Statistical and philosophical issues concerning replicability in clinical research.

David Trafimow*, Hunter A Myüz

Department of Psychology, New Mexico State University, Las Cruces, New Mexico

Accepted on September 12, 2017

Commentary

An important limitation in clinical research concerns replicability issues; unfortunately, much clinical research tends not to replicate [1-11]. Because evidence-based practice depends on having replicable evidence, this is an important problem that needs discussion. Our goal in this short commentary is to explain why we believe that current research practice contributes to replicability problems. We see two categories of concerns: these are statistical and philosophical.

The Statistical Issue

In clinical research, as in much research across the sciences, there is almost universal dependence on null hypothesis significance testing (NHST). In short, NHST includes proposing a null hypothesis (usually that the experimental and control condition population means are the same), setting an alpha level that provides the cutoff for determining "statistical significance" (usually, this is set at 0.05), and computing a p-value. If the p-value is less than the alpha level (e.g., $p<0.05$), the researcher rejects the null hypothesis in favor of the alternative hypothesis. The finding is deemed "statistically significant," and journals likely will publish it. In contrast, if the p-value is greater than the alpha level (e.g., $p>0.05$), the researcher does not reject the null hypothesis. The finding is deemed not to be statistically significant, and journals will be unlikely to publish it. What clinical researchers fail to realize, however, is that although NHST seems reasonable, it is replete with flaws, one of them being that the procedure renders problems replicating statistically inevitable.

To commence, it is important to be conscious of exactly what a p-value is. It is the probability of obtaining the finding, or one more extreme, given that the null hypothesis is true. One consequence of this is that p-values have a sampling distribution just as any other statistic does [12]. If one performed the same experiment many times, and computed a p-value each time, there would be a wide variety of p-values, with a different p-value for each replication. Of course, with real research, there typically is only one experiment, but the notion of a distribution of p-values is nevertheless relevant. To see this, consider that the single experiment a researcher conducts, along with the p-value that the researcher obtains, could be any of the p-values in the distribution of p-values that potentially could have been obtained. Because most p-values, in a distribution of p-values, are unlikely to be below 0.05, it follows that the researcher needs some luck to obtain $p<0.05$. Even if there really is an effect at the population level, obtaining $p<0.05$ is unlikely unless the population effect size is enormously larger than effect sizes typically obtained, the sample size is enormously larger than sample sizes typically obtained, or both. The upshot is that whenever a researcher obtains $p<0.05$, it is likely that some luck was involved. Should clinical researchers expect luck to replicate? Although this will happen some of the time, it often will fail to happen—hence the statistical inevitability of replication problems [13]. Empirical confirmation of this statistical inevitability, which often is termed "statistical regression" or "regression to the mean," was obtained in psychology, where the Open Science Collaboration [14] attempted to replicate many experiments published in top journals, and found that most failed to replicate.

The mathematics of statistical regression are impossible to dispute, but clinical researchers might wonder why a seemingly valid procedure, such as NHST, forces replication problems as indicated in the foregoing paragraph. The answer is that NHST is not a valid procedure. To see why, consider again that the researcher who performs NHST has the goal of rejecting the null hypothesis so she can publish. But to reject the null hypothesis, we would expect to be able to show that the null hypothesis has a small probability of being true, given the obtained data. But remember that the p-value gives the probability of the data given the null hypothesis. And therein lies the problem: the probability of the data given the null hypothesis is not the same quantity as the probability of the null hypothesis given the data. Thus, the p-value is the inverse conditional probability of the one needed and Trafimow [15] has provided an extensive mathematical demonstration that the quantities can be quite different indeed.

But to make the inverse conditional probability issue intuitive, consider the probability that a person is president of the USA, given that the person is a citizen of the USA. This doubtless would be a number close to zero. In contrast, consider the inverse conditional probability that a person is a citizen of the USA, given that the person is president of the USA. Given constitutional requirements, this probability is close to unity (and is unity if the constitution is followed). Thus, we see that inverse conditional probabilities can be quite different from each other. Put more broadly, NHST makes use of a conditional probability—the p-value—that is the inverse of the conditional probability that is needed to validly reject the null hypothesis. Thus, we do not have the case where a valid procedure paradoxically produces replication problems; rather, NHST is logically invalid and so it is not paradoxical that it also produces replication problems.

The Philosophical Issue: Auxiliary Assumptions

Thus far, we have imagined the same experiment performed many times, and showed that NHST renders replication problems statistically inevitable. But matters are worse than that because, in real research, it is impossible to perform exact replications. There may be differences in dates, times, rooms,

and so on. In addition, there may be differences in the way a treatment is employed, the measures used to assess the effect of the treatment, and so on. To gain a broader understanding of the issue, it is useful to consider the role of theory in clinical research.

Consider any clinical theory in the clinical literature. That theory will include nonobservational terms, just as any theory in any area of science does. Even in physics, Newton's famous equation

Force = Mass × Acceleration

Contains "mass" as a nonobservational term that should not be confused with "weight," which is observational. The difference is rendered obvious upon considering that the same object would have the same mass on Earth or Jupiter but would have different weights. Put simply, mass cannot be observed directly whereas weight can be observed directly. Inferences to statements about mass can be made based on weight, providing that the researcher has additional (auxiliary) assumptions enabling her to connect mass and weight. Put more generally, empirical tests of theories necessitate that the researcher form empirical hypotheses containing observational terms. To bridge the gap between nonobservational terms in theories, and observational terms in empirical hypotheses, it is necessary to use auxiliary assumptions, though many researchers are not conscious of this [16, 17].

An implication of the necessity to have auxiliary assumptions, many of which are implicit, is that it is very likely that a replication attempt will involve different auxiliary assumptions than the original study. Consequently, if there is a discrepancy in the findings (and there often is such a discrepancy), it is difficult to determine whether the original study was at fault or whether the replication study was at fault. Worse yet, as we pointed out in the foregoing section, even under ideal conditions of having exact replications, the use of NHST is sufficient—nay, guarantees—replication problems even without considering the issue of auxiliary assumptions. In any event, even if researchers were to stop using NHST, replication failures nevertheless might remain common (though probably less so) due to differences in auxiliary assumptions. Can anything be done about this?

Although there is no single complete solution, we believe that substantial progress can be made by having researchers focus on auxiliary assumptions. Currently, clinical researchers focus on the theory that provides the basis for the treatment, or on details of the treatment itself, and attention to these matters obviously is valuable. However, attention also should be paid to auxiliary assumptions connecting the details of the treatment—as the researcher is planning on implementing it—to the theory from which the treatment is derived. Thus, we emphasize the connection between theory and treatment details (and dependent measure details) as being worthy of attention. We anticipate that increased attention to auxiliary assumptions connecting nonobservational terms in theories, and observational terms in empirical hypotheses pertaining to treatments and measures of the effectiveness of treatments, would pay dividends. For one thing, there doubtless will be many cases where the best way to connect a nonobservational term in a theory to an observational term in an empirical hypothesis is less than perfectly clear. The field could benefit greatly by discussions aimed at bringing such difficulties to light, with discussion devoted to addressing those difficulties. Successful replications should increase because the auxiliary assumptions upon which original studies and replication efforts are based would be clearer. Even where replications are not successful, auxiliary assumptions that are clearly stated could be tested, whereas as matters currently stand, the clarity of auxiliary assumptions often is insufficient for testing.

At the statistical level, the prevalent adherence to NHST in clinical research, with the 0.05 cutoff level for publishing, practically guarantees replication problems. But even if this were not so, the differences in auxiliary assumptions, including implicit as well as explicit ones, would nevertheless render replication difficult. What can be done about it?

At the statistical level, the answer is obvious. Researchers should stop using NHST and journals that publish clinical research should be willing to publish research not involving NHST. To take an example from psychology, the editors of Basic and Applied Social Psychology have banned NHST from the journal [18] and have published several recent articles critical of current statistical and methodological practices [19-25]. Furthermore, the American Statistical Association is moving in this direction, including sponsoring a symposium late in 2017 to discuss moving away from NHST. Clinical research should move in this direction too.

More broadly, however, there remains the issue of auxiliary assumptions. We hope to have made a start in this direction by bringing into sharp focus that many theoretical terms are nonobservational whereas empirical terms necessarily are observational, and auxiliary assumptions render possible traversing the gap. We freely admit that it is not possible to make explicit all auxiliary assumptions that go into a study, but that is no excuse for refusing to attempt to do as well as possible. Making implicit auxiliary assumptions explicit increases the likelihood that researchers will better see the differences between original studies and replication studies. Placing auxiliary assumptions in the foreground, rather than letting them languish in the background, also should lead to better tests of clinical theories and more valid theory-based treatments because the connections between theories and treatments will be better specified. Why not make the effort?

References

1. Ioannidis JP, Allison DB, Ball CA, et al. Repeatability of published microarray gene expression analyses. Nature genetics. 2009:41(2):149-55.
2. Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature reviews Drug discovery. 2011:10(9):712.
3. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012:483(7391):531-3.
4. Landis SC, Amara SG, Asadullah K, et al. Finkelstein R. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012:490(7419):187.

5. Mobley A, Linder SK, Braeuer R, et al. A survey on data reproducibility in cancer research provides insights into our limited ability to translate findings from the laboratory to the clinic. PLoS One. 2013:8(5):e63221.
6. Valentine JC, Biglan A, Boruch RF, et al. Replication in Prevention Science. Prev Sci. 2011:12:103-17.
7. Collins FS, Tabak LA. NIH plans to enhance reproducibility. Nature. 2014:505(7485):612.
8. Morrison SJ. Reproducibility project: cancer biology: Time to do something about reproducibility. Elife. 2014 Dec 10; 3:e03981.
9. Nosek BA, Lakens DD. Registered reports: A method to increase the credibility of published results. Social Psychology. 2014:45(3):137-41.
10. Vasilevsky NA, Brush MH, Paddock H, et al. On the reproducibility of science: unique identification of research resources in the biomedical literature. PeerJ. 2013:1:e148.
11. Perrin S. Preclinical research: Make mouse studies work. Nature. 2014:507(7493):423-5.
12. Cumming G. Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis. Routledge; 2013 Jun 19.
13. Trafimow D, Earp BD. Null hypothesis significance testing and Type I error: The domain problem. New Ideas in Psychology. 2017:45:19-27.
14. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015;349(6251):aac4716
15. Trafimow D. Hypothesis testing and theory evaluation at the boundaries: surprising insights from Bayes's theorem. Psychological review. 2003:110(3):526-35.
16. Duhem P, Scott WT. The aim and structure of physical theory. American Journal of Physics. 1954:22:503.
17. Lakatos I. The methodology of scientific research programmes: Volume 1: Philosophical papers. Cambridge University Press; 1978 May 18.
18. Marks M, Trafimow D. Editorial. Basic and Applied Social Psychology, 2015:37(1):1-2.
19. Grice JW, Cohn A, Ramsey RR, et al. On muddled reasoning and mediation modeling. Basic and Applied Social Psychology. 2015 Jul 4; 37(4):214-25.
20. Kline RB. The mediation myth. Basic and Applied Social Psychology. 2015:37(4):202-13.
21. Pashler H, Rohrer D, Abramson I, et al. A social priming data set with troubling oddities. Basic and Applied Social Psychology. 2016 Jan 2; 38(1):3-18.
22. Tate CU. On the overuse and misuse of mediation analysis: It may be a matter of timing. Basic and Applied Social Psychology. 2015 Jul 4; 37(4):235-46.
23. Thoemmes F. Reversing arrows in mediation models does not distinguish plausible models. Basic and Applied Social Psychology. 2015 Jul 4; 37(4):226-34.
24. Valentine JC, Aloe AM, Lau TS. Life after NHST: How to describe your data without "p-ing" everywhere. Basic and Applied Social Psychology. 2015 Sep 3; 37(5):260-73.
25. Witte EH, Zenker F. Reconstructing recent work on macrosocial stress as a research program. Basic and Applied Social Psychology. 2016 Nov 1; 38(6):301-7.

***Correspondence to:**

David Trafimow
Department of Psychology
New Mexico State University
P.O. Box 30001/MSC 3452, Las Cruces, New Mexico -88003, USA
Tel: 575-646-4023
E-mail: dtrafimo@nmsu.edu