Theme 16 - Research Reproducibility (2016)

Biomedical research is critically dependent on accurate and reproducible research publications. A loss of faith in the scientific literature would not only hinder further research advances and development of new clinical therapies, but it might also undermine public trust and funding. A number of recent articles have raised concerns about research reproducibility, noting that they only rarely involve research misconduct. Instead, problems encountered in replicating the work of others can have multiple causes, ranging from differences in approaches, materials, or scientific rigor, to insufficient information about methods. This case study addresses many of the potential experimental design issues, practices, and pressures that can undermine research reproducibility.

Because this important topic is both broad and provocative, with issues that could be discussed for hours, discussion leaders and participants will need to identify ways to keep the discussion on schedule. Three potential alternative approaches to this are: (a) keep discussion of the entire case concise and well-paced; (b) discuss a selected subset of the sections, labeled by Roman numerals, that are the most relevant to the particular IC and audience, and/or use only selected questions; or, (c) dedicate more than one hour to discussing this case.

The Appendix for this case provides a summary of factors that can weaken research reproducibility. It also lists web-based resources that can help strengthen this crucial foundation of modern biomedical research.

I. Drs. Smith and Garcia have independent labs in the NIH intramural program and are not familiar with each other's research. Dr. Smith's lab is studying a novel protein that they name 'tumorstatin' because they demonstrate that it is a potent inhibitor of tumor cell growth in tissue culture. Independently, Dr. Garcia's lab studies a molecule [which is only later found to be the same protein] that they call 'tumorin' because multiple experiments show that it promotes tumor growth in mice when highly expressed. Each lab feels pressure to publish quickly in a high-profile journal so lab members can obtain jobs, tenure, or grants, and each group submits their paper hastily to the same prominent journal named *High-impact*.

- Do you feel pressure to publish your research rapidly and in high-profile journals?

II. By chance, each investigator is invited by the journal to review the other's paper, not realizing that they are studying the same protein. Based on his extensive experience with animal and clinical studies, Dr. Smith harshly criticizes many perceived technical problems with the tumorin study, including missing controls, failure to randomize animals with observer blinding to avoid bias, failure to handle tissue samples sufficiently carefully in a standardized manner, and using assays now known to be unreliable. He says the paper fails to meet the guidelines from journal editors on an NIH site: http://www.nih.gov/about/reporting-preclinical-research.htm His postdoc co-reviewer criticizes both the use of an antibody in a commercial kit known to have poor specificity and the over-interpretations of microscopy images beyond theoretical limits of resolution. The Editor knows that Dr. Smith is a very tough reviewer and because the other referees were more positive, her letter to Dr. Garcia leaves the door open for resubmission if all

of the concerns can be resolved while continuing to provide exciting new findings for a "clean" complete story.

1. Under what conditions can a postdoc participate in reviewing for a journal?
2. Are all of the issues cited in their review reasonable and based on currently accepted practice?
3. Are there dangers from biased thinking in even the most careful labs to obtain the "right" answer or in trying to "prove" a hypothesis?
4. What differences in standards of research conduct exist between studies to obtain preliminary data to generate hypotheses versus testing a specific hypothesis?
5. What are pros and cons of hypothesis-driven and exploratory research that addresses a question for which any clear answer will be useful?
6. Does human clinical research have similar or additional requirements or considerations?

III. Dr. Garcia is upset by the tough review with seemingly unreasonable demands, and considers quickly submitting the paper to a specialty journal. But her two postdocs realize that their future job prospects would weaken as a result. They argue forcefully that the reviewer was unfair, and say they can quickly complete the experiments to resolve each concern to get the high-visibility publication. Although Dr. Garcia believes that the research findings are valid whether or not they agree with one's hypothesis, she gives her postdocs a free hand because she trusts them, knowing they received many hours of research ethics training. Also, getting this high-visibility publication will strengthen her site visit review next year.

1. Is there an ethical "slippery slope" when a lab tries to obtain specific results for paper acceptance?
2. How can emotional reactions to bad reviews affect subsequent decision making?
3. If only one of several reviewers raises a subtle but potentially important issue, is it acceptable to pull the paper and submit elsewhere, hoping the issue won't be raised in a fresh review?
4. Besides more specialized or less-competitive journals, what are "predatory journals"?
5. How do trainees in your group learn research ethics and best research practices?

IV. Dr. Garcia's journal review of the paper on tumorstatin from Dr. Smith's lab points out that the gels are of very low-quality, suggesting the experiments had not been repeated. She criticizes a graph reporting significance of $P < 0.05$ using an inappropriate statistical test, questions some beautiful images showing huge effects that seem too good to be true compared to the findings in a graph showing a 25% promotional effect, as well as use of only one cell line and an inhibitor with borderline specificity. Besides raising these concerns, she requests access to the primary data to check their validity. She adds that the field usually applies an independent approach to verify surprising findings. The Editor, who is a friend of Dr. Smith and would like to publish the paper, asks whether Dr. Smith can resolve the concerns that are holding up acceptance for publication, including providing primary data when practical.

1. What specific research and ethical issues are raised here, and how important or reasonable is each? For example, is it reasonable for reviewers to request access to primary data?
2. How do you draw the line between appropriate everyday conduct of science, sloppy science, and research misconduct?

3. How important are good reviewers and editors, not only for research reproducibility, but also for avoiding demands for unnecessary experiments?
4. How do personal relationships between authors, reviewers, and editors affect the peer review process?
5. How common is it for researchers to be more critical of work by others compared to their own?

V. Dr. Smith is incensed by the review and believes it came from a biased competitor. He clarifies with his lab that the experiment in question had been performed four times and worked twice, so they can state that they performed four repeats. He suggests that they find another statistical test that supports the "right" answer, and that extra data points be added as needed to achieve statistical significance. He asks his postdocs to find another cell line that gives the same results, another inhibitor, and another assay that can support the claims, with minimum sample sizes to complete their work within the 3-month resubmission deadline. Although a postdoc has lost some of the primary data, they agree to send just enough to satisfy the journal. Their division director is sympathetic to these efforts, and they think they understand him to say: "Because research is often handicapped by imperfect instruments and biological variability, judicious selection of methods and data is sometimes necessary to support visionary ideas and success in our tough field."

1. Which of Dr. Garcia's points are the most important problems, and which are less important?
2. How important is it to preserve original data and why? When should they be shared?
3. Are lab environment and hierarchy important for research reproducibility? How did these differ between the Smith and Garcia labs?

VI. A. With hard work and skillful revisions, each paper is accepted for publication in *High-impact*. When members of the two labs see posters from the other lab at a major conference, they discover to their surprise that their protein sequences are identical. Each group is sure that the other is wrong because they see contradictory effects on tumor cells.

1. Is it possible that both labs are correct? How might this occur, and can you provide any examples?
2. Might local lab environmental or other conditions in their institutions affect the results, such as conditions in their cell culture and animal facilities, different chow, etc.?
3. If you were Dr. Smith or Dr. Garcia, what would you do?
4. What if you were a lab member?

VI. B. Each lab races to repeat/refute the other group's findings, and they request key materials from each other. Dr. Smith provides some missing information but balks at providing their cell lines because these cell lines are widely available. Although a new postdoc in Dr. Smith's lab initially encounters trouble repeating the lab's findings, she is able to do so after guidance from an experienced postdoc. Dr. Garcia hesitates to share their transgenic mice because Dr. Smith may conduct similar follow-up studies, but she provides some additional unpublished information.

1. Have you encountered problems in trying to replicate results from another research group or even from your own?
2. Do authors currently provide sufficiently detailed methods in papers and subsequent access to tools including plasmids, cells used for the experiments, animals, and computer code?
3. How can doing an experiment "the right way" affect results, and more broadly, how can experimental, environmental, and biological variability alter findings and conclusions?

APPENDIX: Factors that Can Compromise Research Reproducibility

Conceptual weaknesses and cognitive bias

- Not distinguishing between exploratory research examining multiple hypotheses/possibilities and testing of a specific hypothesis
- Trying to "prove" and defend a hypothesis rather than trying to answer a question
- Concluding "my experiment worked" if it is the preferred answer
- Lack of concern about approaches that might lead to research misconduct
- Insufficiently rigorous peer reviewers and journal editors

Research background and cultural differences

- Insufficient or ineffective training in responsible conduct of research
- Hierarchy in which the boss/mentor's hypotheses are favored over actual findings
- Cutting corners and sloppy research
- Selective interpretation of data
- Problematic lab culture (social dynamics)

Internal and external pressures

- Perceived need to publish in high-visibility journals
- Needing large numbers of publications, even if in poor or predatory journals
- Requirements by journals for exciting novel results, not negative findings
- Demands for clean, definitive, complete stories with impressive-looking data
- Deciding not to publish unwanted or controversial findings
- Demands from reviewers and editors for specific supportive findings
- Needing to find jobs, get tenure, or keep funding to take care of one's staff

Biological variability

- Local environmental factors: type of housing, water, feed, climate control, physical activity
- Strain, sex, or age of animals or cells
- Effects of microbiome, or undetected infection
- Incomplete penetrance, wide variations of expression or phenotype

Experimental design and performance

- Failure to retain primary data
- Lack or misuse of appropriate controls
- Not considering that effects can be dose-dependent, or differ *in vitro* versus *in vivo*
- Not distinguishing between technical and biological replicates for data points
- Low power (e.g., small n) leading to false-positive results
- Piecemeal add-ons to sample size
- Small effect size
- Insufficient number of repeat experiments
- Exclusion of certain experiments or data points
- No randomization, observer blinding, or checking by independent evaluator(s)
- Faulty use of statistics (failure to correct for multiple variables, over-dependence on $P < 0.05$, selecting a statistical test because it gives a preferred answer)
- Use of only a single approach

- Inconsistent or unreliable sample handling (faulty collection, storage, thawing)
- Poorly performing assays and/or failure to keep up with the newest, best technologies
- Vague or loose outcome definition (especially clinical endpoints)

Technical issues
- Incorrect instrument settings, e.g., background, sensitivity
- Pushing beyond the limits of a technology
- Insufficient antibody validation
- Off-target effects of inhibitors or stimulators
- Complete faith in purchased kits that may have sub-optimal validity
- Non-availability of key reagents/animal models
- Contaminated cell lines or sequence errors in plasmids
- Poor communication or failure to fully assist another lab struggling to reproduce one's finding

Presentation
- Incomplete methods (sloppy or deliberate)
- Lack of availability of primary data, metadata, computer codes, and unique reagents
- Selective presentation of "representative" data
- Not following best practices in the field, e.g., imaging or FACS guidelines, antibody validation, performing RNA interference, etc.

INTERNET RESOURCES

NIH website on research reproducibility: http://www.nih.gov/science/reproducibility/

NIH policy on sharing of unique research materials: https://grants.nih.gov/grants/sharing.htm

Guidelines from journal editors: http://www.nih.gov/science/reproducibility/principles-guidelines.htm

Video reproducibility training modules: https://oir.nih.gov/sourcebook/ethical-conduct/research-ethics/committee-scientific-conduct-ethics-csce/responsible-conduct-research-training/instruction-responsible-0

Reproducibility of data collection and analysis in modern technologies: Potentials and pitfalls
Cell Biology http://videocast.nih.gov/summary.asp?Live=15277&bhcp=1
Structural Biology http://videocast.nih.gov/summary.asp?Live=15910&bhcp=1
Genome Technology http://videocast.nih.gov/summary.asp?Live=16381&bhcp=1