Biometrical Journal

On the role of benchmarking data sets and simulations in method comparison studies

Sarah Friedrich[1,2] 🔵 | Tim Friede[3,4] 🔵

[1]Institute of Mathematics, University of Augsburg, Augsburg, Germany

[2]Centre for Advanced Analytics and Predictive Sciences, University of Augsburg, Augsburg, Germany

[3]Department of Medical Statistics, University Medical Center Göttingen, Humboldtallee, Göttingen, Germany

[4]DZHK (German Centre for Cardiovascular Research), Partner Site Göttingen, Göttingen, Germany

Correspondence
Sarah Friedrich, Institute of Mathematics, University of Augsburg, Universitätsstr. 14, Augsburg, Germany.
Email: sarah.friedrich@math.uni-augsburg.de

Abstract

Method comparisons are essential to provide recommendations and guidance for applied researchers, who often have to choose from a plethora of available approaches. While many comparisons exist in the literature, these are often not neutral but favor a novel method. Apart from the choice of design and a proper reporting of the findings, there are different approaches concerning the underlying data for such method comparison studies. Most manuscripts on statistical methodology rely on simulation studies and provide a single real-world data set as an example to motivate and illustrate the methodology investigated. In the context of supervised learning, in contrast, methods are often evaluated using so-called benchmarking data sets, that is, real-world data that serve as gold standard in the community. Simulation studies, on the other hand, are much less common in this context. The aim of this paper is to investigate differences and similarities between these approaches, to discuss their advantages and disadvantages, and ultimately to develop new approaches to the evaluation of methods picking the best of both worlds. To this aim, we borrow ideas from different contexts such as mixed methods research and Clinical Scenario Evaluation.

KEYWORDS
benchmarking, machine learning, neutral comparison studies, simulation studies

1 | INTRODUCTION

The process of examining a research question empirically consists of several steps ranging from study design and data analysis to the interpretation of the results (Friedrich et al., 2021a). Each of these steps involves decisions to be made: Which trial design is adequate for answering the research question? Which analysis methods are available for the kind of data collected and what has to be taken into account when interpreting the results? The steps of this process have also been discussed in the context of drug development using the so-called Clinical Scenario Evaluation (CSE) (Benda et al., 2010; Dmitrienko & Pulkstenis, 2017; Friede et al., 2010). The CSE framework consists of three core elements: options, assumptions, and metrics. The different options for each step are compared using the respective metrics and taking the underlying assumptions into account. Simulation studies can be used in different stages of this process: to determine an

adequate design including, for example, sample size planning, to inform a subsequent trial of "ideal" parameter settings or expected outcomes (in silico clinical trials) as well as to compare different methods for the statistical analysis, see also Morris et al. (2019) for an overview. A relevant aspect in this context is to distinguish between models and methods. As Morris et al. (2019) put it: "The term 'method' is generic. Most often it refers to a model for analysis, but might refer to a design or some procedure (such as a decision rule)." In this sense, the method comprises questions such as "How to fit a model?" and "How to draw inference?" It is important to keep this in mind when considering the comparison of different methods.

In order to choose the "best" approach for a specific design or data analysis, fair comparisons between existing methods are essential. One can argue that a comparison study will never be completely neutral or fair in practice. In this paper, we therefore adopt the definition of "neutral comparisons" given by Boulesteix et al. (2013), namely, that the focus of the article should be on the comparison itself instead of introducing a novel method, that the authors should be reasonably neutral, and that the study should be designed and evaluated in a rational way. See also Strobl and Leisch (2022) for a similar discussion. Recently, it has been noted in the context of data analysis that there is a tendency to overoptimistic reporting of the performance of new methods and a lack of neutral comparison studies in the literature, see, for example, Boulesteix (2015), Boulesteix et al. (2013, 2017), Van Mechelen et al. (2018), Weber et al. (2019), Buchka et al. (2021), Nießl et al. (2021), Pawel et al. (2022). For example, Boulesteix et al. (2013) found only 12 comparison studies out of 55 articles on supervised classification in a literature search. Neutral comparison studies, however, are essential to guarantee a fair comparison of existing methods across different scenarios, thus allowing an applied researcher to determine the "best" method for her or his situation. Similar criticism can also be formulated for the case when simulations are used in a trial design context. When planning a comparison study, a lot of options exists, see, for example, Nießl et al. (2021) for an overview of design and analysis options. Besides the choice of an adequate design and proper reporting of the results, however, the question arises on what kind of data the methods should be compared. Here, different disciplines have different approaches.

When publishing a paper on statistical methodology, manuscripts usually consist of three major parts: theoretical derivations revealing (often asymptotic) properties of the proposed method, a simulation study investigating the small sample behavior and/or comparing the proposed method to relevant competitors, and a data example demonstrating the application of the proposed method to real-world data. *Biometrical Journal*, for example, explicitly encourages authors to "include a description of the problem and a section detailing the application of the new methodology to the problem."[1] In this "classical" format, the simulation study usually covers a wide range of scenarios, while the application to real-world data is often restricted to a single example data set. Depending on the kind of paper, that is, focusing on data analysis or on trial designs, these data examples can serve different roles. For example, Mütze et al. (2020) use a data set on pediatric multiple sclerosis to demonstrate how this study could have been stopped early, if different monitoring procedures had been used. On the other hand, they also discover scenarios where the observed follow-up time does not provide enough information yet.

In the context of machine learning (ML), particularly supervised learning, another approach is common: The performance of methods is usually compared on so-called benchmark data sets, which serve as gold standard and enable comparison of methods on real-world data. That way, they serve as an important step in the process between method development and clinical use (Friedrich et al., 2021b). Simulation studies, on the other hand, are much less common in most ML applications. In some areas of ML, however, simulations also play a role, for example as digital twins (Batty, 2018). This idea is currently employed in different application areas, ranging from industrial applications (Jiang et al., 2021) to agriculture (Pylianidis et al., 2021) and precision medicine (Voigt et al., 2021). Another exemption is learning from simulated data, see Michoel et al. (2007), Gecgel et al. (2019), Behboodi and Rivaz (2019) for some examples in different application areas.

Sometimes, methods are also compared on several real data sets as well as on synthetic data, see Hothorn et al. (2005) and Bischl et al. (2013) for early examples. Recently, a number of so-called *empirical studies* have been published in statistical papers, see, for example, Stegherr et al. (2021a), Wiksten et al. (2016), Seide et al. (2019), Turner et al. (2021) for examples. These papers demonstrate the method(s) of interest on a variety of real-world data sets, thus not solely relying on simulations.

When comparing these approaches, matters are complicated by the fact that different terms are used in different application areas. While most papers in the context of bioinformatics, ML, and artificial intelligence (AI) talk of benchmarking (e.g. Buchka et al., 2021; Dwivedi et al., 2020; Koch et al., 2021; Raji et al., 2021), the terms "empirical study" (Seide et al., 2019; Stegherr et al., 2021a), "empirical evaluation" (Turner et al., 2021), or "empirical comparison" (Wiksten et al., 2016)

---

[1] https://onlinelibrary.wiley.com/page/journal/15214036/homepage/productinformation.html

are also common. Clark and Handcock (2022), on the other hand, mention neither benchmarking nor empirical study, but describe their approach as "[…] a separate and novel contribution to the assessment on the model classes […] by a pairwise assessment on the population of networks that the research community would choose to fit them on." On the other hand, the term benchmark is also used to refer to an "ideal" method, for example, an approach that has complete information, which would not be available in an actual trial (Mozgunov et al., 2022). This makes systematic reviews of the literature difficult and to the best of our knowledge, no systematic comparison of the different approaches exists to date.

In this paper, we aim to investigate differences and similarities between the approaches, discuss their advantages and disadvantages, and develop a new framework aimed at picking "the best of both worlds." Furthermore, we identify tasks that are necessary to be addressed by the scientific community in order to enable the combination of both approaches on a regular basis.

The paper is organized as follows: In Section 2, we give more formal definitions of the concepts of benchmarking and simulation studies and contrast the pros and cons of the two approaches. We summarize our findings in some recommendations in Section 3 and use these to critically discuss some examples in Section 4. We close with a discussion in Section 5.

2 | DIFFERENTIATING BENCHMARKING AND SIMULATION STUDIES

2.1 | Simulation studies

Simulation studies are a common tool in statistics and complement theoretical derivations of statistical methods. The basic idea is to investigate the behavior of a method when applied to synthetic data, that is, data with known properties (Boulesteix et al., 2020). In particular, simulation studies can serve different purposes: (i) Compare several existing methods to determine which performs "best" in a given scenario, (ii) investigate small sample properties of a method in addition to asymptotic results based on theory, (iii) study the robustness of a method if underlying assumptions are violated, and (iv) support assessments of complex design scenarios and sample size planning for an individual study or a series of experiments. Aspect (iv) is especially relevant in the context of complex study designs such as adaptive designs (Friede et al., 2011, 2020) as well as in the CSE framework (Benda et al., 2010) and is conceptually different from the other approaches: Instead of drawing conclusions for a wide range of applications or settings, the focus here is on one specific study (or a series of studies) and the aim is to find the "best" design for this specific application. It is worth mentioning that simulation studies often reflect a frequentist approach, where the true parameters are fixed but unknown values. In Bayesian statistics, simulation studies are mainly used to analyze frequentist properties of posterior-based decision rules (Morita et al., 2010; Thall & Simon, 1994). Similarly, it is more common to modify the choice of priors in sensitivity analyses (e.g., chapter 6 of Gelman et al., 1995). Thus, simulation studies can be viewed as a model-based approach in the sense that mathematical concepts and models need to be known (or assumed to be known). Based on these models, we can investigate which data we can fit them to and where their limits are.

Recently, Morris et al. (2019) have shown, that simulation studies are often poorly designed, analyzed, and reported. To overcome this issue, they provide recommendations and guidelines for the design, implementation, and reporting of simulation studies. Earlier, Burton et al. (2006) already proposed implementing a protocol for simulation studies and provided a checklist of important considerations for the design of such a study. A detailed explanation aimed at applied researchers is given by Boulesteix et al. (2020). Chipman and Bingham (2022) suggest to employ methods from design and analysis of experiments, such as factorial designs and ANOVA methods, when planning and conducting simulation studies. Following these guidelines can improve the conduct of simulation studies and provide a basis for fair and neutral comparisons based on simulated data.

2.2 | Benchmarking

Benchmarking originates from computer sciences (Raji et al., 2021). According to Xie et al. (2021), a benchmark study is a "systematic comparison between computational methods, in which all of them are applied to a gold standard dataset and the success of their […] predictions are summarized in terms of quantitative metrics […]." As Hothorn et al. (2005) describe it, benchmarking aims at "[measuring] performances in a landscape of learning algorithms." The assessment of an algorithm's quality by means of, for example, cross-validation, started in the 1970s with the pioneering work of Stone

(1974). Later, the focus shifted to comparisons of algorithms rather than performance assessment tasks and benchmarking algorithms on various data sets came up in the 1990s with a "shift from rationalism to empiricism" (Church & Hestness, 2019). Since then, benchmarking has established a tradition in ML, specifically in the context of supervised learning, where competitions such as ImageNet (Deng et al., 2009) fostered the comparison of different methods on a common data set. This trend increased in the last few years due to greater availability of open data. For example, the Neural Information Processing Systems conference (https://neurips.cc/) introduced a new track specifically for data sets and benchmarks in 2021 (Vanschoren & Yeung, 2021). Data repositories such as the UC Irvine Machine Learning Repository (Dua & Graff, 2017) or Kaggle (https://www.kaggle.com/) provide platforms for benchmarking data sets. Such platforms also exist for specific applications. For ECGs, for example, methodological development has been hampered until the recent publication of the PTB-XL data set, which is hosted by PhysioNet (Goldberger et al., 2000; Strodthoff et al., 2021).

In contrast to simulation studies, benchmarking provides a data-driven approach. This approach might often be closer to the questions faced by applied users of statistical models: Given my research question and my data, which is the "best" approach I can choose for an adequate analysis? Moreover, for algorithmic approaches without an underlying mathematical model as in many AI applications, where the focus is primarily on prediction instead of inference, designing a simulation study is not straightforward. In supervised learning, where the data are equipped with labels and thus allow for calculating performance measures, benchmarking provides a comparison between methods that is close to real-world applications. In the context of unsupervised learning, the situation is more involved, since the data do not contain known labels. In the special case of cluster analysis, researchers often use data sets with known labels to evaluate their algorithms, although the true labels are actually unknown in clustering applications (Ullmann et al., 2022). Thus, the role of the test data is not as clear as it is in the supervised learning context (Ullmann et al., 2021) and the choice of performance evaluation methods is more complex. This also enables drawing overoptimistic conclusions in cluster analysis, as demonstrated by Ullmann et al. (2022) on both synthetic and real data. Van Mechelen et al. (2018) provide guidelines for benchmarking in cluster analysis and point out the importance of repositories equipped with metadata to provide a good data basis. This issue is also discussed by Zimmermann (2019), who criticizes a lack of suitable data in unsupervised learning. To reduce the issue of unknown class membership in real data, Van Mechelen et al. (2018) recommend to combine "simulations and empirical data as these may yield complementary information." Zimmermann (2019) also advocates the creation of artificial data as an alternative to real-world data, but stresses the need of realistic data-generating mechanisms, that capture the important properties of real data.

A range of papers discuss issues with overoptimistic benchmarking studies and provide guidelines on the conduct and reporting of fair comparison studies, see Weber et al. (2019), Kreutz (2019), Zimmermann (2019), Van Mechelen et al. (2018), Nießl et al. (2021), Buchka et al. (2021), and the references cited therein. A particularly important aspect are the data resources, that is, availability of relevant real-world data for a given problem. In the context of network analysis, Clark and Handcock (2022) approach the problem of representativity by choosing a population of networks based on publications in a premier journal for social network analyses. This population of networks has successfully completed peer review and can thus be "deemed of sufficient scientific interest" (Clark & Handcock, 2022). In most research areas, however, there is still no gold standard of benchmark data sets. This can lead to other major problems like data leakage, that is, spurious findings that arise as artifacts of the data collection process or preprocessing steps. Kapoor and Narayanan (2022) show that this is a widespread issue and leads to severe reproducibility failures in many different research areas. Another aspect of overoptimistic reporting is that comparisons are usually not based on sound statistical test decisions, see Hothorn et al. (2005) as well as Boulesteix et al. (2015) for a hypothesis testing framework. This also relates to sample size calculations: From a statistical perspective, the benchmarking data sets serve as "cases." Thus, it is important to compare the methods on a sufficient number of cases. This can be calculated in advance using methods of sample size calculation (Boulesteix, 2015).

2.3 | Comparison

As described above, benchmarking and simulation studies provide two different approaches to a similar problem: evaluating the performance of several alternative methods based on data, simulated or real. While simulation studies present a more theoretic approach, where the underlying statistical model and some theoretical concepts of the data-generating process have to be known, benchmarking provides a data-driven approach. We will now compare the two approaches with regard to their respective advantages and disadvantages. A short summary of the findings is presented in Table 1.

A huge advantage of simulation studies is that the "ground truth" is known, although sometimes it cannot be derived analytically but is assessed through simulations (Austin, 2010). Thus, it is possible to accurately investigate proposed meth-

TABLE 1 Strengths and weaknesses of benchmarking and simulation studies. The ✓ should be interpreted as "has a tendency to perform better in this respect" rather than an absolute assessment of suitability.

	Simulation studies	Benchmarking
Ground truth known	✓	×
Unlimited data available	✓	×
Computational cost	×	✓
Closer to reality	×	✓
Data-centric viewpoint	×	✓
Model-based viewpoint	✓	×
Used beyond method comparison	✓	×
Applicable to algorithmic approaches	×	✓

ods with respect to bias, coverage probability, or control of the Type I error, for example, since the underlying true values are known by design. Another advantage is that (practically) as much data as required can be simulated, if sufficient computer power is available. Thus, in contrast to real-world data sets, there are hardly any restrictions on sample size. On the other hand, simulated data might not adequately reflect properties of real-world data and the generalizablity of the results may be limited. This aspect particularly comes into play when studying AI applications, which are often applied to complex, high-dimensional data sets. Adequately capturing the properties of this kind of data, particularly with respect to correlations and interactions between variables, might be difficult in a simulation study. Furthermore, the use of simulations can be limited by computational costs: For computationally expensive methods (e.g., Bayesian approaches), it might not be feasible to conduct a simulation, which requires a large number of simulation runs, in a reasonable amount of time. Moreover, there exists an infinite space of possible parameters and simulation settings. Thus, a simulation study can only ever cover a tiny part of that space. This makes the choice of the data-generating process highly subjective and concerns about the relevancy and plausibility of simulated data for real-world applications are warranted (Boulesteix et al., 2017). Thus, the settings should be chosen reasonably and interpreted with caution (Boulesteix, 2015; Pawel et al., 2022).

Benchmarking, on the other hand, is applied to real-world data and therefore allows an assessment of whether the choice of methods matters in practice. Especially in the context of supervised learning, where the data themselves contain the "truth" and when the focus is on prediction rather than inference, benchmarking is closer to reality than a simulation study. Interestingly, many popular benchmarking data sets originally stem from the statistical literature, see Hothorn et al. (2005) for some examples. As Clark and Handcock (2022) call it, this approach takes a *data-centric* viewpoint as opposed to a model-based viewpoint. In other situations, however, it might matter that the ground truth is not known in a real-world data set, for example, in the context of hypothesis testing. Recently, benchmarking has been criticized for a number of reasons. As Raji et al. (2021) point out, benchmarking data sets often fail to achieve the goal of "generality" they are imagined to possess. Instead, they are "inherently specific, finite and contextual" (Raji et al., 2021). A central issue in this context is validity: How well does the data and the associated evaluation metric represent the given task? Are the questions investigated actually relevant to applicants in the field? Does the benchmark study represent relevant real-world data? See also the discussions in Raji et al. (2021), Koch et al. (2021), and Buchka et al. (2021), as well as Bao et al. (2021) for a practical example. A circumstance adding to this issue is that benchmarking studies often use samples of convenience, that is, data sets most easily accessible for the researchers (Koch et al., 2021; Raji et al., 2021). These might either be widely spread data sets or chosen from a familiar context of the researchers (Buchka et al., 2021). In the latter case, the results might not easily generalize to other situations. As Koch et al. (2021) observe, there is an increasing concentration on fewer and fewer data sets used in benchmarking over time and these have been introduced by just a handful of institutions. Thus, they might not even be neutral but potentially influenced by some objective or even sponsored by a specific firm or institution. As a consequence, benchmarking data sets often possess a poor representation of real-world data. For example, many data sets used for training algorithms in natural language processing are only available in English and the majority of images on ImageNet stem from Western Countries (Raji et al., 2021), thus potentially introducing bias in the ML algorithms. The major point of criticism with respect to overoptimistic benchmarking thus stems from the underlying data, with issues such as representativity, validity, and data leakage. However, as Kapoor and Narayanan (2022) point out, there are also other reasons for overoptimistic findings, such as choosing an evaluation metric that is not ideally suited for the task at hand.

To further compare the approaches, we take the point of view of CSE. The CSE framework consists of three core elements: options, assumptions, and metrics (Benda et al., 2010). The metrics serve as tools for comparing the differ-

TABLE 2 Benchmarking and simulations in light of the CSE framework.

	Assumptions	Options	Metrics
Benchmarking	Data sets	Hyperparameters, comparators, choice of software	Performance measures for specific situation
Simulation	Data-generating process	Simulation setting, comparators, hyperparameters, choice of software	Performance measures for specific situation

TABLE 3 Exemplary metrics for different statistical tasks used in benchmarking and simulation. The table is adapted from Morris et al. (2019). CI = confidence interval, SE = standard error, AIC = Akaike Information Criterion, BIC = Bayesian Information Criterion, MSE = mean-squared error.

Statistical task	Benchmarking	Simulation
Estimation	Empirical SE, length of CI	Bias, SE, MSE, coverage, length of CI
Hypothesis testing	–	Type I error, power
Model selection	AIC/BIC	Sensitivity/ specificity for covariate selection, AIC/BIC
Prediction/classification	Predictive accuracy, i.e., calibration and discrimination	Predictive accuracy, i.e., calibration and discrimination
Study design	–	Sample size, duration, power/precision

ent options (which can be varied by the researcher) given the underlying assumptions (which are fixed but unknown). Some of the assumptions might be informed by previous studies whereas others have to rely on subject-matter knowledge only. Given the uncertainty, it is good practice to vary the assumptions in the sense of sensitivity analyses, for example. When viewing benchmarking and simulations in this framework, we get the following: For benchmarking, the competing "options" include the choice of comparators, the tuning of hyperparameters, as well as the choice of statistical software. The "assumptions" in this setting are the data sets. Similar to CSE, they should span a range of optimistic, realistic, and pessimistic situations. Based on these, the competing options can be compared using various metrics, which depend on the specific situation.

For a simulation study, the competing "options" consist of the various choices related to the simulation setting (simulation scenarios, choice of software, number of simulation runs, etc.) as well as the choice of competitors. The "assumptions" here are those on the underlying data-generating process, that is, the mathematical or statistical model that provides the backbone of the simulation study. The important difference in the "assumptions" between benchmarking and simulation studies is that in the simulation study, the data-generating process is known to and chosen by the researcher. In benchmarking, on the other hand, we only observe a realization of the (unknown, underlying) data-generating process. While there exists a choice with respect to which data sets are being analyzed, the true data-generating process will always remain unknown. However, it is usually not necessary for the data analysis to completely know the data-generating process. In a nonparametric approach, for example, the assumptions on the underlying data distribution are usually rather weak. In a simulation study, in contrast, it is hardly possible not to specify the data-generating process in detail, even if fewer or less stringent assumptions are made in the analysis. For example, Mütze et al. (2017) evaluate permutation approaches under a variety of parametric distributions, which differ with respect to skewness, for example, see table III in Mütze et al. (2017). Similarly, Friedrich et al. (2017c) investigate their wild bootstrap approach for nonparametric data in a simulation study where data are generated according to a variety of parametric distributions, reflecting both ordinal and continuous data settings. The options can again be compared by a variety of metrics, which depend on the specific situation. An overview of these aspects is provided in Table 2. Table 3 provides an overview of different metrics used in simulation and benchmarking for a variety of statistical tasks. As we can see, there is no difference between benchmarking and simulations with respect to prediction or classification, since the true labels are contained in the data. With regard to the other tasks, all metrics applied to benchmarking data sets can also be applied to simulated data. Additionally, simulated data allow to compute metrics on the population level, such as bias or Type I error, which cannot be computed on a real data set.

Finally, it should be noted that in both simulation studies and benchmarking, the choice of the comparators, an adequate study design, and transparent reporting of results and limitations are fundamental. In particular, the chosen "methods" need to be clearly defined, including possible pre-processing steps or parameter tuning, and the latter must be optimized for each method separately (Van Mechelen et al., 2018; Weber et al., 2019).

TABLE 4 Comparison of data integration approaches in mixed methods research and in the context of simulation and benchmarking.

Approaches for integration	Mixed methods research	Simulation and Benchmarking	Examples
Merging data	Combine qualitative data (e.g. texts or images) with quantitative data	Combination of empirical study on several data sets and simulation	Seide et al. (2019): Empirical study on 40 data sets complemented by simulation study
Connecting data	Use information from one data analysis (quantitative or qualitative) to inform a subsequent data collection (qualitative or quantitative)	Simulation study inspired by data example; Plasmode simulations; Reconstruction of data sets; Using simulation results to inform subsequent studies	Friedrich and Friede (2020): Simulation study inspired by COVID-19 data; Franklin et al. (2014): Plasmode simulations; Dormuth et al. (2022): Reconstruction of data sets
Embedding data	Data set of secondary priority is embedded within a larger, primary design	Simulation study with additional analysis of a data example	Friedrich et al. (2017b): Extensive simulations complemented by one exemplary data analysis

3 | RECOMMENDATIONS

Our discussion shows that each approach has its merits and shortcomings. In light of these considerations, we recommend to take a broader point of view and learn from the other discipline, respectively. In the following, we provide three recommendations on how this can be achieved:

Combining simulations and benchmarking
First, we encourage statisticians to perform benchmarking analyses additionally to the traditional simulation study, where possible and feasible. As seen previously, the latter is not always the case: Applying benchmarking in trial design studies, for example, is only feasible if the proposed design results in a shorter observation period compared to the design underlying the data. See Mütze et al. (2020) for an example, where the required target information was not achieved in the data example based on the proposed methods. On the other hand, simulation studies might not be feasible, if the data and models involved are computationally very expensive. Examples include resampling approaches in complex models (Ditzhaus & Friedrich, 2020), model-based recursive partitioning (Huber et al., 2021), as well as many ML methods such as boosting (Klinkhammer et al., 2022; Thomas et al., 2017). Particularly when combining these approaches, computation times and/or memory issues may become problematic.

To the best of our knowledge, a sensible approach for combining the two worlds is missing. Ideas for bridging the gap could be taken from mixed methods research (Creswell & Plano Clark, 2017; Hesse-Biber, 2010), where quantitative and qualitative research methods are combined in order to maximize the strengths and minimize the weaknesses of each type of data. An overview of the approaches in mixed methods research and how they translate to our situation is given in Table 4. Integration, that is, the interaction between the different components of the study, is an essential aspect in mixed methods research (O'Cathain et al., 2010). According to Creswell et al. (2011), there are three core approaches to integrate different forms of data: merging data, connecting data, and embedding data. In the context of simulation studies and benchmarking, "embedding data" can be viewed as the current practice in many statistical manuscripts, where the results of a simulation study (large, primary design) are enhanced by additionally analyzing a data example (secondary priority). Similarly, the recent trend toward combining so-called empirical studies based on several data sets with simulated data can be viewed as merging data: The two types of data (simulated and real) are analyzed separately and the results are combined in a discussion section. The third idea in mixed methods research is connecting data. The aim here is to use the results obtained from one type of data (e.g., qualitative data) to inform a subsequent study (e.g., by developing new items for a quantitative data collection). The order of the two types of data may be reversed here. In the context of simulation studies and benchmarking, several existing approaches fall into this category:

1. Simulating data based on a real data example: Many simulation studies aim to mimic a real data example, see, for example, Friedrich et al. (2017a), Bluhmki et al. (2018), Ohneberg et al. (2019), Friedrich and Friede (2020), Graf et al. (2022) to name just a few. Note, however, that the degree to which the simulated and real data overlap varies greatly: Sometimes simulations are simply using the estimated mean and (co-)variances from the real data, sometimes other

aspects such as length of follow-up or number of observed events are simulated based on observed data. Moreover, new simulation approaches are often inspired by a data example, for which the existing methods are not adequate. Sylvestre and Abrahamowicz (2008), for example, developed a simulation approach for time-dependent covariates based on a permutation approach, see also Sylvestre et al. (2010). Similarly, Crowther and Lambert (2013) extended existing approaches for simulating time-to-event data motivated by a data example for which the existing simulation approaches seemed too simplistic.

2. Reconstructing data sets based on published information: To counter-act the problem of available individual patient data, several approaches have been considered to reconstruct these data based on published information. The method by Guyot et al. (2012), for example, allows for reconstruction of survival data based on published Kaplan–Meier curves. Some examples for method comparisons based on this approach include Royston et al. (2019), Dormuth et al. (2022). A similar approach has been proposed by Bluhmki et al. (2019), who use resampling approaches based on published Nelson–Aalen plots for simulating realistic data.

3. Plasmode simulation studies: Another relevant concept in this context are Plasmode simulations, see, for example, Franklin et al. (2014). Here, the idea is to use part of a real-world data set (e.g., to capture difficult relationships between large numbers of covariates) and to artificially create an outcome (e.g., treatment effect) of the researcher's choice.

As mentioned above, it is possible to reverse the order of data types in mixed methods research. This approach might also be possible in our context, even though it has, to our knowledge, not been implemented yet. Speaking in terms of clinical trials, simulation studies represent prospective, experimental designs, but are conducted under "laboratory conditions." Benchmarking studies, on the other hand, are usually conducted retrospectively, that is, based on already existing data. However, in some situations it is possible to use information obtained from simulations to inform subsequent trials:

1. Simulation-based optimization of designs: Simulation studies can be used to explore settings that are especially relevant for the method under consideration. Afterwards, the method can be verified in benchmark data sets which represent the settings identified by the simulations. This approach is comparable to several existing approaches in different fields. For example, simulation is increasingly used to determine a promising set of input parameters for a biomanufacturing system, for example the production of antibodies for drug development (Wang et al., 2019). Another example are in silico clinical trials, i.e. "trials for pharmacological therapies or medical devices based on modelling and simulation technologies" (Musuamba et al., 2021). Here, the idea is to use individualized simulations to speed up the process of drug development by informing clinical trials beforehand on expected outcomes and possible modifications (Viceconti et al., 2016). In silico clinical trials can thus either complement or replace in vivo clinical trials. An early example for the application of in silico clinical trials is given by Clermont et al. (2004), who investigate the feasibility of this approach in clinical trials of severe sepsis.

2. Empirical comparison of designs: Our second suggestion relates to the aspect of trial designs. As stated above, this is a field that is not yet present in benchmarking experiments, although simulation studies allow to investigate trial designs as well. Thus, one could imagine comparing several design approaches (which proved promising in simulations) in the real world. More specifically, one would conduct, for instance, a prospective, randomized controlled trial, where the "interventions" are different trial designs. One idea in this direction are so-called SWATs (Study Within A Trial) (Clark et al., 2022; National Institute for Health and Care Research, 2022). To date, SWATs are mainly used to evaluate the effectiveness of recruitment strategies, but could potentially be extended to cover more complex design aspects as well. In some situations, this type of evaluation might be unrealistic, since heterogeneity between studies would be too large to ensure comparable results. Similarly, conducting several clinical trials would often be too time consuming. Thus, experiments where this approach might be possible would have to be more homogeneous and/or less time consuming. One example is the context of economics, where experiments are sometimes conducted as business simulation games (Jobjörnsson et al., 2022).

Establish infrastructure, databases and gold standards

It should be noted that this recommendation stretches beyond encouraging the individual user to apply both benchmarking and simulations in his or her next study. In order to adequately address this issue, the whole community is needed. In particular, the necessary infrastructure needs to be established. This starts with providing and extending databases and data repositories that enable large-scale benchmarking studies. To address the issues raised in the context of benchmarking, these data need to be adequately curated, equipped with metadata, and cautiously monitored (Koch et al., 2021; Raji et al., 2021; Strodthoff et al., 2021; Van Mechelen et al., 2018; Zimmermann, 2019). In addition, the community needs to

establish what can be viewed as a "gold standard data set" for a given application. In this context, there is a role to play for the scientific societies in developing guidelines and providing recommendations.

Encourage conduct and publication of comparison studies

Finally, there is also a role to play for scientific journals. Here, publication bias and fear of rejection still provide pressure for publishing "new" and "better" approaches. Moreover, most high-ranking statistical journals do not mention comparison studies in their scopes (Boulesteix et al., 2017). Thus, to enable more comparison studies, publication culture needs to change as well. Authors should be encouraged to conduct and publish neutral comparison studies, while reviewers should acknowledge the value of these studies and consider this in their recommendations. Finally, editorial boards should facilitate publication of neutral comparison studies through corresponding policies, by extending the scope of their journals or through special issues such as this one.

4 | EXAMPLES

In the following, we discuss some exemplary studies with regard to possible improvements. We deliberately chose studies, in which at least one of the authors was involved, since the purpose is not to criticize others but to discuss pros and cons of existing studies in light of the arguments made in this paper.

4.1 | A simulation study

Motivated by an early nonrandomized trial in COVID-19, Friedrich and Friede (2020) investigate the behavior of different causal inference methods in a large simulation study. In terms of Table 4, data are *connected* in this paper. The study can be considered neutral, since no new methodology is proposed and neither of the authors have been involved in developing any of the approaches under consideration. The parameters underlying some simulation scenarios are motivated by the data example, while other scenarios were taken from another paper (Austin, 2007). However, the authors did not accurately follow the recommended ADEMP structure by Morris et al. (2019) or the CSE framework by Benda et al. (2010). An important aspect to note here is that even though the data are artificial, the ground truth is not known in all scenarios. In particular, the "true" causal risk difference is estimated based on simulated counterfactual outcomes in large data sets ($n = 10,000$) and the underlying parameters are iteratively modified, until the desired risk difference is approximately reached (Austin, 2010). Based on these values, the methods are compared with respect to bias, length of confidence intervals, root mean squared error, and coverage probability, since the statistical task is estimation (cf. Table 3). Although the paper is motivated by a real data example, the authors did not include an analysis of this data example in the final manuscript. This was due to the fact that the data set failed to illustrate the methods compared. In particular, some methods investigated in the simulations could not be applied to the data example and all methods essentially came to the same conclusion, see Figure 1. This was due to the major statistical and design issues in the original study that could not be rectified by more elaborate analysis methods. However, this case demonstrates that picking a simple data example can result in misleading conclusions and should thus be avoided. To sum up, Friedrich and Friede (2020) provide an example of a thorough simulation study, but without comparing the methods on real data. Thus, it remains unclear whether the theoretical results observed in the simulations would lead to different conclusions in real-world applications.

4.2 | An empirical study

An example of an empirical study is Stegherr et al. (2021a). Here, estimators typically used in the study of adverse events with varying follow-up times are compared in 17 randomized clinical trials. The properties of these estimators have been analyzed and discussed previously. In particular, a special issue in *Pharmaceutical Statistics* was dedicated to the topic (Kieser, 2016). There, methods were demonstrated on single data examples (Allignol et al., 2016; Bender et al., 2016; Proctor & Schumacher, 2016). Moreover, Stegherr et al. (2021b) compared the methods on artificial data in a simulation study.

　　The aim of this empirical study is to "investigate and demonstrate which biases can occur in practice." Data collection, inclusion criteria, analyses methods, and the set-up of the meta-analysis are explained in Stegherr et al. (2020). The methods are compared to a gold-standard method by investigating the ratios of the probability estimates obtained with
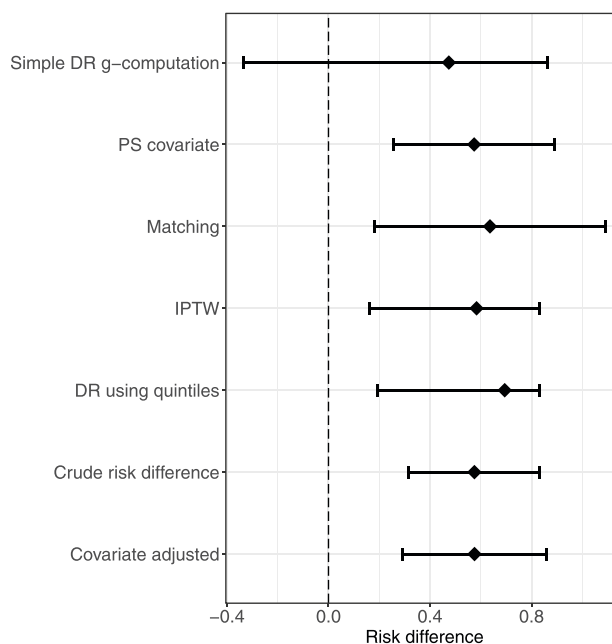
FIGURE 1 Estimated risk differences with 95% confidence intervals obtained by the different methods. For details on the different methods, see Friedrich and Friede (2020).

the different estimators divided by the probability estimates obtained with the gold standard. Thus, the ground truth is assumed to be obtained through the gold standard method in this example, which has implications for the assessment of some properties such as bias.

Due to the opportunistic sample of data sets used in the empirical study, however, generalizability is limited (Stegherr et al., 2021a). In particular, more than two thirds of the trials included in the study stem from oncology and adverse events were heterogeneous due to their backgrounds in different therapeutic areas. Thus, the sample is not representative of clinical studies in general. In order to improve this study and in light of the recommendations above, a database of randomized trials with time-to-event outcomes investigating adverse events would be needed. This, however, brings along issues of data protection, which were addressed in the study by analyzing the data at the respective sponsor's site and only transferring aggregated results, that is, the calculations are done in a distributed fashion.

4.3 | An empirical study complemented by simulations

As a final example, we consider Seide et al. (2019). This empirical study on 40 meta-analyses is complemented by a simulation study, and is thus in line with our advice to combine both approaches. According to Table 4, this study thus *merges* empirical and artificial data.

In particular, an empirical data set of 40 meta-analyses was extracted from recently published reviews in a systematic manner. Similar to Stegherr et al. (2021a), the different methods were compared to a gold-standard approach and the ratios of the obtained point estimates were considered as metrics. Moreover, the length of the empirical confidence intervals was compared on the empirical data. In the simulation study, coverage probabilities could additionally be used as metric, since the ground truth was known in this case. As the authors state "A consideration of all meta-analyses might have led to a more complete picture, but was not feasible with the resources of this project [...]," highlighting again the need for adequate databases such as Cochrane Database of Systematic Reviews.

5 | DISCUSSION

Method comparison studies are an important tool to provide recommendations for both applied and methodological researchers. While applied researchers wonder about the "best" method to pick for their data analysis, method com-

---

Combine benchmarking and simulations whenever possible, e.g.

- Enhance simulation studies through data example(s)

- Conduct simulation studies based on real data

Establish infrastructure, databases and gold standards within community

- Extend databases and data repositories

- Provide curation and meta data

Encourage conduct and publication of neutral comparison studies

- Include neutral comparison studies in journals' scopes

- Acknowledge added value of neutral comparison studies in review process
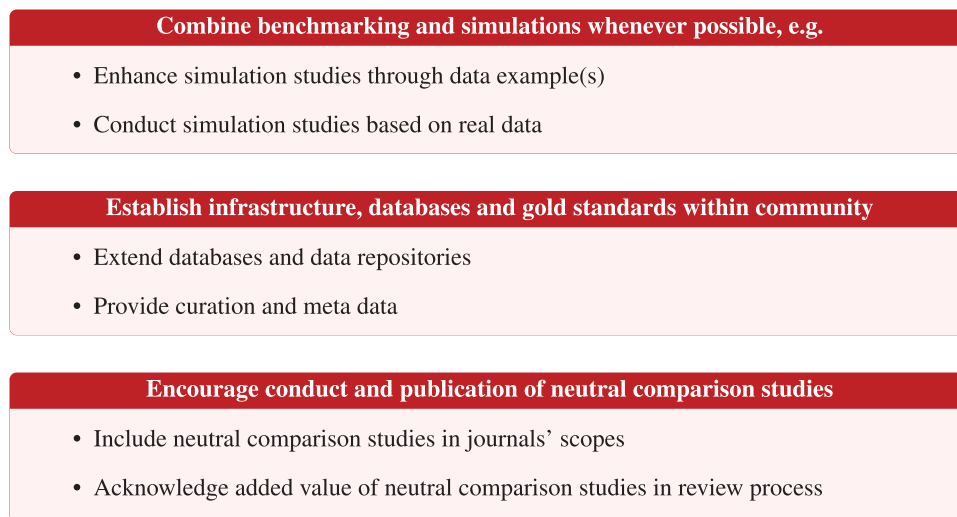

FIGURE 2   Recommendations.

parisons can also help methodological researchers in determining potential for further improvements or identifying limitations of existing methods and thus a need for the development of new approaches. In order to yield valid results, however, method comparison studies need to be conducted in a neutral fashion, not biased toward novel methods. In practice, one might argue that a comparison study can never be entirely neutral (Strobl & Leisch, 2022), or a benchmarking analysis, since in any case there are choices to make (regarding the underlying data-generating process or the data sets). Thus, the term "neutral" in this paper should be interpreted as: "being [...] focused on the comparison of existing methods already described elsewhere rather than on a new prototype method being introduced [...]" (Strobl & Leisch, 2022). In this paper, we have focused on the aspect of the underlying data: These could be real data sets from practical applications or artificial data. The idea for these considerations was born while working on a white paper of the German Consortium in Statistics (DAGStat, www.dagstat.de) on AI (Friedrich et al., 2021a). In this paper, we have introduced the two approaches and discussed their respective advantages and shortcomings. Since no approach is always superior to the other, we recommend to use a combination wherever possible and we have made some suggestions on how that could be achieved, see Figure 2 for a summary.

Some final remarks are in place. First, we have not discussed possible approaches to combining the results obtained on several data sets (real-world or artificial) to come to a final conclusion regarding the "best" method. Here, several approaches exist. Most commonly, the methods are ranked according to their performance and results are presented as summaries of this ranking, see Nießl et al. (2021) for a detailed discussion. As pointed out by Boulesteix et al. (2013), the concepts of meta-analysis could also be extended for the framework of method comparison studies. In this context, it should also be noted that answers like "method A performs universally better than method B" cannot be expected from comparison studies. Instead, one should rather consider which aspects of the underlying data (real or simulated) are associated with the good or bad performance of a method, see Strobl and Leisch (2022) for an extensive discussion of this topic and Varga et al. (2022) for a recent application of this approach. Second, one of the major selling points for simulation studies is that the ground truth is usually known for artificial data. Although this is true in many applications, it cannot always be achieved. In the context of causal inference, for example, the "truth," that is, the true causal effect, is often estimated even in simulations (Austin, 2010; Friedrich & Friede, 2020). The advantage of simulated data is, of course, that very large data sets can be generated on which to estimate the causal risk difference, for example, but this should be kept in mind. Third, as mentioned briefly above, an important aspect in benchmarking studies is the availability of relevant real-world data. Two aspects need to be considered in this context: (1) Availability: The recent push for open science and, as a consequence, data sharing will hopefully continue to improve the availability of data sets, thus enabling more large-scale benchmarking studies. In particular, many journals now require or encourage data sharing. Moreover, platforms like the UCI Machine Learning Repository (Dua & Graff, 2017), Kaggle (https://www.kaggle.com/), and the NIH Data Sharing Repositories (National Library of Medicine, 2022) as well as the R-package **OpenML** (Casalicchio et al., 2019) provide lots of data sets for benchmarking tasks. (2) Quality: It is also important that data are in a standard format and of sufficient quality to make benchmarking possible. This includes, for example, the collection of metadata and a cautious monitoring

of the data quality. As already noted in Section 2.2, missing standards for the underlying data can lead to major problems fueling the reproducibility crisis, such as data leakage (Kapoor & Narayanan, 2022).

ACKNOWLEDGMENTS

CONFLICT OF INTEREST STATEMENT
The authors have declared no conflict of interest.

DATA AVAILABILITY STATEMENT
The COVID-19 data used in Section 4 is provided in Supplementary Table 1 of Gautret et al. (2020).

ORCID
Sarah Friedrich https://orcid.org/0000-0003-0291-4378
Tim Friede https://orcid.org/0000-0001-5347-7441

REFERENCES
Allignol, A., Beyersmann, J., & Schmoor, C. (2016). Statistical issues in the analysis of adverse events in time-to-event data. *Pharmaceutical Statistics*, *15*(4), 297–305.

Austin, P. C. (2007). The performance of different propensity score methods for estimating marginal odds ratios. *Statistics in Medicine*, *26*(16), 3078–3094.

Austin, P. C. (2010). A data-generation process for data with specified risk differences or numbers needed to treat. *Communications in Statistics - Simulation and Computation*, *39*(3), 563–577.

Bao, M., Zhou, A., Zottola, S., Brubach, B., Desmarais, S., Horowitz, A., Lum, K., & Venkatasubramanian, S. (2021). It's COMPASlicated: The messy relationship between RAI datasets and algorithmic fairness benchmarks. *arXiv preprint arXiv:2106.05498*.

Batty, M. (2018). Digital twins. *Environment and Planning B: Urban Analytics and City Science*, *45*(5), 817–820.

Behboodi, B., & Rivaz, H. (2019). Ultrasound segmentation using U-Net: learning from simulated data and testing on real data. In *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE.

Benda, N., Branson, M., Maurer, W., & Friede, T. (2010). Aspects of modernizing drug development using clinical scenario planning and evaluation. *Drug Information Journal*, *44*(3), 299–315.

Bender, R., Beckmann, L., & Lange, S. (2016). Biometrical issues in the analysis of adverse events within the benefit assessment of drugs. *Pharmaceutical Statistics*, *15*(4), 292–296.

Bischl, B., Schiffner, J., & Weihs, C. (2013). Benchmarking local classification methods. *Computational Statistics*, *28*(6), 2599–2619.

Bluhmki, T., Dobler, D., Beyersmann, J., & Pauly, M. (2018). The wild bootstrap for multivariate Nelson–Aalen estimators. *Lifetime Data Analysis*, *25*(1), 97–127.

Bluhmki, T., Putter, H., Allignol, A., & Beyersmann, J. (2019). Bootstrapping complex time-to-event data without individual patient data, with a view toward time-dependent exposures. *Statistics in Medicine*, *38*(20), 3747–3763.

Boulesteix, A.-L. (2015). Ten simple rules for reducing overoptimistic reporting in methodological computational research. *PLOS Computational Biology*, *11*(4), e1004191.

Boulesteix, A.-L., Binder, H., Abrahamowicz, M., & Sauerbrei, W. (2017). On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, *60*(1), 216–218.

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

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

Buchka, S., Hapfelmeier, A., Gardner, P. P., Wilson, R., & Boulesteix, A.-L. (2021). On the optimistic performance evaluation of newly introduced bioinformatic methods. *Genome Biology*, *22*(1), 1–8.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292.

Casalicchio, G., Bossek, J., Lang, M., Kirchhoff, D., Kerschke, P., Hofner, B., Seibold, H., Vanschoren, J., & Bischl, B. (2019). OpenML: An R package to connect to the machine learning platform OpenML. *Computational Statistics*, *34*(3), 977–991.

Chipman, H., & Bingham, D. (2022). Let's practice what we preach: Planning and interpreting simulation studies with design and analysis of experiments. *Canadian Journal of Statistics*, *50*(4), 1228–1249.

Church, K. W., & Hestness, J. (2019). A survey of 25 years of evaluation. *Natural Language Engineering*, *25*(06), 753–767.

Clark, D. A., & Handcock, M. S. (2022). Comparing the real-world performance of exponential-family random graph models and latent order logistic models for social network analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *185*(2), 566–587.

Clark, L., Arundel, C., Coleman, E., Doherty, L., Parker, A., Hewitt, C., Beard, D., Bower, P., Brocklehurst, P., Cooper, C., Culliford, L., Devane, D., Emsley, R., Eldridge, S., Galvin, S., Gillies, K., Montgomery, A., Sutton, C., Treweek, S., & Torgerson, D. (2022). The PROMoting the USE of SWATs (PROMETHEUS) programme: Lessons learnt and future developments for SWATs. *Research Methods in Medicine & Health Sciences*, *3*(4), 100–106.

Clermont, G., Bartels, J., Kumar, R., Constantine, G., Vodovotz, Y., & Chow, C. (2004). In silico design of clinical trials: A method coming of age. *Critical Care Medicine*, *32*(10), 2061–2070.

Creswell, J. W., Klassen, A. C., Plano Clark, V. L., & Smith, K. C. (2011). Best practices for mixed methods research in the health sciences. *Bethesda (Maryland): National Institutes of Health*, *2013*, 541–545.

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research*. Sage Publications.

Crowther, M. J., & Lambert, P. C. (2013). Simulating biologically plausible complex survival data. *Statistics in Medicine*, *32*(23), 4118–4134.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). IEEE.

Ditzhaus, M., & Friedrich, S. (2020). More powerful logrank permutation tests for two-sample survival data. *Journal of Statistical Computation and Simulation*, *90*(12), 2209–2227.

Dmitrienko, A., & Pulkstenis, E. (2017). *Clinical trial optimization using R*. CRC Press.

Dormuth, I., Liu, T., Xu, J., Yu, M., Pauly, M., & Ditzhaus, M. (2022). Which test for crossing survival curves? A user's guideline. *BMC Medical Research Methodology*, *22*(1), 34.

Dua, D., & Graff, C. (2017). *UCI machine learning repository*. https://archive.ics.uci.edu/ml/index.php

Dwivedi, V. P., Joshi, C. K., Laurent, T., Bengio, Y., & Bresson, X. (2022). Benchmarking graph neural networks. *Journal of Machine Learning Research*, *23*, 1–48.

Franklin, J. M., Schneeweiss, S., Polinski, J. M., & Rassen, J. A. (2014). Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational Statistics & Data Analysis*, *72*, 219–226.

Friede, T., Nicholas, R., Stallard, N., Todd, S., Parsons, N., Valdés-Márquez, E., & Chataway, J. (2010). Refinement of the Clinical Scenario Evaluation framework for assessment of competing development strategies with an application to multiple sclerosis. *Drug Information Journal*, *44*(6), 713–718.

Friede, T., Parsons, N., Stallard, N., Todd, S., Marquez, E. V., Chataway, J., & Nicholas, R. (2011). Designing a seamless phase II/III clinical trial using early outcomes for treatment selection: An application in multiple sclerosis. *Statistics in Medicine*, *30*(13), 1528–1540.

Friede, T., Stallard, N., & Parsons, N. (2020). Adaptive seamless clinical trials using early outcomes for treatment or subgroup selection: Methods, simulation model and their implementation in R. *Biometrical Journal*, *62*(5), 1264–1283.

Friedrich, S., Antes, G., Behr, S., Binder, H., Brannath, W., Dumpert, F., Ickstadt, K., Kestler, H. A., Lederer, J., Leitgöb, H., Pauly, M., Steland, A., Wilhelm, A., & Friede, T. (2021a). Is there a role for statistics in artificial intelligence? *Advances in Data Analysis and Classification*, *16*, 823–846.

Friedrich, S., Beyersmann, J., Winterfeld, U., Schumacher, M., & Allignol, A. (2017a). Nonparametric estimation of pregnancy outcome probabilities. *The Annals of Applied Statistics*, *11*(2), 840–867.

Friedrich, S., Brunner, E., & Pauly, M. (2017b). Permuting longitudinal data in spite of the dependencies. *Journal of Multivariate Analysis*, *153*, 255–265.

Friedrich, S., & Friede, T. (2020). Causal inference methods for small non-randomized studies: Methods and an application in COVID-19. *Contemporary Clinical Trials*, *99*, 106213.

Friedrich, S., Groß, S., König, I. R., Engelhardt, S., Bahls, M., Heinz, J., Huber, C., Kaderali, L., Kelm, M., Leha, A., Rühl, J., Schaller, J., Scherer, C., Vollmer, M., Seidler, T., & Friede, T. (2021b). Applications of artificial intelligence/machine learning approaches in cardiovascular medicine: A systematic review with recommendations. *European Heart Journal - Digital Health*, *2*(3), 424–436.

Friedrich, S., Konietschke, F., & Pauly, M. (2017c). A wild bootstrap approach for nonparametric repeated measurements. *Computational Statistics & Data Analysis*, *113*, 38–52.

Gautret, P., Lagier, J.-C., Parola, P., Hoang, V. T., Meddeb, L., Mailhe, M., Doudier, B., Courjon, J., Giordanengo, V., Vieira, V. E., Dupont, H. T., Honoré, S., Colson, P., Chabrière, E., Scola, B. L., Rolain, J.-M., Brouqui, P., & Raoult, D. (2020). Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial. *International Journal of Antimicrobial Agents*, *56*(1), 105949.

Gecgel, O., Ekwaro-Osire, S., Dias, J. P., Serwadda, A., Alemayehu, F. M., & Nispel, A. (2019). Gearbox fault diagnostics using deep learning with simulated data. In *2019 IEEE international conference on prognostics and health management (ICPHM)*. IEEE, 1–8.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*, Chapman and Hall/CRC.

Goldberger, A. L., Amaral, L. A. N., Glass, L., Hausdorff, J. M., PCh, I., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220.

Graf, R., Zeldovich, M., & Friedrich, S. (2022). Comparing linear discriminant analysis and supervised learning algorithms for binary classification—A method comparison study. *Biometrical Journal*.

Guyot, P., Ades, A. E., Ouwens, M. J. N. M., & Welton, N. J. (2012). Enhanced secondary analysis of survival data: Reconstructing the data from published Kaplan-Meier survival curves. *BMC Medical Research Methodology*, *12*(1), 9.

Hesse-Biber, S. N. (2010). *Mixed methods research: Merging theory with practice*. Guilford Press.

Hothorn, T., Leisch, F., Zeileis, A., & Hornik, K. (2005). The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, *14*(3), 675–699.

Huber, C., Benda, N., & Friede, T. (2021). Subgroup identification in individual participant data meta-analysis using model-based recursive partitioning. *Advances in Data Analysis and Classification*, *16*, 797–815.

Jiang, Y., Yin, S., Li, K., Luo, H., & Kaynak, O. (2021). Industrial applications of digital twins. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2207), 20200360.

Jobjörnsson, S., Schaak, H., Musshoff, O., & Friede, T. (2022). Improving the statistical power of economic experiments using adaptive designs. *Experimental Economics*.

Kapoor, S., & Narayanan, A. (2022). Leakage and the reproducibility crisis in ML-based science. *arXiv:2207.07048*.

Kieser, M. (2016). Special Issue:Analysis of adverse event data. *Pharmaceutical Statistics*, *15*(4), 287–289.

Klinkhammer, H., Staerk, C., Maj, C., Krawitz, P. M., & Mayr, A. (2023). A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, *13*, 1076440.

Koch, B., Denton, E., Hanna, A., & Foster, J. G. (2021). Reduced, reused and Recycled: The life of a dataset in machine learning research. *arXiv preprint arXiv:2112.01716*.

Kreutz, C. (2019). Guidelines for benchmarking of optimization-based approaches for fitting mathematical models. *Genome Biology*, *20*(1), 281.

Michoel, T., Maere, S., Bonnet, E., Joshi, A., Saeys, Y., den Bulcke, T. V., Leemput, K. V., van Remortel, P., Kuiper, M., Marchal, K., & de Peer, Y. V. (2007). Validating module network learning algorithms using simulated data. *BMC Bioinformatics*, *8*, S2.

Morita, S., Thall, P. F., & Müller, P. (2010). Evaluating the impact of prior assumptions in Bayesian biostatistics. *Statistics in biosciences*, *2*(1), 1–17.

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

Mozgunov, P., Paoletti, X., & Jaki, T. (2022). A benchmark for dose-finding studies with unknown ordering. *Biostatistics*, *23*(3), 721–737.

Musuamba, F. T., Rusten, I. S., Lesage, R., Russo, G., Bursi, R., Emili, L., Wangorsch, G., Manolis, E., Karlsson, K. E., Kulesza, A., Courcelles, E., Boissel, J.-P., Rousseau, C. F., Voisin, E. M., Alessandrello, R., Curado, N., Dall'ara, E., Rodriguez, B., Pappalardo, F., & Geris, L. (2021). Scientific and regulatory evaluation of mechanistic in silico drug and disease models in drug development: Building model credibility. *CPT: Pharmacometrics & Systems Pharmacology*, *10*(8), 804–825.

Mütze, T., Konietschke, F., Munk, A., & Friede, T. (2017). A studentized permutation test for three-arm trials in the "gold standard" design. *Statistics in Medicine*, *36*(6), 883–898.

Mütze, T., Salem, S., Benda, N., Schmidli, H., & Friede, T. (2020). Blinded continuous information monitoring of recurrent event endpoints with time trends in clinical trials. *Statistics in Medicine*, *39*(27), 3968–3985.

National Institute for Health and Care Research. (2022). *Studies within a trial (SWAT) and studies within a review (SWAR)*. https://www.nihr.ac.uk/documents/studies-within-a-trial-swat/21512?pr=

National Library of Medicine. (2022). *NIH Data Sharing Repositories; National Library of Medicine; National Institutes of Health; U.S. Department of Health and Human Services*. https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., & Boulesteix, A.-L. (2021). Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *12*(2), e1441. https://doi.org/10.1002/widm.1441

O'Cathain, A., Murphy, E., & Nicholl, J. (2010). Three techniques for integrating data in mixed methods studies. *BMJ*, *341*, c4587.

Ohneberg, K., Beyersmann, J., & Schumacher, M. (2019). Exposure density sampling: Dynamic matching with respect to a time-dependent exposure. *Statistics in Medicine*, *38*(22), 4390–4403.

Pawel, S., Kook, L., & Reeve, K. (2022). Pitfalls and potentials in simulation studies. *arXiv:2203.13076*.

Proctor, T., & Schumacher, M. (2016). Analysing adverse events by time-to-event models: The CLEOPATRA study. *Pharmaceutical Statistics*, *15*(4), 306–314.

Pylianidis, C., Osinga, S., & Athanasiadis, I. N. (2021). Introducing digital twins to agriculture. *Computers and Electronics in Agriculture*, *184*, 105942.

Raji, I. D., Bender, E. M., Paullada, A., Denton, E., & Hanna, A. (2021). AI and the everything in the whole wide world benchmark. *arXiv preprint arXiv:2111.15366*.

Royston, P., Choodari-Oskooei, B., Parmar, M. K. B., & Rogers, J. K. (2019). Combined test versus logrank/Cox test in 50 randomised trials. *Trials*, *20*(1), 172.

Seide, S. E., Röver, C., & Friede, T. (2019). Likelihood-based random-effects meta-analysis with few studies: Empirical and simulation studies. *BMC Medical Research Methodology*, *19*(1), 16.

Stegherr, R., Beyersmann, J., Jehl, V., Rufibach, K., Leverkus, F., Schmoor, C., & Friede, T. (2020). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY): Rationale and statistical concept of a meta-analytic study. *Biometrical Journal*, *63*(3), 650–670.

Stegherr, R., Schmoor, C., Beyersmann, J., Rufibach, K., Jehl, V., Brückner, A., Eisele, L., Künzel, T., Kupas, K., Langer, F., Leverkus, F., Loos, A., Norenberg, C., Voss, F., & Friede, T. (2021a). Survival analysis for AdVerse events with VarYing follow-up times (SAVVY)—Estimation of adverse event risks. *Trials*, *22*(1), 420.

Stegherr, R., Schmoor, C., Lübbert, M., Friede, T., & Beyersmann, J. (2021b). Estimating and comparing adverse event probabilities in the presence of varying follow-up times and competing events. *Pharmaceutical Statistics*, *20*(6), 1125–1146.

Stone, M. (1974). Cross-Validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 111–133.

Strobl, C., & Leisch, F. (2022). Against the "one method fits all data sets" philosophy for comparison studies in methodological research. *Biometrical Journal*.

Strodthoff, N., Wagner, P., Schaeffter, T., & Samek, W. (2021). Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, *25*(5), 1519–1528.

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.

Sylvestre, M.-P., Evans, T., MacKenzie, T., & Abrahamowicz, M. (2010). *PermAlgo: Permutational algorith to generate event times conditional on a covariate matrix including time-dependent covariates*, R package version 1.2.

Thall, P. F., & Simon, R. (1994). Practical Bayesian guidelines for phase IIB clinical trials. *Biometrics*, *50*(2), 337–349.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., & Hofner, B. (2017). Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, *28*(3), 673–687.

Turner, S. L., Karahalios, A., Forbes, A. B., Taljaard, M., Grimshaw, J. M., & McKenzie, J. E. (2021). Comparison of six statistical methods for interrupted time series studies: Empirical evaluation of 190 published series. *BMC Medical Research Methodology*, *21*(1), 134.

Ullmann, T., Beer, A., Hünemörder, M., Seidl, T., & Boulesteix, A.-L. (2022). Over-optimistic evaluation and reporting of novel cluster algorithms: An illustrative study. *Advances in Data Analysis and Classification*.

Ullmann, T., Hennig, C., & Boulesteix, A.-L. (2021). Validation of cluster analysis results on validation data: A systematic framework. *WIREs Data Mining and Knowledge Discovery*, *12*(3), e1444.

Van Mechelen, I., Boulesteix, A.-L., Dangl, R., Dean, N., Guyon, I., Hennig, C., Leisch, F., & Steinley, D. (2018). Benchmarking in cluster analysis: A white paper. *arXiv preprint arXiv:1809.10496*.

Vanschoren, J., & Yeung, S., (Eds.). (2021). *Proceedings of the neural information processing systems track on datasets and benchmarks.* https://datasets-benchmarks-proceedings.neurips.cc/paper/2021

Varga, A. N., Morel, A. E. G., Lokkerbol, J., van Dongen, J. M., van Tulder, M. W., & Bosmans, J. E. (2022). Dealing with confounding in observational studies: A scoping review of methods evaluated in simulation studies with single-point exposure. *Statistics in Medicine*, *42*(4), 487–516.

Viceconti, M., Henney, A., & Morley-Fletcher, E. (2016). In silico clinical trials: How computer simulation will transform the biomedical industry. *International Journal of Clinical Trials*, *3*(2), 37–46.

Voigt, I., Inojosa, H., Dillenseger, A., Haase, R., Akgün, K., & Ziemssen, T. (2021). Digital twins for multiple sclerosis. *Frontiers in Immunology*, *12*, 669811.

Wang, B., Xie, W., Martagan, T., Akcay, A., & Corlu, C. G. (2019). Stochastic simulation model development for biopharmaceutical production process risk analysis and stability control. In *2019 winter simulation conference (WSC)* (pp. 1989–2000). IEEE.

Weber, L. M., Saelens, W., Cannoodt, R., Soneson, C., Hapfelmeier, A., Gardner, P. P., Boulesteix, A.-L., Saeys, Y., & Robinson, M. D. (2019). Essential guidelines for computational method benchmarking. *Genome Biology*, *20*(1), 125.

Wiksten, A., Rücker, G., & Schwarzer, G. (2016). Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Statistics in Medicine*, *35*(15), 2503–2515.

Xie, C., Jauhari, S., & Mora, A. (2021). Popularity and performance of bioinformatics software: The case of gene set analysis. *BMC Bioinformatics*, *22*(1), 191.

Zimmermann, A. (2019). Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey. *WIREs Data Mining and Knowledge Discovery*, *10*(2), e1330.