Clinical Chemistry 63:5
000–000 (2017)

Perspectives

The Reproducibility Wars: Successful, Unsuccessful, Uninterpretable, Exact, Conceptual, Triangulated, Contested Replication

John P.A. Ioannidis[1,2,3,4*]

The recent publication of first results from the "Reproducibility Project: Cancer Biology" has stirred debate. The project synopsis put together by Nosek and Errington *(1)* tried to describe carefully what replication means, how to judge whether "same" (or different) results emerge in a replication experiment, and how to interpret divergent results in original vs reproducibility studies. However, multiple other commentators on these first reproducibility studies have enhanced our uncertainty. Almost every commentator reached a somewhat different conclusion. Reproducibility of inferences has been dismal.

Several authors of the original papers along with other commentators have questioned the reproducibility effort. These retorts typically defend the original findings, interpreting the replications as more successful than unsuccessful. They question replications on multiple fronts, e.g., inappropriate statistical methods or poor experimental competence. They lament the inappropriate shaming consequences, when poorly done replication efforts tarnish great scientists. They worry that reproducibility checks destroy discovery and stall efforts to translate promising research. They wonder whether we should waste money on replication.

The reproducibility wars are not exclusive to laboratory science. Last year, a similar debate erupted with a Technical Comment exchange on the respective reproducibility project on psychology *(2)*, although the available data were far more extensive (100 experiments instead of just 5 preliminary ones) and had a massive participation of top psychologists (270 scientists and their teams) trying to reproduce results. Some famous psychology academics nevertheless concluded that their field had no reproducibility problem and reproducibility was misleading business. This position is immediately suspect. If psychological science is so perfect, how can it be that 270 of the best psychologists in the world working under optimal conditions of openness and most rigorous protocols and methods could get everything so badly wrong? If the most closely controlled and scrutinized corpus of experiments ever done on psychological science was so flawed, what should we expect of the rest of the field? Similar debates have also emerged recently in clinical medicine for efforts trying to promote openness and transparency with full availability of raw data and protocols *(3)*. Openness could allow reanalyses of clinical data to become routine. Reaction has ranged from thoughtful suggestions on how to improve the process, skepticism about harming clinical research, and emotional tirades against "research parasites." Defenders of the dysfunctional status quo have included several powerful academic and research figures.

A similar defense of the status quo is also emerging in laboratory science. Previous reproducibility efforts have suggested reproducibility rates of <25% for top studies in basic and preclinical biomedical research *(4)*. The main criticism of the largest efforts to date was that they were run by the industry without open data. The Reproducibility Project: Cancer Biology overcame these limitations, yet criticism intensified.

Debate arose even for a successful replication. Cimetidine did have an effect against lung cancer reproduced in mouse xenografts, but results were not formally statistically significant at $P = 0.05$. The effect size is what matters; it was smaller in the replication than in the original but still sizeable. CIs vastly overlapped. Of course, it would have been nice to have nominal statistical significance, but this is less essential. A debate nevertheless arose whether the multiplicity correction used by the replicators in their statistical analysis was inappropriate. Although multiplicity correction applied to the original study, it might be unnecessary in the focused replication. There is some partial merit to this argument. However, the protocol and the analysis plan of the replication had been available in public and shared with the original authors for comments, editing, and corrections before running any experiments. The original authors did not correct the analysis plan at that time. When top scientists seek post hoc revisions of the analysis plan to get a formally statistical result, this makes me pause. Perhaps it

[1] Stanford Prevention Research Center, Department of Medicine, [2] Department of Health Research and Policy, Stanford University School of Medicine, [3] Department of Statistics, Stanford University School of Humanities and Sciences, and [4] Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford.
* Address correspondence to the author at: Stanford University, Medical School Office Building, Room ×306, 1265 Welch Rd., Stanford, CA 94305. E-mail: jioannid@stanford.edu.
Received January 28, 2017; accepted February 3, 2017.
Previously published online at DOI: 10.1373/clinchem.2017.271965

provides an explanation why post hoc significance chasing is so rampant in the biomedical literature that an implausible 96% of published papers with *P* values have significant results *(5)*.

For 2 of the 5 original experiments, the reproducers could not get the whole chain of laboratory experimentation to work as in the original publications. For example, in an effort to reproduce whether an antibody to CD47 has antitumor activity in mice, the 95% CIs of whatever outcome was measurable showed gross divergence in the original and reproducibility experiments. Point estimates differed almost 50-fold. However, an unanticipated problem had ensued. Tumors generally grew slowly and even spontaneously regressed in the control group, a phenomenon not seen in the original publication. The original authors claimed that the reproducers would have been welcome to visit their laboratory to master the technique. In another topic, mutations in *PREX2* (phosphatidylinositol-3,4,5-trisphosphate dependent Rac exchange factor 2) increased melanoma growth in the original study, but the reproducers found that cells with and without mutations both formed tumors very fast. Apparently, the cells no longer behaved as in the original experiments.

The editor of *Elife* called these 2 papers "uninterpretable". Anyone working in laboratory knows how difficult it is to make some methods work, when the sole reference is the short, elliptical Methods section of a peer-reviewed paper. Grapevine rumors spread in corridors, in unofficial e-mail chats, and in meeting breaks when multiple scientists and laboratories fail to make a method work. Currently, it is notoriously difficult to publish somewhere that "I have not been able to make this laboratory method work" in the peer-reviewed literature, although it is one of the most common frustrating experiences of any laboratory scientist. Sometimes, one option is indeed to visit the laboratory where the method originated. However, this is not always realistic or desirable, for many reasons. Moreover, such an approach of science-by-visitation assumes that the scientific literature is more of an advertisement rather than a solid record that allows reproducing investigative procedures in sufficient detail. If efforts in reproducible research could simply improve the recording and utility of Methods sections so that they could be meaningful and functional, this would already be a major advancement.

The fiercest debates arose when the reproducers could carry out all the original experiments as planned, but still could not reproduce the original results with CIs that excluded chance. The original author questioned the competence of the reproducers and their materials and complained about the damage done to the translational effort. Other scientists seconded this, saying that they had been able to reproduce the results in other publications. I took a closer look at these other "successful" reproductions. They typically pertained to very different experiments. For example, the claimed reproduction in the news coverage at *Nature* applied to hepatocellular carcinoma, while the original study evaluated prostate carcinoma. Other replications involved different tumors, different experimental systems, different antitumor peptides, different experimental conditions, and so forth.

This situation is extremely common in laboratory medicine. There is actually a strong disincentive toward reproducing something in the very same way as originally done. There is pressure to have something new to say, to do things differently. This leads to conceptual rather than exact replication. Then, when the pieces of evidence on the same theme are substantially different, the successful scientist weaves a common story, a consistent narrative. Triangulation has thus become the art of building biological fairy tales. I do acknowledge that conceptual replication and triangulation can be useful in some situations. However, they have a major drawback: almost anything can fit into a triangulation narrative by invoking some speculative "biological plausibility" as the connecting glue.

It is likely that most of these conceptual and triangulation links are overstated leaps of faith. Otherwise, it is very difficult to explain why we have so many successful narratives in the basic sciences, but very few of these proposed discoveries eventually work in humans. Moreover, a published conceptual replication with a different design and/or experimental conditions does not say how many laboratories have tried and how many different designs and experimental conditions failed and remain unpublished. In the current environment where there is florid selective reporting and chasing of significant results, it is unknown whether 0, 3, 10, or 100 experiments and variants have been tried and failed for each successful publication. The approach of the reproducibility project to preregister the replication effort is thus essential. There is a preregistered detailed protocol and even a preregistered report. Essentially, one writes the paper (without the exact numbers) before running the experiments. This approach safeguards that selective reporting bias will not be an issue.

We are still scratching the surface of reproducibility in laboratory biomedical research. We clearly need to learn more. No matter if replications fail or succeed, we have a lot to learn from them. It is not about shaming and tarnishing reputations. It is about whether our observations are solid and eventually can have practical value. For those who question how much we can spend on replication, the answer is probably a thousand-fold more, easily. The entire Reproducibility Project: Cancer Biology undertaking costs $2 million. Waste in biomedical research is estimated to be tens, perhaps even hundreds of billions, annually. To do better, insights on reproducibility will be crucial. Laboratory research is of tremendous

importance. We should not drown its excellence in a sea of irreproducible results.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures or Potential Conflicts of Interest: *Upon manuscript submission, all authors completed the author disclosure form. Disclosures and/or potential conflicts of interest:*

Employment or Leadership: None declared.
Consultant or Advisory Role: J.P.A. Ioannidis, Center for Open Science and Science Exchange Reproducibility Initiative (unpaid).
Stock Ownership: None declared.
Honoraria: None declared.
Research Funding: None declared.
Expert Testimony: None declared.
Patents: None declared.

References

1. Nosek BA, Errington TM. Making sense of replications. Elife 2017;6:e23383.
2. Open Science Collaboration. Estimating the reproducibility of psychological science. Science 2015;349:aac4716.
3. Emmert-Streib F, Dehmer M, Yli-Harja O. Against dataism and for data sharing of big biomedical and clinical data with research parasites. Front Genet 2016;7:154.
4. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res 2015;116:116–26.
5. Chavalarias D, Wallach JD, Li AH, Ioannidis JP. Evolution of reporting P values in the biomedical literature, 1990–2015. JAMA 2016;315:1141–8.