HHS Public Access

Author manuscript

Lab Anim (NY). Author manuscript; available in PMC 2020 May 20.

Published in final edited form as:

Lab Anim (NY). 2018 July ; 47(7): 175–177. doi:10.1038/s41684-018-0088-6.

Getting personal with the "reproducibility crisis": interviews in the animal research community

Ben G. Fitzpatrick[1,2,*], **Elena Koustova**[3], **Yun Wang**[1]

[1]Tempest Technologies, Los Angeles, California, United States.

[2]Department of Mathematics, Loyola Marymount University, Los Angeles, California, United States.

[3]Office of Translational Initiatives and Program Innovations, National Institute on Drug Abuse, National Institutes of Health, Bethesda, Maryland, United States.

To the Editor:

Reproducibility problems are prevalent in research involving animal models, and publications on methodological pitfalls of animal experimentation are copious. Kilkenny et al.[1] found reporting and robustness problems in the statistical analysis of nearly 60% of published animal research surveyed, with over 60% using inefficient experimental designs. Button et al.[2] estimated that roughly 50% of published neuroscience studies have power less than 20%. The study of Freedman et al.[3] indicates that experimental design and statistical analysis are factors in over 50% of irreproducible preclinical research. In addition to statistical concerns, inadequate animal study design tops the list of reproducibility offenders. Careful attention to the experimental design process can improve the quality of the experiment and hence the harm-benefit balance for laboratory animals. Toward that end, the National Institute on Drug Abuse (NIDA), part of the National Institutes of Health (NIH), supports a number of projects in experimental design for the animal studies that provide the foundation for therapeutic discovery and development. Our company, Tempest Technologies, was awarded a Small Business Innovation Research grant to develop online experimental design tools for animal researchers. As part of the tool development process, we interviewed biomedical researchers, policy-makers, funders, and regulators to learn about the challenges of animal studies. At a high level, our findings corroborate those reported in Baker[4], but our personal conversations offer insights into the design and analysis of animal experiments and the concerns of stakeholders.

Interviewing animal research stakeholders about reproducibility

During the summer months of 2016, Tempest personnel contacted over 1000 scientists, program officials, regulators, and administrators in academia, government, industry, and private foundations; 131 accepted our invitation for an interview.

* fitzpatrick@tempest-tech.com.

Eighty three in-person interviews were conducted at southern California universities and businesses and three professional conferences, and 48 interviews were conducted with web or phone conferences. Each discussion was conducted with a single interviewee and we used an open-ended qualitative interview approach, beginning with questions about reproducibility, following with questions about experimental design.

To protect the privacy of our interviewees, we do not use any attributions. We caution that our sample is a convenience sample rather than a random sample, and that our open-ended interview process is an exploratory study rather than a confirmatory one. We also note that our interviews are exempt from human subjects protections under 45 CFR 46.101(b)(2) from 45 CFR part 46 requirements.

Over 44% of interviewees were biomedical faculty from academic institutions, and 26% were university graduate students and postdocs. Just over 11% were professionals in the biomedical industry, and another 10% were government scientists, program managers, and administrators. The remaining interviewees were Institutional Animal Care and Use Committee (IACUC) staff, consulting biostatisticians, and private foundation personnel.

Findings from the research community

Our overarching observation was that research reproducibility is a real concern among the stakeholders we interviewed. For a few interviewees, the Amgen[5] and Bayer[6] results demonstrate that, in the long run, science works as it should. Many noted the urgency to translate research into human health, as well as the lost effort chasing false leads. Those flawed results may have a long lifetime. For example, Begley and Ellis[5] reported that the articles that they were unable to reproduce had already accumulated 248 citations.

The term “reproducibility” itself was felt to be ambiguous and could take several different meanings. Some interpret it as exact replication of an experiment within a lab; others, replication in a different lab based on exact protocols (or just a publication). It could also refer to transitioning to different experimental conditions or even different model organisms. Some industry and academic clinical researchers prefer the term “robustness,” connoting a result that is consistent across a spectrum of experimental conditions and animal models. Goodman et al.[7] propose a specific terminology, including reproducibility of methods, results, and inferences, as well as generalizability and robustness as extensions. Most interviewees interpreted reproducibility as Goodman’s “results reproducibility,” meaning obtaining the same results in an independent study with closely matched procedures.

Beyond these high-level observations, most of the reproducibility concerns can be categorized as follows:

Category 1: Reproducibility and institutional culture.

The most commonly cited contributor to irreproducibility was the pressure to publish in high-impact venues, compounded by the additional pressure to generate grant funding. Many interviewees noted that the research community’s preference for papers in *Science* or *Nature* was a problem, and a number of senior researchers from academic institutions and

industry expressed strong disdain for the reliability of results in these two journals while continuing to note the weight that publishing therein carries for tenure, promotion, and academic stature.

Closely related was the need to demonstrate positive results. Students and postdocs noted pressure to “find something” when a study produces a negative result. A number of academic scientists felt that highly publicized rates of failures to reproduce studies are over-estimates of the actual outcomes in academic labs: unpublished negative-result studies are not included in reproducibility estimates.

Trust between research communities was also a concern. Industry scientists emphasized the importance of negative results and the potentially enormous financial consequences of pursuing a poor lead too far. They were quite pessimistic about reproducibility: several placed the chances of being able to reproduce published studies at 10% or less. They also uniformly trusted industry research over the academic literature. Many academic researchers working on clinical trials or translation to human subjects concurred with this opinion. Several industry researchers felt lab auditing by the NIH would be necessary for improvement, with compliance failures leading to funding penalties. One academic researcher has even hired a “rigorologist” to monitor lab processes continuously and report any issues. Some academic researchers expressed negative opinions on industry’s ability to perform replications properly, suggesting that industry-conducted reproducibility studies could not be trusted.

Another common theme involved lab and project scale. Several senior scientists from academia, government, and industry suggested that academic labs without sufficient supervision of junior personnel were likely to produce substandard research. A number of interviewees thought that principal investigators with multiple large grants would be unable to oversee their labs properly.

In the conversations with IACUC personnel, both staff and faculty, the common concern over reproducibility and its impact on animal welfare was expressed, as well as the support for the NIH’s recent initiatives in improving rigor in animal experiments. These interviewees saw a need for much stronger support of researchers in designing animal experiments.

Category 2: Complexity of biomedical experiments.

The pace of technological progress was also a concern for multiple interviewees. To succeed, researchers must master several subdisciplines and new, evolving, and complex equipment. Maintaining expertise requires a large staff of students, postdocs, or external collaborations, and intra-team communication can be perceived as a strain. A few researchers pointed out that results obtained using technologies unique to a single lab are at high risk of being irreproducible.

An additional challenge with advances in measurement technology is that of data quantity. The ability to measure many things simultaneously creates new opportunities for developing complex scientific hypotheses, but this in turn can lead to difficulties in statistical data analysis as well as difficulties in maintaining appropriate expertise and protocol discipline.

Category 3: Scientific hypotheses and data analysis.

Two important issues for reproducibility are the scientific hypothesis of a study and the statistical analysis of collected data to examine the hypothesis. Formulation of a clear hypothesis that can be approached with a statistical test is a key ingredient of most empirical scientific studies. These concepts may seem to be on the opposite ends of an experiment's "work flow," but in discussions we found some coupling of the two, particularly in light of negative results.

Academic biostatisticians we interviewed lamented that quite a bit of their "business" involved requests to "save" negativeoutcome experiments. One postdoc referred to "mission creep:" given a negative result, can the data be used to examine a different question? Slight modifications of the hypothesis may lead to a statistically significant, and hence publishable, outcome for the experiment. This mission creep can be a perfectly legitimate activity in the process of scientific investigation. However, as a number of interviewees noted, scientists must distinguish between exploratory and confirmatory results. Therapeutic development, in the opinion of these interviewees, requires much more effort on the confirmatory side, while basic research sometimes blurs the two.

A number of researchers also pointed to the phenomenon of "*p*-hacking," in which the researcher analyzes their data in different ways until a desired "$p<0.05$" result is obtained. One interviewee relayed the story of a clinical study in which a "nearly significant" result led an investigator to drop one patient and enroll one more to obtain a "$p<0.05$" outcome.

Some interviewees noted the related but perhaps more problematic behavior of "cherry picking" from multiple studies and reporting only the successful one. One interviewee relayed a failure to reproduce a published result. The lead author of the study later confessed that they had attempted the experiment six times and only reported results of the single significant one.

Category 4: Experimental design considerations.

With few exceptions, interviewees involved in animal research expressed various levels of discomfort with the statistical aspects of experimental design and analysis. Again, the academic biostatisticians with whom we spoke noted that few researchers involve statisticians at the design stage, a situation that often creates difficulties later on at the analysis stage. Many more of the academic researchers developed design plans without statisticians than with them. Some researchers who do regularly collaborate with statisticians expressed surprise that others do not do so. One private foundation supporting biomedical research has formed a panel of experts to advise its research community on experimental design and statistical analysis.

Despite IACUC requirements for statistical justification of animal use, power computations for sample size determination were viewed as problematic. As noted above, Button et al. found that many experiments in neuroscience suffered from a lack of power due to small sample sizes. Interestingly, the use of power as a sample size justification was largely viewed by our interviewees as a necessity for grant applications, but the most common approach in the lab is experiencebased: choose a sample size that worked in previous

experiments. Interviewees using this approach were typically more senior (or the graduate students of a senior investigator) with a substantial knowledge of the literature and record of experimentation. These scientists were somewhat skeptical of statistical approaches to sample size determination, and the utility and relevance of these computations was a topic of serious debate. Many if not most academic researchers viewed power computations as a "necessary evil" to satisfy reviewers.

Practical issues with "power calculation" sample size determination involve existing equipment and personnel. For example, the number of available cages may limit the number of animals that can be included in a study. Or if the lab has a centrifuge that only hosts 28 test tubes, then a sample size of 32 becomes a burden. Such constraints were often cited as important problems in sample size determination.

A more difficult issue to resolve in applying statistical arguments to determine appropriate sample size is that of scientifically or clinically relevant effect size. Several researchers suggested that the treatment levels chosen should be large enough to create "obvious" effects and that statistics should not be necessary to detect a significant treatment effect. Some of these researchers justified their positions using graphical aids from their manuscripts papers to demonstrate "visual significance." From a pure discovery point of view, this line of thought may very well be reasonable. Such an approach can be viewed as proof-of-concept: there are treatment levels that are very likely to generate a sizeable outcome effect. Translation from basic research to human health, however, requires deeper exploration of treatment levels and dose-response relationships, and industry scientists and clinical researchers were not typically swayed by "obvious" treatment arguments.

Some researchers, typically those closer to translation and clinical research, suggested that pilot studies to investigate the dose-response relationship were necessary to get preliminary estimates of the effect sizes one might expect. This position was reinforced quite strongly by the biostatisticians we interviewed, who noted that the absence of such prior information creates a serious difficulty for understanding clinically or scientifically relevant effect sizes. Without a strong sense of the minimal effect size to be detected, statisticians also warned of the possibility of "power hacking," an analog of p-hacking: choosing an effect size that leads to a pre-determined sample size having the appropriate power. Lenth[8] notes the problems of "canned" relative effect sizes without careful regard for scientifically important effects.

Interestingly, while nearly every researcher we interviewed implements some form of randomization, we found a number of interviewees to be critical of blinding. Landis et al.[9] and several references therein point to blinding as a potential source of bias in animal experiments. Our interviewees had several problems with blinding. One is the potential variability introduced by multiple personnel handling animals. Another observation we heard many times is that the experimentalist can detect the treated subjects due to the very nature of the response to treatments. Researchers studying behavior seemed especially likely to view blinding as pointless for this reason. Cost in additional lab personnel necessary to implement blinded experiments was a common problem cited by academics. But industry scientists, even those at relatively small companies, viewed the expense as worth it, and contract research organization personnel indicated that blinding is part of their practice.

Discussion

The challenge of improving reproducibility in biomedical research involves many stakeholders. The incentives and culture of academic research push for rapid publication that is in conflict with carefully conducted and documented experiments and data analysis. Industry research objectives, especially within the structure of the FDA approval process, provide stronger incentives for reproducible research outcomes. While changing the culture of academic research is at best a very longterm process, the elements of the industry research process, such as experimental design and protocol registration, could be adopted by funding agencies or animal welfare organizations to support rigorous planning and design of animal experiments. Institutional Review Boards and IACUCs may need to take a more active oversight role.

The outcomes of these interviews have led us to agree with Peng[10], who recommends "massive-scale education efforts" for scientists. Those researchers with well-established collaborations with statisticians were more confident in their prospective experimental designs as well as their data analysis. Computer-based tools for experimental design, in paradoxical distinction with data analysis software, remain the purview of expert users. On the training front, we noted two key issues from discussions with interviewees. First, translation of concepts from the scientific domain of inquiry into the appropriate statistical language is surprisingly difficult. Second, design matters such as the difference between a "statistically significant" and a "scientifically relevant" result or effect are challenges poorly addressed in statistical coursework. Improving the statistical rigor and robustness of preclinical research involving animals necessarily means more focus on these issues.

Acknowledgements

We would like to thank Dr. Irina Sazonova and Mr. Victor Prikhodko of NIDA for their support of this project. We would also like to thank the I-Corps™ at NIH instructors, especially Dr. Aileen Huang-Saad and Mr. Todd Morrill, for their criticism, guidance, and encouragement as we conducted our interviews and processed our findings. Finally, we are very thankful to all the interviewees who generously shared their time and their insights and helped us understand their challenges in experimental design and reproducibility.

References

1. Kilkenny C et al. PLoS One 4 (11), e7824 (2009). [PubMed: 19956596]
2. Button KS Nat. Rev. Neurosci 14, 365–376 (2013). [PubMed: 23571845]
3. Freedman LP, Cockburn IM & Simcoe TS PLoS Biol. 13 (6), e1002165 (2015). [PubMed: 26057340]
4. Baker M Nature 533, 452–454 (2016). [PubMed: 27225100]
5. Begley CG & Ellis L Nature 483, 531–533 (2012). [PubMed: 22460880]
6. Prinz F, Schlange T & Asadullah K Nat. Rev. Drug Discov 10, 712–713 (2011). [PubMed: 21892149]
7. Goodman SN, Fanelli D & Ioannidis JPA Sci. Transl. Med 8 (341), 1–7 (2016).
8. Lenth R Am. Stat 55 (3), 187–193 (2001).
9. Landis SC et al. Nature 490, 187–191 (2012). [PubMed: 23060188]
10. Peng R Signif. 12 (3), 30–32 (2015).