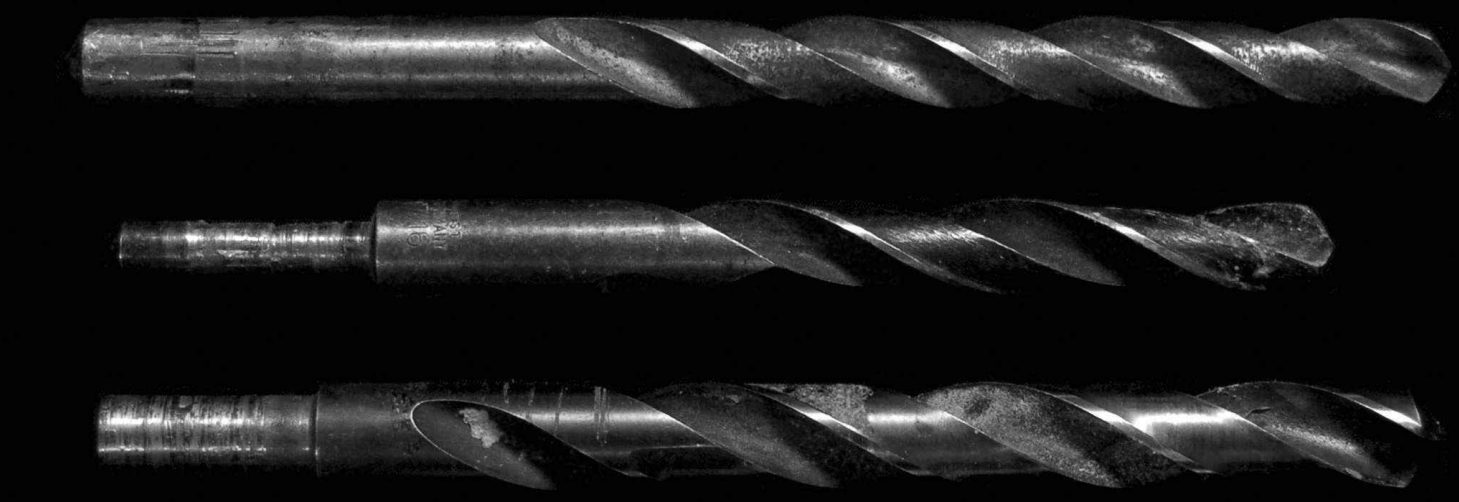
A replication crisis in methodological research?



Statisticians have been keen to critique statistical aspects of the "replication crisis" in other scientific disciplines. But new statistical tools are often published and promoted without any thought to replicability. This needs to change, argue **Anne-Laure Boulesteix**, **Sabine Hoffmann**, **Alethea Charlton** and **Heidi Seibold**

Imagine you need to take a drug. A new drug is available that has been investigated mainly through *in vitro* experiments (i.e., in test tubes, rather than in living organisms). It was shown to improve survival in a few patients selected by the pharmaceutical company. Would you feel safe taking this drug? Probably not, especially if you are a statistician. You would ask why no pre-registered randomised clinical trial was conducted to investigate the efficacy and safety of the drug.

Now imagine you need to use a statistical method for some data analysis. One of the available methods, a new method, was investigated mainly through simulations (i.e. using synthetic data sets). It was shown to be more efficient than other statistical methods in a few example data sets selected by the developer of the method. Would you feel confident using it? Weirdly, if you are a statistician, you probably would.

Statisticians are among the first to call for more rigour in clinical trials and other applied fields of statistics. Yet, in their own methodological research, statisticians commonly make claims on the performance and utility of methods based merely on theory, limited simulations, or arbitrarily selected real data examples. In the current replication crisis in science, statisticians caution against questionable research practices in fields like psychology, biology and medicine. Yet the same questionable practices should also be avoided in the development and reporting of new statistical methods.

Sins

Table 1 sets out "seven sins of methodological research", inspired by a recent *Significance* article by Held and Schwab.[1] These practices include "fishing expeditions" (i.e. running numerous different analyses in the hope that one will yield good results), then "selectively reporting" the good results while leaving the others in the metaphorical "file drawer". In some cases, this "file drawer problem" affects whole projects, whose results are deemed unexciting and are therefore not published

Table 1: The seven sins of methodological statistical research.

The seven sins of methodological research	Further reading
Fishing expeditions/selective reporting	Jelizarow *et al.*[2]; Hutson[3]
Publication bias	Boulesteix *et al.*[4]
Lack of neutral comparison studies	Boulesteix *et al.*[5, 6]
Lack of replication studies	Liu and Meng[7]
Poor design of comparison studies	Keogh and Kasetty[8]; Boulesteix *et al.*[6]; Christodoulou *et al.*[9]
Lack of meta-analyses	Gardner *et al.*[10]
Lack of reporting guidelines	

at all, further exacerbating so-called "publication bias".

In an intriguing example of how fishing expeditions and selective reporting work, Jelizarow *et al.*[2] showed that they could make a new discriminant analysis method *seem* better than existing methods, simply by picking the best results using different data sets, method variants and pre-processing approaches. In reality, the new method was no better than those already in use. Such problems are not limited to classical statistical methods. We see the same issues in machine learning and artificial intelligence.[3]

While most scientists would agree that selective reporting is bad practice, many (ourselves included) can stumble into this pitfall subconsciously, with no intention to "cheat", encouraged by the fact that new techniques are introduced in the scientific literature using only examples where they seem to work perfectly. In a survey of papers on new techniques, for example, we found that all – without exception – were claimed to perform better than existing competitors.[11]

Clearly, methodological results are affected by something akin to publication bias. However, discussing publication bias, which has attracted a lot of attention in medical and social sciences since the 1950s, seemed to be surprisingly taboo in methodological research until we tried to define the concept in this context.[4]

Our contention is that, as a result of publication bias and fishing expeditions, the scientific literature is rife with statistical methods that supposedly perform better than all other methods – but which are never compared to other methods except by their (potentially biased) inventors.

Comparisons

The replicability of methodological research findings has, to our knowledge, never been systematically investigated, which is somewhat unexpected given the many empirical studies devoted to replicability in other scientific fields in the last decade. It is not hard to imagine, however, that claims about the superiority of new methods over existing ones may be overly optimistic and not replicable. Such concerns could be put to rest if the statistical community were to conduct more neutral comparison studies, meaning studies that are not conducted with the aim of demonstrating the superiority of a particular (new) method, and are authored by researchers who are, on the whole, equally familiar with the various proposed methods.

The STRATOS initiative (stratos-initiative.org) is, we believe, a step in the right direction, aiming to provide guidance for the statistical analysis of observational medical studies. STRATOS emphasises the importance of comparison studies performed by groups of experts from different "statistical schools". There are also efforts such as OpenML (openml.org), which tries to tackle this issue in machine learning by opening up results of thousands of machine learning benchmarks to the public and allowing anyone to add their own results.

However, the pressure on researchers to publish in journals, and the reluctance of journals to accept the results of neutral comparison studies, remains a crucial obstacle. Contrast this with clinical research, where clinical trials are considered important pieces of scientific work, even if the treatment approach has been described elsewhere before. If statistical methods were treated like drugs, there would be a strong demand for neutral and well-planned comparison studies: patients would refuse a drug that has not been reliably proven to be better, so why are we using statistical methods based on the results of one (potentially biased) study?

What is the role of replication studies? We all agree that they are needed in applied research, but does this also hold true for methodological research? The goal of such studies would be to confirm the results of previous methodological papers, using, say, alternative simulation designs, other real data sets and a different implementation. Such formal replication studies are rare to non-existent in methodological research. Would they be deemed non-innovative and not

Anne-Laure Boulesteix is a professor of biometrics in the Institute for Medical Information Processing, Biometry and Epidemiology at LMU München, Germany.

Sabine Hoffmann is a postdoctoral researcher in the Institute for Medical Information Processing, Biometry and Epidemiology at LMU München, Germany.

Anne-Laure Boulesteix @BoulesteixLaure · Jun 1

I'm looking for examples of statistical methods that were 1) first deemed promising but not properly evaluated, 2) used in practice, 3) finally identified as flawed/misleading, thus questioning the results previously obtained with it. Any idea?

48 31 89

Figure 1: Co-author's tweet, asking for examples of widely used, but ultimately flawed, statistical methods.

worthy of publication by most renowned statistics journals?

The pitfalls of existing methods are often discovered accidentally and demonstrated in the scientific literature many years after the original publication. This may result in flawed methods becoming widely used and, in the worst cases, accumulating years of potentially misleading results. The numerous reactions from the statistical community to a tweet on this general issue (see Figure 1) suggest that this is perceived to be a huge problem, and one prominent example is the so-called magnitude-based

Alethea Charlton is a student assistant in the Institute for Medical Information Processing, Biometry and Epidemiology at LMU München, Germany.

Heidi Seibold is a postdoctoral researcher at LMU München, Bielefeld University and Helmholz Zentrum München, Germany.

inference method, which was widely used in sport statistics but eventually found to be flawed.[12]

Design

There is a clear need for more neutral comparisons and replications of methodological statistical research, but how should such studies be performed?

In many fields related to statistics, such as computational biology and bioinformatics, scientists can rely on a wide body of literature offering guidance on how to perform comparison studies. But, surprisingly, the design of comparison studies of statistical methods has hardly been addressed, even though the design of experiments is an intrinsically statistical issue.

Research on the appropriate design of benchmark studies, regardless of whether based on simulation[13] or on real data,[6] is still in its infancy. We can learn a lot from the world of clinical trials in this respect. For example, the calculation of the required number of real data sets and strategies to avoid bias (e.g., when handling missing values) are concepts that are relevant in both clinical and benchmarking studies. Thus far, however, the focus has been almost entirely on the world of clinical trials.

Another key concept, virtually unused in the field of methodological research, is meta-analysis. Well established in health and social sciences, meta-analyses systematically collate all existing research on a specific topic and are considered to be the highest level of evidence. A few first attempts to summarise existing methodological literature have emerged, including a formal meta-analysis of methods for the assessment of a certain type of software in computational biology/bioinformatics,[10] and a systematic review of the performance of machine learning versus logistic regression.[9] However, despite these important initial steps, quantitative or systematic reviews on the performance of methods described in the *methodological* statistical literature are extremely rare, and what is more, how they should be performed is unclear.

Lastly, we should not fall at the final hurdle: reporting, another important issue related to replicability. Appropriate reporting has been the subject of much conversation over the past decade, in fields ranging from randomised clinical trials to prediction models relying on artificial intelligence in health science. To date, however, no guidance is available for reporting of methodological statistical research. Our personal experience is that critical information is often missing in methodological research articles, such as the exact way in which the simulated data were generated. This incomplete reporting impedes understanding of the advantages, limitations and expected performance of the methods considered, not to mention potentially rendering studies impossible to reproduce. A recommended approach in this context is the publication of analysis codes, which allow readers to reproduce a study's analysis at the click of a mouse. This is already required by some journals (such as the *Biometrical Journal*) and will hopefully become more common in the coming years.

To sum up, we argue that statisticians can learn a great deal from clinical research (and other fields) about comparison studies, reporting and research synthesis, and that these valuable lessons should be applied to methodological statistical research, ultimately leading towards more evidence-based statistical practice. After all, improving the replicability of methodological research is an important step in improving research quality across all fields that apply statistical methods. ■

Acknowledgements

We thank the German Research Foundation (DFG; individual grants BO3139/4-3 and BO3139/7-1 to ALB) and the German Federal Ministry of Education and Research (BMBF; grant no. 01IS18036A) for funding, and the Twitter community who helped improve our paper with valuable comments and literature recommendations (see bit.ly/2EfmHyz and bit.ly/3jgu4EL).

References

1. Held, L. and Schwab, S. (2020) Improving the reproducibility of science. *Significance*, **17**(1), 10–11.

2. Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K. and Boulesteix, A. L. (2010) Over-optimism in bioinformatics: An illustration. *Bioinformatics*, **26**(16), 1990–1998.

3. Hutson, M. (2018) Artificial intelligence faces reproducibility crisis. *Science*, **359**, 725–726.

4. Boulesteix, A. L., Stierle, V. and Hapfelmeier, A. (2015) Publication bias in methodological computational research. *Cancer Informatics*, **14**, 11–19.

5. Boulesteix, A. L., Hable, R., Lauer, S. and Eugster, M. J. (2015) A statistical framework for hypothesis testing in real data comparison studies. *American Statistician*, **69**(3), 201–212.

6. Boulesteix, A. L., Wilson, R. and Hapfelmeier, A. (2017) Towards evidence-based computational statistics: Lessons from clinical research on the role and design of real-data benchmark studies. *BMC Medical Research Methodology*, **17**(1), 138.

7. Liu, K. and Meng, X. L. (2016) There is individualized treatment. Why not individualized inference? *Annual Review of Statistics and Its Applications*, **3**, 79–111.

8. Keogh, E. and Kasetty, S. (2003) On the need for time series data mining benchmarks: A survey and empirical demonstration. *Data Mining and Knowledge Discovery*, **7**(4), 349–371.

9. Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., van Calster, B., *et al.* (2019) A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, **110**, 12–22.

10. Gardner, P. P., Watson, R. J., Morgan, X. C., Draper, J. L., Finn, R. D., Morales, S. E. and Stott, M. B. (2019) Identifying accurate metagenome and amplicon software via a meta-analysis of sequence to taxonomy benchmarking studies. *PeerJ*, **7**, e6160.

11. Boulesteix, A. L., Lauer, S. and Eugster, M. J. (2013) A plea for neutral comparison studies in computational sciences. *PLoS ONE*, **8**(4), e61562.

12. Sainani, K. (2018) The problem with "magnitude-based inference". *Medicine & Science in Sports & Exercise*, **50**(1), 2166–2176.

13. Morris, T. P., White, I. R. and Crowther, M. J. (2019) Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**(11), 2074–2102.

There is a clear need for more neutral comparisons and replications of methodological statistical research, but how should such studies be performed? Surprisingly, the design of comparison studies of statistical methods has hardly been addressed

Matt Artz/Unsplash.com