CellPress

Neuron
Perspective

Rigor or Mortis: Best Practices for Preclinical Research in Neuroscience

Oswald Steward[1,*] and Rita Balice-Gordon[2,*]
[1]Reeve-Irvine Research Center, Departments of Anatomy & Neurobiology, Neurobiology & Behavior, and Neurosurgery, University of California Irvine School of Medicine, 837 Health Science Road, Irvine, CA 92697-4265, USA
[2]Neuroscience Research Unit, Pfizer, Inc., 610 Main Street, 5th floor, Cambridge, MA 02139, USA
*Correspondence: osteward@uci.edu (O.S.), rita.balice-gordon@pfizer.com (R.B.-G.)
http://dx.doi.org/10.1016/j.neuron.2014.10.042

Numerous recent reports document a lack of reproducibility of preclinical studies, raising concerns about potential lack of rigor. Examples of lack of rigor have been extensively documented and proposals for practices to improve rigor are appearing. Here, we discuss some of the details and implications of previously proposed best practices and consider some new ones, focusing on preclinical studies relevant to human neurological and psychiatric disorders.

Introduction

This issue of *Neuron* focuses on translational neuroscience and features extraordinary advances in preclinical research that point the way to new therapies for devastating neurological and psychiatric disorders that affect millions. Against this background of accomplishment, however, is the increasing documentation that many published findings are not replicated when there are explicit attempts to do so (Prinz et al., 2011). This includes many promising preclinical findings in different fields. For example, one study in oncology research reports failure to replicate 90% of the tested papers (Begley and Ellis, 2012). In one example in neuroscience, 70 different drugs reported to prolong life in a mouse model of ALS had no significant effect in 221 separate replication experiments involving over 18,000 mice (Scott et al., 2008). Also, in an explicit replication project in spinal cord injury, supported by the National Institute of Neurological Disorders and Stroke (NINDS), only about 10% of the target studies were fully replicated, although there were some partial replications (see Steward et al., 2012 for an interim summary).

In June, 2012, the NINDS sponsored a workshop entitled "Optimizing the Predictive Value of Preclinical Research." This workshop was held because of the growing perception of problems with the basic experimental design of preclinical studies relevant to neurological disorders, a concern about insufficient rigor in experimental execution, and a concern about failure to replicate promising preclinical findings. The purpose of the workshop was to define the problems and possible solutions. One suggested action plan was to develop a consensus short list of standards for preclinical studies in general, and a more specific and perhaps more extensive list for studies involving particular approaches, model systems, or disorders. A first step toward this goal came with the publication of "A call for transparent reporting to optimize the predictive value of preclinical research" (Landis et al., 2012). Based on this, the NIH issued a call for publishers to follow best practices related to reproducibility (http://www.resourcenter.net/images/cSE/files/2014/NIHPrinc.pdf).

Subsequent reports have further documented examples of lack of reproducibility and examples of common practices that lack rigor (Begley, 2013; Howells et al., 2014). These reports have triggered further discussions about changes in grant review criteria on the front end (Tabak, 2014; Wadman, 2013) and review criteria for publications on the back end (Howells et al., 2014; McNutt, 2014). Building on these illustrations of the problems and other suggested guidelines (Henderson et al., 2013; Kilkenny et al., 2010), here we discuss several best practices to improve rigor in preclinical research in neuroscience, focusing on transparency and thoroughness in reporting considerations of experimental design, data analysis and statistics, data inclusion/exclusion, data management, publication, and resource sharing.

What Is a Preclinical Study?

A definition proposed at the NINDS workshop in 2012 was that a preclinical study is one that tests a biological concept in an animal model of a human disorder. However, in biotech/pharma, "preclinical" typically means everything prior to human biology validation studies and/or phase 1 human safety trials, i.e., everything done in cells and animals. Preclinical studies that are conducted in vitro also can suffer from a lack of rigor and thus reproducibility. Thus, we favor the broader definition, which encompasses molecular and in vitro studies that provide mechanistic understanding of disease pathophysiology to in vivo studies using animal models of human neurological or psychiatric disorders. Indeed, validation of a novel molecular disease mechanism in vivo has become a virtual prerequisite for publication in "high profile" journals. Why is this issue especially important for preclinical research? In the normal course of science, many basic discoveries turn out to be wrong. Science is generally self-correcting and moves on (but see Ioannidis, 2012 for an opinion on why this may not always be the case). In the case of translational work whose goal is to accelerate development of therapeutic interventions, promising findings are widely reported in the lay press and may lead to the rapid dedication of significant resources to accelerate translation to the clinic. Thus, time and dollars can be wasted before there is time for the self-correction process to occur. A failure to replicate in this context takes a toll on the public, shaking people's faith in science as fundamental to improving health. It is for these reasons that

failure to replicate preclinical research focused on translating basic science to the clinic has a high potential impact. The large number of failed clinical trials for neurological and psychiatric disorders is often discussed in terms of lack of "construct validity," i.e., that animal models do not adequately represent the pathophysiology of human neurological disorders. These are complicated issues that require detailed consideration of mechanisms (Hyman, 2012; Nestler and Hyman, 2010; Willner and Mitchell, 2002). A consideration of these is beyond the scope of this Perspective, but these issues are discussed elsewhere in this issue (Pankevich et al., 2014).

Experimental Design Considerations

Good Project Management

Designing preclinical experiments is multifaceted, similar in important ways to designing clinical trials, but often involves less formal planning. Clinical trials always predefine the experimental and control groups, number of subjects, methods of assignment to groups, blinding, primary and secondary outcome measures, and stopping points. In contrast, most papers reporting preclinical studies are collections of data generated at different times and carried out with different degrees of preplanning that are bundled together and presented as if the whole thing was a linear plan from concept to conclusions. Typically, the situation is less organized. The story may start with a novel discovery followed by pilot experiments and then more defined experiments and finally end with a test of concept in an animal model of a neurological or neuropsychiatric disorder. Individual labs may not have experience in the different levels of analysis (from molecule to animal model). When such studies are published, there is an implicit assumption (and sometimes explicit statement) that the findings might be the first step in moving the work from the bench to the bedside. But the key prerequisites for this are that the preclinical findings are robust and replicable. Although the early stages of research on a project are often nonlinear, unplanned, and serendipitous, the pivotal data in a preclinical study should be gathered with formal planning to maximize rigor.

The elements to be planned in a preclinical study are similar to what is done in a clinical trial. Determination of endpoint sensitivity, and thus the required "n" of subjects to adequately test a hypothesis, random assignment of subjects to treatment groups, experimenter blinding, and inclusion of positive and negative control groups, are de rigueur in preclinical and clinical studies alike. Careful consideration of design parameters in advance of experimental execution has several benefits. Decision trees that include key go/no-go benchmarks, and the parameters that need to be met for each, not only can help clarify but can also capture thinking in a particular information landscape. Decision points can be revisited as information changes or new data are obtained, and if these are captured in written or electronic lab notebooks, become living documents of critical thinking and iterative hypothesis refinement. Additional benefits include optimization of resource allocation and use, including experimenter time, as well as providing a roadmap to assembling key findings for a publication. A key benefit is that time devoted to experimental design and decision making criteria can highlight ambiguity in hypothesis testing and raise awareness of, and thus reduce unconscious bias contributing to, "testing to a foregone conclusion." With the understanding that most papers include serendipitous findings and data from nonlinear and unstructured assessments, a best practice would be for the Methods section of manuscripts to specifically identify the data sets from preplanned analyses that include the elements below and thus were executed with a high level of rigor.

Pre-experiment Power Calculations

One recurring theme at the 2012 Workshop as well as in other recent summaries of best practices like the ARRIVE guidelines (https://www.nc3rs.org.uk/arrive-guidelines; Kilkenny et al., 2010) was the importance of power analyses to determine the number of subjects required for hypothesis testing to a prespecified level of confidence. Power calculations include consideration of endpoint sensitivity, expected data variability, possible effect size, and desired level of confidence, underscoring the need for deep understanding of these parameters prior to experiment execution. Most institutional animal care and use committees (IACUCs) also require high-level power analyses for studies involving live animals, but the same considerations apply for in vitro studies.

Carrying out meaningful power analyses is not trivial, especially for novel endpoints without prior data needed to predict effect size and calculate expected standard deviation. Also, for IACUCs, power calculations may be used as a means to minimize the number of animals used in a particular experiment, and researchers may be tempted to reduce the number of animals to speed completion of a study. The consequence is that a study with significant results may still have very low statistical power. Indeed, one analysis suggests that low statistical power is an endemic problem in neuroscience (Button et al., 2013). Another recent review on rigor in translational medicine opines that, "It is better to fund a large study that gives useful and reliable data than to fund any number of smaller studies that might appear to provide a reduction in animal numbers but in fact provide data of limited use" (Howells et al., 2014). Larger numbers of animals means higher costs for fewer experiments, and balancing this against "overall impact" is something that funding agencies will need to consider going forward.

Power calculations require a determination of "n" and there is not universal agreement on what "n" means in different neuroscience subdisciplines. A casual perusal of a recent issue of this journal revealed that cellular experiments use, for example, n of spines or synapses, n of neurons or n of slices from the same or different animals. In vivo physiology experiments use n of cells, rarely n of animals, while behavioral experiments use an n of animals.

An important consideration is that elements in a single animal are related samples because they come from an organism that has a life experience that might affect all elements of the sample. But there are situations where numbers of animals are intentionally kept small for ethical reasons (for example, studies involving primates). There are statistical approaches to this problem, for example, using a repeated-measures statistic design where different samples from a single subject (for example different sections) are treated as a repeated measure (Darian-Smith et al., 2014). Enhancing rigor for studies with small numbers of animals that would otherwise be underpowered may require

developing new statistical tools with nested designs. While there is room for differences of opinion and practice, an explicit statement that explains the definition of and rationale for n should be included in the methodology section of manuscripts.

Defining Starting and Endpoints of a Preclinical Study

As discussed above, most preclinical studies differ from clinical trials in that they often start with a serendipitous finding, move to preliminary studies to work out parameters, define conditions, and optimize testing protocols and the sensitivity of endpoints of interest, and then to formally planned tests of concept. It is acceptable (and may be necessary) to include data from pilot experiments in the final report, but it is important to distinguish between "preliminary" and "final" experiments because of the probable difference in formal planning. For example, studies of disease models where pathophysiology evolves over months to years, or models of brain or spinal cord injury, often include data from experiments carried out over prolonged time intervals, sometimes several years. In may be appropriate to include animals from preliminary experiments in the final analysis, but it is important to explain how the data were compiled. Conditions can change over time, especially when measures depend on personnel whose skills may evolve (up or down!) or when there is personnel turnover. The basis for inclusion/exclusion of data sets in the published report should be made explicit, ideally as part of the pre-execution experimental plan. Methodology and analysis sections of manuscripts should distinguish between preliminary and final experiments and describe which of the former are included in the analyzed data.

Random Assignment to Groups

A particularly problematic execution consideration in preclinical experiments is random assignment of subjects to treatment groups, a sine qua non in clinical trials. On the surface, this would seem easy to achieve in preclinical experiments involving animals, but there are practical complications. For example, surgical procedures required for creating models of disorders (spinal cord injury, TBI, stroke, and jugular vein cannulation for IV drug studies) are complicated and time consuming. Consequently, there is a limit to the number of animals that can be prepared on a given day. Careful planning is required to ensure that subjects prepared at different times are distributed across groups, and the best practice would be to do this in a random fashion. Random assignment is not always feasible or practical, however. Consider the situation where an experiment involves creating a spinal cord injury, TBI, or stroke in animals, some of which will receive transplants of cells that must be expanded in vitro for some time prior to transplantation. If a lab has the capacity to create injuries in ten rats on a given day, and there are two treatment groups (cells and no cells), then n = 5 per group for that day. This would almost certainly not pass a power analysis for any except the most robust and sensitive endpoints. The obvious theoretical solution is to create injuries on several different days and then combine data from the different days, but this would require staging the cell culture procedure so that cells would be available on the many different days that they were needed. Some labs may have the capacity (both technical and financial) to do this; others may not. It is possible (likely?) that many studies in the literature take the more practical approach of preparing a batch of cells (or other complicated treatment), then preparing the group that receives transplants on one day and the control groups on different days. While these practical considerations are reasonable, a best practice is to explicitly describe these in a manuscript methodology section. Another example of a complication is that preparation of a treatment (cells for example) may take time, and after preparation, the cells may remain viable for only a defined time period. To avoid deterioration of the cells, it may be necessary (or at least optimal) to prepare the transplant group at one time and the control at a different time. This strategy does not invalidate the study, but it is a critical best practice to be completely transparent about the procedure in the methods section of a manuscript and consider caveats. As a side note, when the viability of a treatment may change over time (as in cells with limited viability), it is important to keep track of the order in which subjects are treated so that potential loss of effect can be detected. The issue of record keeping is discussed further below.

Blinding as a Practical Standard

A key but quite problematic element of experimental execution is blinding the experimenter to subject groups/treatments. One issue is when blinding should come into play. Slavish adherence to blinding is not desirable. In the exploratory phase of a research project, blinding disables the most powerful analytical tool that we possess—our brains. The problem with blinding during initial analysis is that it is often unclear what the "phenotype" of a positive (or negative) result will be. For example, if an intervention is designed to enhance axon growth, this could be manifest in many different ways. Failing to analyze anatomical material with knowledge of groups may cause one to miss important effects or to design quantitative analyses that will not detect differences among treatment groups. The blind analysis takes place after one has an idea what one is trying to assess. One can code and shuffle slides for blind analysis or pass the analysis to someone who remains blind.

There are also reasons for at least one member of the research team to be unblinded during behavioral testing. For example, if some animals begin to exhibit symptoms of pain, wasting, or any of a host of abnormal behaviors, it is prudent to figure out whether these cases are from one treatment group. Such events may call for early termination of an experiment or modification of the experimental plan.

There is consensus that blinding is important to reduce possible bias, but blinding can be harder to achieve than one might guess, and sometimes cannot be achieved at all. For example, for studies that involve surgical procedures, details of the procedure may allow animal identification in later testing. Postoperative care may also introduce identifiers. Best practice would be to have the individuals who deliver treatments blind with respect to the test versus control (vials marked Drug A versus Drug B, for example). This may be impossible, however, for treatments that involve surgical delivery of cells or that require different surgical protocols. When the staff who deliver treatments cannot be kept blind, then it is important that treatment and testing be done by different staff. It is also useful to have a separate person compile and analyze data. Obviously, it is difficult and probably impossible for a single individual or a small lab with few personnel to carry out a truly blind experiment involving an injury or similar intervention.

It is also sometimes impossible to be blind; the treatment itself may lead to changes that allow identification of groups. Treatments may cause physical signs (e.g., treatment with Rolipram causes porphyrin to accumulate in the eyes), changes in general behavior (hypo- or hyperactivity), changes in weight, or changes in physical appearance (e.g., rats treated with the P2X7 receptor antagonist Brilliant Blue G are blue for a few days after treatment). The procedures used to achieve blinding, description of factors such as the ones above that make it impossible to achieve blinding, and the interpretive caveats resulting from lack of blinding should be included in methodology sections of papers.

Appropriate Expertise with Assays and Assessments

Preclinical studies often involve analyses at multiple levels, sometimes ranging from genetic/molecular/cellular to behavior. The starting point for a preclinical study may be at a molecular/cellular level and then proceeds through animal models of disorders. Few labs are expert in all level of analyses in a typical preclinical study, and there can be lack of appreciation of limitations and caveats. This is a broad and complicated topic but some key points are worth noting. All assessments should include appropriate positive and negative controls. Defining what controls are sufficient is not always straightforward. Also, it is important to distinguish between "assays" that have positive and negative standards linked to defined physical parameters and "assessments" that have no external standards. Most functional/behavioral tests have no external standards, so positive and negative controls are valuable. It is also worthwhile to distinguish between assessments that have a reproducible quantitative readout versus those that depend on judgment (albeit by highly trained experts). In the latter, reliability over time and interrater reliability are important. Acceptance of claims is influenced by the experience of a lab, and historical reliability is a factor. For example, if a given lab has published numerous studies using a particular injury model, and the outcome measures in control groups are comparable over time, one is more confident of a report of a treatment effect than is the case from a lab without prior experience.

The issue of sufficiency of expertise raises the question of whether some preclinical research should be done on a collaborative or contract basis in labs with extensive experience. This was the rationale behind the NINDS-supported FORE-SCI contracts as well as some other more recent initiatives (Tabak, 2014). It may be useful to consider new paradigms involving multiple investigators and core labs focused on particular disorders, or fee-for-service organizations expert in a particular approach, as is commonly done in biopharma. In academia, such activities would require completely different funding paradigms than currently exist.

Record Keeping for Methodological Details

Scientists generally keep good records of outcome data but are often less compulsive about records of methodological details that can be critically important in terms of replicability. Examples are too numerous to cover but include such things as timing of surgical procedures, time of day and order of behavioral testing, timing between functional/behavioral testing, and other events including postoperative care (for example, bladder expression for spinal cord injured animals). A similar long list could be made for experiments involving cells in culture, which can be highly sensitive to very slight changes in culturing conditions, causing them to respond differently in the same assay (Bissell, 2013). Keeping track of conditions that are not actually part of the formal experiment can allow detection of variables that influence outcome. In preclinical studies, there is no such thing as collecting too much information.

Data Analysis and Statistical Considerations

Establishing a statistical analysis plan, including interim analyses and futility assessments, how data will be tested across treatment groups for significance, and rules for data exclusion, prior to initiation of a study, are less common in preclinical experiments than in clinical studies and drug trials, where these are laid out in expansive protocols and typically reviewed at several organizational levels. Description of analysis parameters in advance of experimental execution has several benefits. Indeed, it is difficult to envision a scenario where this would not benefit scientific rigor and replicability and reduce bias. The statistical analysis plan should be described in the methodology section of manuscripts. In terms of "n" for statistical analyses, it is important to distinguish between "technical replicates" and "biological replicates." An example of a technical replicate is repeating assays on a single set of samples. An investigator may prepare multiple gels from the same sample and then quantify a given band across gels yielding a measure of central tendency and variance. Gene array analysis often involves repeated amplification of RNA from a single sample, which is a technical replicate. Technical replicates serve as internal checks on reliability of assays in an experiment. One could argue that multiple slices from a single animal should be considered a type of technical replicate because the different samples are not independent, but this is not standard practice. Biological replicates involve different independent starting samples from different subjects. For additional considerations on statistical methods for experimental biology, see Vaux (2012) and Vaux et al. (2012). A best practice is to describe the statistical treatment of technical and biological replicates in the methodology section of manuscripts and define the meaning of error bars in figure legends (the latter has been a requirement for *Nature* since 2004).

One consideration involves compiling groups by combining data generated at different times: when an interim analysis approach is used, or in experiments with difficult procedures or low yield, data from subjects tested at different times and as part of different experimental groups are often combined. Many preclinical studies, for example, studies in animal models of neurodegenerative disorders like Alzheimer's disease, or brain injury studies, involve multiple groups or comparisons over months to years. Defining starts and stops is especially important when each individual experiment involves compiling groups by combining data generated at different times. There are well-accepted statistical approaches for comparing groups over time (e.g., repeated-measures ANOVA) and well-defined post hoc tests allow comparisons at particular time points, if the overall ANOVA is significant. Carrying out multiple t tests is problematic, however, because there is a high overall risk of type I error. This risk is generally underappreciated in in vitro and in vivo preclinical studies alike.

There are other less recognized issues with multiple comparisons that come up in preclinical studies. Consider, for example, a study that tests whether a particular intervention improves function after spinal cord injury. Because patients could benefit from improvements in any of a number of different functions, it is common to assess multiple endpoints in preclinical models. The same is true of clinical studies but clinical trials need to define a primary endpoint on which the trial stands or falls. A common measure in rodent SCI studies is hindlimb motor function, which is often assessed using different standardized tests, for example, the BBB scale and subscale and ladder beam. One might also measure sensation by assessing pain and temperature sensitivity and touch (three or more different tests), as well as bladder function. Within each data set, corrections for multiple comparisons can and should be applied, but the more complicated issue is how to deal with the fact that there are multiple unrelated analyses, each of which carries its own risk of type I error. A further consideration is how to interpret an outcome that is common: a significant effect on one endpoint, but not others. On the one hand, the effect could seem (and be) highly significant in terms of indicating a particular outcome that may point to future therapeutic benefit in patients. It would be a huge accomplishment if an intervention enabled recovery of bladder function following spinal cord injury even if other functions were not improved. But if this is the only significant finding in a study with multiple analyses, the possibility of a type I error is high. This problem is confounded by the "file drawer" phenomenon, where the overall study may include analyses that are not reported because of the lack of a significant difference.

The example above considers functional analyses of living animals, and although the data can be analyzed in many ways, new data cannot be generated once the animals are dead, which limits analyses to the assessments that were preplanned. The situation is different for anatomical analyses. For example, the initial plan might be to assess lesion size and tissue sparing, but the results might suggest additional analyses such as assessments of white matter sparing, neuron or oligodendrocyte loss, astroglial reactivity, or immunostaining for particular systems (e.g., 5HT) or particular cell types (e.g., inflammatory cells). While it may be justified to test alternative hypotheses, it is also important to avoid "significance chasing," i.e., doing different analyses until one shows a significant difference. For preclinical studies, all analyses that were carried out in a particular study should be reported (although it is not required to present all data); and it should be explicitly stated which analyses were preplanned and which were not.

Reporting all analyses does not solve the basic problem, however, and to our knowledge, there are no well-accepted statistical approaches to deal with multiple, unrelated analyses across groups of subjects (but see section below regarding data mining and archiving for retrospective analysis). One way to increase the comfort level is to report the statistical power of significant findings. This is now being required in at least one neuroscience journal (*eNeuro*). Another way is to repeat the experiment. If the same pattern of results is reproducible (self-replication), this test-retest bolsters the confidence level in study conclusions. Here, however, it is important that the endpoints be truly independent. The methodology section of manuscripts should indicate when a test-retest approach was used, i.e., that one experiment with x endpoints was completed and analyzed, and then a different experiment with the same endpoints was executed.

There are well-accepted best practices for reporting statistical results accurately. Differences are either statistically significant or not. There is no meaning to the phrase "nearly significant." On the other side, it is also not appropriate to say "groups are the same" just because they are not significantly different. For nonsignificant differences, a measure of central tendency, variance, statistical test, and p value should be reported with the statement that data did not differ significantly. While combining data from different experimental groups that do not differ significantly is common practice, this could be questionable depending on the situation. When this is done, descriptive statistics for each group should be reported as well as pooled values.

Data Inclusion and Exclusion Considerations

One class of preclinical study involves surgical or other interventions to model neurological disorders (e.g., spinal cord injury, TBI, stroke, and some models of neurodegenerative disorders, such as MPTP treatment to model aspects of Parkinson's disease). Outcomes are typically compared between groups treated in different ways with the assumption that the injuries are comparable. Importantly, however, surgical procedures are variable and there can be accidental injury, excessive bleeding, or errors in lesion production. A common (and acceptable) practice is to exclude animals based on predefined exclusion criteria. It is reasonable, for example, to exclude subjects based on adverse events during a surgery; the practice, however, is often to keep the animal and decide later based on early results. This is especially true when there has been extensive preoperative testing of subjects. The decision is not straightforward. It is unfortunate to waste time and money by discarding an animal that could provide useful data. On the other hand, it may compromise the experiment to include an animal in which procedural error causes an out of range lesion.

Sometimes decisions to exclude animals are based on functional parameters observed during the early postinjury period. This is probably acceptable when exclusions are determined prior to the time that treatments are delivered, but this may not be possible for treatments delivered during the early postinjury period. In some cases, it may be reasonable to compose groups based on functional or other outcome measures, but again, this is acceptable only before treatments are initiated.

In preclinical studies of neurological or psychiatric disorders, animals' health may be compromised by the model. Animals may die over the course of the study or have to be euthanized because of animal welfare concerns (deteriorating health, pain, unacceptable levels of disability). Indeed, for some degenerative disorders (animal models of ALS or Huntington's disease), lifespan is an outcome measure. If there is attrition, the question arises of how to deal with missing data. For example, if an animal dies suddenly midway through an experiment, should the data from that subject be removed from all analyses or included up to the point of death? The decision can dramatically affect interpretations (Couzin-Frankel, 2013). There is room here for differences of opinion, but it is critical to be completely transparent about how this issue was handled. We recommend that best

practice involves reporting attrition as a standard for preclinical studies because treatment may affect overall health either positively or negatively, which skews data.

Data Management

The best practices we outline above in experimental design, execution, analysis, and resource sharing will probably not impact rigor and reproducibility unless best practices in data management are also implemented. Data management includes recording key experimental design and execution parameters, rigor in archiving raw data, and curation of the process of turning raw data into a summary figure and thus a conclusion. There are published best practices for maintaining lab notebooks available from NIH (https://www.training.nih.gov/assets/Lab_Notebook_508_(new).pdf), and the Howard Hughes Medical Institute (http://www.hhmi.org/grants/office/scimgmt.html), as well as recommendations for data recording (González-Beltrán et al., 2012).

It is worth reinforcing that a complete experimental record includes not only experimental design parameters, details of execution of different from standard protocols, a description of reagents and animal care and use, analysis processes and macros and statistical procedures, but also how a particular conclusion was reached. Experienced staff and trainees in most labs, including those of the authors, often maintain lab notebooks that include some of this information, but rarely all of it. But it is also all too common to maintain day-to-day notes on a pad of lined paper—or even scraps of paper—rather than in a lab notebook. Developing lab standards for indexing and maintaining key information, regardless of format, is an obvious best practice. Standards should be sufficient to ensure that all relevant information supporting a conclusion can readily be located at any point during and after study completion.

All scientists would agree that raw data should be archived, but archiving practices vary widely. Many labs have electronic systems, e.g., lab servers or one or more external hard disks, to archive many types of data. Every PI thinks it trivial to put their hands on raw data generated by lab personnel, until they actually attempt to do so. Few of us think of archiving software to read proprietary file formats that rapidly become outdated—until we need to access that key data. In addition to raw data, execution protocols, analysis macros and other procedures, and statistical analysis protocols should also be archived. The file structure and naming convention of relevant documents should be obvious to colleagues, explained and cross-referenced in lab notebooks. A best practice would be to think through what information would be required by someone outside of the lab to replicate the experiment and adopt processes to facilitate recording and archiving this information.

Most scientists would also agree that raw data should never leave the lab, at least not before archiving, but we are all lax about this, sometimes with serious consequences. We often overlook the high failure rate of laptops and external hard drives that most of us use to be productive outside of standard hours, and we have all experienced the consequences of hardware failure. Not only should data on laptops, lab computers, and equipment be archived, but raw data should be backed up in at least one other location in case of hardware failure. Archiving and backups should occur with a periodicity that mirrors the generation of new information, in most cases daily or weekly, rather than only when a figure or a presentation is prepared.

It is worth noting that the Federal Information Management and Security Act (FISMA) defines rules for data security and backup. FISMA currently applies to all federal contractors, but it is conceivable that this requirement will be extended to other forms of Federal funding.

An often overlooked step between experiment execution and conclusion that is essential to document and archive is the process of turning raw data into a summary figure with quantification that supports a particular conclusion. Even if primary data are readily accessible, it is often unclear how those data were turned into the data included in a published figure, and who was responsible for this key activity. A best practice would be to include this information in lab notebooks or other formats, such that these steps can be readily identified and repeated if needed. The principles we wish to highlight here as best practices are that all researchers share the responsibility for developing processes to maintain the integrity of the primary data and to identify mechanisms to facilitate the storage and retrieval of data and analyses that turn data into conclusions.

Minimizing Bias

> Bias is unintentional and unconscious. It is defined broadly as the systematic erroneous association of some characteristic with a group in a way that distorts a comparison with another group … The process of addressing bias involves making everything equal during the design, conduct and interpretation of a study, and reporting those steps in an explicit and transparent way (Ransohoff and Gourlay, 2010).

There is no way to completely eliminate bias. Investigators do the work that they do because they suspect or believe that their approach holds promise. One can also be overtly biased in attitude or approach. Everyone suffers from these forms of bias in one way or another; the important thing is to minimize the impact of personal bias on experimental results. Experimental techniques may also be biased in ways that are not immediately recognizable. For example, it is recognized that simple counts of elements in histological preparations can be biased. Counts of any element (cells for example) in a histological section depend on where the counts are made and the size of the element being counted (Guillery, 2002). It is for this reason that "unbiased sampling" techniques have been developed, commonly grouped under the heading "stereology." In preclinical research involving surgical procedures, there can be unrecognized procedural bias. For example, there may be a tendency to be more careful with the "experimental" group during the surgery, especially if these have had prior costly treatments. If the order of surgery is nonrandom, fatigue may affect the surgeon's performance over the course of a day. A best practice is to explicitly report approaches used to reduce/control for unintended bias in Methods sections, but in the end, it may be impossible to eliminate all sources of unintended bias.

Publication and Reporting Considerations

The common practice of reviewers of requesting additional experiments presents different challenges. We (the authors) have both given and received such requests, and it is true that the additional experiments that reviewers request generally strengthen a paper. At the same time, this is an example of a "perverse incentive," because the results of additional experiments may determine whether or not a paper will be found acceptable. A draconian solution would be something like a one-strike rule; a paper would be considered based on the data available or it would be rejected. Additional data would not be requested or accepted. There are obvious downsides to this, including the lost opportunity to improve upon the science. We propose that at a minimum, a best practice would be to include a statement in the methodology section of a manuscript indicating which experiments were done in response to a request for additional data. Another interesting idea involves "registered reports" in which it is the experimental approach that is reviewed and accepted or rejected by a journal and published regardless of the outcome. This is a new publishing initiative at the journal *Cortex* (Chambers, 2013). One purpose is to address the "file drawer" problem for negative results, but with the added benefit that all aspects of the experiment are preplanned with rigor in much the same way as a clinical trial.

Many of the issues and best practices we raise here will require more expansive methodology and reagent sections in published manuscripts. These would require review that is as comprehensive as for other parts of the manuscript. One approach would be for journals to relax word and page limits for print and online articles. An alternative approach would be to have an abbreviated methods section in the print version and a more compete version in the online and PDF versions. In those journals that permit supplemental information, for particularly novel or complex experimental procedures, authors could be requested to provide links to detailed experimental protocols. An effort to comprehensively describe reagents and methods in papers would be the single simplest step to facilitate replication. Because reviewers often are asked to concentrate on the broad impact, novelty, and general interest of a manuscript, it may be useful to consider implementation of a two-phase review. Details of methodology would be considered separately if a paper was considered potentially acceptable based on general considerations defined by journal policy. Reviewers could then specify whether or not they reviewed for methodological details. Some journals are now including reviewer checklists, which define areas to be addressed, and at least one journal, *Science*, now utilizes a separate expert review of statistical methods prior to final acceptance.

Resource Sharing

A problematic issue in terms of replication is the availability of specialized reagents and animal models to individuals who wish to replicate published results. Many journals and granting agencies have well-articulated policies on resource sharing. While requests are rarely denied outright by authors, many languish in inboxes or voicemail. The issue is much more complex for reagents and resources from companies that are not commercial products, such as proprietary antibodies or agents that are under development such as human stem cells. Negotiation of MTAs or other collaborative agreements often delay and sometimes prevent the replication of key findings and require an investment of time and energy that working scientists may consider disproportionate to their value.

An extreme measure would be for funding agencies and journals to retract funding or refuse to accept papers if authors are unable or unwilling to provide key tools required for replication. But this does not take into consideration the complexity of MTA negotiations and potential complications for papers using reagents from companies that are not commercial products. Having such a requirement would impede or prevent some collaborations between the academic and private sector, and requiring that a paper be rejected for these reasons is not desirable. Clearly, one rule will not fit all, and this topic deserves further consideration by academic and industrial neuroscientists alike.

A moderate option would be to require manuscripts to list those resources that will be made available upon request and the author point of contact. Reviewers could be asked to review this information and determine whether the resources are sufficient for replication or, if the situation warrants, an exemption from the general requirement. It is clear that these recommendations would have a positive impact on data reproducibility, but it is equally clear that maintaining resources and in particular animal lines could prove practically and financially burdensome to authors. One initiative worth noting is from NeuroInformatics Framework Resource Identification Initiative to establish a framework for discoverability of key reagents (https://www.force11.org/Resource_identification_initiative). Consideration of a time limit for resource availability postpublication, earmarking of grant funds to support resource sharing as a budget line item, depositing animal lines at a commercial vendor prior to or within 3 months of publication, among other practicalities, would address some of the challenges presented by these solutions. Although this landscape is quite complicated, we propose a best practice of stating specifically in a manuscript which resources would be readily available and which have access restrictions.

Data Sharing

A particularly important but often overlooked resource is the raw data collected in experiments. Making raw data available upon request, for example, to enable an independent statistical assessment or for statistical comparison with replication data, is as valuable an experimental resource as a construct encoding a novel fluorescent indicator or transgenic mouse line. There would be widespread agreement that raw data should be archived, and most, but admittedly not all, data are now electronically archived in formats that are widely accessible (e.g., spreadsheets, images, dot-cvs or text files, etc.) and easily disseminated upon request. Archiving data for a decade as a best practice balances the desire to archive data for future use and the practical burdens of storage. Interesting experiments are being launched in terms of data archiving and analysis in the area of spinal cord injury. One approach is the development of a reporting standard called "Minimal Information about a Spinal Cord Injury experiment" (MIASCI). The idea is that standard data elements would be collected as part of every spinal cord injury experiment and reported to an online database (Lemmon et al., 2014). Another approach involves collection of raw data

of all types from investigators who are willing to contribute and compiling the data into a database called "Visualized Syndromic Information and Outcomes for Neurotrauma-SCI" (VISION-SCI; Nielson et al., 2014). The database allows for syndromic analysis using advanced statistical methods, including principal component analysis (PCA), where composite PC scores can be used for hypothesis testing regarding injury severity or treatment condition on the entire PC outcome measure, as opposed to traditional methods testing single measures at a time. This approach allows for a more complete assessment of recovery on the complex interactions of multiple measures simultaneously (Ferguson et al., 2013; Nielson et al., 2014).

Reporting Negative Results and Replication Studies

In an optimal scientific ecosystem of rigorous experimental design/execution/statistical analysis and open resource and raw data sharing, how might we as neuroscientists disseminate results on lack of replication? In considering publication of negative results in general, it is worth distinguishing between an exact replication and a retest of concept. The latter is probably much more common than the former. An exact replication is one in which an attempt is made to duplicate the conditions of the original experiment in all respects, and obviously this is impossible. Duplication requires extensive communication with the original investigators, because Methods sections are at best abbreviated summaries of what the original investigators considered to be the critical methodological details, and reagents and animals used may or may not be sufficiently similar to replicate findings, among other differences. A retest of concept involves testing the same intervention, but not necessarily under exactly the same experimental conditions. One wonders how often rumors about lack of replication involve instead lack of conceptual validation. It would be great to have a forum for making results of replication available to the broader community, but quality control of replication studies is as problematic as for any other type of study. A recent meeting of the European College of Neuropharmacology network focused on preclinical data reproducibility in the context of R&D organizations. This group is surveying members in an attempt to compile a list of papers that at least one R&D group had tried to replicate without success. The goal is to acknowledge challenges in replicating data and develop a forum for sharing these results. In the end, one comes back to the traditional model of peer review, and with that, one might as well just publish the data in the usual way. Despite expectations to the contrary, it is not difficult to publish negative results of replications in regular journals, as evidenced by one example that every replication carried out as part of the NINDS Facilities of Research Excellence-Spinal Cord Injury was accepted in a peer-reviewed journal.

In terms of approaches to replication and visibility for replication efforts and failures to replicate, there are a number of pilot initiatives. For example, Science Exchange has announced plans for commissioned replication studies (https://www.scienceexchange.com/). Also, *eLife* has announced that they will consider articles reporting replication studies for key cancer papers (in collaboration with Center for Open Science and Science Exchange), and *eNeuro*, the new open access journal of the Society for Neuroscience, will also consider replication studies and reports of negative results.

Table 1 lists the best practices we have discussed here. As best practices are defined and adopted, it will be important to develop toolkits for training at all levels. This does not just apply to early-stage investigators. Many of the practices that are perilous are just now being revealed, and senior investigators are no more knowledgeable about these than a first year graduate student. A starting point to raise the bar for scientific rigor is to train the next generation of scientists in these and other best practices. It will be important to develop new toolkits for students and postdocs that might include didactic lecture-based classes. While most graduate classes include a discussion of research articles, there is rarely discussion of overall rigor. Professional societies, academic programs, and educators could develop and broadly disseminate polished webinars with training modules. Increasing rigor for established investigators will probably require other approaches, and the main "training" may come from the school of natural consequences, such as enforcement of new guidelines by funding agencies on the input side and journals on the output side.

Relationship between the "Rs": Enhanced Rigor Does Not Guarantee Reproducibility

Current attention on lack of experimental rigor resulted from increasing awareness of lack of reproducibility in preclinical research. The assumption is that the former is the cause of the latter, but this may not be true. Results from biological experiments are influenced by a host of variables, not all of which are actually under the control of the experimenter. This is especially true of assessments of function using behavioral tests, which are often considered to be the most important outcome measure in preclinical studies of neurological disorders. Even if strain, gender, and age are held constant, animal behavior in a given setting depends on a host of variables including prior training, housing conditions, testing conditions, the way the investigators handle the animals, time of day, and, yes, maybe even phase of the moon. A recent study even indicates that responses in tests of pain sensitivity vary depending on the gender of the handlers and even whether handlers wear clothing previously worn by someone of the opposite sex (Sorge et al., 2014).

Much of the discussion about why particular studies are not replicated focus on the argument that the replication study did not exactly replicate the conditions of the original experiment. This is especially true with complicated modern techniques that require years of experience to master (for a viewpoint on this, see Bissell, 2013). The failure of a less experienced lab to replicate the findings may be due more to the inexperience than the veracity of the original claim. This may be true, but in terms of preclinical studies that aspire to provide translatable insights into human disorders, a treatment effect that depends so critically on the exact experimental conditions is highly unlikely to be translatable to the highly variable human situation. The bottom line is that the more a particular result depends on the exact experimental conditions, the less likely it is to be replicable even if the original experiment was done with the highest level of rigor.

Conclusion

So how bad is it that there is a lack of replication in preclinical studies (or, what about "mortis")? In the authors' opinion, it is

Table 1. A Primer of Best Practices to Enhance Rigor and Reproducibility

Topic	Best Practice	Benefits
Experimental Design	Describe experiment planning in manuscript Methods section, including: ● Power calculations (endpoint sensitivity, variability, effect size, desired level of confidence, definition and rationale for n). ● Inclusion/exclusion of data sets, description of pilot, and final data sets included in analyses. ● Random assignment to treatment groups, description of exceptions. ● Procedures to achieve blinding, exceptions to blinding, and resulting interpretive caveats. ● Details of reagents and assays sufficient to facilitate independent replication. ● Positive and negative controls.	Capture thinking in incomplete information landscape. Iterative hypothesis refinement. Deep understanding of assessments in advance of execution. Reduce testing to foregone conclusion. Optimize resource allocation and use. Create roadmap to assembling publication.
Analysis and Statistics	Describe statistical analysis plan in manuscript Methods section, including: ● Methods to test for significance. ● Interim analyses, futility assessments. ● Data inclusion/exclusion, attrition. ● Statistical treatment of technical and biological replicates. ● Test-retest approaches. ● Statement of central tendency, variance, statistical test, and p value for significant and nonsignificant differences. ● Descriptive statistics for groups as well as pooled values.	Enhance awareness of and reduce sources of potential unconscious bias. Minimize type 1 error.
Data Management	Develop lab standards for indexing and maintaining information, including: ● Recording of key experimental design and execution parameters. ● Archiving raw data and at least one backup with appropriate frequency. ● Curation of process from raw data to summary figure to conclusion.	Ensure all information supporting a conclusion can be located during and after study completion.
Resource Sharing	Include lists of resources in manuscripts that will be made available and point of contact for requests. Indicate time limit for resource availability, if any. Include budget line item to support resource sharing in funding applications. Deposit animal lines at commercial vendor within 3 months of publication. Provide raw data upon request.	Simplify sharing of reagents, protocols, raw data to facilitate replication, interpretation of data. Help distinguish lack of conceptual validation versus lack of replication. Enable meta-analyses and data basing.
Publication and Reporting	Provide comprehensive review checklist for methodology, reagents, and resource sharing. Two-stage review: if manuscript meets general journal criteria (novelty, impact, general interest), initiate second stage of review for technical merit including details relating to rigor.	Raise awareness of key metrics for determining rigor. Facilitate replication of key findings.

critical to have a laser focus on exactly what did not replicate and why. As discussed above, most preclinical studies start with novel basic biological findings that form a foundation that is then built upon to translate the findings to an animal model relevant to one or more aspects of a neurological or psychiatric disorder. This is a tall order for most CNS diseases, and most of these attempts will fail, especially if the results depend critically on the exact details of the experiment. Phenomena that are not robust against minor variations in conditions will be hard to translate into therapeutic interventions. We have the best chance of learning from failures to replicate if both the original and replication studies are carried out using practices that assure the highest level of rigor.

The fact that scientists at all levels and in different types of organizations continue to take long shots on this goal is a good thing, even in the face of the inevitable failure—including failures

of our own making. We owe it to the society that supports us to learn as much as possible from these failures.

ACKNOWLEDGMENTS

O.S. was the recipient of NIH Contracts N01-NS-3-2353 and HHSN27120080039C as part of the replication project called "Facilities of Research Excellence-Spinal Cord Injury (FORE-SCI). O.S. is one of the co-founders of the company "Axonis," which holds options on patents relating to PTEN deletion and axon regeneration. R.B.-G. declares no competing financial interests.

REFERENCES

Begley, C.G. (2013). Six red flags for suspect work. Nature *497*, 433–434.

Begley, C.G., and Ellis, L.M. (2012). Drug development: Raise standards for preclinical cancer research. Nature *483*, 531–533.

Bissell, M. (2013). Reproducibility: The risks of the replication drive. Nature *503*, 333–334.

Button, K.S., Ioannidis, J.P., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S., and Munafò, M.R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nat. Rev. Neurosci. *14*, 365–376.

Chambers, C.D. (2013). Registered reports: a new publishing initiative at Cortex. Cortex *49*, 609–610.

Couzin-Frankel, J. (2013). When mice mislead. Science *342*, 922–923, 925.

Darian-Smith, C., Lilak, A., Garner, J., and Irvine, K.A. (2014). Corticospinal sprouting differs according to spinal injury location and cortical origin in macaque monkeys. J. Neurosci. *34*, 12267–12279.

Ferguson, A.R., Irvine, K.A., Gensel, J.C., Nielson, J.L., Lin, A., Ly, J., Segal, M.R., Ratan, R.R., Bresnahan, J.C., and Beattie, M.S. (2013). Derivation of multivariate syndromic outcome metrics for consistent testing across multiple models of cervical spinal cord injury in rats. PLoS ONE *8*, e59712.

González-Beltrán, A.N., Yong, M.Y., Dancey, G., and Begent, R. (2012). Guidelines for information about therapy experiments: a proposal on best practice for recording experimental data on cancer therapy. BMC Res. Notes *5*, 10.

Guillery, R.W. (2002). On counting and counting errors. J. Comp. Neurol. *447*, 1–7.

Henderson, V.C., Kimmelman, J., Fergusson, D., Grimshaw, J.M., and Hackam, D.G. (2013). Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med. *10*, e1001489.

Howells, D.W., Sena, E.S., and Macleod, M.R. (2014). Bringing rigour to translational medicine. Nat. Rev. Neurol. *10*, 37–43.

Hyman, S.E. (2012). Revolution stalled. Sci. Transl. Med. *4*, 55cm11.

Ioannidis, J.P.A. (2012). Why science is not necessarily self-correcting. Perspect. Psychol. Sci. *7*, 645–654.

Kilkenny, C., Browne, W.J., Cuthill, I.C., Emerson, M., and Altman, D.G. (2010). Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. J Pharmacol Pharmacother *1*, 94–99.

Landis, S.C., Amara, S.G., Asadullah, K., Austin, C.P., Blumenstein, R., Bradley, E.W., Crystal, R.G., Darnell, R.B., Ferrante, R.J., Fillit, H., et al. (2012). A call for transparent reporting to optimize the predictive value of preclinical research. Nature *490*, 187–191.

Lemmon, V.P., Ferguson, A.R., Popovich, P.G., Xu, X.M., Snow, D.M., et al. (2014). Minimum information about a spinal cord injury experiment: a proposed reporting standard for spinal cord injury experiments. J. Neurotrauma *31*, 1354–1361.

McNutt, M. (2014). Reproducibility. Science *343*, 229.

Nestler, E.J., and Hyman, S.E. (2010). Animal models of neuropsychiatric disorders. Nat. Neurosci. *13*, 1161–1169.

Nielson, J.L., Guandique, C.F., Liu, A.W., Burke, D.A., Lash, A.T., Moseanko, R., Hawbecker, S., Strand, S.C., Zdunowski, S., Irvine, K.A., et al. (2014). Development of a database for translational spinal cord injury research. J. Neurotrauma *31*, 1789–1799.

Pankevich, D.E., Altevogt, B.M., Dunlop, J., Gage, F.H., and Hyman, S.E. (2014). Improving and accelerating drug development for nervous system disorders. Neuron *84*, this issue, 546–553.

Prinz, F., Schlange, T., and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? Nat. Rev. Drug Discov. *10*, 712.

Ransohoff, D.F., and Gourlay, M.L. (2010). Sources of bias in specimens for research about molecular markers for cancer. J. Clin. Oncol. *28*, 698–704.

Scott, S., Kranz, J.E., Cole, J., Lincecum, J.M., Thompson, K., Kelly, N., Bostrom, A., Theodoss, J., Al-Nakhala, B.M., Vieira, F.G., et al. (2008). Design, power, and interpretation of studies in the standard murine model of ALS. Amyotroph. Lateral Scler. *9*, 4–15.

Sorge, R.E., Martin, L.J., Isbester, K.A., Sotocinal, S.G., Rosen, S., Tuttle, A.H., Wieskopf, J.S., Acland, E.L., Dokova, A., Kadoura, B., et al. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. Nat. Methods *11*, 629–632.

Steward, O., Popovich, P.G., Dietrich, W.D., and Kleitman, N. (2012). Replication and reproducibility in spinal cord injury research. Exp. Neurol. *233*, 597–605.

Tabak, L. 2014. Enhancing reproducibility and rigor of research findings. Peer Review Notes Parts 1–3, http://public.csr.nih.gov/aboutcsr/NewsAndPublications/PeerReviewNotes/Pages/Peer-Review-Notes-May-2014Part2.aspx.

Vaux, D.L. (2012). Research methods: Know when your numbers are significant. Nature *492*, 180–181.

Vaux, D.L., Fidler, F., and Cumming, G. (2012). Replicates and repeats—what is the difference and is it significant? A brief discussion of statistics and experimental design. EMBO Rep. *13*, 291–296.

Wadman, M. (2013). NIH mulls rules for validating key results. Nature *500*, 14–16.

Willner, P., and Mitchell, P.J. (2002). The validity of animal models of predisposition to depression. Behav. Pharmacol. *13*, 169–188.