Received: 15 August 2017 | Revised: 20 October 2017 | Accepted: 22 October 2017

DOI: 10.1002/bimj.201700129

LETTER TO THE EDITOR

Biometrical Journal

On the necessity and design of studies comparing statistical methods

In data analysis sciences in general and in biometrical research particularly, there are strong incentives for presenting work that entails new methods. Many journals require authors to propose new methods as a prerequisite for publication, as this is the most straightforward way to claim the necessary novelty. The development of new methods is also factually often a sine qua non condition to be recruited as a faculty member or to obtain personnel funding from a methods-oriented research agency, not least because it noticeably increases the chance to get published as outlined above. Thus, in statistical research and related methodology-oriented fields such as machine learning or bioinformatics, the well-known adage "publish or perish" could be translated into "propose new methods or perish."

Such a research paradigm is not favorable for studies that aim at meaningfully comparing alternative existing methods or, more generally, studies assessing the behavior and properties of existing methods. Yet, given the exponential increase in the number and complexity of new statistical methods being published every year, the end users are often at a loss regarding what are the "optimal" or even "appropriate" methods to answer the research question of interest given a particular data structure. It becomes more and more difficult to get an overview of existing methods, not to mention the overview of their respective performances in different settings (Sauerbrei, Abrahamowicz, Altman, Le Cessie, & Carpenter, 2014).

Moreover, it is well known that studies comparing a suggested new method to existing methods may be (strongly) biased in favor of the new method. This is a consequence of various factors starting with the authors' better expertise on the new method compared to the competing methods. Another factor is the combination of publication pressure (publish or perish) and *publication bias*—in the sense that a new method performing worse than existing ones has (severe) difficulties to get published (Boulesteix, Stierle, & Hapfelmeier, 2015). This may lead to simulation designs that might be—intentionally or unintentionally—biased. Note that not only empirical evaluations but also theoretical properties suggesting the superiority of a method under particular assumptions may be in principle potentially affected by this kind of bias. Deriving theoretical results for statistical approaches relevant in practice is extremely difficult and possible only under strong assumptions (Picard & Cook, 1984). We speculate that authors assessing the theoretical properties of their new method tend to make assumptions that are rather favorable for the new method—also a form of bias.

In contrast, *neutral* comparison studies, as defined by Boulesteix, Wilson, and Hapfelmeier (2017a), are dedicated to the comparison itself: they do not aim to demonstrate the superiority of a particular method and are thus not designed in a way that may increase the probability to observe incorrectly this superiority. Furthermore, they involve authors who are, as a collective, approximately equally competent on all considered methods. Neutral comparison studies can be thus considered as unbiased. Yet, in practice, such neutral comparison studies may be very time consuming and difficult to both organize and perform. The need to ensure "equal competence" on all methods being compared may exclude some more complex (but perhaps more suitable) approaches or require a close collaboration among many experts.

According to their official scope, most high-ranking statistical journals mainly focus on the development of new methods and on innovative applications, while comparison studies are not mentioned. These reasons, combined with the difficulties of conducting neutral comparison studies as outlined above, may explain the relative paucity of papers focusing on the comparison of existing methods. Most papers published in statistical journals suggest new methods (this term is used here; it includes "relevant" modifications of existing methods). Many of these new methods are not extensively compared with other methods by other researchers than their developers, except perhaps in a later paper, by the same or other authors, often aiming to demonstrate that the new approach is superior.

For many (if not most) data analysis problems, there is no lack of available methods and no need for new methods. In fact, the multiplicity of possible data analysis approaches is even an issue on its own as recently illustrated by Silberzahn and Uhlmann (2015). Whereas such "embarras du choix," related to the "multiplicity of perspectives"—including but not limited to model selection criteria—described by Gelman and Hennig (2017), is not bad per se and one should not attempt to eliminate it by formulating strict guidelines, it is not clear how one should deal with multiple approaches in practice. It is principally recommended to apply several analysis approaches to the data, but there is no consensus on how the multiple results should be reported. Moreover, the possibility of obtaining different results with different approaches raises concerns about "fishing for

significance" (Boulesteix, Hornung, & Sauerbrei, 2017b; Ioannidis, 2005; Simmons, Nelson, & Simonsohn, 2011). The fast-paced development of new methods and lack of neutral comparison studies tends to aggravate rather than solve these problems.

The problems described above are not new, potentially suggesting that the situation cannot easily be changed. Yet, research outside the field of computational sciences provides evidence that a different approach is feasible. Let us consider this situation in light of a keen analogy with clinical research: imagine that medical researchers spend most of/all their time developing new therapies that are not evaluated in clinical trials. Imagine that performing clinical trials for comparison is considered noninnovative research, not worth funding and not worth publishing in high-ranking journals. Imagine that nobody cares about the way clinical trials are performed, their design, their biases, their reliability, the interpretation of their results, etc. This clearly unacceptable situation would be somewhat similar to the (partial) lack of interest of the statistical research community in systematic (simulation-based) comparisons of existing statistical methods.

Moreover, it is not clear how these comparison studies should be performed and reported: more (meta)research is needed. While the design of unbiased clinical trials has been an active research topic for decades, there is no consensus on what makes a reliable comparison study in applied statistics. Which designs are most appropriate? What are typical sources of potential biases and how can they be avoided? How can the results be interpreted without the tendency for overinterpretation? Which mixture of simulated and real data should be used? How should real data be selected? How should simulated data be generated in a realistic way inspired from real datasets? Sometimes the complexity of the relevant data structure may require development and validation of new algorithms for data generation and/or comparison of the computational efficiency and accuracy of the alternative algorithms (Sylvestre & Abrahamowicz, 2008). What parameters and assumptions should be varied across the simulated scenarios? What range of sample sizes should be assessed? How can we assess the practical relevance of simulation results, which depends on the real-life plausibility of the simulation scenarios? How can an acceptable neutrality of the authors team be achieved and how can non-neutrality (the analogon of "conflicts of interest" in clinical research) be disclosed? Which "competing methods" should be considered? We need to recognize that there is no agreement among experts on the "state-of-the-art" methods for many topics relevant in practice. Consequently, in comparison studies, researchers have the freedom to decide the competing method(s) and the performance criteria used to compare them. For an example of discussions regarding the difficulties of making a decision about these issues, see a simulation study comparing the use of fractional polynomials and spline-based procedures in multivariable models with continuous variables (Binder, Sauerbrei, & Royston, 2013). In conclusion, beyond the daunting task of getting familiar with several different methods, researchers performing extensive comparison studies thus also face the daunting task of designing their study on their own, without much relevant literature to rely on.

As a consequence of the lack of comparison studies, end-users' decisions for or against application of particular methods are often consciously or subconsciously driven by arguments that are to some extent independent of the performance of the method, such as the charisma and marketing strategy of its developers, its use in similar previous studies, the method's fancy name that is easy to remember when heard at a conference, or the availability of user-friendly software. Since there is often not much "evidence" about relative advantages and disadvantages of the competing available methods, the specific choice cannot be "evidence based." In an ideal world, a group of "suitable experts" would carefully compare relevant methods and provide guidance about their strengths and weaknesses in various situations (usually performance and usability depends on several criteria) and the data analyst could choose the method based on the specific aims of the study and the knowledge, or assumptions, about the structure of the data at hand. Guidance for practice, however, is missing for many methodological issues and a data analyst facing a practical problem has neither the required time nor the required expertise to conduct meaningful comparison studies, especially as many papers presenting new methods are not reproducible. Fortunately, the importance to make research more reproducible has been recognized in the last decade and improvements have started. Important steps include trial registration, data sharing, reporting guidelines, and a few journals such as the *Biometrical Journal* explicitly encouraging authors to submit together with their paper the data and analysis codes for the purpose of reproducibility (Hofner, Schmid, & Edler, 2016). Note that some users do have the time and the expertise to conduct such comparison studies, but their results more often remain confidential because they do not have enough time to publish them in a meaningful form (i.e., with enough details and enough investigations in various settings) and/or (as mentioned above) because of the lack of publication channels for such studies.

In this context, to improve comparison studies of statistical methods and their reproducibility we consider it desirable to (i) reinforce the status of neutral comparison studies and studies evaluating the behaviors of existing methods in the scientific community with the aim to create incentives to perform such studies; (ii) develop research activities dedicated to what we could call "comparology," that is, research on how to reliably assess statistical methods—in analogy to the active research field devoted to clinical trial methodology; and (iii) derive reporting guidelines to increase transparency of comparison studies, similar to existing guidelines for many types of studies in the health sciences (Simera et al., 2010). We believe that statistical journals should contribute to implementing changes in this direction by encouraging (high quality) submissions reporting comparison

studies or addressing their methodology in a broad sense. The recently launched STRATOS Initiative[1] (Sauerbrei et al., 2014) aims at addressing related issues in the field of statistical methods for observational studies in the health sciences. The initiative realized quickly that a special "Simulation Panel," dedicated to the methodology of comparison studies, is needed.

Anne-Laure Boulesteix[1]
Harald Binder[2]
Michal Abrahamowicz[3]
Willi Sauerbrei[2]
for the Simulation Panel of the STRATOS Initiative

[1]*Institute for Medical Information Processing, Biometry and Epidemiology (IBE), LMU Munich, Marchioninistr. 15, 81377 Munich, Germany*
[2]*Institute for Medical Biometry and Statistics, Faculty of Medicine and Medical Center, University of Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany*
[3]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, 1020 Pine Avenue West, Montreal, QC H3A 1A2, Canada*

Correspondence
Anne-Laure Boulesteix, Institute for Medical Information Processing, Biometry and Epidemiology (IBE), LMU Munich, Marchioninistr. 15, 81377 Munich, Germany.
Email: boulesteix@ibe.med.uni-muenchen.de

REFERENCES

Binder, H., Sauerbrei, W., & Royston, P. (2013). Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: A simulation study with continuous response. *Statistics in Medicine*, *32*, 2262–2277.

Boulesteix, A.-L., Hornung, R., & Sauerbrei, W. (2017b). On fishing for significance and statisticians degree of freedom in the era of big molecular data. In M. Ott, W. Pietsch, & J. Wernecke (Eds.), *Berechenbarkeit der Welt?* (pp. 155–170). Wiesbaden: Springer.

Boulesteix, A.-L., Stierle, V., & Hapfelmeier, A. (2015). Publication bias in methodological computational research. *Cancer Informatics*, *14*(Suppl 5), 11–19.

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

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society A*, *180*, 967–1033.

Hofner, B., Schmid, M., & Edler, L. (2016). Reproducible research in statistics: A review and guidelines for the *Biometrical Journal*. *Biometrical Journal 58*, 416–427.

Ioannidis, J. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Picard, R. R., & Cook, R. D. (1984). Cross-validation of regression models. *Journal of the American Statistical Association*, *79*(387), 575–583.

Sauerbrei, W., Abrahamowicz, M., Altman, D. G., Le Cessie, S., & Carpenter, J. (2014). STRengthening Analytical Thinking for Observational Studies: The STRATOS initiative. *Statistics in Medicine*, *33*(30), 5413–5432.

Silberzahn, R., & Uhlmann, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, *526*(7572), 189–191.

Simera, I., Moher, D., Hirst, A., Hoey, J., Schulz, K. F., & Altman, D. G. (2010). Transparent and accurate reporting increases reliability, utility, and impact of your research: Reporting guidelines and the EQUATOR network. *BMC Medicine*, *8*, 24.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Sylvestre, M. P., & Abrahamowicz, M. (2008). Comparison of algorithms to generate event times conditional on time-dependent covariates. *Statistics in Medicine*, *27*(14), 2618–2634.

[1] The international STRengthening Analytical Thinking for Observational Studies (STRATOS) Initiative (http://stratos-initiative.org) aims to provide accessible and accurate guidance documents for relevant topics in the design and analysis of observational studies.