodies. Nature, 521, 274–276.
17. Showalter, M.R. *et al.* (2017) Replication study: the common feature of leukemia-associated IDH1 and IDH2 mutations is a neomorphic enzyme activity converting alpha-ketoglutarate to 2-hydroxyglutarate. Elife, 6, e26030.
18. Shan, X. *et al.* (2017) Replication study: inhibition of BET recruitment to chromatin as an effective treatment for MLL-fusion leukaemia. Elife, 6, e25306.
19. Aird, F. *et al.* (2017) Replication study: BET bromodomain inhibition as a therapeutic strategy to target c-Myc. Elife, 6, e21253.
20. Kandela, I. *et al.* (2017) Replication study: discovery and preclinical validation of drug indications using compendia of public gene expression data. Elife, 6, e17044.
21. Mantis, C. *et al.* (2017) Replication study: coadministration of a tumor-penetrating peptide enhances the efficacy of cancer drugs. Elife, 6, e17584.
22. Horrigan, S.K. *et al.* (2017) Replication study: melanoma genome sequencing reveals frequent PREX2 mutations. Elife, 6, e21634.
23. Horrigan, S.K. (2017) Replication study: the CD47-signal regulatory protein alpha (SIRPa) interaction is a therapeutic target for human solid tumors. Elife, 6, e18173.
24. Nosek, B.A. *et al.* (2017) Making sense of replications. Elife, 6, e23383.
25. Mullard, A. (2017) Cancer reproducibility project yields first results. Nat. Rev. Drug Discov, 16, 77.
26. Lowe, D. (2017) A first look at reproducibility in cancer biology. Sci. Transl. Med. January 19.
27. Baker, M. *et al.* (2017) Cancer reproducibility project releases first results. Nature, 541, 269–270.
28. de Vrieze, J. (2017) 'Replication grants' will allow researchers to repeat nine influential studies that still raise questions. Sci Insider. July 11.