Against the "one method fits all data sets" philosophy for comparison studies in methodological research

Strobl, Carolin ; Leisch, Friedrich

DOI: https://doi.org/10.1002/bimj.202200104

Received: 1 April 2022 | Revised: 12 August 2022 | Accepted: 20 August 2022

DOI: 10.1002/bimj.202200104

Biometrical Journal

DISCUSSION

Against the "one method fits all data sets" philosophy for comparison studies in methodological research

Carolin Strobl[1] | **Friedrich Leisch**[2]

[1]Department of Psychology, University of Zurich, Zurich, Switzerland

[2]Institute of Statistics, University of Natural Resources and Life Sciences, Vienna, Austria

Correspondence
Prof. Dr. Carolin Strobl, Chair for Psychological Methods, Department of Psychology, University of Zurich, Binzmühlestr. 14/Box 27, Zurich 8050, Switzerland.
Email: carolin.strobl@uzh.ch

Abstract

Many methodological comparison studies aim at identifying a single or a few "best performing" methods over a certain range of data sets. In this paper we take a different viewpoint by asking whether the research question of identifying the best performing method is what we should be striving for in the first place. We will argue that this research question implies assumptions which we do not consider warranted in methodological research, that a different research question would be more informative, and how this research question can be fruitfully investigated.

KEYWORDS
comparison studies, metalearning, simulation studies

1 | AIM

Substantive researchers today are faced with a multitude of analysis methods readily available in software, and need to make a decision about which method they should use for analyzing their data to answer their current research question. In order to facilitate this decision, this special issue calls for neutral methods comparison studies. While one might—and we will—argue below that an entirely neutral comparison study is impossible, in the call for papers for this special issue, neutral comparison studies are defined as being (i) focused on the comparison of existing methods already described elsewhere rather than on a new prototype method being introduced; (ii) authored by a group of researchers who are (ideally) approximately equally familiar with all the compared methods.

In this sense, the main focus in the discussion of neutral comparison studies seems to be that statistical or machine learning methods, which are considered as competitors in the "race" for best performance on a sample of data sets, should be treated fairly. As a means to this end, Boulesteix et al. (2013) as well as the call for papers for this special issue have emphasized how important it is that the authors of the comparison study have equal expertise for each of the investigated methods. Otherwise one method might "win the race" because it has been tuned more skillfully than the others.

Many comparison studies aim at identifying a single or a few "best performing" methods over a certain range of data sets. Performance here typically refers to prediction accuracy, which is often sensibly assessed on a held-back test sample or by means of cross validation (cf., e.g., the studies investigated by Boulesteix et al., 2013, or the studies of Fernández-Delgado et al., 2014; Olson et al., 2017; Wu et al., 2017; or Palotti et al., 2019; Cosenza et al., 2020; or Zöller and Huber, 2021, to name just a few examples). Other studies provide collections of benchmark data sets for very specific application domains like Wu et al. (2018). For such specific data collections, the search for methods that, for example, minimize the

maximum possible loss or prove most robust against departures from assumptions on average can make a lot of sense, see, for example, Royston and Parmar (2020) as an example for the latter.

As a scientific community, we should strongly support more neutral comparison studies. Unfortunately, they are sometimes hard to publish. Moreover, because they are easier to design, good comparison studies may be easier to achieve for rather specific tasks rather than very general tasks. Methodological metastudies could then be used to generalize from specific tasks to an overall performance, but these would probably be even harder to publish.

As a coauthor of the paper "The support vector machine under test" (Meyer et al., 2003), for example, one of the authors of this discussion (FL) can state that this was indeed a neutral comparison study of a wide range of regression and classification methods, although the title suggests a different story. We did not invent a new support vector machine (SVM), we integrated an existing SVM library into R (R Core Team, 2022) because we needed it for a different project. The only reason to choose a nonneutral title for the paper was that SVMs were the "hot and new" method at the time and we thought that with this title, the paper would be easier to publish and get some citations (which turned out to be true). Otherwise the paper suffers from the same problem as all general comparisons: the data sets used are popular data sets from the UCI repository (Dua & Graff, 2019). For which tasks these are representative is hard to say.

In this paper, we take a different viewpoint by asking whether the research question of identifying the best performing method is what we should be striving for in the first place. We will argue that this research question itself implies assumptions which we do not consider warranted in methodological research. We will critically discuss these assumptions, argue that a different research question would be more informative, and consider how this research question can be fruitfully investigated.

2 | BACKGROUND

In methodological research, comparison studies are often published either with the explicit or implicit aim to promote a new method by means of showing that it outperforms existing methods. In this case, typically, the authors of the study are the authors of the new method, or are at least "fans" of the new method—and Boulesteix et al. (2013) plausibly assume that they are not treating all methods equally. This might come into play at the stage of defining the quality criterion for the study, the stage of tuning model parameters, but also at the stage of interpreting the results. As Boulesteix et al. (2013) summarize: "Given the same quantitative outputs, the impression of the reader can be affected by the choice of the vocabulary in the results section, by graphical representation, or by the choice of the main quantitative criterion used to compare the methods." Yousefi et al. (2009) also show that methods comparisons can be biased by reporting only favorable results, that is, by cherry-picked data sets.

In experimental psychology or medicine, this kind of effect would be considered as experimenter bias. One general cure for experimenter effects is blinding. To some degree, the idea of blinding could be incorporated in methodological research. For example, at the stage of interpreting the study results, the researcher could be kept blind with respect to the labeling of which results were produced by which method. Only after the description of the results has been finalized would the labels be added. While this would avoid overoptimistic reporting, in practice, it would be difficult to prohibit any modifications on the text after unblinding. At other stages of methodological research, blinding is even more difficult or impossible to achieve, in particular for the model tuning stage, where it is unavoidable that the researchers conducting the analyses know which method or algorithm they are using.

Hence, comparison studies in papers presenting a new method should simply be taken for what they really are: proof that the authors were able to find at least one or more data sets (or simulation designs) where the new method outperforms older methods. Otherwise they would most likely not have been able to publish the paper. To some extent, the quality of the comparison study also depends on the review process, that is, which data sets and methods reviewers requested for a revised version of the manuscript (which is, however, unknown to the reader).

An existing setup that would, in principle, ensure that "equal opportunities" for all methods are machine learning challenges, where a training data set is provided and several individual or teams of researchers submit their fitted models, which are then evaluated typically on a held-back test data set. This setup gives every individual or team the chance to equip their favorite method as best as they can. At the same time, we would argue that the comparison of methods on one or more real data sets with the aim of identifying a clear winner is an ill-posed question in the first place—and believe that the fact that, among a group of top performing methods, the declared winner varies from competition to competition or from study to study supports our view.

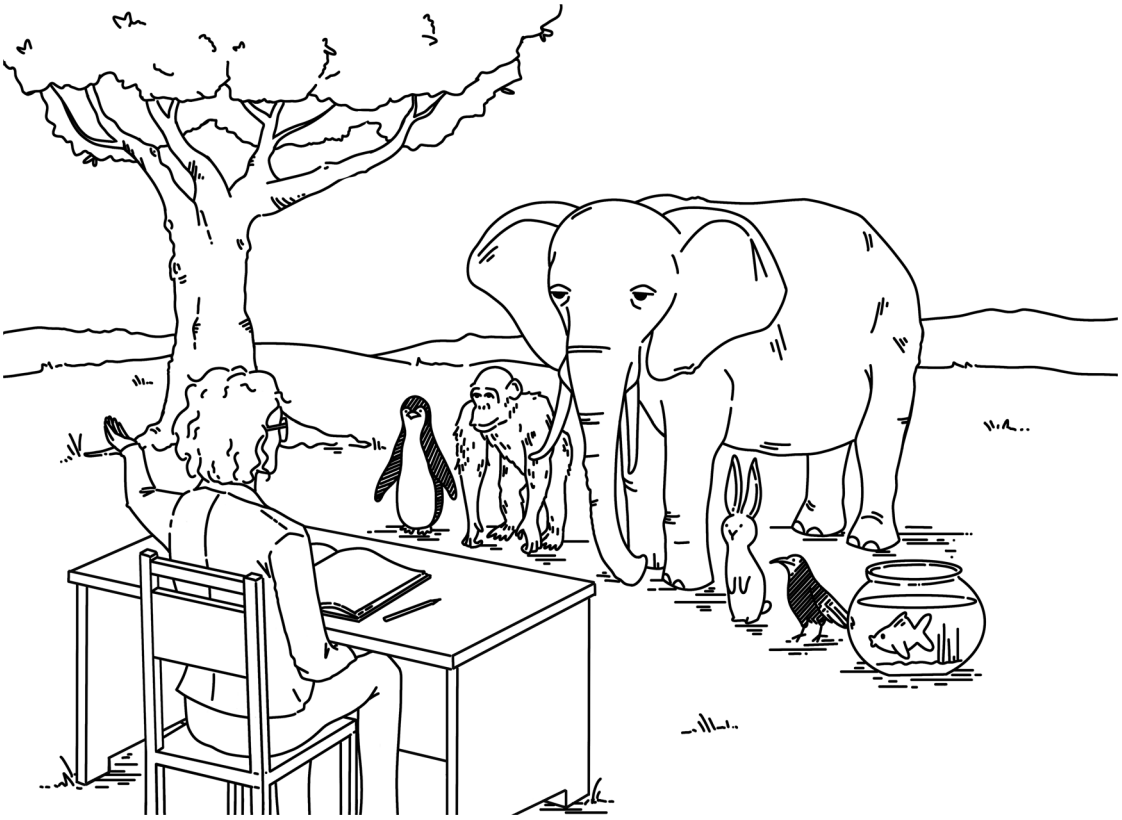

FIGURE 1 "Climb the tree". Drawing by Alexandra Kalberer

3 | UNDERLYING ASSUMPTIONS

In the comparison of methods, we are actually facing an encounter of one or more methods, on the one hand, and one or more data sets, be they real or simulated, on the other hand. The method that performs best with respect to some outcome measure (such as prediction accuracy or power) after averaging (or using another form of aggregation) over the data sets is considered the winner. This means that explicitly or implicitly, it is assumed that declaring an overall winner method was a sensible research question.

We can use a sports tournament as a metaphor for this setup, where the methods are the contestants and the data sets are the different sports disciplines on which we measure the contestants' performance. Maybe you have previously seen one of those comic drawings caricaturing a "fair" tournament, where an ape, a fish, and a few other animals are all given the same task: "Climb the tree" (cf. Figure 1). Obviously, this task is easier for the ape than for the fish. Another task, like: "Swim to the other side of the pool," would show quite a different performance ranking of the contestants.

From this metaphor, it becomes obvious that who will be declared the winner of the contest depends upon several aspects, namely (a) the capabilities of the contestants on each task, (b) the particular tasks chosen for the contest, and (c) how the results for the different tasks are aggregated (e.g., with a weighted or unweighted average) to form the final result.

While aspect (a) does contain the capabilities of the contestants (in our case the performance of the methods), that is, what most readers would probably agree is the core of what we would like to find out—from the metaphor, it becomes obvious that performance on a single task is a function of contestant (method) × task (data set) properties, rather than a function of a contestant's (method's) properties alone.

From the metaphor, it is also easy to imagine, now focusing on aspect (b), that the competition between ape and fish will lead to quite different results if the tasks selected for the tournament are climbing a tree and climbing a rope versus swimming through a pool and swimming upriver. In the same way, the result of a method comparison study will depend on the tasks (data sets) which the methods are supposed to tackle (cf., e.g., Choi & Lee, 2017; Kwon & Sim, 2013; Macià et al., 2013; Oreski et al., 2016).

Which tasks to select also strongly depends on which properties of the methods one is interested in. Breiman (2001) described the fundamentally different modeling cultures of statistics and machine learning. Two decades ago statisticians

were mostly interested in model assumptions, such that mathematical proofs of certain model characteristics like consistency of parameter estimation can be derived. In machine learning, on the other hand, predictive performance was the main goal to achieve, mostly "proven" by comparison studies. Since then a lot has changed: statisticians now happily use methods developed by machine learners to fulfill statistical goals and vice versa. If the goal of fitting a model is to understand the relationships between predictor variables and response, it does not help to use a black box model even if it wins many competitions.

Now one might argue that the selection of the data sets could be done in a more fair way, for example, by means of drawing a random sample of data sets. Boulesteix et al. (2013) argue along these lines: "Considering the high variability of relative performance of methods across data sets and the moderate number of data sets considered in each study [...], a comparison study based on different data sets may obviously yield substantially different results. [...] Therefore, it is important to make a selection of data sets that is 'as representative as possible' to cover the domain of interest. At best, the data sets are chosen from a set of data sets representing the domain of interest using standard sampling methodology."

From this statement, it becomes evident that the underlying view is that there was a population of data sets for each domain of interest, and that we could draw data sets as samples from this domain with standard sampling schemes, such as fully random or stratified sampling. However, in order to be actually able to draw a random sample, we would first have to define from what population exactly the data sets should be drawn. What could this population of data sets for the domain of interest be? All data sets in the world? All data sets within the discipline of biology or psychology? All data sets with certain characteristics, such as a binary response variable, in the discipline of biology or psychology? How could we possibly define this population, when, to name just a few examples, the boundaries between disciplines are overlapping (where do we place neuroscience in relation to biology and psychology?), the definition of the breadth of a domain is tedious (should we include only studies measuring depression by means of a certain standardized scale, or only based on psychiatric diagnosis, or both), no data bases of all data sets within one domain exist (even in those disciplines that have picked up quickly on open science standards, such as psychology), and new data are being collected right at this moment?

As statisticians, we should be aware that without being able to even define the population, how could we possibly be able to draw a representative sample of datasets? Or if we look at a convenience sample, such as those data sets available in some online repository, how could we possibly judge whether they are a representative sample from our unknown data set population? For the preopen-science era, would not it even be more realistic to assume that those data sets that were donated for repositories, at a time when this was the exception rather than the rule, might be those that were particularly boring, difficult, or otherwise nonrandomly selected?

If our only aim was that the comparison study is fair and not biased in favor of a certain method, one could argue that using data sets from a randomly chosen repository was "random enough" in an epistemic sense, since at least the authors of the comparison study would not know beforehand which method will perform particularly good on this particular sample of data sets. But as soon as comparison studies explicitly or implicitly create the impression that they could answer the question which method was the best in general, fairness is not enough. For answering that question, the result, that is, which method wins, would have to be consistent across different samples from our population of data sets and hence generalizable to our population of data sets—which, as we have argued above, is not the case because results based on different data set samples are not consistent and the population of data sets from which we could randomly sample to tackle consistency and generalizability is undefined.

4 | ARE WE ASKING THE WRONG QUESTION?

So possibly, we should admit that the research question "What is the best method in general" is ill-posed and proceed to a research question that might at first sight seem less ambitious, but could be handled in a more founded way, and thus, prove very useful. This question could be "Why is it that in study A method X won, but in study B it was method Y?" Meaning: "What are the properties of the real or simulated data used in each study that are associated with the different performance of the methods?"

For example, the question "What works better, a regression tree or linear regression?," can only be sensibly answered by a clear "That depends." If the data generating process (DGP) is linear (by simulation design or real but close to linear), a linear model will always outperform a decision tree. On a jump function DGP, however, the tree will outperform the linear model (note that any resemblance to the above metaphor on ape and fish is fully intended). Does it make sense to compare the two on a linear DGP, find that the linear model wins, and conclude that the linear model will outperform

the tree on all data sets yet to come? Obviously not, because it is the combination of method × DGP that drives prediction performance (and as a side remark: also prediction stability, as shown by Philipp et al., 2018).

Beside linearity (or linear separability in classification problems), other relevant data set characteristics may include the number and type of potential predictor variables, the amount of noise variables among them, the degree and pattern of correlations between predictors, the number of observations, and so on. In meta-analytic terms, one could think about the data set properties as metavariables, which we could and should incorporate like in metaregression. Just like we might find—and this could be very informative—that substance X works better than Y only in studies where they were applied orally (as opposed to, e.g., intravenously), we can think about the data set characteristics as metavariables that can provide us with additional information to motivate or empirically validate hypothesis about the type of data sets that a method is particularly suited for.

In machine learning, the field of "metalearning" (or more specifically "algorithm recommendation") pursues exactly this aim: to predict which algorithm will perform best on a particular data set based on its characteristics (Kalousis & Hilario, 2000; Lemke et al., 2015). In order to be able to make such predictions, the properties of several data sets and the performance of several candidate methods on these data sets are recorded together in a database (Alcobaça et al., 2020). For comparing statistical or psychometric methods by means of simulation studies, Skrondal (2000) has also coined the term "metamodel" to describe the effects of factors from the simulation design on method performance.[1]

Several studies, including Kwon and Sim (2013), Eugster et al. (2014), Oreski et al. (2016), and Choi and Lee (2017), have illustrated that the predictive performance of statistical and machine learning methods depends on data set characteristics. Like in Eugster et al. (2014), we argue that it can be very informative to investigate, rather than ignore, these dependencies. While some of the data set characteristics, such as the number of potential predictor variables, is visible "from the outside" of an empirical data set, others, like the true form of the association between predictors and response, are not. The latter type of data set properties can only be known and systematically varied in simulation studies, as we will now further discuss.

5 | SIMULATION VERSUS REAL DATA

Once we acknowledge that it is not the properties of the methods alone, but the combination of properties of the methods and properties of the data sets that determine performance, it is straightforward that a good study design, from which we can actually learn in which situations which methods perform particularly well, is one that systematically varies the data set properties.

The most obvious way to do this is through simulation studies, where it is possible to systematically vary and cross different aspects of the DGP as experimental factors. One major advantage of this approach is that, in addition to those data set characteristics that are visible "from the outside," in a simulation study, we also have full knowledge of all other data set properties, including the true functional form of the association between predictors and response as well as any associations between predictors (which may also be hard to identify "from the outside" if they are, e.g., nonlinear).

Ideally, and with simulation studies this is possible, we should investigate the effects of data set characteristics on method performance in a hypothesis driven way. Like in the empirical sciences, where, for example, biological knowledge can help us hypothesize which treatment will work particularly well for a certain disease, we should explicitly hypothesize and explain our a priori expectations based on mathematical or computational properties of the methods. Just like we consider it good scientific practice for empirical research, we should discuss the results of methods comparisons in the light of these hypotheses, and also clearly distinguish between findings that agree with or contradict our a priori hypotheses and any additional, exploratory findings. This can help the readers distinguish between results which the simulation study was particularly well designed to discover and results that should be further investigated in additional studies.

With respect to our argumentation above, simulation studies have the conceptual advantage that the population from which the sample of data sets used in the comparison study is drawn is clearly defined. This is not the case when a single or a sample of real data sets is used. As a consequence, it can happen that the method that performed best on certain data sets does not perform equally well on other data sets, because the method is well suited or able to adapt well to the specific properties of the data sets used in the comparison study, but not to the properties of other data sets. In this sense, the

[1] Necessitated by the computational restrictions at that time, Skrondal (2000) also argues that interactions between factors in simulation experiments could safely be ignored, which we do not agree with. Whenever possible, a full factorial design with more than two levels per factor should be used in simulation studies in order to capture any unforeseen nonmonotone and interaction effects.

results of comparison studies based on real data sets do not necessarily generalize well to other data sets. This is related to the replication crisis in the substantive sciences, where, aggravated by p-hacking and similar practices, reported findings may rely too strongly on random patterns in a particular sample, so that they do not generalize to new samples (see also Hullman et al., 2022). When we apply methods to synthetic data sets that were systematically created in a simulation study, however, the DGP—that is, the population from which we draw the samples—is entirely known. With a sufficient number of replications, we can easily rule out that the results would have come out differently had we drawn yet another sample (i.e., simulated yet another individual data set) from the population (i.e., from the DGP). Still, we should be aware that—just like in real data studies—the results are potentially only valid for the scenarios considered in the particular study. Unlike a mathematical proof, a simulation experiment cannot make universally valid claims. A clearly motivated simulation design can, however, illustrate such fundamental properties of methods that a qualified reader will be able to accept the generality of certain results, for example, that a linear DGP is easier to grasp for a linear model than for a tree.

Of course, there are many other things one can criticize about simulation studies, in particular, that they can be too simplistic and unrealistic. This concern is particularly valid when a simulation study is conducted under "textbook conditions," such as all distributional assumptions being met, no missing values, and so on. To address this concern, however, the DGP can easily be extended to less idealistic scenarios, such as different and nonstandard shapes of distributions, different amounts and patterns of missing values, and so on. Another approach to make simulated data more realistic, which is recommended, for example, by Burton et al. (2006), is to use a real data set as the motivating example and simulate data such that they "closely represent the structure of this real data set." In particular, they suggest that the predictor variables could be taken from the real data, which preserves all their properties, including their correlation structure, and only the response variable could be generated based on the desired functional form.

If you believe that simulation can never adequately capture the complexities of real data, the next candidate approach would be to start with real data, but systematically alter them into what could be termed "systematically modified data" in order to be able to vary certain data set properties. This is not possible for all data set characteristics, but for some can be done quite easily, for example, by adding extra noise variables or replacing observed values by missing values. Boulesteix et al. (2020), for example, systematically add measurement error to the predictor variable of an empirical data set to illustrate how the ordinary least squares (OLS) regression coefficient estimate is affected by this. While this approach may be more realistic, it also goes along with a loss of information about and control over the DGP.

Another candidate approach would be to use a convenience sample of real data sets, with the caveats already outlined above, including the fact that "what these evaluations tell us about the methods' accuracy is relevant to the considered specific real data example(s), but may not be relevant to other settings" (Boulesteix et al., 2020). Therefore, we argue that—especially, but not only—when real data are used in methods comparison studies, data set characteristics should also be recorded and investigated in the spirit of metalearning, rather than making overgeneralizing claims in the currently still prevalent "one method fits all data sets" philosophy.

Again, using substantive empirical research as a metaphor, the three approaches discussed above—systematic simulation, systematic modification of real data sets, and unmodified use of real data sets—correspond to the different approaches available for empirical research, together with their known pros and cons: lab experiments (highly controlled, warrant causal interpretations of results, but may be far from realistic), field experiments (take what is there, which is not exactly known but realistic, and vary systematically what you can), and observational studies (take what is there and report what you can, do not warrant causal claims without strong additional assumptions).

The strongest case for a new method could probably be made by means of a threefold comparison based on theoretical properties, empirical performance in simulation studies, and empirical performance on real data sets. While the performance on real data will be important to convince prospective users of the practicality of the method, theoretical arguments and systematic simulation studies can provide a deeper understanding of its behavior.

6 | OTHER ASPECTS OF GOOD METHODOLOGICAL RESEARCH PRACTICE

While we argue that methodological comparison studies should be enriched with considerations of data set properties in the spirit of metalearning, we agree with many other demands that have been voiced to promote good research practices in our field. For example, we fully agree that precautions against overoptimistic reporting of methods comparisons must be kept up (see, e.g., Boulesteix et al., 2013; Jelizarow et al., 2010; Yousefi et al., 2009). We also see the previously described parallels between empirical and methodological research with respect to the pressure to find significant gains of a new treatment or method, and the risk of negative research findings ending up unpublished in a file drawer (Boulesteix et al.,

2013). Just like the empirical sciences, we should therefore discuss both technical and systemic means and incentives to counter such issues (see, e.g., Leising et al., 2021).

We believe that shifting the focus of methods comparisons away from finding an overall winner toward more differentiated results incorporating data set characteristics can add to this process in two ways.

First, we believe that it would paint a more realistic picture of what our methods have to offer for substantive researchers—our "customers"—when we provide them with more information about which methods they can expect to perform well in their particular kind of data situation. With this information, they can make an informed choice on the method that, based on methodological considerations and empirical results, is expected to perform best in their empirical study and include this method, for example, as part of their preregistered analysis plan—rather than running a whole collection of methods that have performed well in one or the other comparison study, and then being at risk for cherry picking when having to decide which results to present in the final substantive publication.

Second, for us as methodologists to accept that there is no universally best method, but a variety of situations where one or the other method is particularly suited could take a lot of pressure out of the publication process for presenting and comparing methods. If being the overall winner is no longer the aim, it is no longer necessary to tweak the tuning parameters, data set choices, and so on. Stressing our favorite metaphor one last time, this would mean that the ape could get the price for best climber, the fish could get the price for best swimmer—and all creatures could live happily ever after…

ACKNOWLEDGMENT

The authors would like to thank two anonymous reviewers and the associate editor for suggesting valuable additions and clarifications to our first draft, as well as Alexandra Kalberer for creating the beautiful drawing for Figure 1.

Open access funding enabled and organized by Projekt DEAL.

ORCID

Carolin Strobl https://orcid.org/0000-0003-0952-3230

REFERENCES

Alcobaça, E., Siqueira, F., Rivolli, A., Garcia, L. P. F., Oliva, J. T., & de Carvalho, A. C. P. L. F. (2020). MFE: Towards reproducible meta-feature extraction. *Journal of Machine Learning Research*, *21*(111), 1–5.

Boulesteix, A.-L., Groenwold, R. H., Abrahamowicz, M., Binder, H., Briel, M., Hornung, R., Morris, T. P., Rahnenführer, J., & Sauerbrei, W. (2020). Introduction to statistical simulations in health research. *BMJ Open*, *10*(12), e039921.

Boulesteix, A.-L., Lauer, S., & Eugster, M. (2013). A plea for neutral comparison studies in computational sciences. *PLoS One*, *8*(4), e61562.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *16*(3), 199–231.

Burton, A., Altman, D. G., Royston, P., & Holder, R. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*, 4279–4292.

Choi, Y., & Lee, H. (2017). Data properties and the performance of sentiment classification for electronic commerce applications. *Information Systems Frontiers*, *19*, 993–1012.

Cosenza, D. N., Korhonen, L., Maltamo, M., Packalen, P., Strunk, J. L., Næsset, E., Gobakken, T., Soares, P., & Tomé, M. (2020). Comparison of linear regression, k-nearest neighbour and random forest methods in airborne laser-scanning-based prediction of growing stock. *Forestry: An International Journal of Forest Research*, *94*(2), 311–323.

Dua, D., & Graff, C. (2019). *UCI machine learning repository*. University of California, Irvine, School of Information and Computer Sciences. http://archive.ics.uci.edu/ml

Eugster, M., Leisch, F., & Strobl, C. (2014). (Psycho-)analysis of benchmark experiments – A formal framework for investigating the relationship between data sets and learning algorithms. *Computational Statistics & Data Analysis*, *71*, 986–1000.

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*(90), 3133–3181.

Hullman, J. R., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning. *ArXiv*, abs/2203.06498.

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

Kalousis, A., & Hilario, M. (2000). Model selection via meta-learning: A comparative study. In *Proceedings 12th IEEE Internationals Conference on Tools with Artificial Intelligence ICTAI 2000*, pp. 406–413.

Kwon, O., & Sim, J. M. (2013). Effects of data set features on the performances of classification algorithms. *Expert Systems with Applications*, *40*(5), 1847–1857.

Leising, D., Thielmann, I., Glöckner, A., Gärtner, A., & Schönbrodt, F. D. (2021). Ten steps toward a better personality science – How quality may be rewarded more in research evaluation. *Personality Science*, *3*, e6029.

Lemke, C., Budka, M., & Gabrys, B. (2015). Metalearning: A survey of trends and technologies. *Artificial Intelligence Review*, *44*(1), 117–130.

Macià, N., Bernadó-Mansilla, E., Orriols-Puig, A., & Kam Ho, T. (2013). Learner excellence biased by data set selection: A case for data characterisation and artificial data sets. *Pattern Recognition*, *46*(3), 1054–1066.

Meyer, D., Leisch, F., & Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, *55*, 169–186.

Olson, R. S., Cava, W. L., Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2017). PMLB: A large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, *10*(36), 1–13.

Oreski, D., Oreski, S., & Klicek, B. (2016). Effects of dataset characteristics on the performance of feature selection techniques. *Applied Soft Computing*, *52*, 109–119.

Palotti, J., Mall, R., Aupetit, M., Rueschman, M., Singh, M., Sathyanarayana, A., Taheri, S., & Fernandez-Luque, L. (2019). Benchmark on a large cohort for sleep-wake classification with machine learning techniques. *npj Digital Medicine*, *2*(50), 1–9.

Philipp, M., Rusch, T., Hornik, K., & Strobl, C. (2018). Measuring the stability of results from supervised statistical learning. *Journal of Computational and Graphical Statistics*, *27*(4), 685–700.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Royston, P., & Parmar, M. (2020). A simulation study comparing the power of nine tests of the treatment effect in randomized controlled trials with a time-to-event outcome. *Trials*, *2*1(315).

Skrondal, A. (2000). Design and analysis of monte carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*(2), 137–167. PMID: 26754081.

Wu, D., Jennings, C., Terpenny, J., Gao, R. X., & Kumara, S. (2017). A comparative study on machine learning algorithms for smart manufacturing: Tool wear prediction using random forests. *Journal of Manufacturing Science and Engineering*, *139*(7), 071018.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., & Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning. *Chemical Science*, *9*, 513–530.

Yousefi, M. R., Hua, J., Sima, C., & Dougherty, E. R. (2009). Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, *26*(1), 68–76.

Zöller, M.-A., & Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research*, *70*, 409–472.

How to cite this article: Strobl, C., & Leisch, F. (2022). Against the "one method fits all data sets" philosophy for comparison studies in methodological research. *Biometrical Journal*, 1–8. https://doi.org/10.1002/bimj.202200104