Master's Thesis

Confirmatory studies in methodological statistical research: concept and illustration

Author

Felix Julian David Lange

Supervisor

Prof. Dr. Anne-Laure Boulesteix

Department of Statistics

Ludwig-Maximilians-Universität München

Munich, December 19, 2022

Abstract

Hypothesis-generating, exploratory research and hypothesis-testing, confirmatory research are both essential to progress in science. However, failing to separate the two types of research can lead to non-replicable results when exploratory findings are misperceived or intentionally presented as confirmatory. To transparently conduct strictly confirmatory analyses, the practice of publicly registering research plans before the data analysis has become increasingly popular. This process is called pre-registration. For a number of applied research fields and study types, templates to aid researchers in specifying sufficiently detailed plans are available. In the context of methodological statistical research, the exploratory-confirmatory distinction has received little attention in the scientific literature so far. Consequently, there is no guidance available regarding the pre-registration of methodological research in particular. To address this gap, this thesis proposes an approach for a strictly confirmatory real-data study in this field and provides a corresponding pre-registration template for comprehensively planning such a study. The suggested approach is illustrated with a large-scale benchmark experiment, and its results more or less confirm the findings of an existing simulation study. Specifically, the illustration indicates that random forests (a) require more events per variable (EPV) than logistic regression to realize their predictive performance potential and (b) are highly optimistic even when generated with a large number of EPV. It also demonstrates how pre-registration can prevent over-optimistic results, thereby suggesting that the adoption of the proposed approach could lead to more credible methodological statistical research.

Contents

List of Figures

List of Tables

1 Introduction

In the last two decades, it has been repeatedly demonstrated that the findings of many studies cannot be replicated in applied research disciplines such as psychology (Open Science Collaboration 2015), medicine (Ioannidis 2005) and economics (Camerer et al. 2016). Among other things, publication bias, questionable research practices like p-hacking and analytical flexibility contribute to the non-replicability of research findings (Munafò et al. 2017).

In connection with those reasons, the distinction between exploratory and confirmatory research has received increasing attention in recent years. The former is also referred to as hypothesis-generating, data-contingent research or by the term postdiction, and the latter is also called hypothesis-testing, data-independent research or prediction. Exploratory research does not necessarily involve a specified hypothesis and seeks to identify patterns in observed data, thereby generating hypotheses for future studies. These hypotheses can then be tested on new data with a confirmatory analysis, which requires the a priori specification of a clear hypothesis. To make progress in science, both of these stages are essential (Wagenmakers et al. 2012; Nosek et al. 2018; Nilsen et al. 2020; Schwab and Held 2020). However, null hypothesis significance testing and p-values are only valid and retain their diagnosticity for purely confirmatory research (Wagenmakers et al. 2012; Nosek et al. 2018).

In practice, many researchers blur the line between exploratory and confirmatory analyses. This can be problematic, because it opens up the possibility of exploratory findings being misperceived or presented as results from a confirmatory analysis, which can happen unintentionally because of cognitive biases or deliberately by exploiting the analytical flexibility to present a desired result. Whether intentional or not, presenting an exploratory finding as confirmatory leads to overconfidence in the result and increases the chance of it being a false positive result that cannot be replicated (Nosek et al. 2018; Button et al. 2013).

To clearly and transparently distinguish the two types of research, thereby ensuring the purely confirmatory nature of a study, the public registration of analysis plans prior to collecting or accessing data has been suggested (Wagenmakers et al. 2012; Nosek et al. 2018). This practice is called pre-registration and has become increasingly popular in recent years (Simmons et al. 2021). Together with clarifying the distinction between exploratory and confirmatory research, pre-registration addresses questionable research practices like

p-hacking, selective reporting of desired results and HARKing ("hypothesizing after the results are known", Kerr 1998; Hardwicke and Ioannidis 2018).
It has recently been suggested that issues such as publication bias, selective reporting and fishing for significance also affect methodological statistical research, leading to over-optimistic results that are not replicable (Boulesteix et al. 2020). Therefore, it could be argued that clearly distinguishing the two research stages through pre-registration would also be useful in this field. In fact, in the context of real-data benchmark experiments, the adoption of pre-registration and study protocols has been identified as potentially helpful in reference to clinical research where those elements are already standard practice (Boulesteix et al. 2017). For both clinical trial protocols and pre-registration documents for applied research, various templates exist for different study types to aid researchers in specifying all relevant details.
However, while there have been some pre-registered methodological studies,[1] neither the exploratory-confirmatory distinction nor the concept of a confirmatory real-data study have been explored in the context of methodological statistical research. Consequently, no template is available for the pre-registration of this kind of research in particular.
Therefore, the aim of this thesis is to explore, conceptualize and illustrate the idea of a deliberately confirmatory real-data study in the field of methodological statistics. As the central part of the presented approach, this thesis proposes the pre-specification of a comprehensive research protocol, similar to those used in clinical trials, and suggests a template for such a protocol. This study protocol template shall serve as the starting point of a guideline for confirmatory studies in methodological statistical research. Furthermore, the included illustration aims to provide new confirmatory insights on the sample size needed for binary prediction models from a large-scale benchmark experiment.
This thesis is structured as follows. Chapter 2 outlines the initial concept of a confirmatory real-data study in methodological statistical research after reviewing related existing concepts and by discussing relevant aspects. In Chapter 3, the previously detailed approach is applied to a research question regarding the events per variable (EPV) in binary classification and the results of that study are reported. Then, in Chapter 4, the initial study concept is reflected upon in the context of the application and limitations are discussed. Lastly, Chapter 5 summarizes the findings of this thesis.

[1] See https://preregister.science for recent examples.

2 Concept

This chapter outlines the initial concept of a confirmatory study in methodological statistical research. Firstly, Section 2.1, provides a brief overview of pre-registration and connections to clinical research, as both serve as the basis and inspiration for the concept. Section 2.2 follows with a discussion of various aspects of a confirmatory study and its associated research protocol. The components of the proposed study protocol and the concept in general are summarized in form of a template in Section 2.3.

2.1 Background: pre-registration and clinical research

The important clear distinction between exploratory, hypothesis-generating and confirmatory, hypothesis-testing research is often not observed in practice. Pre-registration is recognized as an effective tool to facilitate this distinction and is characterized by the time-stamped registration of a study prior to collecting or accessing data (Nosek et al. 2018). The registration is realized by archiving a document on a public independent registry such as the Open Science Framework, which can be used to register research from all disciplines. The contents and level of detail in pre-registration documents can vary, ranging from just a basic study design to comprehensive research protocols (Munafò et al. 2017).
By publicly registering hypotheses, study design, methods and analysis plan before the beginning of a study, pre-registration addresses several different questionable research practices (QRPs; Hardwicke and Ioannidis 2018; Munafò et al. 2017). These practices include exploiting the researcher degrees of freedom to achieve statistical significance (p-hacking; Simmons et al. 2011), selective reporting of desired results and HARKing ("hypothesizing after the results are known", Kerr 1998). In the context of methodological statistics, specifically real-data benchmark experiments, examples of QRPs are the post hoc exclusion of certain benchmarked methods or datasets and performing many different benchmark variations in the hope of finding the superiority of a particular method (Boulesteix et al. 2017; Nießl et al. 2022a). Whether researchers engage in these practices intentionally or not, pre-registration allows for an accessible assessment of their extent by others who can compare the published analyses and results to the publicly archived study plan, provided the pre-registered study protocol is sufficiently detailed. Thus, one would expect pre-registration to reduce the presence of the mentioned QRPs.

Closely related to pre-registration are Registered Reports (RRs), a publishing initiative launched in 2013 by the journal *Cortex* (Chambers 2013) and referred to as "reviewed pre-registration" (van 't Veer and Giner-Sorolla 2016). As of September 2022, RRs have been adopted by over 300 journals,[1] including interdisciplinary ones that may be suitable for confirmatory methodological statistics studies (Nießl et al. 2022a).
In the RRs format, the conventional publication process with peer review is split into two stages and pre-registration is integrated. In the first stage, before the study is conducted, a detailed research proposal and analysis plan is peer-reviewed. Following this, accepted proposals are offered in-principle acceptance, guaranteeing the publication of the results if the authors adhere to their pre-registered protocol and interpret the findings appropriately. Then, the authors conduct the study and subsequently submit a final manuscript for the second stage peer review. In case of a positive evaluation regarding protocol adherence and results interpretation, the RR article is then published (Hardwicke and Ioannidis 2018).
The RRs publishing model is an extension of pre-registration and, therefore, also mitigates the previously mentioned issues addressed by pre-registration. Additionally, RRs prevent publication bias since the publishing decision is made before results are known and thus not influenced by the nature of the results (Munafò et al. 2017). Furthermore, this adaptation of pre-registration has the added benefit of facilitating outside feedback and discussion by peers in the early stages of the research process.
While the concept proposed in this thesis is based around the idea of pre-registration in general, the suggested contents, possibly with some minor adjustments, should also be sufficient for a first-stage submission of a RR article.

In clinical research, explicitly confirmatory studies have been established and deemed necessary for decades. In the traditional classification of clinical trials into four phases, there is a clear distinction between the exploratory Phase I and Phase II trials and the confirmatory trials in Phase III and Phase IV (Umscheid et al. 2011; Sedgwick 2014).
Essential for those confirmatory studies, motivated by apparent publication bias and in an effort to increase transparency in research involving human subjects, pre-registration has become standard practice for clinical trials in the last 20 years (Munafò et al. 2017; De Angelis et al. 2004). Pre-registration in this context refers to the registration of a trial before patient enrollment and is required by law in many countries (Nosek et al. 2018; Regulation (EU) No 536/2014). Public trial registration is also a requirement for publication in leading biomedical journals (De Angelis et al. 2004) and has been one of the principles in the Declaration of Helsinki since 2013 (World Medical Association 2013).

[1] See https://www.cos.io/initiatives/registered-reports for a list of journals that publish RRs to different extents.

While the entry in a trial registry provides a brief summary, the study protocol serves as the detailed description and is the foundation of a clinical study (Chan et al. 2013a). Not only is the required information included in a trial registration not as detailed as in a study protocol, certain key aspects may not be covered at all in that brief summary. For example, the widely adopted WHO Trial Registration Data Set listing the minimum required 24 registration items does not require researchers to specify the analysis plan at the time of registration (World Health Organization 2018).
Furthermore, it should be noted that, while both basic trial registration and study protocol are compiled before the start of a trial, the pre-registration of the latter has not been nearly as common in the past compared to the former. Recent analyses of randomized clinical trials published in 2012 and 2016 found that most available protocols were made public a long time after the start of the trial (Campbell et al. 2022) and that the vast majority of available protocols is not dated prior to patient enrollment (Spence et al. 2019). This sub-optimal practice is mentioned here to note a critical aspect where the later proposed concept goes beyond its clinical inspirations, in this case by adhering to comprehensive pre-registration as previously described.

Although the conceptualization of a confirmatory study in methodological statistical research is original, transferring concepts from clinical to statistical research is not a novel approach. Boulesteix et al. (2017) suggest that adopting established clinical practices such as sample size calculations, strict inclusion criteria and trial protocols could improve real-data benchmark experiments common in statistical research. The ideas suggested by Boulesteix et al. (2017) are discussed in more detail in various parts of the following section 2.2 along with other considerations.

Due to their established and frequent use, clinical trial protocols and their contents have been the topic of guidelines for decades (Tetzlaff et al. 2012). To improve the quality of trial protocols, the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) statement and checklist were published in 2013 (Chan et al. 2013a). This guideline provides an evidence-based minimum set of 33 items and is endorsed as an international standard (Rivera et al. 2020). In addition to guidelines, an abundance of clinical protocol templates from universities, research organizations, government agencies and journals is available online.
Along with the growing popularity of pre-registration in non-clinical disciplines, several pre-registration templates have also been proposed and are available online (Stewart et al. 2020; see also https://osf.io/zab38/wiki/home/ for a variety of templates). Some of them are generic and applicable in any discipline, while others are specific to certain

fields or study types. Moreover, there are templates specifically for studies using pre-existing data (van den Akker et al. 2019; Mertens and Krypotos 2019). Although developed with psychological research in mind, these templates in particular are worth mentioning given that real-data benchmark experiments are generally conducted with already existing datasets.
The presented concept of a confirmatory study in methodological statistical research is heavily built on the ideas of pre-registration and detailed study protocols as well as existing work such as guidelines or templates.

2.2 A confirmatory real-data study in methodological statistical research: aims and considerations

General considerations and aims

The purpose of a confirmatory study, by definition, is to evaluate pre-specified, so-called a priori, hypotheses. To that end, one must generally have some information forming the basis of these hypotheses, either theoretical in nature or from previous analyses. Without enough information about a phenomenon or method, it may be advisable to first conduct an exploratory study, similar to the process in clinical research. However, when the end goal is a confirmatory study, one could also streamline this two-stage process by splitting data into two subsets. The first subset of data would be analyzed to generate hypotheses and, before testing them with the second subset, pre-registration would ensure the distinction between the two analyses (Nosek et al. 2018; Wagenmakers et al. 2012).

In methodological computational research, a common type of real-data study is a benchmark experiment that analyzes and compares a set of methods with respect to some performance metric. Such a study could certainly be conducted in a confirmatory fashion using pre-registration. However, the scope of the concept discussed here and the intended use of the proposed template is much broader. The presented approach is intentionally rather general to be utilizable for all kinds of hypotheses and all research with at least one hypothesis that will be evaluated using real data. This includes, for example, comparison studies with hypotheses about criteria other than standard performance measures or the investigation of a single method with regard to some pre-defined metric. Independently of the specific hypotheses, by combining a detailed protocol with pre-registration, the approach is designed to address QRPs and maximize transparency and reproducibility.
One study type for which the suggested approach may be particularly suitable is replication studies whose primary purpose is to confirm previous methodological results. Although considered important, they are still rare in statistical research (Boulesteix et al. 2020).

Additionally, while replication studies are meaningful on their own, pre-registration and adherence to a detailed protocol would further increase their credibility and impact.

As implied in the previous section, the impact and effectiveness of pre-registration depends on the comprehensiveness of the archived document. For any field and type of research, McPhetres (2020) describes what an effective pre-registration should contain and accomplish. He states that it should:

> a) Restrict as many RDFs [researcher degrees of freedom] as possible
> b) Detail all aspects of a study's method and analysis
> c) Detail information on decisions made during the planning stages
> d) Specify how the results will be used and interpreted (McPhetres 2020, p. 4)

These four aspects have guided the design of the pre-registration template for confirmatory methodological research suggested in this thesis with regard to its general contents. Additionally, they should serve as general guidance when filling the template with the details of a specific study and determining whether the provided description needs further refinement.
Regarding the level of detail in the described analysis plan, the standard general-purpose pre-registration form by the Open Science Framework states that the author should ask themselves whether there is "enough detail provided to run the same analysis again with the information provided" (Bowman et al. 2020, p. 8). This may be used as a guiding question for all parts of the protocol to ensure it is specific enough and therefore limit the researcher degrees of freedom as much as possible.
Analytic flexibility and lack of detail in pre-specified statistical analysis plans are common issues in clinical trials (Kahan et al. 2020). To address this problematic practice, Kahan et al. (2020) outline a five-point framework for effective analysis pre-specification with the goal of limiting p-hacking. The five covered aspects can easily be adapted for real-data methodological studies; in fact, the original wording is already only somewhat specific to clinical trials. Thus, this framework may also be used as guidance in the protocol writing process.

The overall structure of the suggested protocol template is adapted from existing guidelines and templates such as SPIRIT 2013 to incorporate previous design research and best practices from clinical trials and other applied research fields. Some aspects of those guidelines do not need to be adapted due to the fact that methodological statistical research does not directly involve human subjects and mostly deals with pre-existing, anonymized data. Examples of such aspects are the extensive safety and ethics considerations critical

in clinical trial protocols. Furthermore, there are some important characteristics unique to methodological real-data studies that additionally must be addressed in a research protocol. These will be discussed in the following sections.

Dataset selection and number of datasets

In their analogy to clinical trials, Boulesteix et al. (2017) point out that in real-data benchmark experiments in statistical research datasets play the role of trial participants. While different in nature, the selection of the statistical units is an important aspect in both study types and should reflect the intended area of application and study objectives. Similar to the practice in clinical trials, Boulesteix et al. (2017) advocate for the use of strict inclusion and exclusion criteria in benchmark studies as well as the precise reporting of them. Such criteria could, for example, reference the number of observations, the type of the outcome, the subject matter of the datasets or their source. Strict dataset selection criteria clearly define the scope of a study, and transparently reporting them allows others to fairly assess the generalizability of findings (Boulesteix et al. 2017; Couronné et al. 2018). As mentioned, the post hoc exclusion of certain analyzed datasets in methodological studies can be problematic as it can lead to biased results (Macià et al. 2013). The use of inclusion and exclusion criteria alone does not fully address this issue since the criteria could be adjusted, even if just slightly, after seeing the results. However, by reporting the precise dataset selection criteria in a pre-registration, the a posteriori tuning of these criteria can be avoided.
Given the large impact of the choice of datasets on results (Nießl et al. 2022a,b; Boulesteix et al. 2013), the concept suggested in this thesis addresses the QRP of post hoc exclusions even further to transparently ensure the integrity of a study. It is proposed that the entire dataset selection process, including the check of exclusion criteria, be conducted prior to pre-registration and that the complete list of selected datasets be included in the pre-registration document. Besides making the specific analyzed datasets entirely transparent before their analysis, this approach also allows the study's researchers to check whether the number of datasets is large enough as part of the planning of the study. If the dataset selection is insufficient, this is realized prior to any analyses, which means the researchers can then consider adjustments to the design or inclusion criteria of the study without undermining its confirmatory nature.

The number of studied datasets is another aspect for benchmark experiments as it has, for example, been suggested that method comparison studies are often underpowered due to small numbers of considered datasets (Boulesteix et al. 2013; Boulesteix et al. 2015a). While the proposed concept of a confirmatory methodological study is not limited to large-scale studies with many datasets, an underpowered confirmatory study might not be able to fulfill its intended purposes.

For real-data comparison studies using statistical tests, Boulesteix et al. (2015a) recommend taking power considerations into account during planning to determine an adequate number of datasets. To calculate the number of datasets needed to achieve a given power, they suggest a simple method that is similar to formulas common in clinical trials. Depending on the existing literature for the selected methods and measures, one may need to conduct a preliminary pilot study to obtain the standard deviation estimate necessary for the power calculation.
For some methodological studies, the number of available datasets could also be limited and further sampling might not be practical. In those cases, one may still calculate the power that can be achieved with that limited fixed number of datasets, unless a necessary pilot study would reduce the sample size too much.
Whether the number of studied datasets is based on statistical considerations or practical constraints, its determination or a rationale for it should be documented in the preregistration.

Prior knowledge, transparency, and neutrality

Real-data studies in methodological research generally involve pre-existing datasets. This circumstance poses a risk to the clear distinction between exploratory and confirmatory research that pre-registration intends to protect. After all, this distinction and thus the possibility of completely confirmatory testing rely on the assumption that hypotheses and analysis plans are formulated blind to the data that will be used in the study. Damaged blinding due to prior knowledge can occur in different ways, for example, as a result of reading previous studies or having analyzed some of the data (Nosek et al. 2018). Nosek et al. (2018) note that this includes situations where the prior knowledge of certain data concerns different outcomes than the ones of interest in the planned study.
In the context of methodological statistics, examples of this include knowing how certain methods conceptually similar to the studied methods have performed on the same datasets or how the studied methods have performed on the selected datasets with respect to a different measure. In both situations, the prior data knowledge might influence one's decisions regarding the planned analysis. To parts of methodological research, such as machine learning research, the described issue is arguably particularly relevant in view of the frequent use of the same benchmarking datasets and suites (Friedrich and Friede 2022; Bischl et al. 2021).
Furthermore, the risk of harmful prior knowledge increases with the number of benchmark experiments one conducts and reads about. This is especially the case if these past studies are concentrated on certain research fields, something that seems likely given usually focused research interests.

Since complete blindness is unlikely, researchers should transparently report their prior knowledge in a pre-registration to maximize the validity of the analyses and inferences (Nosek et al. 2018). This includes published and unpublished own work with the selected datasets as well as previous studies by others using the data that one has encountered.

Neutrality is a related aspect that is particularly important for methodological research in general and thus for confirmatory methodological studies. For comparisons of methods, the importance of and need for neutral studies has been repeatedly emphasized in light of possible bias in benchmark experiments that are conducted to demonstrate the performance of a newly developed method (Boulesteix et al. 2013; Weber et al. 2019; Buchka et al. 2021). Boulesteix et al. (2013, p. 8) define neutral comparison studies as benchmark experiments that focus on the comparison itself and are conducted by "reasonably neutral" authors in a rational way. Reasonably neutral authors are ones that do not have a preference among the studied methods and are equally familiar with all of the methods (Boulesteix et al. 2017). These criteria can easily be adapted for confirmatory methodological studies in general. However, Boulesteix et al. (2017) point out that these criteria are difficult to fulfill in practice and that non-neutrality may arise subconsciously.
Especially given the reality that perfect neutrality is rarely possible, it is essential to be transparent about the level of neutrality and familiarity with the studied methods. As with many other pre-registration aspects intended to achieve maximum transparency, a neutrality statement allows others to independently assess the credibility of a study. Additionally, it helps reviewers of the analysis plan determine whether neutrality is a substantial issue and if, for example, blinding strategies need to be implemented.

Planned analyses, contingencies and sensitivity analyses

The statistical analysis plan is arguably the most significant part of any pre-registration and responsible for limiting the researcher degrees of freedom. Besides the previously mentioned overall guidance regarding this aspect, some more content-specific considerations regarding this part are presented here.
First, it should be noted that, although frequentist significance testing and p-values seem to be the most common types of inferences in benchmark experiments, the proposed concept of a confirmatory study is not limited to them. In accordance with the set goal of broad applicability, the template is suitable for analyses using all kinds of information, such as Bayes factors, credible intervals or effect sizes. Whatever inference criteria are employed, the associated cut-off values or decision rules must be included in the analysis plan.
Furthermore, the analysis plan should contain details on the preprocessing, for example, how missing values in the individual datasets will be treated or variables will be trans-

formed. Additionally, one should specify contingencies for common issues and backup analysis plans, if applicable. Common relevant issues for methodological studies include missing values in the evaluation measures, non-convergence of algorithms in general (Boulesteix et al. 2017) and unmet assumptions of statistical analytic techniques. Pre-specifying and justifying analysis options for such issues ensure the integrity of the study, even if something does not go according to the original plan.

An illustration by Nießl et al. (2022a) shows that depending on design and analysis plan choices, benchmark studies can lead to different results. Therefore, it is important to investigate the robustness of the results to changes in the made choices by considering alternative analysis strategies and reporting the corresponding results. In the pre-registration template, alternative analysis strategies are to be included in the section for sensitivity analyses which should be part of every confirmatory methodological study in some form. Given that the performance of a method is dependent on dataset properties (Strobl and Leisch 2022), this thesis also recommends to specify an investigation of the results with regard to dataset characteristics in that section.
In addition to the plan for evaluating confirmatory hypotheses together with a sensitivity analysis, one may also include planned exploratory studies in the pre-registration.

Reproducibility, software and data sharing

As previously specified, the information in the pre-registration should, at minimum, be precise to a degree that it enables others to reproduce the study. For statistical research, this relates especially to the used software and analysis code, all of which should be published. Since the specific implementation of a statistical method could make a difference and implemented default model parameters, for example, can vary, the pre-registration should list the specific software packages (with version numbers) to aid reproductions.
Furthermore, the protocol template requires users to include a data sharing plan. Ideally, all results data and, if possible, the analyzed datasets are openly accessible. Aside from exact reproductions, others could then try alternative analysis strategies or sensitivity analyses beyond those in the original publication to further test the robustness of the results (Hoffmann et al. 2021; Nießl et al. 2022a). Additionally, open results data allows exploratory analyses by others which could form the basis for a different confirmatory study.

Deviations from the plan and reporting of results

Since it is impossible to anticipate all contingencies and deviations from the study plan do occur in practice, it must be addressed how they should be handled. For adjustments

to the protocol after the initial pre-registration but before the start of the study, one can simply update the registration and, if applicable, explain the changes. Reasons for this kind of adjustment include results from planned pilot studies or feedback from others (of course pilot studies and reviews could also happen prior to the initial pre-registration). All updates and amendments to the protocol should be transparently logged, for example in the appendix of the most recent version. Deviations from the protocol that occur while the study is being conducted should be clearly indicated and justified in the reporting of the results without exception, for example in a dedicated section to facilitate the assessment of their extent.

Reporting in general is a crucial factor for the replicability of findings, and the lack of guidance in that regard for methodological statistical research has been pointed out as problematic (Boulesteix et al. 2020). While a comprehensive reporting guideline is beyond the scope of this thesis, the following principles related to pre-registration should increase the reporting quality and transparency. Firstly, to address the QRP of selective reporting, all pre-specified analyses and measures, including planned sensitivity and exploratory analyses, must be reported (Nosek et al. 2018). Secondly, confirmatory analyses and exploratory analyses should be reported in separate sections to clearly distinguish them. Lastly, a link to the registry entry should be provided to make readers aware of the pre-registration. Additionally, one may want to include the entire protocol in an appendix of the report, especially if one wants to refer to it in the text.

Regarding the publication of results, the pre-registration template has a section to specify a dissemination policy. Aside from plans for a traditional peer-reviewed publication, this section should clarify whether the results will be reported in the absence of acceptance by a journal and, if so, where, when and in what form. This is an important point to consider given that not all studies will be published in journals, for instance due to publication bias, which is a phenomenon also present in methodological research (Boulesteix et al. 2015b, 2020).

2.3 Suggested pre-registration protocol template

Section	Expected information (if applicable)
0. Administrative information	a) Study title, author, affiliation of involved people, funding b) Pre-registration link, protocol version and date
1. Introduction	
2. Background	a) Previous work/existing literature b) Description of the studied method(s) c) Description of the collected/involved measure(s) (e.g., performance measures)
3. Study rationale, objectives, research questions and hypotheses	a) Overall study rationale and aims b) Primary objectives, research questions and hypotheses c) Secondary objectives, research questions and hypotheses d) Exploratory objectives and research questions
4. Datasets	a) Population, sources, inclusion and exclusion criteria b) Selection process and its results
5. Prior knowledge and neutrality	a) Known prior work based on the selected datasets, the analyzed measures in that work and its relation to the planned study b) Prior knowledge about the datasets themselves c) Neutrality statement regarding the investigated methods
6. Benchmark experiment plan	a) Benchmark design b) Preprocessing procedure (e.g., handling of missing values or certain kinds of variables in the individual datasets) c) Method implementations and configurations
7. Analysis plan	a) Confirmatory analyses (for each hypothesis separately) - Operationalization of hypothesis and evaluation metric - Statistical techniques to evaluate hypothesis - Inference criteria b) Sample size considerations c) Contingencies and backup plans (e.g., for missing values in the evaluation metric, outliers or assumption violations) d) Alternative analysis strategies and sensitivity analyses e) Exploratory and other planned analyses
8. Software	
9. Dissemination	a) Dissemination plan b) Availability of code, data and materials
References	
Appendix	a) List of selected datasets b) Protocol amendment history

Table 2.1: Suggested study protocol template for the pre-registration of a confirmatory real-data study in methodological statistical research.

3 Illustration

In this chapter, the previously described concept of a confirmatory real-data study in methodological statistical research is illustrated using a large-scale benchmark experiment. This example study investigates the connection between prediction performance and the number of events per variable for two binary classification methods: logistic regression and random forests. Firstly, in Section 3.1, the corresponding research protocol is provided, employing the template outlined in Section 2.3. Then the deviations from the protocol that arose while conducting the study are detailed in Section 3.2. In Section 3.3, the results of the planned and unplanned analyses are presented. Lastly, the findings as well as some limitations of the study are discussed and summarized in the Section Section 3.4.

3.1 Study protocol

Administrative information	
Title	The connection between the number of events per variable and prediction performance: a large-scale real-data study comparing logistic regression and random forests
Author	Felix Julian David Lange
Contributors and their affiliation	Felix Julian David Lange, LMU Munich
Protocol version	1.1
Version date	November 21, 2022

3.1.1 Introduction

Binary classification and the prediction of binary outcomes are common tasks in many applied research fields. In the development of binary prediction models, a frequently considered concept is the number of events per variable (EPV) (Ogundimu et al. 2016). The number of EPV is defined as the ratio of the number of observations in the minority class of the outcome variable to the number of predictor variables (i.e., the degrees of freedom

needed to represent them; van Smeden et al. 2019). Previous studies on the topic of EPV are mostly (a) simulation studies, (b) focused on regression modeling and (c) not examining prediction performance.
This protocol is for a large-scale real-data benchmark experiment that investigates the predictive performance of logistic regression and random forests in relation to the number of EPV. The study aims to confirm two results from a simulation study by van der Ploeg et al. (2014) by analyzing 75 real datasets, and its basic design is as follows. During the benchmark, for each dataset, 24 data subsets with different numbers of EPV (ranging from 5 EPV to 500 EPV) will be sampled. A standard logistic regression model using all available variables and random forests with default hyperparameters will then be generated for every subset. The prediction performance of the two methods will be primarily measured by the area under the receiver operating characteristic curve (AUC) and through repeated 5-fold cross-validation. In addition to the primary confirmatory analyses, the robustness of the results with respect to design and analysis choices and dataset characteristics will be examined.

3.1.2 Background

Previous work/existing literature

The concept of EPV in the context of logistic and Cox regression analyses has been extensively studied through simulations. For these modeling techniques, several rules of thumb have been suggested, recommending how many EPV should be available to develop a prediction model. The most widely adopted minimal sample size criterion, especially in clinical research, is the lower limit of 10 EPV (Ogundimu et al. 2016; van Smeden et al. 2019). Peduzzi et al. (1996), who studied the effect of the number of EPV on the regression coefficients, proposed this rule. However, in recent years, the validity of this rule of thumb and the usefulness of EPV criteria in general have been questioned (van Smeden et al. 2016, 2019).
The only known study investigating statistical techniques other than regression models in this context is by van der Ploeg et al. (2014) and serves as the basis for the planned benchmark. Van der Ploeg et al. (2014) compared the predictive performance of logistic regression, classification and regression trees, support vector machines, neural nets and random forests in relation to the EPV in three simulated datasets. They used AUC as a performance measure and examined the optimism of the generated models (i.e., the difference between mean apparent and mean validated AUC). Van der Ploeg et al. (2014) conclude that, in comparison with logistic regression, modern machine learning techniques like support vector machines, neural nets and random forests require considerably more EPV to reach a stable AUC and small optimism. Additionally, they note that the differ-

ence in absolute performance between random forests and simple logistic regression models is marginal in their examples.

Description of the studied methods

Logistic regression Logistic regression is a commonly used, standard approach to analyze binary variables and models the following conditional probability of one of the two classes of a binary variable of interest Y:

$$P\left(y_i = 1 \mid x_{i1}, \ldots, x_{ik}\right) = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})} \quad ,$$

where $x_{i1}, \ldots, x_{ik}$ are the explanatory variables and $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients. The coefficients are the parameters of the model and estimated through maximum likelihood estimation (Fahrmeir et al. 2021). The number of parameters is also used in the calculation of the number of EPV (van Smeden et al. 2019). Predicted probabilities from a model can be used to classify observations by setting the probability threshold that must be exceeded to assign an observation to the modeled class. This study uses the common default threshold of 0.5 (Couronné et al. 2018).

Random forest Random forest is an ensemble learning technique that was introduced by Breiman (2001). The algorithm involves growing a large number of decision trees based on bootstrap samples and aggregating their results. Compared to a single classification tree, this aggregation of many trees leads to a reduction in variance. When growing one of the classification and regression trees of the ensemble, the splitting of nodes is based on purity and only a random selection of features is considered for each split. This aspect of the random forest method reduces the correlation between individual trees, which further improves the variance reduction (Hastie et al. 2009).
The random forest algorithm has several hyperparameters that can be tuned, such as the number of features considered at each split or the minimum number of observations in terminal nodes (Couronné et al. 2018). However, even without hyperparameter tuning, random forests perform well on prediction problems (Hastie et al. 2009), though the generated models are somewhat hard to interpret (Couronné et al. 2018).

Description of the collected/involved measures

AUC The area under the receiver operating characteristic (ROC) curve (AUC) is a measure of discriminative ability (i.e., how well a model can distinguish between the observations in the two outcome classes). A model with perfect discriminative ability has an AUC of 1, while an uninformative model has an AUC of 0.5 (Steyerberg 2019). The AUC

can also be interpreted as the probability that a random observation with the outcome has a higher estimated outcome probability than a random observation without the outcome (Hanley and McNeil 1982).

Accuracy The accuracy of a prediction model is another discrimination measure and defined as the proportion of correctly classified observations (Metz 1978). As a result of that definition, accuracy values range from 0 to 1 with higher values indicating better predictive performance.

Brier score The Brier score is a measure of overall model performance and incorporates calibration aspects in addition to discrimination aspects. It is a quadratic scoring rule and defined as the mean squared difference between the true outcome class (0 or 1) and the predicted probability. Consequently, Brier scores range from 0 to 1, and a lower Brier score signifies better predictive performance (Steyerberg 2019).

3.1.3 Study rationale, objectives, research questions and hypotheses

Overall study rationale and aims

Complementary to the existing simulation studies, this study intends to provide real-data evidence on the connection between prediction performance and the number of EPV. Its overall aim is to confirm two results from van der Ploeg et al. (2014) using a large number of real datasets.

Primary objectives, research questions and hypotheses

This study has two research questions and hypotheses, both are considered primary. To clearly distinguish them in the protocol as well as the eventual results, they will be referred to as first and second hypothesis, respectively. The first research question is whether random forests require more EPV than logistic regression models to achieve a stable predictive performance. Van der Ploeg et al. (2014) conclude from their simulations that "RF [random forests] need far more events per variable to achieve a stable AUC-value than ... LR [logistic regression]" (p. 13). It is expected that this is also the case for real datasets. To assess this prediction, the following hypotheses will be used:

H_0: Random forests need the same number of EPV as logistic regression to achieve a stable predictive performance.

H_1: Random forests need more EPV than logistic regression to achieve a stable predictive performance.

The second objective of the study is to investigate the optimism of random forest models which is defined as the difference between the performance on the training data and the test set performance. Van der Ploeg et al. (2014) remark that for their three simulated datasets "the optimism of the RF [random forest] models remained high [≥ 0.01] ... even at a large number (over 200) of events per variable" (p. 7) and "did not even converge towards zero at the largest number of events per variable that we evaluated" (p. 9). To confirm this finding in this study, the following hypotheses will be used:

H_0: The optimism of the random forest models generated at 500 EPV is equal to or smaller than 0.01.

H_1: The optimism of the random forest models generated at 500 EPV is larger than 0.01.

Exploratory objectives and research questions

At the time of writing, there are no plans for analyses unrelated to the two previously outlined confirmatory research questions. However, the planned sensitivity analyses described in Section 3.1.7 are exploratory in nature, as their purpose is to examine the impact of certain dataset characteristics on the results and no specific hypotheses regarding those associations are made in advance.

3.1.4 Datasets

Population, sources, inclusion and exclusion criteria

The planned benchmark experiment concerns binary classification, and thus, the dataset population of interest is the set of datasets with binary target variables. The source of the datasets used in this study is the OpenML platform (Vanschoren et al. 2013), where users can upload all kinds of datasets and their machine learning results. As of September 2022, the OpenML database includes 22,160 public datasets, which can be filtered by status ("active"/"verified", "deactivated" or "in preparation") and various dataset characteristics. The origin of the available datasets varies. Some are previously unpublished datasets; others are re-uploads of datasets from other sources, such as the UCI Machine Learning Repository (Dua and Graff 2017) or Kaggle.com. Resulting from uploads of identical duplicates, transformed datasets or sampled datasets, the database often has multiple versions of the same original dataset.
To select suitable datasets, the following eligibility criteria were defined based on other classification benchmarks (Bischl et al. 2021; Grinsztajn et al. 2022) as well as the design of the planned study.

Inclusion criteria:

a) Active (i.e., "verified") OpenML datasets without missing values
b) At least 625 EPV with a minimum of two feature variables
c) Fewer than 10,000 features

Exclusion criteria:

d) Cannot be loaded from OpenML
e) Duplicate of or overlap with another OpenML dataset
f) Missing values have been removed or imputed or rows have been removed in other non-random ways
g) Too little information is available for a reliable assessment
h) Requires taking time or space dependency between observations into account (time-series, stream-like or spatial data) or requires grouped sampling (grouped observations)
i) Artificial datasets as well as simulated datasets that are not connected to a real-world application
j) Fewer than two categorical or numeric features
k) No apparent target variable

The absence of missing values was required, as the treatment of missing data was not meant to be addressed in this study to remove additional complexity with regard to the preprocessing. The minimum number of EPV is 625 so that with 5-fold cross-validation, the training data will always have at least 500 EPV, as this is roughly the largest number of EPV in the study by van der Ploeg et al. (2014).[1] It was determined to limit the number of features to 10,000 to avoid computer memory issues. Moreover, this limit only excluded nine datasets, seven of which also meet at least one of the exclusion criteria.
Datasets where missing values had been removed or imputed were excluded since the inclusion of those datasets would be equivalent to indirectly addressing missing values. Criterion f) also prevents subjective data subsets. Additionally, datasets without sufficient information on their content were excluded, as these datasets cannot be reliably assessed with regard to the other criteria. This information requirement does not mean that there must be a scientific reference or source included in the OpenML dataset description, just that there is enough information to make the selection decision. Criterion h) was added as dependent observations would violate logistic regression assumptions and grouped observations would require more complex, non-standard cross-validation procedures.

[1] Originally, 1,000 EPV were planned as a minimum requirement. However, it was quickly realized that this would reduce the set of eligible datasets too much.

Selection process and its results

The dataset selection process can be broken down into two steps. First, the list of datasets is reduced in `R` through automated checks of eligibility. Second, the remaining datasets are manually reviewed for exclusion, a process that is documented in a spreadsheet. Following this manual selection, it is also determined and noted whether the selected datasets require some form of preprocessing.
In the case of this study, the automated part in `R` reduced the number of potential datasets J from 22,160 to 752. After the manual eligibility check of those, 128 suitable datasets were left. However, this selection included a cluster of datasets containing the same type of data. Specifically, 39 datasets had QSAR (quantitative structure–activity relationship) data with molecular fingerprints as features.[2] Without going into more detail, including this group would clearly be an overrepresentation of this one specific type of data. Therefore, the study only includes one of these 39 datasets, specifically the one with the most observations. Thus, the final dataset selection for this study was a group of 90 datasets. Of the 90 datasets, 15 were used in a pilot study, leaving 75 datasets for the main benchmark experiment.[3] Figure 3.1 shows the dataset selection process in the form of a flowchart.

3.1.5 Prior knowledge and neutrality

Known prior work based on the datasets

The only known work that incorporates a dataset selection approach similar to the one described above is a study by Couronné et al. (2018). In their large-scale benchmark experiment, they compared the prediction performance of logistic regression and random forests using 243 real datasets from the OpenML database. However, due to different inclusion criteria and the continued growth of the OpenML database since 2016,[4] the overlap between their selection and the 90 datasets selected for this study is small: Only ten datasets are included in both the study by Couronné et al. (2018) and this one. Moreover, of those ten, nine are subject to preprocessing in this study and, consequently, might differ substantially from the unprocessed datasets analyzed in Couronné et al. (2018).
As in the planned benchmark, Couronné et al. (2018) evaluated prediction performance using AUC, accuracy and Brier score. Therefore, it is known how both investigated methods perform over a large selection of real datasets from OpenML. However, this knowledge does not affect the study described here since the absolute performance values are not the focus of the planned benchmark and the overlap between the analyzed dataset selections is insignificant.

[2] In the dataset selection spreadsheet, the datasets belonging to this cluster can easily be identified: all their names start with "QSAR-TID".
[3] Lists of the two dataset groups can be found in Table A.1 and Table A.2 in the protocol appendix.
[4] Couronné et al. (2018) selected the datasets for their study in October 2016.

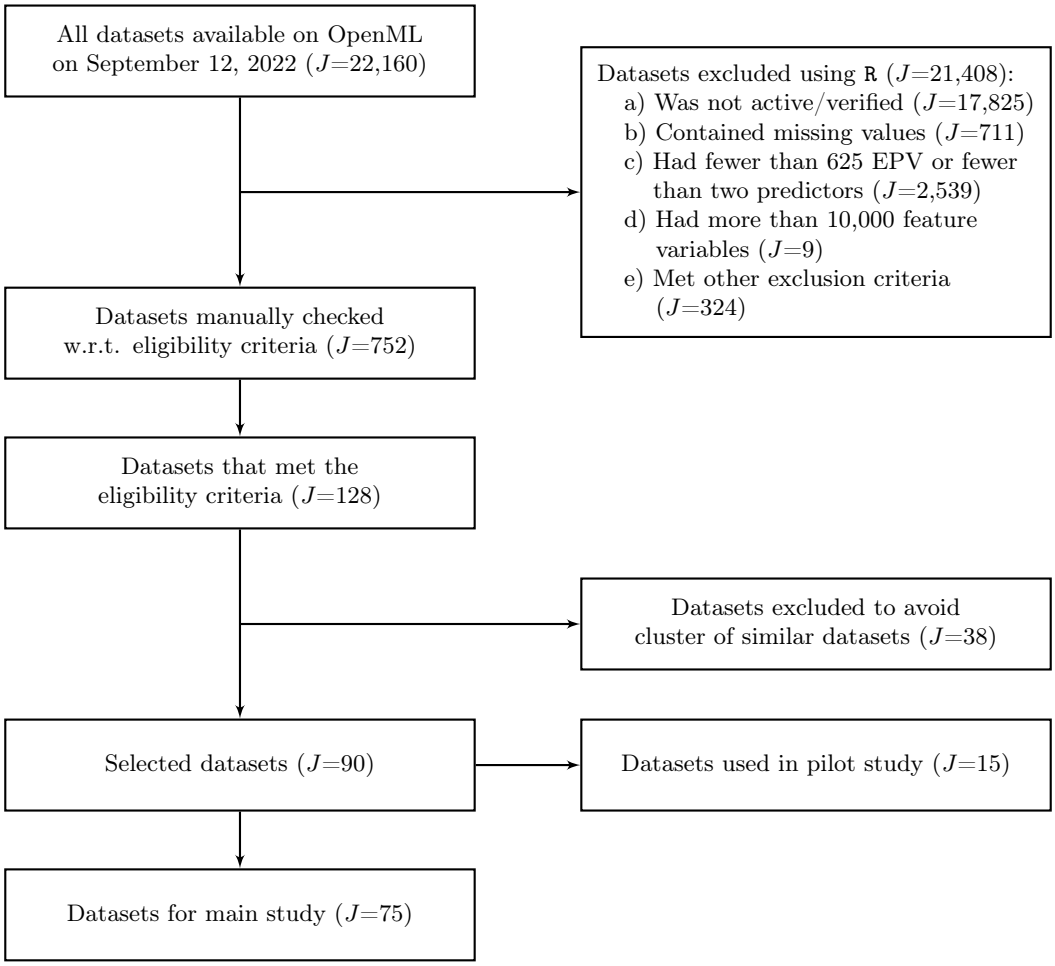

Figure 3.1: Dataset selection flowchart. Regarding the exclusions using `R`, the order in which the exclusion reasons are listed is also the order in which they were determined.

Prior knowledge about the datasets themselves

The author of this protocol has not analyzed any of the datasets that will be used for the main benchmark experiment.[5] However, during the dataset selection process, in order to check the eligibility criteria and determine if preprocessing is needed, the datasets were inspected to a certain extent. These superficial examinations were limited to extracting outcome class sizes and determining the type of data in variables. Furthermore, even within those limitations, the inspections never went on longer than necessary to make the decision at hand.

[5] 15 of the 90 datasets were used in a pilot study that is described in Section 3.1.7, leaving 75 datasets for the main benchmark experiment. None of these 75 datasets have been analyzed.

Neutrality statement for the investigated methods

Regarding the two methods investigated in this study, the author has no personal preference or conflict of interest and is more or less equally familiar with both of them. Additionally, the author has no publications on either method but has previously used logistic regression models and random forests on individual datasets.

3.1.6 Benchmark experiment plan

Benchmark design

To evaluate logistic regression and random forests for different dataset sizes, this study employs repeated stratified cross-validation (CV) and training data subsets. For each analyzed dataset, the benchmark experiment involves the following steps. First, in the first layer of sampling, stratified 5-fold CV is repeated 10 times, resulting in 50 iterations. Then, for every first-layer iteration (i.e., every 80/20 train-test split), the training data is sampled into subsets corresponding to different numbers of EPV (second layer of sampling). Specifically, 24 training data subsets for numbers of EPV $\in \{5, 10, 20, 30, 40, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500\}$ are created using stratified random sampling. The subsets are sampled in a nested way such that one subset contains all the previous smaller subsets. Figure 3.2 shows the two layers of sampling for one dataset. Each training subset as well as the full training data is then used to train a logistic regression model and a random forest model. The resulting 25 models for each method for this one CV fold are evaluated using the same test data.

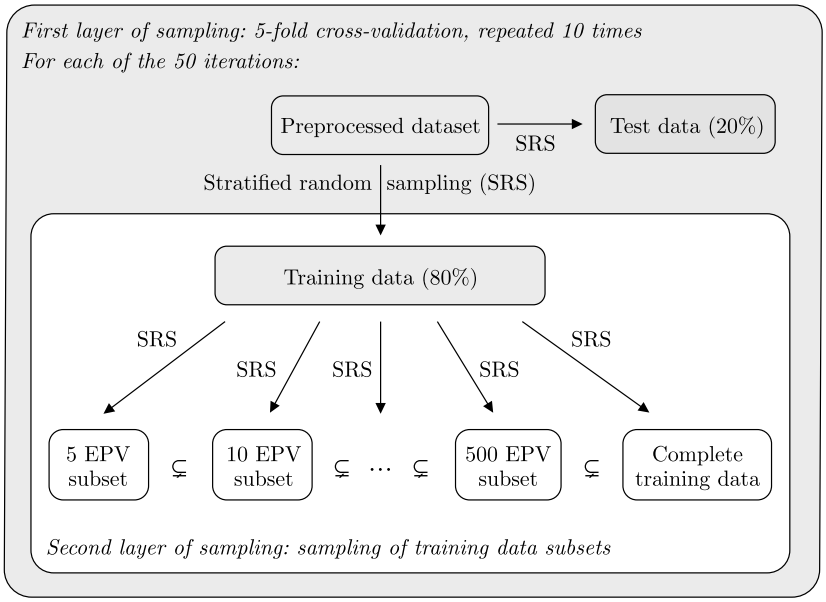

Figure 3.2: Visualization of the two layers of sampling used for a given dataset in the study.

Over all CV iterations for one dataset, this two-layer sampling procedure therefore leads to 50 (iterations) $*$ 25 (training sets) = 1,250 models for each method. For every model, the three mentioned performance measures — AUC, accuracy and Brier score — are calculated using the corresponding training and test data, resulting in six performance values for each trained model.

Preprocessing procedure

Before the benchmark experiment, the datasets are preprocessed to a certain degree. Related to and in preparation for this preprocessing, some issues connected to the `mlr3` package employed for the benchmark must be addressed. The `mlr3` setup works by converting the OpenML datasets into `mlr3` task objects using the metadata supplied by OpenML. On these `mlr3` tasks, the machine learning algorithms are then applied. However, for some datasets the OpenML metadata is insufficient or incorrect, which will cause errors when trying to convert them into `mlr3` tasks or issues after the conversion during the benchmark. To avoid such problems, the following preparations must be made: (a) if `mlr3` cannot determine the task type for a dataset, determine the task type; (b) if `mlr3` cannot identify a default target variable, identify the target and (c) correct the target type, if necessary (some binary classification tasks are recognized as regression tasks because the target is wrongly encoded as numeric).

Regarding the actual preprocessing, the following steps are performed as necessary:

1. Remove certain variables (e.g., IDs and other character variables, redundant targets)
2. Correct feature types for incorrectly encoded features
3. Turn all categorical variables into dummy variables
4. Remove constant feature variables
5. Remove sparse feature variables (less than 10% of values belonging to the minority class for dummy variables and less than 10% non-zero values for numeric features)
6. Transform the target to a binary variable (based on median for regression tasks and majority class vs. rest for multiclass classification tasks)
7. Sample the maximum number of feature variables p possible while ensuring at least 625 EPV

Step 3 was incorporated to facilitate the calculation of the number of parameters and to ensure the same feature encoding across all datasets. Step 5 was included to reduce the risk

of separation and constant features in the samples, especially considering the low number of observations in the smaller training subsets. The decision to include non-binary problems and dichotomize their targets in step 6 was made to increase the number of available datasets.[6] Furthermore, it is argued that the apparent issues with dichotomization (e.g., loss of information, discriminative ability and statistical power) are not that relevant to the primary research questions investigated in this study.
Table A.3 in the protocol appendix summarizes the preprocessing outcomes for the 90 selected datasets.

Method implementations and configurations

For the logistic regression models, a simple model with all available predictors is fit using the standard implementation in the `glm()` function from the `R stats` package (R Core Team 2022). More complex models with, for example, interaction terms, quadratic terms or splines are not considered, as this would require careful individual modeling for each dataset which is not feasible for a large number of datasets.
The random forest models are fit using the implementation in the `ranger R` package and its default parameters (Wright and Ziegler 2017). No hyperparameter tuning is performed as this also seems unfeasible given the 1,250 random forest models for each dataset.

3.1.7 Analysis plan

First hypothesis

Operationalization of hypothesis and evaluation metric To operationalize the hypothesis that random forests need more EPV than logistic regression to achieve a stable predictive performance, it first has to be defined what is meant by "stable predictive performance". In the paper by van der Ploeg et al. (2014) from which this hypothesis originates, the stability (or plateauing) of prediction performance in terms of AUC seems to have been assessed visually using learning curves. Furthermore, they also used the word "good" instead of "stable" to describe the desired predictive accuracy and assumed that prediction performance monotonically increases with sample size. Taking these aspects into consideration for the planned study, it was determined that a model has achieved a stable or good predictive performance if its AUC is within 5% of the maximum achievable AUC. Connecting this criterion to the number of EPV, the relevant evaluation metric for the first hypothesis is the minimum number of EPV for each method at which the AUC of the generated model is at least 95% of the maximum achievable AUC.

[6] Excluding datasets with non-binary targets would have reduced the number of available datasets from 90 to 37.

In the context of the benchmark design described above, the evaluation metric is formalized as follows. For CV iteration $i = 1, \ldots, 50$ of a given dataset j, let $AUC_{ij}^{test}(n)$ be the test data AUC of the model generated using the training data subset with n EPV, and let $MaxAUC_{ij}^{test}$ be the test data AUC of the model generated using the full training data. The minimum number of EPV at which a method achieves a good predictive performance is then defined for iteration i of dataset j as

$$(EPV_{min})_{ij} = \min\left\{n \in EPV \mid AUC_{ij}^{test}(n) \geq 0.95 \cdot MaxAUC_{ij}^{test}\right\} \quad ,$$

where $i = 1, \ldots, 50$, $j = 1, \ldots, J$, and $EPV = \{5, 10, 20, 30, 40, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500\}$.
Therefore, EPV_{min} corresponds to the number of EPV of the smallest training data subset for which the defined criterion is met.
Let $(EPV_{min}^{RF})_{ij}$ and $(EPV_{min}^{LR})_{ij}$ denote this minimum number of EPV for random forests and logistic regression, respectively. To evaluate whether random forests need more EPV than logistic regression, the ratio between the EPV_{min} for the two methods $(EPV_{min}^{RF})_{ij}/(EPV_{min}^{LR})_{ij}$ is calculated for every iteration. The ratios are then aggregated over all 50 CV iterations using the geometric mean to estimate the ratio for a given dataset j:

$$\begin{aligned}\widehat{EPV_{min}Ratio}_j = \overline{EPV_{min}Ratio}_j &= \left(\prod_{i=1}^{50} \frac{(EPV_{min}^{RF})_{ij}}{(EPV_{min}^{LR})_{ij}}\right)^{\frac{1}{50}} \\ &= \frac{\left(\prod_{i=1}^{50}(EPV_{min}^{RF})_{ij}\right)^{\frac{1}{50}}}{\left(\prod_{i=1}^{50}(EPV_{min}^{LR})_{ij}\right)^{\frac{1}{50}}} = \frac{\overline{(EPV_{min}^{RF})}_j}{\overline{(EPV_{min}^{LR})}_j} = \frac{\widehat{(EPV_{min}^{RF})}_j}{\widehat{(EPV_{min}^{LR})}_j}\end{aligned}$$

This ratio is used to evaluate the hypothesis. If the ratio is exactly 1, random forests and logistic regression models need the same number of EPV to achieve a good predictive performance. If it is > 1, random forests need more EPV, and if it is < 1, logistic regression models need more EPV. Therefore, the null hypothesis is that the population mean ratio ($\mu_{RF/LR} = \mu_{RF}/\mu_{LR}$) is equal to 1. Since ratios are inherently asymmetric and generally not normally distributed, the statistical analysis is performed on the (natural) log-transformed ratios, leading to the following final hypotheses:

$$\begin{aligned}H_0\text{: } &\log(\mu_{RF}/\mu_{LR}) = \log(\mu_{RF}) - \log(\mu_{LR}) = \log(1) = 0 \\ H_1\text{: } &\log(\mu_{RF}/\mu_{LR}) = \log(\mu_{RF}) - \log(\mu_{LR}) > \log(1) = 0\end{aligned}$$

Statistical techniques to evaluate hypothesis After applying the natural log-transformation above, the null hypothesis can be tested with a one-sided t-test for differences between paired measurements, assuming that the log differences are approximately normally distributed. With d_j denoting the difference for dataset j, the sample mean of the log differences $\bar{d}$ is calculated as follows:

$$\bar{d} = \frac{1}{J}\sum_{j=1}^{J} d_j = \frac{1}{J}\sum_{j=1}^{J} \log((\widehat{EPV_{min}^{RF}})_j) - \log((\widehat{EPV_{min}^{LR}})_j)$$

In addition to the t-test, the distributions of the log differences as well as the minimum numbers of EPV will be visualized with boxplots. Furthermore, summary statistics such as mean, median, quartiles and standard deviations will be computed for both the log-transformed and the untransformed minimum numbers of EPV.

Inference criteria The inference will be based on the p-value and confidence interval of the paired t-test using a nominal significance level of $\alpha = 0.05$. As the hypothesis above is one of two hypotheses tested in this study, a multiple test adjustment must be performed. To control the family-wise error rate, the method by Bonferroni (1936) will be applied. Even though the Bonferroni correction is rather conservative, it was chosen because it can be easily incorporated into sample size calculations (Vickerstaff et al. 2019). As a result of adjustment, the first hypothesis will be tested at $\alpha = 0.05/2 = 0.025$.

Second hypothesis

Operationalization of hypothesis and evaluation metric For the second hypothesis, the optimism of a model is defined as the difference between the prediction performance on the training data and the test set performance, both measured in terms of AUC. As stated in the hypothesis, only the random forest models trained on 500 EPV are examined for this research question.
For CV iteration $i = 1, \ldots, 50$ of a given dataset j, let $AUC_{ij}^{train}(500)$ and $AUC_{ij}^{test}(500)$ be the train and test data AUC of the model generated using the training data subset with 500 EPV. The optimism of the 500 EPV random forest model is then calculated as follows:

$$RFopt_{ij}^{500} = AUC_{ij}^{train}(500) - AUC_{ij}^{test}(500) \quad , \; i = 1, \ldots, 50 \; , \;\; j = 1, \ldots, J$$

To estimate the optimism for a given dataset j, the values for the 50 CV iterations are aggregated:

$$\widehat{RFopt}_j^{500} = \frac{1}{50} \sum_{i=1}^{50} RFopt_{ij}^{500} \quad , \; j = 1, \ldots, J \; .$$

These J values are used to evaluate the null hypothesis that the population mean optimism of random forest models μ_{RFopt} is equal to or smaller than 0.01 against the alternative hypothesis:

$$H_0\text{: } \mu_{RFopt} \leq 0.01$$
$$H_1\text{: } \mu_{RFopt} > 0.01$$

Statistical techniques to evaluate hypothesis Under the assumption of approximate normality, the above hypothesis can be tested using a one-sample one-sided t-test. The sample mean optimism $\overline{RFopt}$ is given by

$$\overline{RFopt} = \frac{1}{J} \sum_{j=1}^{J} \widehat{RFopt}_j^{500} \quad .$$

Additionally, the distribution of the J optimism values will be visualized in a boxplot, and summary statistics such as mean, median, quartiles and standard deviation will be computed.

Inference criteria The inference will be based on the p-value and confidence interval of the t-test. Like the first hypothesis, the second hypothesis will be tested at the significance level $\alpha = 0.05/2 = 0.025$ as a result of the Bonferroni correction.

Sample size considerations

The sample size for this study (i.e., the number of datasets) was not planned before the dataset selection and was more determined by the number of available datasets and the eligibility criteria. The goal was to include as many datasets as possible in the benchmark. That meant including every eligible dataset while ensuring that the inclusion and exclusion criteria fit the study's objectives.
Although the sample size of the study was not determined through power considerations and is rather fixed, power calculations can be useful in estimating the power of the study. For the paired t-test used to evaluate the first hypothesis of this study, the power $1 - \beta$ of the test to detect a given log difference Δ can easily be calculated for a given number of datasets J, a given standard deviation σ and the determined significance level $\alpha = 0.025$ (Boulesteix et al. 2015a).

Given that no studies with the exact same evaluation metric or hypothesis exist, a small pilot study was conducted to obtain a rough estimate of the standard deviation. To that end, 15 of the 90 selected datasets were randomly sampled and analyzed as planned. The number of datasets in the pilot study was chosen more or less arbitrarily with the intent to balance the reduction of datasets available for the main study and the reduction in precision of the pilot study variance estimate. The pilot study yielded the standard deviation estimate $\hat{\sigma}_{pilot} = 0.96$ for the log difference. To adjust and inflate this imprecise estimate, the so-called upper confidence limit (UCL) approach by Browne (1995) was employed. Specifically, the upper confidence levels 80% and 95% for the estimated standard deviation were considered based on previous studies (Whitehead et al. 2015). Lastly, the scenario of the standard deviation being twice as high as the estimate from the pilot study was included in the power considerations.
Using these four standard deviation estimates, Figure 3.3 illustrates the power calculations for various values of J and Δ. It shows that the study should generally have sufficient power. With the 75 available datasets, the t-test might not be able to detect a rather small log difference of 0.405 (indicating that random forests need 50% more EPV than logistic regression). However, an effect of 0.693 (corresponding random forests needing twice as many EPV as logistic regression) is detected with at least 80% power under all considered standard deviation scenarios. Should the true effect be even larger, as van der Ploeg et al. (2014) suggest, it should be detected with near 100% power under the assumption that the standard deviation is within the considered range.
The protocol appendix includes analogous plots for the second hypothesis (see Figure A.1).

Contingencies and backup plans

This section addresses two issues that might arise during the study: missing values and unmet assumptions of the chosen statistical techniques. The issue of missing values is, in theory, twofold in this study. Firstly, there is the possibility of missing values in the collected performance measures (AUC, accuracy, Brier score). Secondly, missing values may also arise in the calculated evaluation metrics, either because of the mentioned missing AUC performance values or by design. Due to the fact that only the evaluation metrics are interpreted in this study and considering that this interpretation is quite removed from the absolute performance values, only missing values in the evaluation metrics ($RFopt_{ij}^{500}$ and $(EPV_{min})_{ij}$) will be explicitly imputed. Missing AUC performance values will therefore only be addressed indirectly.
It should be noted that the $RFopt_{ij}^{500}$ values will only be missing if the corresponding performance values are missing, but the $(EPV_{min})_{ij}$ values can also be missing if no performance values are missing, namely when no training data subset generates a model with

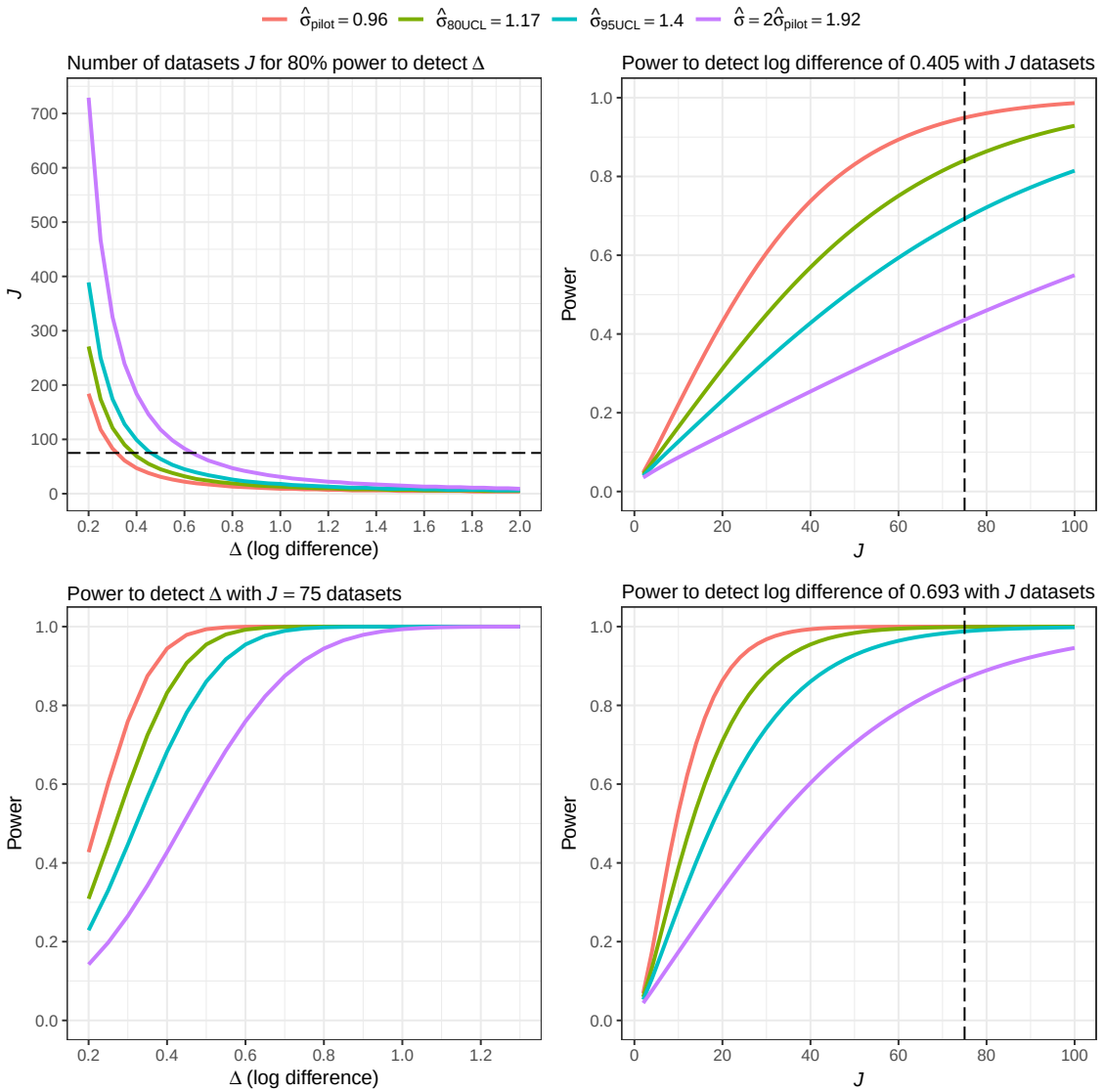

Figure 3.3: Power plots for first hypothesis, so Δ refers to the log difference between the needed numbers of EPV for random forests and logistic regression. The four colored lines represent four standard deviation estimates (or scenarios) based on the pilot study results. The dashed black line indicates the number of datasets that will be analyzed in the main benchmark experiment (75).

an AUC within 5% of the $MaxAUC$. For both evaluation metrics, a 20%-threshold rule is defined to deal with possible missing values. Previous studies have applied similar rules for the imputation of performance values (Bischl et al. 2013; Herrmann et al. 2021). In this study, the 20%-threshold rule is specified as follows. If the values for $(EPV_{min})_{ij}$ or $RFopt_{ij}^{500}$ are missing in less than 20% of the 50 CV iterations, the missing values are imputed using the mean value of the remaining iterations. If the values for $(EPV_{min})_{ij}$

or $RFopt_{ij}^{500}$ are missing in more than 20% of the 50 CV iterations, the missing values are imputed using the worst possible value. For $RFopt_{ij}^{500}$, the "worst possible value" is defined as 0.5, representing the scenario of a perfect $AUC_{ij}^{train}(500)$ of 1 minus an $AUC_{ij}^{test}(500)$ of 0.5 that corresponds to random prediction. For $(EPV_{min})_{ij}$, the "worst possible value" is the number of EPV in the full training data of iteration i (i.e., the number of EPV used to generate the $MaxAUC_{ij}^{test}$ in the first place).

The possibility of statistical assumptions of the employed t-tests not being met is addressed with the following back up plans. For the t-tests, the assumption of approximate normality is particularly critical and will be checked with Shapiro-Wilk tests and Q-Q plots. In case of a violation, the non-parametric Wilcoxon signed-rank test will be performed as an alternative.

Alternative analysis strategies and sensitivity analyses

Two types of sensitivity analyses will be performed for the results of the benchmark. Firstly, the connection between dataset characteristics and the results will be explored for both research questions. For that purpose, the 75 analyzed datasets are grouped by certain dataset characteristics. The subgroups will then be re-analyzed separately and differences in the results will be visualized. Table 3.1 lists the dataset meta-features and corresponding categories or cut-off values that are used to define subgroups of datasets.

Meta-feature	**Categories/cut-offs**
Original task type prior to dichotomization	Regression tasks, multiclass classification tasks, binary classification tasks
Types of features	Only numeric features, only binary features, features of both types
Number of observations (n)	5,000, 10,000, 30,000, 100,000
Number of features (p)	3, 5, 10, 20
Total number of EPV (EPV^{tot})	700, 1,250, 5,000
Events fraction (pct^{evt})	0.2, 0.4, 0.5

Table 3.1: Considered dataset meta-features and chosen meta-feature values to define dataset subgroups for the first type of sensitivity analysis.

For the numerical meta-features, the cut-offs are based on the distribution of the respective feature and all subgroups contain at least ten datasets. The subgroups based on the categorical meta-features also contain at least ten datasets, except for the subgroup with the originally multiclass datasets (only has nine datasets).
The results of the subgroup analysis for each of the two research questions will be visualized in boxplots, similar to the ones shown in Couronné et al. (2018, Fig. 5).

The second type of performed sensitivity analysis examines which results alternative analysis strategies would have obtained. It is only conducted for the first research question as the operationalization of that hypothesis was heavily based on several design and analysis choices. The decisions made and the alternative options considered in the sensitivity analysis are specified in Table 3.2.

Design or analysis choice	**Default**	**Alternative options**
Performance measure	AUC	Accuracy, Brier score[7]
Performance threshold in evaluation metric (stability threshold)	95%	90%, 99%
Handling of missing values in evaluation metric	20%-threshold rule	Worst, mean, weighted
Aggregation of evaluation metric values across CV iterations (within a dataset)	Geometric mean	Median
Aggregation of evaluation metric values across datasets	Geometric mean	Median

Table 3.2: Considered design and analysis choices and corresponding options for the second type of sensitivity analysis. 'Default' refers to the choices made for this study (i.e., the ones used in the primary, confirmatory analyses).

While most of the listed options are fairly self-explanatory, the alternative strategies for handling missing values are briefly described in the following. For a given CV iteration

[7]Since the perfect Brier score is 0 (and not 1, which is the perfect value for AUC and accuracy), the evaluation metric for the first hypothesis is defined slightly differently when the Brier score is used. The definition of the evaluation metric for accuracy and Brier score can be found in Section A.4 in the protocol appendix.

with a missing minimum number of EPV, the "worst" approach means always imputing the number of EPV of the full training data (i.e., the highest number of EPV possible for a dataset). The "mean" option corresponds to replacing the missing value with the mean of the other iterations. If the specified threshold is not achieved in any iteration, the "mean" approach also fills in the "worst" value. The "weighted" imputation method combines the previous two options by weighing them according to the proportion of CV iterations requiring imputation (see Nießl et al. 2022a for a similar approach).
Similar to Nießl et al. (2022a), all combinations of the options listed in Table 3.2 will be examined, resulting in $3 * 3 * 4 * 2 * 2 = 144$ combinations and aggregated results for the first research question. Each choice's impact on the aggregated minimum numbers of EPV for random forests and logistic regression will be assessed using scatterplots where the 144 different results will be visualized. Additionally, boxplots grouped by the considered options will be generated to compare the ranges of the results.

Exploratory and other planned analyses

Besides the mentioned analyses, statistics and visualizations, one additional analysis is planned at the time of writing. Following van der Ploeg et al. (2014), learning curves will be generated that show the connection between the prediction performance or the optimism (on the y-axis) and the number of EPV (on the x-axis).

3.1.8 Software

The study is conducted using `R` (version 4.2.1; R Core Team 2022), and the key `R` packages are noted here. The OpenML datasets are accessed via the `OpenML` package (version 1.10.0; Casalicchio et al. 2019). The benchmark experiment is performed using the `mlr3` package (version 0.14.0; Lang et al. 2019), and within the `mlr3` framework, the random forests are fit using the `ranger` package (version 0.14.1; Wright and Ziegler 2017). The `tidyverse` packages (Wickham et al. 2019) are used for data manipulation and visualization.
A comprehensive list of all packages and versions used for the dataset selection, the benchmark experiment and the analysis can be found in the session information that will be provided in the electronic appendix after the completion of the study.

3.1.9 Dissemination

Dissemination plan

At the time of writing, there are no plans to publish the results of the planned study in a scientific journal or elsewhere.

Availability of code, data and materials

All datasets considered, selected and analyzed in this study are publicly available on OpenML.org and can easily be downloaded, either directly from the website or through programming interfaces. Individual datasets are uniquely identified by their "data.id", which are included in the lists of selected datasets in the appendix as "Data ID".
All code used for the dataset selection, the benchmark experiment and the analysis will be provided in the electronic appendix along with instructions on how the results can be reproduced. All results data and the spreadsheet documenting the check of the eligibility criteria will also be provided.

3.2 Deviations from the study protocol

With regard to the primary confirmatory analysis of the two pre-specified hypotheses, there were no deviations from the study protocol (though the alternative, non-parametric test that was planned as a contingency had to be performed). However, the sensitivity analyses were conducted slightly differently than originally planned to present a more comprehensive assessment of the results' robustness. Firstly, Spearman's correlation coefficients were computed to evaluate the relationship between the metrics and the dataset characteristics from an additional angle. Secondly, the examination of alternative analysis strategies was expanded regarding the considered alternative options. In addition to the three previously specified performance thresholds (90%, 95%, and 99%), two further thresholds, 92.5% and 97.5%, were included as alternative options. Furthermore, dataset characteristics were incorporated in this analysis by considering dataset subgroups, again similar to Nießl et al. (2022a). For each of the four numerical meta-features listed in the protocol (Table 3.1), the 75 datasets were split into two subgroups based on the median of the meta-feature, resulting in eight subgroups. Overall, these adjustments to the considered analysis options increased the option combinations from 144 to 2,160 (9 dataset selections (all datasets and eight subgroups) $*$ 3 performance measures $*$ 5 performance thresholds $*$ 4 imputation methods $*$ 2 $*$ 2 aggregation methods). The third and last change to the sensitivity analyses was the addition of a boxplot that shows more directly how specifically changes in the performance measure and threshold would have influenced the results of the confirmatory analysis of the first hypothesis.
Finally, as part of the exploratory analyses, comparisons between logistic regression and random forest with respect to absolute performance values are presented in reference to Couronné et al. (2018), which were also not planned in the protocol.

3.3 Results

The following presents the results of the planned and additional analyses, roughly in the same order as their descriptions in the analysis plan. First, the results of the primary confirmatory analysis of the two pre-specified hypotheses are described in Section 3.3.1. The sensitivity of these results to dataset characteristics and study design choices is examined in Section 3.3.2. Lastly, in Section 3.3.3, the exploratory findings from the additional planned and unplanned analyses are presented.

3.3.1 Confirmatory analyses

First hypothesis

The evaluation of the hypothesis that random forests require more EPV than logistic regression to reach stable or good predictive performance yielded the following results. To come within 5% of the AUC performance of the respective full training dataset model, logistic regression models needed 7.18 EPV on average (geometric mean [GM]; first quartile [Q1]: 5.55, median: 6.51, third quartile [Q3]: 8.93), while random forest models required 19.41 EPV on average (GM; Q1: 6.03, median: 16.32, Q3: 46.49). Consequently, over the 75 datasets, the geometric mean of the ratio between the minimum numbers of EPV for random forests and logistic regression was 2.70 (Q1: 1.00, median: 2.36, Q3: 6.32). For 51 of the 75 datasets (68%), random forests had a higher average minimum number of EPV than logistic regression, while the opposite was true for 16 datasets (21.33%). For the remaining eight datasets (16.67%), random forest and logistic regression models required the same number of EPV to surpass the 95%-threshold, namely the smallest possible value of 5 EPV.
In all CV iterations of all datasets at least one of the 24 EPV subsets came within 5% of the $MaxAUC_{test}$, meaning that there were no missing values in the evaluation metric and, thus, no imputation was necessary.

As specified in the protocol, the natural logarithm of the aforementioned ratio, the so-called log difference, serves as the basis for the statistical test of the first hypothesis. The arithmetic mean of the log difference between the minimum EPV numbers for random forests and logistic regression was 0.995 (Q1: 0.000, median: 0.860, Q3: 1.843). Figure 3.4 shows the distribution of the mentioned metrics for the 75 analyzed datasets, with the lower-right panel showing the log difference values. The reported summary statistics and the other boxplots in Figure 3.4 clearly demonstrate that the range and variability of values in general differs greatly between the two investigated modeling approaches.

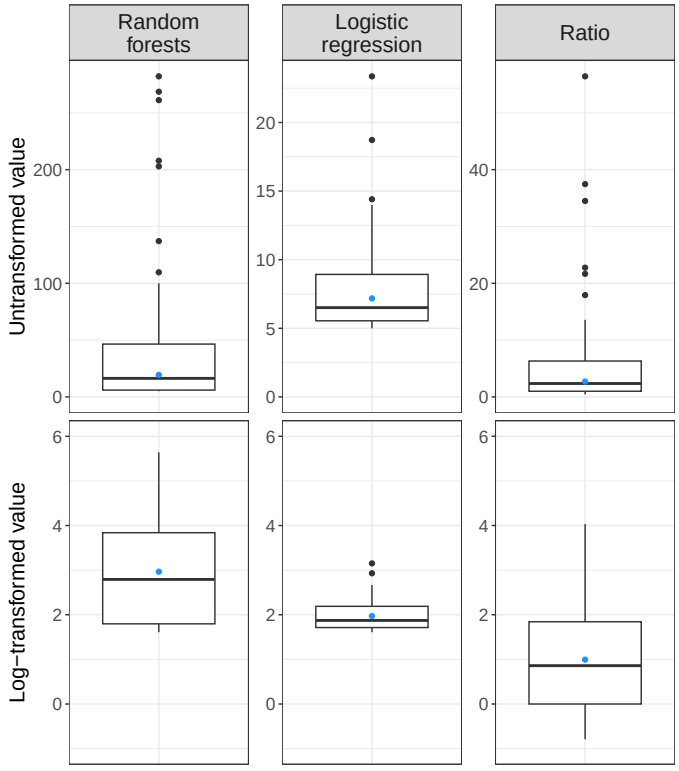

Figure 3.4: Minimum numbers of EPV for random forests and logistic regression as well as the ratio between them for the 75 analyzed datasets. The blue points show the geometric means in the first row and the arithmetic means in the second row.

In order to determine the appropriate statistical test, the normality of the distribution of the log differences was examined.[8] The Shapiro-Wilk test yielded a significant p-value ($p < 0.001$), so the null hypothesis that the population is normally distributed was rejected. Therefore, as specified in the protocol, the non-parametric Wilcoxon signed-rank test for paired samples was performed. The result of this test was a significant p-value ($p < 0.001$), providing clear evidence against the hypothesis that random forests need the same number of EPV as logistic regression to achieve a stable predictive performance.[9]

Second hypothesis

The second hypothesis concerned the optimism of random forest models, which was defined as performance on the training data minus performance on the test data, both measured in terms of AUC. It was hypothesized that the optimism of random forest models trained using 500 EPV is larger than 0.01.

[8] A histogram with a density curve as well as the Q-Q plot for the log difference values is included in Appendix B (Table B.1).

[9] The planned paired t-test was also performed, and its result would have led to the same conclusion (p-value < 0.001, 95%-CI for the log difference: $[0.775, +\infty]$).

Across the 75 datasets, the (arithmetic) mean optimism of the 500 EPV random forests was 0.107 (Q1: 0.008, median: 0.0637, Q3: 0.141). The boxplot in Figure 3.5 shows the distribution of the optimism values.

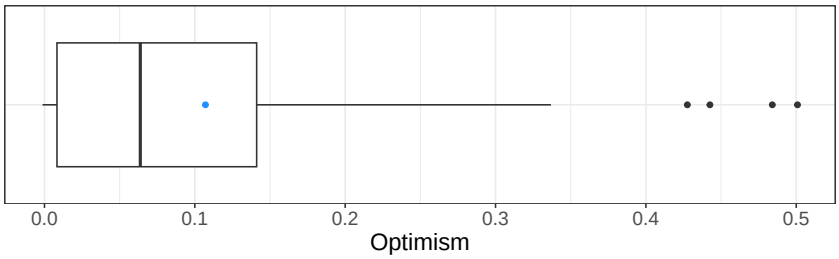

Figure 3.5: Mean optimism of random forests trained using 500 EPV for the 75 analyzed datasets. The blue point shows the arithmetic mean.

The inspection of the Q-Q plot and a histogram (see Appendix B for both) made clear that the distribution of the optimism values is not approximately normal. The Shapiro-Wilk test also had a p-value smaller than 0.001. Consequently, the planned non-parametric alternative, the Wilcoxon signed-rank test, was conducted again, yielding a significant p-value ($p < 0.001$). Therefore, just as with the first hypothesis, the statistical test confirmed the second pre-specified hypothesis of this study.[10]

3.3.2 Sensitivity analyses

Sensitivity of results to dataset characteristics

To examine the robustness of the results from the previous two subsections to dataset characteristics, the 75 datasets were divided into subgroups based on six meta-features: Original task type prior to dichotomization, types of features, number of observations (n), number of features (p), total number of EPV (EPV^{tot}) and events fraction (i.e., percentage of observations in the minority outcome class, pct^{evt}).

Figure 3.6 illustrates the variation in the values of the tested metrics between the subgroups defined by the first two categorical meta-features.

For both confirmatory hypotheses, the subgroup results differ from the results in Section 3.3.1 in roughly the same way. Compared to the analysis with all datasets, random forests and logistic regression seem to perform more similarly in terms of required EPV on tasks whose target did not have to be dichotomized or that had binary features. Moreover, for datasets with at least some binary variables, the median log difference is close to 0, which would be equivalent to random forest and logistic regression models needing the

[10] The planned one-sample t-test was also performed, and its result would have led to the same conclusion (p-value < 0.001, 95%-CI for the optimism value: $[0.083, +\infty]$).

same number of EPV. For the same subgroups, the optimism of the random forests is also lower compared to the original results. The analysis of the dichotomized regression tasks alone does not produce noticeably different metrics than the one with all 75 datasets, but slightly higher log differences and optimism values would be reported if only transformed multiclass classification tasks or datasets with exclusively numeric features were analyzed. Of course, the fact that the separate analysis of the subgroups affects the log difference and optimism values in a similar manner may also be caused by confounding.

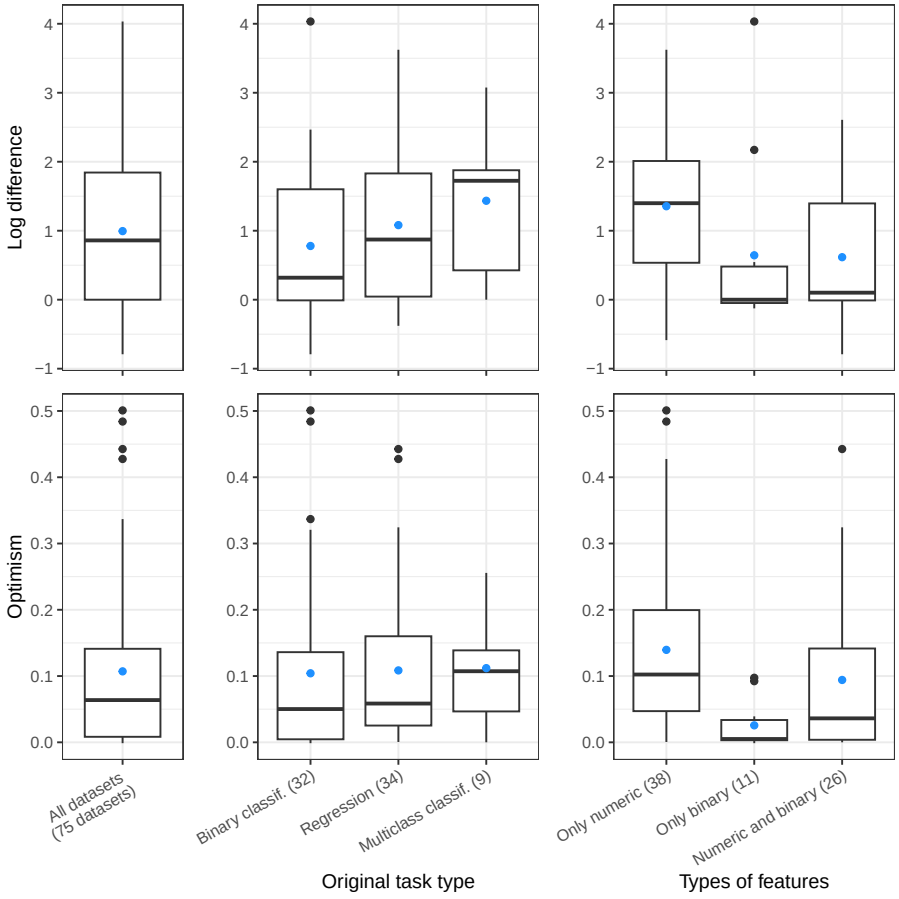

Figure 3.6: Mean log difference and optimism for subgroups of the 75 analyzed datasets based on categorical meta-features. The blue points show the overall arithmetic means.

To investigate the connection between the results and the remaining four numeric meta-features, dataset subgroups were defined based on different cut-off values for each meta-feature. These cut-off values were documented in the protocol and chosen in such a way that each subgroup had at least ten datasets. The boxplots in Figure 3.7 display the values of the tested metrics across the datasets in a given subgroup. For each meta-feature cut-off value t, this figure shows the values for the datasets below the cut-off and

for the remaining datasets. Additionally, the boxplots from Section 3.3.1 are included in the rightmost position of each panel, and the histograms in the bottom row show the distribution of the meta-features' values (log scale).

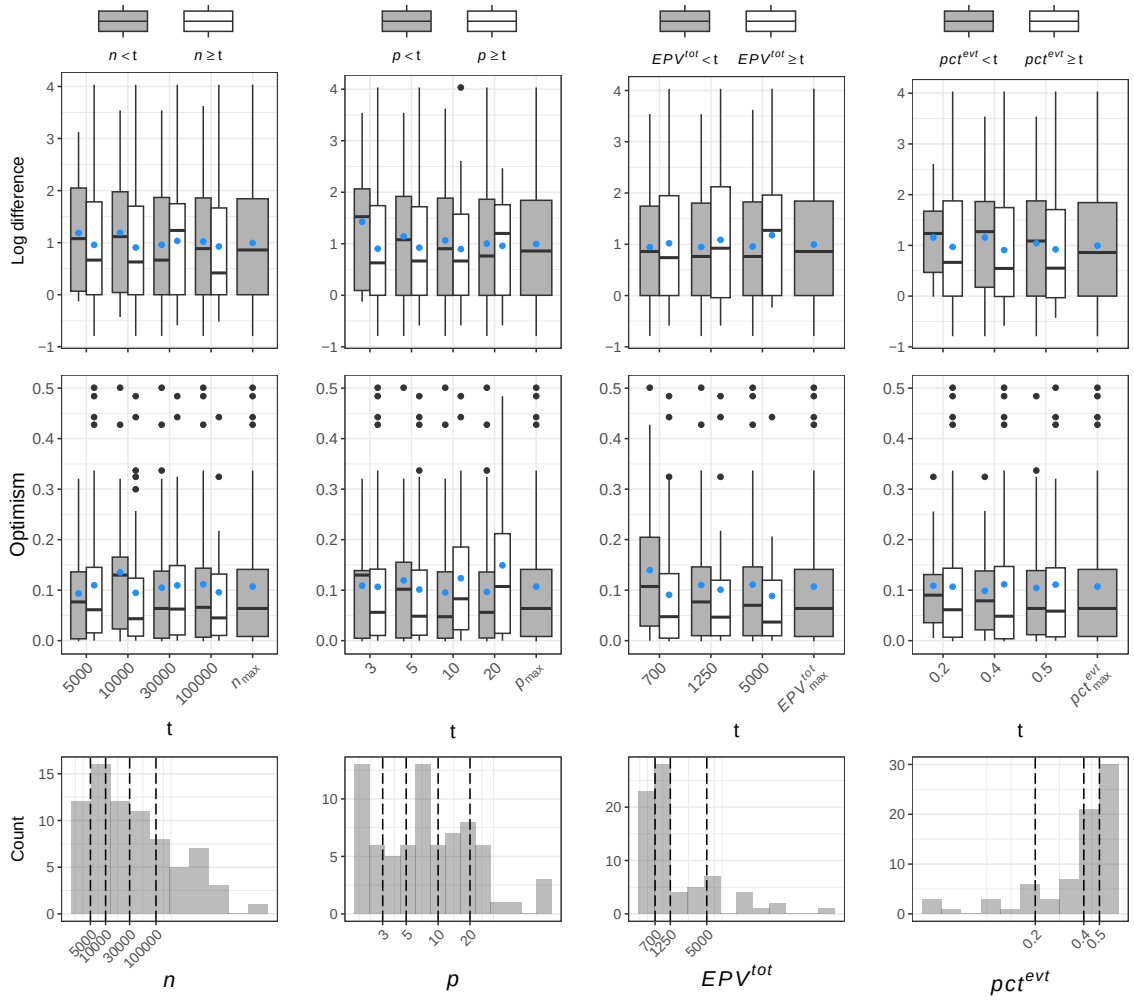

Figure 3.7: Mean log difference (top) and optimism (middle) for subgroups of the 75 analyzed datasets based on numerical meta-features. The blue points show the arithmetic means. On the bottom, histograms for the numerical meta-features are displayed, with vertical lines for the cut-off values defining the subgroups (x-axis is on a logarithmic scale).

Overall, for both metrics separately, the interquartile ranges only vary slightly across the different dataset groups with few exceptions, though there is some variation in the subgroup medians and, to a lesser extent, in the subgroup means. For the log differences, the most notable disparities in distributions or means between two subgroups are at the cut-off value 0.2 for the events fraction meta-feature (pct^{evt}) and at the cut-off value 3 for the number of features (p). The boxplots for these variables show that random forests

require more EPV than logistic regression when the percentage of observations in the minority outcome class or the number of features is lower. For the optimism of random forest models, one can infer from the plots a small negative association with the total number of EPV of a dataset.

To assess the associations between the two metrics and meta-features more precisely, Spearman's correlation coefficients were computed in addition to the previous planned analyses. Table 3.3 contains those coefficients along with the p-value of the corresponding correlation test.

Meta-feature	Log difference		Optimism	
	Spearman's ρ	Spearman's ρ p-value	Spearman's ρ	Spearman's ρ p-value
n	0.006	0.957	-0.032	0.784
p	-0.071	0.546	0.029	0.804
p_{num}	0.038	0.744	**0.230**	**0.047**
p_{bin}	-0.205	0.078	**-0.254**	**0.028**
EPV^{tot}	0.025	0.831	-0.170	0.145
pct^{evt}	-0.143	0.220	-0.064	0.585

Table 3.3: Correlations between the dataset meta-features and the log difference and optimism. p_{num} and p_{bin} refer to the number of numeric and binary features, respectively. Correlations with p-values < 0.05 are written in bold.

The results of the correlation analysis corroborate the conclusions drawn from the plots. Most of the considered associations are very small, and only the meta-features related to the types of variables are significantly correlated with one of the metrics. Aside from the significant correlations, small associations exist with the number of total EPV and the events fraction. In view of these findings, it is concluded that the result of the first confirmatory hypothesis is not very dependent on certain dataset characteristics, at least not on those considered in this section. The confirmatory analysis of random forest model optimism may be slightly more sensitive to meta-features, though probably also only to a rather limited extent.

Sensitivity of results to design and analysis choices

To evaluate the sensitivity of the results from another perspective, alternatives to the decisions made during the study design or analysis were systematically investigated. To this end, a set of considered alternative options for various study design choices was determined. The different strategies examined in the sensitivity analysis are then all possible combinations of those options.

As mentioned, the set of considered options was expanded from the one specified in the study protocol by considering different dataset subgroups and two additional performance thresholds. Table 3.4 contains the updated list of considered options with those used in the confirmatory analysis denoted as the default option.

Design or analysis choice	**Considered options**
Dataset selection	All 75 datasets (default), eight subgroups ($<$ or $\geq$ median of n, p, EPV^{tot}, pct^{evt})
Performance measure	AUC (default), accuracy, Brier score
Performance threshold in evaluation metric (stability threshold)	90%, 92.5%, 95% (default), 97.5%, 99%
Handling of missing values in evaluation metric	20%-threshold rule (default), worst, mean, weighted
Aggregation of evaluation metric values across CV iterations (within a dataset)	Geometric mean (default), median
Aggregation of evaluation metric values across datasets	Geometric mean (default), median

Table 3.4: Updated list of considered design and analysis choices and corresponding options for the second type of sensitivity analysis. 'Default' refers to the options used in the primary, confirmatory analyses.

This set of options resulted in $9*3*5*4*2*2 = 2{,}160$ combinations (i.e., considered design and analysis strategies). For each combination, the same analysis process that was used for the confirmatory analysis of the first research question in Section 3.3.1 was completed: (1) for each dataset and method, compute the evaluation metric EPV_{min} for all 50 CV iterations; (2) impute missing values, if necessary; (3) aggregate the 50 evaluation metric values for each dataset and (4) aggregate those average values across all the analyzed datasets, resulting in a single average minimum number of EPV each method needed to surpass the performance threshold. Therefore, the sensitivity analysis described in this section examined 2,160 possible results from which one could have been reported as the primary, overall study result if a researcher had made the corresponding choices from the specified options. Accordingly, the following analysis also demonstrates how variable the results of this study would have been without the pre-specification of an analysis plan.

The aggregated minimum numbers of EPV for the two modeling approaches (i.e., the result of each of the 2,160 analysis strategies) were visualized in scatterplots (see Appendix B). It is apparent from these plots that primarily the choices of performance measure and threshold are systematically associated with the results. Since the dataset selection did not seem to noticeably influence the results, only the 240 combinations involving all datasets were analyzed further to obtain more straightforward visualizations. The aforementioned connection between a given performance measure-threshold combination and the estimated required numbers of EPV across all datasets is clearly visible in Figure 3.8. As expected, a higher stability threshold generally led to a higher estimate, especially within a given performance measure. Moreover, the use of Brier scores resulted in much higher estimates compared to AUC and accuracy for all considered thresholds.

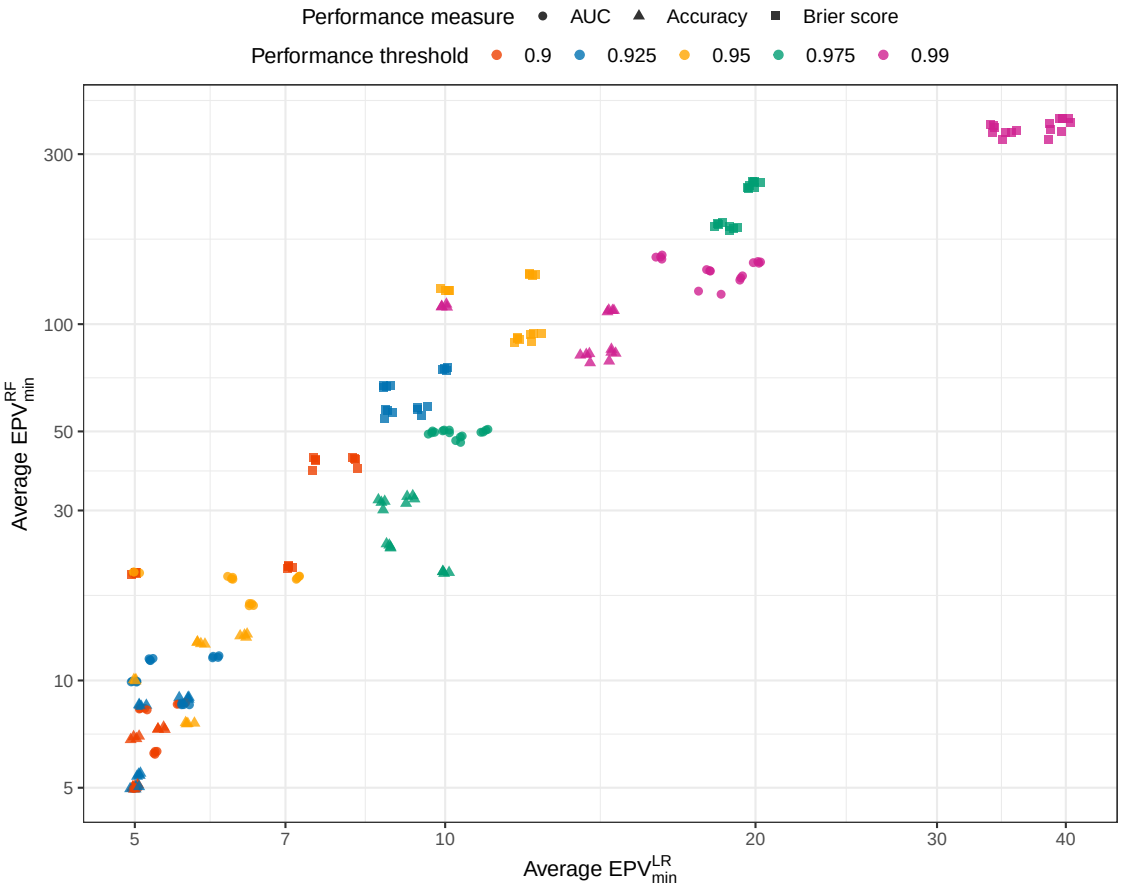

Figure 3.8: Scatterplot of the aggregated minimum EPV results of the 240 considered analyses, colored by performance threshold and with shapes reflecting the performance measure. For both axis, a logarithmic scale is used. Slight jitter was added to visually separate overlapping points.

Similar observations can be made with Figure 3.9, where one can see the variation in the ratios, indicating by what factor random forests required more EPV, across the possible analysis options if one choice is fixed. For example, even when a hypothetical researcher restricts their analysis to AUC performance, they could conclude that random forest models need nine times as many EPV as logistic regression models or need the same number just depending on the other analysis choices.
Furthermore, the boxplots in Figure 3.9 also confirm the impression from the scatterplots included in the appendix that the imputation method alone did not particularly impact the results. The same was generally true for the aggregation methods, though the choice of the method for averaging across datasets did have some effect.

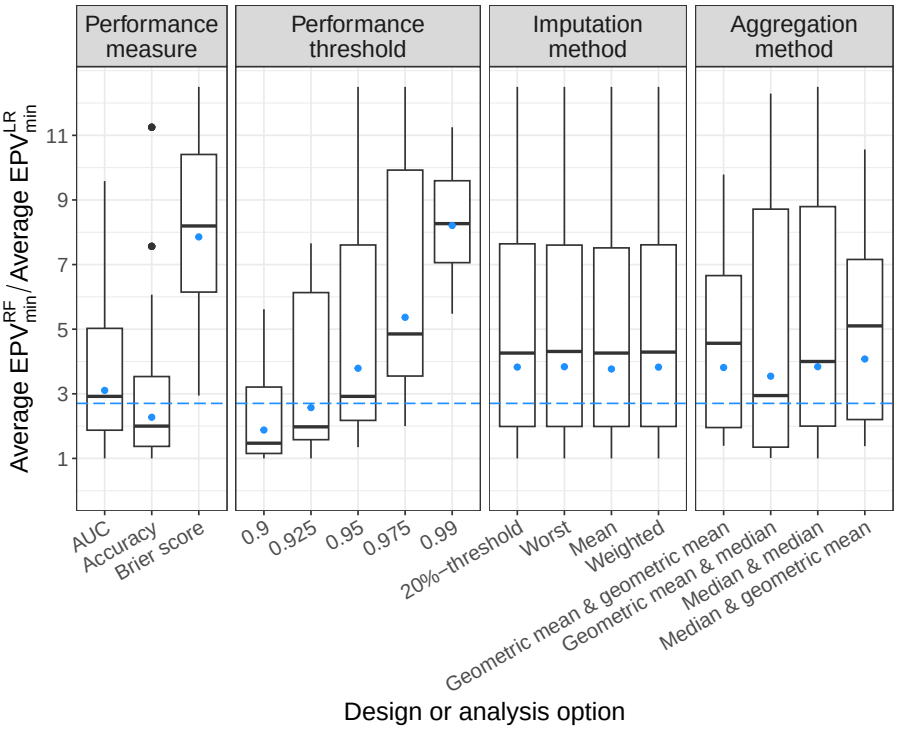

Figure 3.9: Boxplots by design or analysis choice for the aggregated results of the 240 considered analyses, fixing one option at a time. The blue points show geometric means, and the dashed blue line indicates the geometric mean of the actual confirmatory analysis.

The second panel of Figure 3.9 shows that the results vary much less for the highest and lowest stability threshold than for the thresholds 92.5%, 95% and 97.5%. This is because, for the middle three thresholds, the differences between possible results for the three measures are quite large.[11] Given this interaction and how the other considered choices have

[11] This aspect can be observed in Figure B.7 in Appendix B, which contains plots similar to the one in Figure 3.9 but stratified by both performance measure and thresholds.

rather minor impacts, the following final sensitivity analysis was performed. Keeping all choices other than performance measure and threshold fixed at their specified "defaults" (see Table 3.4), the calculation of the log differences from Section 3.3.1 was repeated for the 15 combinations of the two choices to assess their influence directly on the metric used in the statistical test.

Thus, the boxplots in Figure 3.10 show the same kind of boxplot as in the lower-right panel in Figure 3.4 but for all considered measure-threshold combinations. The boxplot from the actual confirmatory analysis from Section 3.3.1 is the third one from the left in Figure 3.10, and the arithmetic mean of the log differences for a given measure-threshold choice is represented by the blue point. It is clear that, depending on the choice, vastly different results could have been reported. Using "90% of *MaxAccuracy*" as the indicator for a stable or good prediction performance would have led to a mean log difference of 0.33, which is equivalent to random forest requiring 39% more EPV than logistic regression on average. Meanwhile, choosing a threshold that is harder to surpass, such as 97.5% or 99%, together with the Brier score would have resulted in a mean log difference of 2.28, suggesting that random forests need almost ten times as many EPV ($\exp(2.28) = 9.78$).

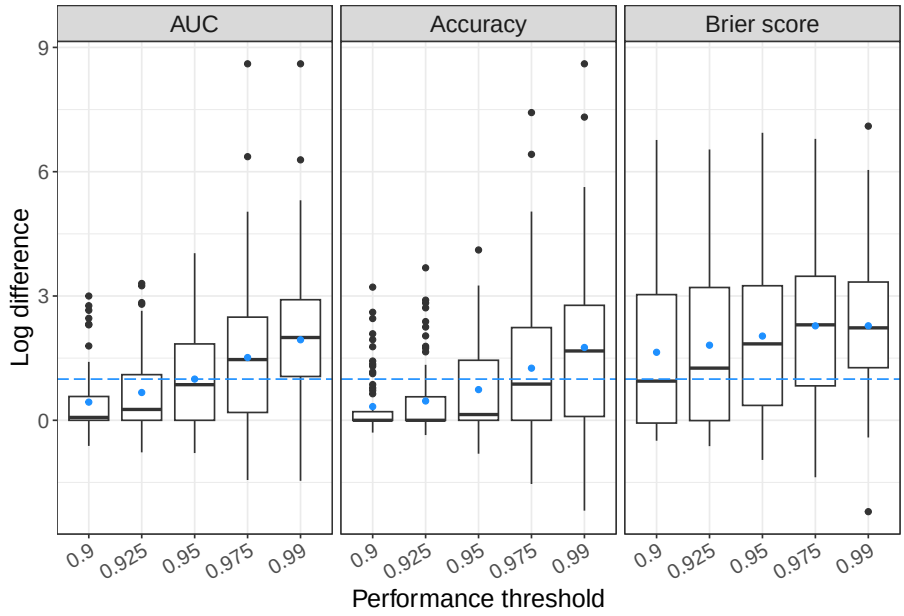

Figure 3.10: Log differences for all 15 performance measure-threshold combinations while leaving the other analysis options at their defaults (20%-threshold rule imputation and geometric mean aggregation). The blue points show arithmetic means, and the dashed blue line indicates the arithmetic mean of the actual confirmatory analysis.

Overall, regarding design and analysis choices, it can be concluded that the result regarding the first hypothesis from Section 3.3.1 is especially sensitive to changes in the performance measure and the stability threshold. The decisions made regarding datasets, missing values imputation and values aggregation, on their own, only have a small effect on the result, if they have one at all. Finally, it should be pointed out that none of the 2,160 considered analysis strategies led to the average required number of EPV being higher for logistic regression than for random forests. At most, both modeling techniques needed the same number of EPV, namely the smallest possible value of exactly 5 EPV. This was the case for 160 of the 2,160 combinations, 12 of which were part of the 240 that did not involve dataset subgrouping.

Important context for many of the previous findings and something that can partly explain them is the extent to which there are missing values in the evaluation metric (i.e., the number of CV iterations where no subset meets the performance requirement). Table B.1 provides a detailed overview of this aspect for the 15 measure-threshold combinations (see Appendix B). Noteworthy in this regard is that logistic regression models reached the given threshold of the given measure in almost all iterations, while random forests failed to do so for up to 31% of them. Secondly, in addition to the general expected increase in missing values when the threshold is raised, there was a clear disparity between the different performance measures. While using AUC and accuracy essentially only led to missing values when a 97.5% or 99% threshold was chosen, evaluations with respect to the Brier score required imputation even when only 90% of the goal performance needed to be achieved and in far more iterations in general. Given that imputation likely results in a drastically higher EPV values, especially if the total EPV of a dataset is large, these aspects can somewhat clarify why the choice of the Brier score would have resulted in so distinctly different results.

3.3.3 Additional exploratory and unplanned analyses

Learning curves for predictive performance and optimism

To provide an additional perspective on the relation between the prediction performance or the optimism and the number of EPV, so-called learning curves were generated. These learning curves are conceptually similar to those presented in van der Ploeg et al. (2014), though in their visualization, they did not show data aggregated across multiple datasets and used a different reference point for their relative performance evaluation.
Figure 3.11 shows the first pair of learning curves, which concerns the AUC performance of the training subset model relative to the performance of the model trained on the full training data of the same iteration. Whereas in the previous analyses, this metric of relative

performance was used to define stability, the plot here indicates what level of performance in relative terms the two methods achieved on average for a certain training set size. It should be noted that while outliers are not displayed for a more concise visualization, they clearly influenced the arithmetic mean values that are included in the plot as a line. A few observations can be made from that plot. Firstly, the results from the confirmatory analyses are somewhat visible in that logistic regression models reached the 95%-threshold for the median between 5 and 10 EPV (actual estimated geometric mean value from Section 3.3.1: 7.18 EPV) and random forests did so somewhere between 10 and 20 EPV (actual estimated value: 19.41 EPV). Secondly, it is also quite obvious that the interquartile range differed greatly between the two methods. While applying logistic regression resulted in relative performance values between 99% and 100% for more than 50% of the datasets at 75 EPV, the random forests only came reasonably close to such a distinction when they were trained with at least 350 EPV. These much longer boxes of the boxplots at all training set sizes illustrate that random forest models were generally more variable. Overall, Figure 3.11 provides additional evidence from a different, visual perspective that logistic regression models reached their predictive potential with much fewer observations.

Figure 3.11: Learning curves and boxplots showing the mean percentage of the test AUC of the complete training set models ($MaxAUC$) that was achieved at different training set sizes. Note that outliers are not displayed for a more concise visualization.

Plotted in Figure 3.12, the second pair of learning curves relate to the second hypothesis of this study and show the mean AUC optimism values for the 75 datasets for each training dataset size. As in the previous plot, there is a clear disparity in variability between values for the two methods. The mean optimism of random forest models decreased slowly

with increasing training observations, and the extreme values for a few datasets heavily influenced the mean estimate. However, when considering the median optimism, the overall trend was only slightly different. In fact, even when using all available training data, which corresponds to 10,276 EPV on average, the median optimism of the random forest models was not below 0.04.

Regarding the optimism of logistic regression models, van der Ploeg et al. (2014) reported that in their three simulations, the modeling approach needed approximately 55–127 EPV, 18–23 EPV and 14–28 EPV, respectively, to achieve a mean optimism below 0.01. The results from the current study overlap at least in the first of those ranges, as a mean optimism below 0.01 for the 75 analyzed datasets was reached somewhere between 75 and 100 EPV. When examining the median optimism, the condition was fulfilled with even fewer observations, somewhere between 40 and 50 EPV.

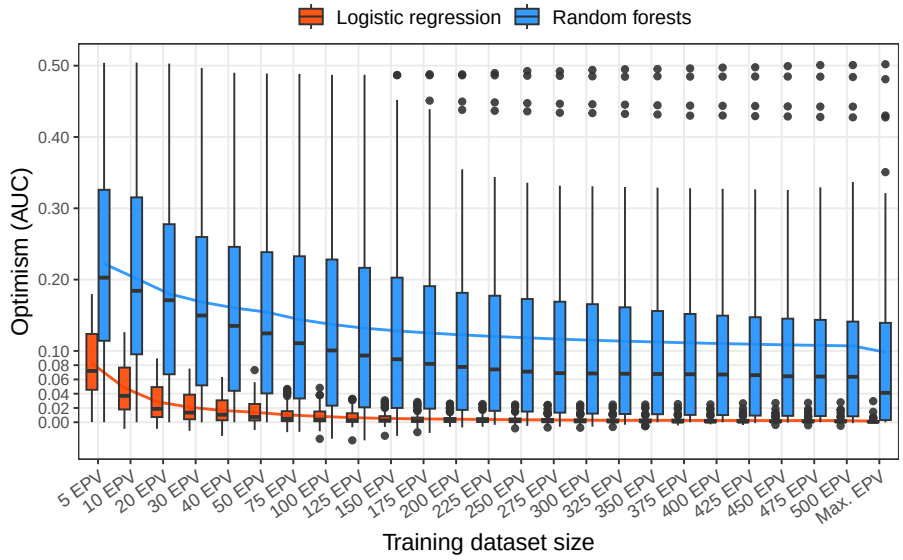

Figure 3.12: Learning curves and boxplots showing the optimism of logistic regression models and random forests in terms of AUC at different training set sizes.

In the third and final plot with learning curves, in Figure 3.13, the mean absolute test set performance in terms of AUC is visualized for each EPV subset size. It shows that, on average, random forests outperformed logistic regression for every training dataset size. Besides this observation, logistic regression models reached a stable mean AUC much quicker. However, the average predictive performance of random forests continually improved with more observations, resulting in increasingly larger differences between the mean AUC values of the two methods. The differences in performance between random forests and logistic regression on the same datasets are presented in more detail in the following section.

Figure 3.13: Learning curves showing the mean AUC test set performance of logistic regression and random forests at different training set sizes.

Evaluation and comparison of absolute performance measures

In this study, the predictive performance of models was intentionally evaluated in relative terms. However, as an additional analysis, the absolute values for AUC, accuracy and Brier score were also examined and compared between the two methods in reference to Couronné et al. (2018). In their benchmark experiment with 243 datasets, they used a setup similar to this study, also investigating logistic regression and random forests in their default configurations and with respect to the three mentioned performance measures. Therefore, the following analysis was performed to see whether their results could be replicated with the 75 datasets analyzed in this study. For this direct comparison, only the test set performance of the models trained with the complete training data was considered and the paired differences for each dataset were calculated as random forest model performance minus logistic regression model performance. Therefore, for AUC and accuracy, a positive difference indicates that random forests perform better, whereas for the Brier score, a negative difference signifies the same.

Table 3.5 lists the results from both studies in the form of means and standard deviations. While both methods performed worse in this study than in Couronné et al. (2018) across all measures, the general direction of the results did not change, with random forests clearly outperforming logistic regression for the 75 analyzed datasets. In fact, for AUC and accuracy, the paired performance differences between the modeling approaches were even noticeably larger compared to Couronné et al. (2018). Furthermore, their finding that "the differences in performance tend to be larger for auc [AUC] than for acc [accuracy] and brier [Brier score]" (p. 7) is also true for the results of this study.

		Couronné et al. (2018, p. 7)		This study	
		μ	σ	μ	σ
AUC	RF	0.867	0.147	0.795	0.172
	LR	0.826	0.149	0.731	0.166
	RF − LR	0.041	0.088	0.064	0.107
Accuracy	RF	0.854	0.134	0.791	0.154
	LR	0.826	0.135	0.753	0.142
	RF − LR	0.029	0.067	0.039	0.064
Brier score	RF	0.102	0.080	0.133	0.084
	LR	0.129	0.091	0.158	0.77
	RF − LR	-0.027	0.054	-0.025	0.048

Table 3.5: Performances of random forests (RF) and logistic regression (LR) in the benchmark experiment by Couronné et al. (2018) and the presented study as well as the differences between their performances (RF − LR).

To examine the differences for the training subsets as well and to illustrate their trend with increasing training set size, the plot in Figure 3.14 was generated. At 5 EPV, the difference in the AUC performance of the two methods is rather small (below 0.02), even smaller for the other two measures, and may not be considered significant. By 50 EPV, random forests outperform logistic regression models by over 0.04 in AUC, over 0.02 in accuracy and under -0.01 in Brier score on average. Beyond 225 EPV, the differences only changed marginally for all three measures.

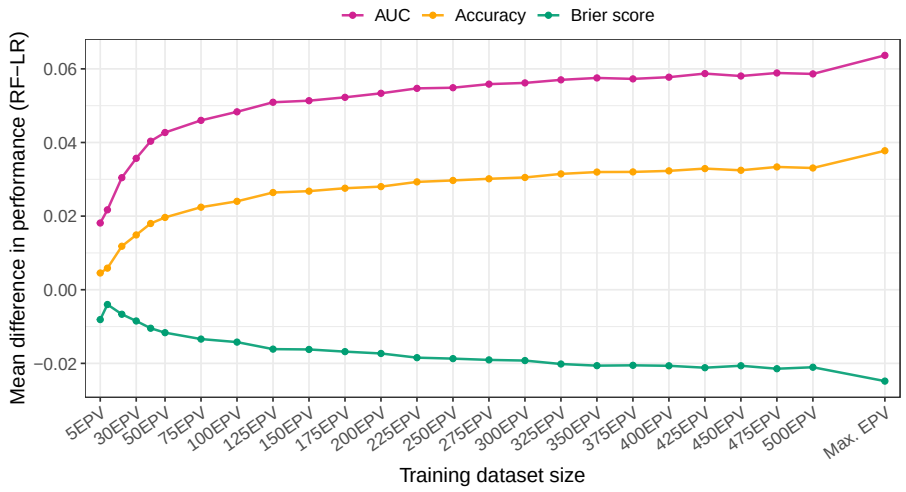

Figure 3.14: Mean differences in performance between random forests and logistic regression at different training set sizes.

3.4 Discussion and conclusion

The presented study was intended to provide insight on the connection between prediction performance and the number of EPV for two binary classification methods, logistic regression and random forests, and aimed to confirm two specific hypotheses. The first hypothesis was that random forests need more EPV than logistic regression to achieve a stable predictive performance, and the second hypothesis was that random forest models are highly optimistic (optimism > 0.01) even if they are generated using a large number of EPV. To evaluate these hypotheses, this study applied the methods to 75 real datasets in an elaborate benchmark experiment. The following briefly summarizes and contextualizes the main results of the study and discusses some limitations.

This study confirmed both the first and second hypotheses, as the results of the analyzed data indicate clear and definite support for each of them. With respect to the first one, the factor by which random forests require more EPV than logistic regression to achieve a stable predictive performance was estimated to be 2.7. The mean optimism of random forest models trained using 500 EPV, which is the quantity of interest for the second hypothesis, was estimated in this study to be 0.107.

A thorough sensitivity analysis regarding dataset characteristics as well as design and analysis choices showed that the mentioned EPV ratio between random forests and logistic regression is primarily dependent on the chosen performance measure and the threshold used to indicate stability. By varying these two parameters within certain considered options, the result of the primary analysis hypothetically could have been a factor between 1.39 and 9.78, validating the reported conclusion of the actual confirmatory analysis. Furthermore, the sensitivity analysis suggests that dataset characteristics such as the number or percentage of observations in the minority outcome class had little to no impact on the result or estimated quantity for either hypothesis.

As a byproduct of the evaluation of the first hypothesis, the minimum number of EPV needed for stable prediction performance was estimated to be 7.18 for logistic regression and 19.41 for random forests.

The primary reference point for this study was the work by van der Ploeg et al. (2014), which is the only known study investigating the relation between the number of EPV and prediction performance for both logistic regression and random forests. Specifically, the hypotheses tested in the presented study were based on two main results of van der Ploeg et al.'s simulation study to assess their validity on a large number of real datasets. Therefore, the confirmation of both of them here means that the presented results are in line with the corresponding conclusions of van der Ploeg et al. (2014). However, while the studies are in agreement regarding the overall conclusion, there are some differences in the specific

estimated quantities. Van der Ploeg et al. (2014) reported that "a stable AUC was reached by LR [logistic regression] at approximately 20 to 50 events per variable" (p. 1), which is considerably higher than the estimate from the presented study (7.18 EPV), although the gap may just be the result of differences in the study design. Besides the apparent contrast between a study with three simulated datasets and one analyzing 75 real datasets, the concept of prediction performance stability was evaluated slightly differently. Van der Ploeg et al. (2014) seem to have assessed a method's AUC stability visually based on learning curves for each dataset and additionally required the AUC optimism of a given model to be <0.01 to consider its performance stable. Meanwhile, the presented study evaluated the concepts of performance stability and model optimism separately and assessed the former quantitatively using a clearly defined measure of relative performance.

Regarding other research on the topic of EPV, it can be noted that the minimum EPV estimate for logistic regression complies with the widely adopted rule of thumb of 10 EPV that is used as a minimal sample size criterion for the development of prediction models (Ogundimu et al. 2016; van Smeden et al. 2019). However, it should also be noted that, in recent years, the validity of this rule of thumb and the usefulness of EPV criteria in general have been questioned (van Smeden et al. 2016, 2019). It has been suggested to avoid such simple rules altogether and instead use a multi-criteria approach to calculate the minimum necessary sample size (Riley et al. 2018). Therefore, the results from this study should not be translated into generalized rules of thumb either. Rather, they could provide another perspective and orientation in conjunction with other considerations.

As an exploratory analysis, random forests and logistic regression were also compared with respect to their absolute performance, and in the 75 analyzed datasets, the former outperformed the latter on average across all considered performance measures (AUC, accuracy and Brier score). This result is contrary to the corresponding finding by van der Ploeg et al. (2014), who suggest that the difference in predictive performance between the two methods is small. However, it is in line with the result of the similarly designed large-scale real-data benchmark experiment by Couronné et al. (2018), though in the presented study, both methods consistently exhibited slightly worse performances compared to their benchmark. This difference may simply be caused by the different dataset selections. Furthermore, the reason for the slightly different results could also be that in the presented benchmark, preprocessed datasets were analyzed, while Couronné et al. (2018) used unprocessed datasets. As part of the preprocessing, the set of features was reduced for 61 of the 75 datasets, which could have led to a loss of information, contributing to the worse average performances.[12]

[12] It should also be noted that there is a small overlap between the dataset selections of the two studies: Ten datasets were included in both benchmark experiments, nine of which were subject to preprocessing in the presented study.

Many limitations exist in the presented study, beginning with the fact that the analyzed datasets all come from a single database (OpenML) and thus do not represent a random selection from the population of interest (i.e., all datasets with a binary outcome variable). The sampling of datasets is a general issue in benchmark experiments, which is why their results should be interpreted as conditional on the given dataset selection (Boulesteix et al. 2017). In the case of this study, the selection consists of a rather large number of datasets from various subject areas, though it also includes datasets with non-binary targets. However, as part of the selection process, datasets that had an apparent overlap in data with another suitable dataset were excluded. Therefore, the selected datasets may be assumed to be sufficiently independent.

Further limitations contributing to a possibly reduced generalizability of the results stem from the intentionally made design decisions, the first one being the minimum dataset size (625 EPV with two features) required by the chosen study design. Secondly and central to the analysis, the definition of prediction stability might not be considered ideal for the investigated hypothesis. Under the assumption that prediction performance monotonically increases with sample size, it was determined that a model has achieved a stable predictive performance when its performance is at least 95% of the maximum achievable one (i.e., the performance of the respective full training dataset model). Not only was the threshold of 95% more or less chosen arbitrarily, even with a different threshold, a more intuitive alternative interpretation of the defined metric might be that a model has reached or come within a reasonable margin of the predictive performance potential of the corresponding method. Using this interpretation, the metric and results still provide valuable insights; they just would not be directly connected to the work by van der Ploeg et al. (2014). However, since they only assessed the stability visually, some quantitatively calculable metric had to be defined to test the hypothesis.

Finally, limitations arise from the preprocessing decisions, particularly the dichotomization of regression targets at the median and the sampling of features. Both of these choices were made deliberately to increase the number of included datasets and the likely resulting loss of information was considered acceptable since the absolute performance of the methods was not the focus of the study. However, two other potential issues are associated with these decisions. The splitting of numeric targets at the median, which was chosen to maximize the number of events, resulted in 31 datasets with an events fraction between 0.47 and 0.5 (in addition to the nine datasets meeting this condition that had binary or multiclass original targets). While an analysis differentiating between the three types of the untransformed targets did not yield a noteworthy result, this very high concentration at just a few values could have influenced the performed analyses, most directly probably the sensitivity analyses regarding the events fraction meta-feature. Moreover, the propor-

tion of datasets with an events fraction in the mentioned range is very likely much higher in this study than in the population of all binary classification datasets, possibly reducing the generalizability of results further.
The potential issue with the reduction of the feature set through sampling is that it might have reduced some datasets to such a high degree that they no longer resemble the original underlying real-world application. It is difficult to assess when this might have been the case; however, for the 41 datasets where features needed to be sampled to arrive at the necessary number of EPV, the set of features was reduced by about 75% (calculated from the number of non-sparse features).

Acknowledging and within these limitations, the presented study contributes further evidence on the relationship between the number of EPV and the predictive performance of logistic regression and random forests. In the context of research on the topic of EPV in general, it is the first benchmark experiment with a large number of real datasets and, thus, provides insights from a different perspective than the many existing simulation studies in this area.
Besides confirming the general direction of two results from previous simulations, the study also illustrates the effects design and analysis choices can have on results through sensitivity analyses.
Future research could explore the topic of prediction stability further, both in connection to the EPV and in general. Additionally, it might be interesting to investigate the research questions of this study for other machine learning approaches or with other benchmark designs. Finally, more pre-registered confirmatory studies with many datasets should be conducted to assess the validity of results from other simulation studies, especially if those studies led to rules of thumb that are now widely used in practical applications.

4 Discussion

In the following, the proposed concept of a confirmatory study in methodological statistics and several associated limitations are discussed. First, in Section 4.1, the practical application of the outlined study concept and suggested template is reflected upon in the context of the presented illustration and beyond. Then, in Section 4.2, limitations of preregistration in general and the concept specifically are described and possible directions for future research are suggested.

4.1 Reflections on the practical application of the concept

When following the strictly confirmatory study approach outlined in this thesis, one is faced with a rather intimidating task: determining and precisely specifying every aspect of a study before any data exploration. This is arguably not how most people commonly approach research projects. The reason why it is so intimidating is the implied magnitude of the made decisions which are meant to be final, since any substantial deviation later on could be viewed as an intentional attempt to influence the results. Accordingly, such a task requires careful consideration of many things, including eventualities that might arise during the study. Consequently, it takes a considerable amount of time, though it has been argued that a detailed plan can save time in later stages of the research process (Lindsay et al. 2016). This was true for the presented illustration. Particularly time-consuming was the dataset selection process, which took the majority of the time spent on the study protocol. Among other aspects, this fact resulted from the choice of the source from which the datasets were selected, OpenML. While the database may seem appealing with over 22,000 available datasets and meta-feature filtering capabilities, it is not without flaws. The quality varies heavily among the datasets and corresponding available metadata, even though the datasets are referred to as "verified" by OpenML. This makes a fully automatic selection process unreliable, resulting in a large number of datasets that must be manually checked for eligibility (in the case of the illustration: 752). Of course, in a context other than this thesis, the dataset selection would likely not be determined by one person, not only to make the process more efficient but also to avoid unwarranted exclusions.
Deviations from the protocol are an inherently practical issue and likely unavoidable to a certain extent even with the most meticulous planning. In the illustration, the devia-

tions were small and only resulted in a more comprehensive sensitivity analysis, although this does not mean that this will be generally the case. Reviews of pre-registered studies from different applied research fields consistently found that deviations, both disclosed and undisclosed, are a common occurrence (Claesen et al. 2021; Ofosu and Posner 2021; Heirene et al. 2021). However, like in the illustration, deviations, even if undisclosed, are not necessarily attempts to exploit researcher degrees of freedom (Claesen et al. 2021). Nonetheless, reporting them in their entirety is considered essential to the approach proposed in this thesis.
One of the side benefits of pre-registration is that the detailed plan reminds researchers what the purpose and hypothesis of the study is and guides the reporting of the results. Additionally, it is meant to help them in avoiding unintentional QRPs. This was found to be the case when conducting the presented study. Without the protocol, one's focus might have unintentionally shifted to the exploratory results or one might have excluded some analyses that were deemed useful before analyzing the data.
As a final meta-observation with respect to the illustration, it should be noted that it did not reveal any critical aspects missing from the suggested template. Therefore, this thesis argues that the template provides a good structure and guide to comprehensively plan a confirmatory methodological study and, thus, is serving its intended purpose.

Taking a more general perspective, there are several issues that one should be aware of when conducting a real-data confirmatory study based on the suggested approach and that one may use to evaluate the practicality of the proposed concept.
A key practical consideration possibly damaging the concept's usefulness, especially compared to the more common pre-registration of studies that collect new data, is that researchers likely already have access to the datasets they plan to analyze. This opens up the possibility of analyzing the dataset prior to pre-registration and makes it difficult for others to assess whether that is the case. The pre-registration of this kind of data and possible solutions to this issue have been discussed in the context of applied research where the associated analysis is referred to as "analysis of preexisting data" or "secondary data analysis" (Mertens and Krypotos 2019; Weston et al. 2019; van den Akker et al. 2019). However, the most widely applicable solution is to be as transparent as possible about one's knowledge about the data (Nosek et al. 2018). This was considered in the creation of the suggested template, which has a section dedicated to declaring prior knowledge.
Another issue related to dataset access that is specific to the proposed approach concerns the dataset selection process. To address the QRP of post hoc dataset exclusions, which is particularly relevant in the context of real-data methodological studies, it was decided that the entire dataset selection process, including the check of exclusion criteria, must be conducted prior to pre-registration. This can require the researcher to access datasets to

some degree which could potentially be problematic, although it is possible to maintain the validity of the pre-registration by accessing the datasets only superficially and only as much as necessary to make a selection decision. Since it is then not only evident that the analyzed data existed prior to pre-registration but also that it has already been accessed, this aspect expects significant trust from other researchers. To reduce the risk of harmful prior knowledge, one could potentially delegate the check of the eligibility criteria to an independent group of people that is not involved in the planning of the study.
Finally, there is the issue of using the same datasets in multiple studies, which poses a threat to the intended confirmatory nature of the studies that builds on the assumption that hypotheses and analysis plans are formulated blind to the analyzed data. Naturally, the risk of harmful prior knowledge due to overlapping dataset selections increases with the number of benchmark experiments one conducts. Whether this will be a problem in practice is uncertain, but it is easy to imagine scenarios where it might, for example, if the total number of datasets available for a certain setting or subject matter is small.

4.2 Limitations of the concept and directions for future research

Since pre-registration is a central component of the concept proposed in this thesis, some limitations are shared by both. First, neither pre-registration in general nor confirmatory methodological studies specifically lead to definitive results (Nosek and Lakens 2014). Not only can even adequately-powered studies produce a Type I or Type II error, the results of real-data benchmark experiments are always dependent on the set of analyzed datasets. Secondly, as noted in Section 2, pre-registration of confirmatory studies is only effective if the pre-registered document is comprehensive and sufficiently restricts the researcher degrees of freedom. A recent comparison of two types of pre-registration found that a more structured format was associated with better specificity (Bakker et al. 2020). The template suggested in this thesis is already quite structured, but future versions could provide more detailed descriptions and example answers. Another way for researchers to ensure that their plan is specific enough is to perform the described analysis on a mock dataset (Wagenmakers et al. 2018) or, in the case of methodological studies, on a set of datasets not selected for the actual study.
Lastly, both pre-registration in general and the suggested template, even if extremely specific, cannot prevent misguided research questions, severely flawed study designs or poor statistical practices. They do, however, make such issues transparent and detectable (Nosek et al. 2018). Moreover, if these flaws are unintentional, feedback from others on the protocol can help to correct them before the study is conducted, which is preferable to conducting a poor study that is then heavily criticized afterwards.

The following describes limitations specific to the suggested concept and template and presents ideas for possible future research.
The most obvious one is that the scope of the thesis and thus the template was intentionally restricted to real-data studies. Consequently, the other common empirical study type in methodological research, simulation studies, does not entirely fit within the suggested approach. The adjustment necessary to accommodate them might not be that extensive, but the two study types were considered distinct enough to warrant this separation to keep the template as straightforward as possible. Moreover, in the context of simulation studies in methodological research, some literature related to study protocols already exists. In earlier work, Burton et al. (2006) and Smith and Marshall (2010) emphasize the importance of simulation protocols and discuss considerations that must be made in that regard. More recently, Morris et al. (2019) proposed a structure for simulation study plans and advocated writing a protocol before the code. This structured approach has, in turn, been employed in two pre-published protocols by Kipruto and Sauerbrei (2022) and Pawel et al. (2022), with the latter explicitly referencing the distinction between exploratory and confirmatory findings. These existing works should be considered when designing a pre-registration template for confirmatory simulation studies in the future. Lastly, the issues regarding data access mentioned in the previous section are even more relevant for simulated data because there is even less or no separation between the researcher and the data compared to analyses of existing datasets.
The issues regarding the data access during the planning of the study could also be considered a limitation of the template in its current form. It could be argued that this unclear, blurry line between the researcher and the data undermines the goal of transparency since it is impossible to tell from the outside to what degree the data has been accessed even if the researcher tries to describe it. As suggested in the previous section, a solution to this problem could be to delegate the dataset selection process to uninvolved people. Therefore, a possible adjustment to the template could be to add a subsection for the specification of such a practice. Alternatively, one could also invert the relevant design decision that the dataset selection must be completed before the pre-registration, though this would essentially make post hoc dataset exclusions possible again.
Another limitation may stem from the fairly broad definition of a real-data confirmatory study in methodological statistics as any study with at least one pre-specified hypothesis that will be evaluated using real data. The choice of this general scope was also intentional to not preemptively restrict the applicability of the approach from its first version, knowing that it could always be narrowed in later iterations. However, it could be argued that the current definition and scope is too broad and unspecific to be useful in practice. Therefore, possible future directions could also be to define a more precise framework for confirmatory

studies or a precise general framework to distinguish between exploratory and confirmatory research in methodological statistics. Either case could involve specifying mandatory template contents or requirements a study would need to meet to be considered confirmatory. In theory, this could ensure that only high-quality studies fall under this label. Possible considerations in this context could be minimum requirements with respect to the number of datasets, the neutrality of the study, the amount of existing previous work or the planned sensitivity analysis. One could also narrow the defined scope by limiting the acceptable inference techniques or making the dissemination of the results regardless of the study outcome mandatory, similar to clinical trials.
Regardless of the specific adjustments or future direction in this area, input from a number of methodological researchers is likely required to fine-tune the initial concept and template proposed in this thesis.

5 Conclusion

The aim of this thesis was to explore, conceptualize and illustrate the idea of a deliberately confirmatory real-data study in the field of methodological statistics. To this end, an approach constructed by adapting a combination of the ideas behind pre-registration and clinical trial protocols to the context of methodological statistical research was proposed in this thesis. As the central part of this approach, a template was suggested to aid in the pre-specification of a comprehensive research protocol. This protocol template was designed after considering several aspects of real-data methodological studies and built upon existing templates and guidelines from applied research fields.

The protocol template and deliberately confirmatory research approach proposed in this thesis were illustrated using a study that investigated the predictive performance of logistic regression and random forests in relation to the number of events per variable (EPV). In a large-scale benchmark experiment involving 75 datasets, models were trained with data subsets corresponding to different numbers of EPV and the stability of their predictions was evaluated using a pre-defined relative performance metric. The results of the study indicate strong support for the following two specified confirmatory hypotheses:

1. Random forests need more EPV than logistic regression to achieve a stable or good predictive performance, or in other words, to realize their predictive performance potential.
2. Random forest models are highly optimistic (indicated by an optimism > 0.01), even if they are generated using a large number of EPV, such as 500 EPV.

Therefore, the presented study provides new confirmatory evidence on the relationship between the number of EPV and the predictive performance of logistic regression and random forests. In the context of research on the topic of EPV in general, it is the first large-scale real-data benchmark experiment and, thus, provides insights from a different perspective than the many existing simulation studies in this area. However, one must be cognizant of the fact that the confirmatory conclusions from the study, like most findings from real-data benchmark experiments, are conditional on the set of analyzed datasets. Furthermore, the presented study has several limitations due to design, preprocessing and analysis decisions. Besides the primary results, the benchmark analysis also illustrated how important the

public pre-specification of hypotheses and plans is for studies that are meant to be confirmatory. In the presented illustration, choosing a performance measure or stability criterion after the analysis, something that is not noticeable without pre-registration, could have resulted in very different, possibly over-optimistic estimates. Thus, the study provides a fitting motivating example of using pre-registration to limit researcher degrees of freedom and the central role pre-registration has in the proposed approach for confirmatory studies. After reflecting upon the suggested pre-registration template in the context of the application, it was determined that no critical elements were missing from it. Therefore, it is argued that the template provides a good structure and guide to comprehensively plan a confirmatory methodological study and thus serves its intended purpose.
To ensure the wide applicability of the proposed approach and template, the concept of a confirmatory real-data study was intentionally defined without many restrictions as any study with at least one pre-specified hypothesis that will be evaluated using real data. Considered advantageous, it can therefore be used for all kinds of studies and hypotheses involving real datasets. Although the approach may be particularly suitable for studies like the presented illustration where the primary purpose is to replicate (i.e., confirm) previous methodological results. Moreover, with minor modifications, the template can even be used to plan exploratory research. However, possible future adjustments to the initial concept also include a much more precise and narrow definition of a confirmatory methodological study and the specification of requirements a study must meet to be considered confirmatory. Those requirements could be related to aspects such as the number of datasets, the neutrality of the study, the amount of existing previous work or the dissemination of the results.

In conclusion, the work in this thesis represents the first exploration and conceptualization of the idea of a deliberately confirmatory study in the context of methodological statistical research. The clearer distinction between exploratory and confirmatory research in practice that is implied by this idea would be a significant step towards more credible, less overly optimistic and more replicable methodological statistical research. While more input from others is needed to comprehensively evaluate and fine-tune the initial concept and the suggested pre-registration template, this thesis argues that the provided work is a good starting point for future meta-research on this topic.

References

Al-Jundi, A. and Sakka, S. (2016). Protocol writing in clinical research. *Journal of Clinical and Diagnostic Research*, 10(11):ZE10–ZE13.

Austin, P. C. and Steyerberg, E. W. (2017). Events per variable (EPV) and the relative performance of different strategies for estimating the out-of-sample validity of logistic regression models. *Statistical Methods in Medical Research*, 26(2):796–808.

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., and Wicherts, J. M. (2020). Ensuring the quality and specificity of preregistrations. *PLOS Biology*, 18(12):e3000937.

Bischl, B., Casalicchio, G., Feurer, M., Gijsbers, P., Hutter, F., Lang, M., Gomes Mantovani, R., van Rijn, J., and Vanschoren, J. (2021). OpenML benchmarking suites. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62.

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

Boulesteix, A.-L., Hoffmann, S., Charlton, A., and Seibold, H. (2020). A replication crisis in methodological research? *Significance*, 17(5):18–21.

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

Boulesteix, A.-L., Stierle, V., and Hapfelmeier, A. (2015b). Publication bias in methodological computational research. *Cancer Informatics*, 14s5.

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

Bowman, S. D., DeHaven, A. C., Errington, T. M., Hardwicke, T. E., Mellor, D. T., Nosek, B. A., and Soderberg, C. K. (2020). OSF Prereg template. MetaArXiv. doi:10.31222/osf.io/epgjd.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Browne, R. H. (1995). On the use of a pilot sample for sample size determination. *Statistics in Medicine*, 14(17):1933–1940.

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

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

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., and Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280):1433–1436.

Campbell, D., McDonald, C., Cro, S., Jairath, V., and Kahan, B. C. (2022). Access to unpublished protocols and statistical analysis plans of randomised trials. *Trials*, 23(1).

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

Chambers, C. D. (2013). Registered reports: A new publishing initiative at *Cortex*. *Cortex*, 49(3):609–610.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., and Etchells, P. J. (2014). Instead of "playing the game" it is time to change the rules: Registered reports at *AIMS Neuroscience* and beyond. *AIMS Neuroscience*, 1(1):4–17.

Chan, A.-W., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., Hróbjartsson, A., Mann, H., Dickersin, K., Berlin, J. A., Doré, C. J., Parulekar, W. R., Summerskill, W. S. M., Groves, T., Schulz, K. F., Sox, H. C., Rockhold, F. W., Rennie, D., and Moher, D. (2013a). SPIRIT 2013 statement: Defining standard protocol items for clinical trials. *Annals of Internal Medicine*, 158(3):200.

Chan, A.-W., Tetzlaff, J. M., Gotzsche, P. C., Altman, D. G., Mann, H., Berlin, J. A., Dickersin, K., Hrobjartsson, A., Schulz, K. F., Parulekar, W. R., Krleza-Jeric, K., Laupacis, A., and Moher, D. (2013b). SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*, 346:e7586.

Claesen, A., Gomes, S., Tuerlinckx, F., and Vanpaemel, W. (2021). Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science*, 8(10).

Couronné, R., Probst, P., and Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC Bioinformatics*, 19(1).

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J. P. M., Schroeder, T. V., Sox, H. C., and Van Der Weyden, M. B. (2004). Clinical trial registration: A statement from the international committee of medical journal editors. *New England Journal of Medicine*, 351(12):1250–1251.

Dua, D. and Graff, C. (2017). UCI machine learning repository. https://archive.ics.uci.edu/ml. Accessed on November 3, 2022.

Eugster, M. J. A. (2011). *Benchmark experiments: A tool for analyzing statistical learning algorithms*. PhD thesis, LMU Munich.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2021). *Regression*. Springer Berlin Heidelberg, second edition.

Friedrich, S. and Friede, T. (2022). On the role of benchmarking data sets and simulations in method comparison studies. arXiv. doi:10.48550/arXiv.2208.01457.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36.

Hardwicke, T. E. and Ioannidis, J. P. A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour*, 2(11):793–796.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York.

Heirene, R., LaPlante, D., Louderback, E. R., Keen, B., Bakker, M., Serafimovska, A., and Gainsbury, S. M. (2021). Preregistration specificity & adherence: A review of preregistered gambling studies & cross-disciplinary comparison. PsyArXiv. doi:10.31234/osf.io/nj4es.

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., and Boulesteix, A.-L. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3).

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., and Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. *Royal Society Open Science*, 8(4).

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (1998). E9: Statistical principles for clinical trials. https://www.ema.europa.eu/en/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf. Accessed on November 3, 2022.

International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (2016). Integrated addendum to ICH E6(R1): Guideline for good clinical practice E6(R2). https://database.ich.org/sites/default/files/E6_R2_Addendum.pdf. Accessed on November 3, 2022.

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, 294(2):218.

John, L. K., Loewenstein, G., and Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5):524–532.

Kahan, B. C., Forbes, G., and Cro, S. (2020). How to design a pre-specified statistical analysis approach to limit p-hacking in clinical trials: the pre-SPEC framework. *BMC Medicine*, 18(1).

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217.

Kipruto, E. and Sauerbrei, W. (2022). Comparison of variable selection procedures and investigation of the role of shrinkage in linear regression-protocol of a simulation study in low-dimensional data. *PLOS ONE*, 17(10):e0271240.

Kiyonaga, A. and Scimeca, J. M. (2019). Practical considerations for navigating registered reports. *Trends in Neurosciences*, 42(9):568–572.

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., and Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, 4(44):1903.

Lindsay, D. S., Simons, D. J., and Lilienfeld, S. O. (2016). Research preregistration 101. *APS Observer*, 29(10):14–16.

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

McPhetres, J. (2020). What should a preregistration contain? PsyArXiv. doi:10.31234/osf.io/cj5mh.

Mertens, G. and Krypotos, A.-M. (2019). Preregistration of analyses of preexisting data. *Psychologica Belgica*, 59(1):338–352.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

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

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1(1).

National Research Council (2010). *The Prevention and Treatment of Missing Data in Clinical Trials*. National Academies Press, Washington, DC.

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., and Boulesteix, A.-L. (2022a). 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.

Nießl, C., Hoffmann, S., Ullmann, T., and Boulesteix, A.-L. (2022b). Explaining the optimistic performance evaluation of newly proposed methods: a cross-design validation experiment. arXiv. doi:10.48550/arXiv.2209.01885.

Nilsen, E. B., Bowler, D. E., and Linnell, J. D. C. (2020). Exploratory and confirmatory research in the open science era. *Journal of Applied Ecology*, 57(4):842–847.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606.

Nosek, B. A. and Lakens, D. (2014). Registered reports. *Social Psychology*, 45(3):137–141.

Ofosu, G. K. and Posner, D. N. (2021). Pre-analysis plans: An early stocktaking. *Perspectives on Politics*, pages 1–17.

Ogundimu, E. O., Altman, D. G., and Collins, G. S. (2016). Adequate sample size for developing prediction models is not simply related to events per variable. *Journal of Clinical Epidemiology*, 76:175–182.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

Pawel, S., Kook, L., and Reeve, K. (2022). Pitfalls and potentials in simulation studies. arXiv. doi:10.48550/arXiv.2203.13076.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., and Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology*, 49(12):1373–1379.

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

Regulation (EU) No 536/2014 of the European Parliament and of the Council of 16 April 2014 on clinical trials on medicinal products for human use, and repealing Directive 2001/20/EC (2014). *Official Journal*, L 158:1–76.

Riley, R. D., Snell, K. I. E., Ensor, J., Burke, D. L., Jr., F. E. H., Moons, K. G. M., and Collins, G. S. (2018). Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. *Statistics in Medicine*, 38(7):1276–1296.

Rivera, S. C., Liu, X., Chan, A.-W., Denniston, A. K., and Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *BMJ*, 370:m3210.

Schwab, S. and Held, L. (2020). Different worlds confirmatory versus exploratory research. *Significance*, 17(2):8–9.

Sedgwick, P. (2014). What are the four phases of clinical research trials? *BMJ*, 348:g3727–g3727.

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

Simmons, J., Nelson, L., and Simonsohn, U. (2021). Pre-registration: Why and how. *Journal of Consumer Psychology*, 31(1):151–162.

Smith, M. K. and Marshall, A. (2010). Importance of protocols for simulation studies in clinical drug development. *Statistical Methods in Medical Research*, 20(6):613–622.

Spence, O., Hong, K., Uba, R. O., and Doshi, P. (2019). Availability of study protocols for randomized trials published in high-impact medical journals: A cross-sectional analysis. *Clinical Trials*, 17(1):99–105.

Stewart, S., Rinke, E. M., McGarrigle, R., Lynott, D., Lunny, C., Lautarescu, A., Galizzi, M. M., Farran, E. K., and Crook, Z. (2020). Pre-registration and registered reports: a primer from UKRN. UK Reproducibility Network.

Steyerberg, E. W. (2019). *Clinical Prediction Models*. Springer International Publishing.

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

Sutton, A., Cruz, M. C. G. D. L., Leaviss, J., and Booth, A. (2017). Searching for trial protocols: A comparison of methods. *Research Synthesis Methods*, 9(4):551–560.

Tetzlaff, J. M., Chan, A.-W., Kitchen, J., Sampson, M., Tricco, A. C., and Moher, D. (2012). Guidelines for randomized clinical trial protocol content: a systematic review. *Systematic Reviews*, 1:43.

Umscheid, C. A., Margolis, D. J., and Grossman, C. E. (2011). Key concepts of clinical trials: A narrative review. *Postgraduate Medicine*, 123(5):194–204.

van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., Hall, A. N., Kosie, J. E., Kruse, E. T., Olsen, J., Ritchie, S. J., Valentine, K. D., van 't Veer, A. E., and Bakker, M. (2019). Preregistration of secondary data analysis: A template and tutorial. PsyArXiv. doi:10.31234/osf.io/hvfmr.

van der Ploeg, T., Austin, P. C., and Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, 14(1).

van Smeden, M., de Groot, J. A. H., Moons, K. G. M., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., and Reitsma, J. B. (2016). No rationale for 1 variable per 10 events criterion for binary logistic regression analysis. *BMC Medical Research Methodology*, 16(1).

van Smeden, M., Moons, K. G. M., de Groot, J. A. H., Collins, G. S., Altman, D. G., Eijkemans, M. J. C., and Reitsma, J. B. (2019). Sample size for binary logistic prediction models: Beyond events per variable criteria. *Statistical Methods in Medical Research*, 28(8):2455–2474.

van 't Veer, A. E. and Giner-Sorolla, R. (2016). Pre-registration in social psychology—a discussion and suggested template. *Journal of Experimental Social Psychology*, 67:2–12.

Vanschoren, J., van Rijn, J. N., Bischl, B., and Torgo, L. (2013). OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60.

Vickerstaff, V., Omar, R. Z., and Ambler, G. (2019). Methods to adjust for multiple comparisons in the analysis and sample size calculation of randomised controlled trials with multiple primary outcomes. *BMC Medical Research Methodology*, 19(1).

Wagenmakers, E.-J., Dutilh, G., and Sarafoglou, A. (2018). The creativity-verification cycle in psychological science: New methods to combat old idols. *Perspectives on Psychological Science*, 13(4):418–427.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., and Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6):632–638.

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

Weston, S. J., Ritchie, S. J., Rohrer, J. M., and Przybylski, A. K. (2019). Recommendations for increasing the transparency of analysis of preexisting data sets. *Advances in Methods and Practices in Psychological Science*, 2(3):214–227.

Whitehead, A. L., Julious, S. A., Cooper, C. L., and Campbell, M. J. (2015). Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for

the external pilot and main trial for a continuous outcome variable. *Statistical Methods in Medical Research*, 25(3):1057–1073.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., and Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686.

World Health Organization (2018). International standards for clinical trial registries (version 3.0). https://apps.who.int/iris/bitstream/handle/10665/274994/9789241514743-eng.pdf. Accessed on November 3, 2022.

World Medical Association (2013). WMA Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA*, 310(20):2191–4.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Wu, Y.-C. and Lee, W.-C. (2014). Alternative performance measures for prediction models. *PLoS ONE*, 9(3):e91249.

A Protocol appendix

A.1 Selected datasets

Data ID	Name	Original task type	n	pct^{evt}	n^{evt}	p
405	mtp	regr	4,450	0.5	2,220	3
558	bank32nh	regr	8,192	0.5	4,096	6
1502	skin-segmentation	binclassif	245,057	0.21	50,859	3
3277	QSAR-TID-10980	regr	5,766	0.5	2,883	4
40668	connect-4	multclassif	67,557	0.34	23,084	29
41167	dionis	multclassif	416,188	0.01	2,469	3
41168	jannis	multclassif	83,733	0.46	38,522	54
42395	SantanderCustomer Satisfaction	binclassif	200,000	0.1	20,098	32
42468	hls4ml_lhc_jets_hlf	multclassif	830,000	0.2	167,851	16
42721	Airlines_DepDelay_1M	regr	1,000,000	0.41	405,990	13
43174	superconduct	regr	21,263	0.5	10,547	16
43377	Pulsar-Dataset-HTRU2	binclassif	17,897	0.09	1,639	2
43450	Milan-Airbnb-Open-Data-(only-entire-apartments)	regr	9,322	0.49	4,604	7
43635	League-of-Legends-Diamond-Games-(First-15-Minutes)	binclassif	48,651	0.49	24,062	14
43904	law-school-admission-bianry	binclassif	20,800	0.32	6,694	10

Table A.1: Dataset characteristics for the 15 datasets used in the pilot study (Data ID: OpenML data ID, Original task type: task type before the dichotomization during preprocessing (regression, binary classification, multiclass classification), n: number of observations, pct^{evt}: percentage of observations in the minority outcome class, n^{evt}: number of observations in the minority outcome class, p: number of features).

Data ID	Name	Original task type	n	pct^{evt}	n^{evt}	p
3	kr-vs-kp	binclassif	3,196	0.48	1,527	2
44	spambase	binclassif	4,601	0.39	1,813	2
189	kin8nm	regr	8,192	0.5	4,096	6
273	IMDB.drama	binclassif	120,919	0.36	43,779	13
287	wine_quality	regr	6,497	0.2	1,277	2
308	puma32H	regr	8,192	0.5	4,096	6
351	codrna	binclassif	488,565	0.33	162,855	8
354	poker	binclassif	1,025,010	0.5	511,308	15
416	yprop_4_1	regr	8,885	0.5	4,420	7
422	topo_2_1	regr	8,885	0.5	4,420	7
507	space_ga	regr	3,107	0.5	1,553	2
537	houses	regr	20,640	0.5	10,317	8
574	house_16H	regr	22,784	0.5	11,347	16
953	splice	binclassif	3,190	0.48	1,535	2
959	nursery	binclassif	12,960	0.33	4,320	6
1120	MagicTelescope	binclassif	19,020	0.35	6,688	10
1216	Click_prediction_small	binclassif	1,496,391	0.04	66,781	3
1433	svmguide1	binclassif	7,089	0.44	3,089	4
1461	bank-marketing	binclassif	45,211	0.12	5,289	8
1475	first-order-theorem-proving	multclassif	6,118	0.42	2,554	4
1481	kr-vs-k	multclassif	28,056	0.16	4,553	7
1489	phoneme	binclassif	5,404	0.29	1,586	2
4134	Bioresponse	binclassif	3,751	0.46	1,717	2
4534	PhishingWebsites	binclassif	11,055	0.44	4,898	7
4545	OnlineNewsPopularity	regr	39,644	0.49	19,562	31
23517	numerai28.6	binclassif	96,320	0.49	47,662	21
40672	fars	multclassif	100,968	0.42	42,116	35
40996	Fashion-MNIST	multclassif	70,000	0.1	7,000	11
41027	jungle_chess_2pcs_raw_endgame_complete	multclassif	44,819	0.49	21,757	6
41081	SVHN	multclassif	99,289	0.19	18,960	30
41082	USPS	multclassif	9,298	0.17	1,553	2
41142	christine	binclassif	5,418	0.5	2,709	4
41150	MiniBooNE	binclassif	130,064	0.28	36,499	50
41159	guillermo	binclassif	20,000	0.4	8,003	12
41161	riccardo	binclassif	20,000	0.25	5,000	7
41163	dilbert	multclassif	10,000	0.2	2,049	3
41214	freMTPL2freq	regr	678,013	0.05	34,060	16
41228	Klaverjas2018	binclassif	981,541	0.46	453,202	96

Continued on next page

Table A.2: Dataset characteristics for the 75 datasets used in the main study.

Continued from previous page

Data ID	Name	Original task type	n	pct^{evt}	n^{evt}	p
41990	GTSRB-HueHist	multclassif	51,839	0.06	3,000	4
42092	house_sales	regr	21,613	0.5	10,749	17
42208	nyc-taxi-green-dec-2016	regr	581,835	0.49	286,219	4
42225	diamonds	regr	53,940	0.5	26,955	18
42477	default-of-credit-card-clients	binclassif	30,000	0.22	6,636	10
42570	Mercedes_Benz_Greener_Manufacturing	regr	4,209	0.5	2,103	3
42571	Allstate_Claims_Severity	regr	188,318	0.5	94,159	103
42572	Santander_transaction_value	regr	4,459	0.5	2,229	3
42688	Brazilian_houses	regr	10,692	0.5	5,346	8
42726	abalone	regr	4,177	0.5	2,081	3
42769	Higgs	binclassif	1,000,000	0.47	470,080	28
42876	WorkersCompensation	regr	100,000	0.5	49,996	8
42903	physicochemical-protein	regr	45,730	0.5	22,861	9
43090	30mlday	regr	300,000	0.5	150,000	23
43093	MiamiHousing2016	regr	13,932	0.5	6,957	11
43140	ACSPublicCoverage	binclassif	1,138,289	0.3	338,456	18
43141	ACSIncome	regr	1,664,500	0.5	829,343	10
43144	SGEMM_GPU_kernel_performance	regr	241,600	0.5	120,800	24
43355	Brilliant-Diamonds	regr	119,307	0.5	59,499	19
43390	Churn-for-Bank-Customers	binclassif	10,000	0.2	2,037	3
43437	Gender-Recognition-by-Voice	binclassif	3,168	0.5	1,584	2
43459	Metro-Manila-Flood-Landscape-Data	regr	3,510	0.42	1,473	2
43527	Malware-Analysis-Datasets-PE-Section-Headers	binclassif	43,293	0.04	1,725	2
43534	Production-cross-sections-of-Inert-Doublet-Model	regr	50,625	0.5	25,312	5
43546	AqSolDB-A-curated-aqueous-solubility-dataset	regr	9,982	0.5	4,991	7
43617	Medical-Appointment	binclassif	61,214	0.21	12,868	20
43622	Binary-Dataset-of-Phishing-and-Legitimate-URLs	binclassif	11,000	0.5	5,500	8

Continued on next page

Table A.2: Dataset characteristics for the 75 datasets used in the main study.

Continued from previous page

Data ID	Name	Original task type	n	pct^{evt}	n^{evt}	p
43745	Delinquency-Telecom-Dataset	binclassif	209,593	0.12	26,162	25
43837	New-Delhi-Rental-Listings	regr	17,890	0.48	8,565	13
43838	Municipal-Debt-Risk-Analysis	binclassif	138,509	0.46	63,971	13
43846	400k-NYSE-random-investments--financial-ratios	binclassif	405,258	0.35	140,818	22
43849	2018-Airplane-Flights	regr	9,534,417	0.5	4,767,198	8
43873	sarcos	regr	44,484	0.5	22,242	21
43892	national-longitudinal-survey-binary	binclassif	4,908	0.38	1,853	2
43926	ames_housing	regr	2,930	0.5	1,463	2
43963	CPS1988	regr	28,155	0.49	13,847	6
44027	year	regr	515,345	0.47	244,074	90

Table A.2: Dataset characteristics for the 75 datasets used in the main study (Data ID: OpenML data ID, Original task type: task type before the dichotomization during preprocessing (regression, binary classification, multiclass classification), n: number of observations, pct^{evt}: percentage of observations in the minority outcome class, n^{evt}: number of observations in the minority outcome class, p: number of features).

Data ID	Number of features							
	removed	as dummies	constant	sparse	non-sparse	p	p_{num}	p_{bin}
44	0	57	0	19	38	2	2	0
189	0	8	0	0	8	6	6	0
273	0	1,001	0	988	13	13	0	13
287	0	11	0	0	11	2	2	0
308	0	32	0	0	32	6	6	0
351	0	8	0	0	8	8	8	0
354	0	75	0	60	15	15	0	15
405	0	202	0	10	192	3	3	0
416	0	251	39	152	60	7	7	0
422	0	266	5	35	226	7	7	0
507	0	6	0	0	6	2	2	0
537	0	8	0	0	8	8	8	0
558	0	32	0	0	32	6	6	0

Continued on next page

Table A.3: Preprocessing outcomes for all 90 selected datasets.

Continued from previous page

Data ID	Number of features							
	removed	as dummies	constant	sparse	non-sparse	p	p_{num}	p_{bin}
574	0	16	0	0	16	16	16	0
953	1	227	0	47	180	2	0	2
959	0	19	0	0	19	6	0	6
1120	1	10	0	0	10	10	10	0
1216	8	3	0	0	3	3	3	0
1433	0	4	0	0	4	4	4	0
1461	0	42	0	20	22	8	3	5
1475	0	51	0	0	51	4	4	0
1481	0	34	0	4	30	7	0	7
1489	0	5	0	0	5	2	2	0
1502	0	3	0	0	3	3	3	0
3277	1	1,024	0	894	130	4	0	4
4134	0	1,776	0	1,127	649	2	2	0
4534	0	38	0	5	33	7	0	7
4545	6	54	0	2	52	31	26	5
23517	0	21	0	0	21	21	21	0
40668	0	84	0	55	29	29	0	29
40672	0	338	0	303	35	35	5	30
40996	0	784	0	105	679	11	11	0
41027	0	6	0	0	6	6	6	0
41081	0	3,072	0	0	3,072	30	30	0
41082	0	256	0	0	256	2	2	0
41142	0	1,636	25	18	1,593	4	4	0
41150	0	50	0	0	50	50	50	0
41159	0	4,296	15	31	4,250	12	12	0
41161	0	4,296	13	20	4,263	7	7	0
41163	0	2,000	0	0	2,000	3	3	0
41167	0	60	6	0	54	3	3	0
41168	0	54	0	0	54	54	54	0
41214	1	43	0	27	16	16	6	10
41228	3	96	0	0	96	96	0	96
41990	0	256	0	13	243	4	4	0
42092	2	102	0	85	17	17	12	5
42208	3	507	0	503	4	4	1	3
42225	0	23	0	5	18	18	6	12
42395	1	200	0	0	200	32	32	0
42468	0	16	0	0	16	16	16	0
42477	1	82	0	46	36	10	6	4

Continued on next page

Table A.3: Preprocessing outcomes for all 90 selected datasets.

Continued from previous page

Data ID	Number of features							
	removed	as dummies	constant	sparse	non-sparse	p	p_{num}	p_{bin}
42570	1	555	12	440	103	3	0	3
42571	1	1,037	0	934	103	103	14	89
42572	1	4,991	256	4,359	376	3	3	0
42688	0	48	0	34	14	8	3	5
42721	0	799	0	786	13	13	3	10
42726	0	9	0	0	9	3	3	0
42769	0	32	0	4	28	28	24	4
42876	4	11	0	3	8	8	5	3
42903	0	9	0	0	9	9	9	0
43090	1	60	0	37	23	23	14	9
43093	1	28	0	13	15	11	9	2
43140	0	123	0	105	18	18	2	16
43141	2	285	0	275	10	10	2	8
43144	3	26	0	2	24	24	0	24
43174	0	81	0	0	81	16	16	0
43355	3	31	0	12	19	19	1	18
43377	0	8	0	0	8	2	2	0
43390	3	11	0	0	11	3	2	1
43437	0	20	0	0	20	2	2	0
43450	4	109	0	62	47	7	4	3
43459	0	4	0	0	4	2	2	0
43527	1	4	0	0	4	2	2	0
43534	7	5	0	0	5	5	5	0
43546	5	23	0	3	20	7	7	0
43617	6	136	0	111	25	20	1	19
43622	3	11	0	1	10	8	7	1
43635	4	14	0	0	14	14	14	0
43745	3	32	0	7	25	25	25	0
43837	4	22	0	8	14	13	7	6
43838	2	23	0	10	13	13	11	2
43846	3	72	0	50	22	22	18	4
43849	7	19	0	11	8	8	2	6
43873	6	21	0	0	21	21	21	0
43892	1	72	0	53	19	2	1	1
43904	0	17	0	7	10	10	5	5
43926	0	321	2	230	89	2	0	2
43963	0	8	0	2	6	6	2	4
44027	0	90	0	0	90	90	90	0

Table A.3: Preprocessing outcomes for the 90 datasets selected for the study. The information in columns 2-7 follows the order of the preprocessing procedure described in Section 3.1.6 (p_{num}: number of sampled numeric features, p_{bin}: number of sampled binary features).

A.2 Power plots for second hypothesis

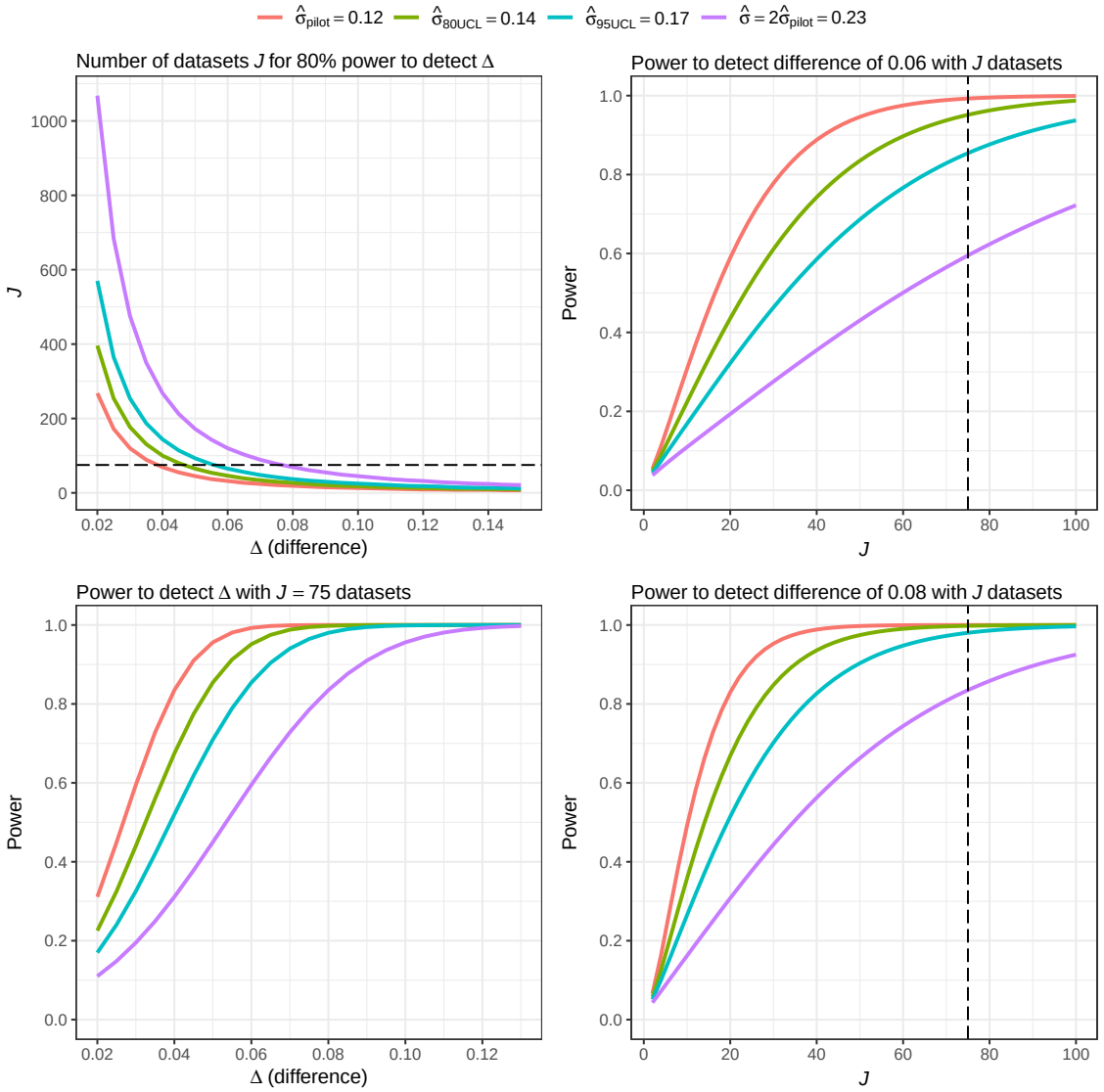

Figure A.1: Power plots for second hypothesis, so Δ refers to a difference in mean random forest optimism. The four colored lines represent four standard deviation estimates (or scenarios) based on the pilot study results. The dashed black line indicates the number of datasets that will be analyzed in the main benchmark experiment (75).

A.3 Protocol amendment history

Change	Rationale	Affected protocol section(s)
Minimum required total number of EPV was reduced from 1,000 EPV to 625 EPV.	An early assessment of the datasets available on OpenML showed that requiring every dataset to have at least 1,000 EPV in total, as originally intended, would lead to a significantly smaller set of suitable datasets.	4. Datasets

Table A.4: Changes compared to the previous version of the study protocol.

A.4 Definition of the evaluation metric for the first hypothesis for accuracy and Brier score

For CV iteration $i = 1, \ldots, 50$ of a given dataset j, let $Acc_{ij}^{test}(n)$ and $BS_{ij}^{test}(n)$ be the test data accuracy and Brier score of the model generated using the training data subset with n EPV, and let $MaxAcc_{ij}^{test}$ and $MaxBS_{ij}^{test}$ be the test data accuracy and Brier score of the model generated using the full training data. Using threshold $t = \{0.9, 0.95, 0.99\}$, the minimum numbers of EPV at which a method achieves a good predictive performance with respect to accuracy and Brier score are then defined for iteration i of dataset j as

$$(EPV_{min}^{Acc})_{ij} = \min\left\{n \in EPV \mid Acc_{ij}^{test}(n) \geq t \cdot MaxAcc_{ij}^{test}\right\} \quad \text{and}$$

$$(EPV_{min}^{BS})_{ij} = \min\left\{n \in EPV \mid BS_{ij}^{test}(n) \leq \frac{1}{t} \cdot MaxBS_{ij}^{test}\right\} \quad ,$$

where $i = 1, \ldots, 50$, $j = 1, \ldots, J$, and $EPV = \{5, 10, 20, 30, 40, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, 400, 425, 450, 475, 500\}$.

B Additional figures and tables

B.1 For the confirmatory analyses

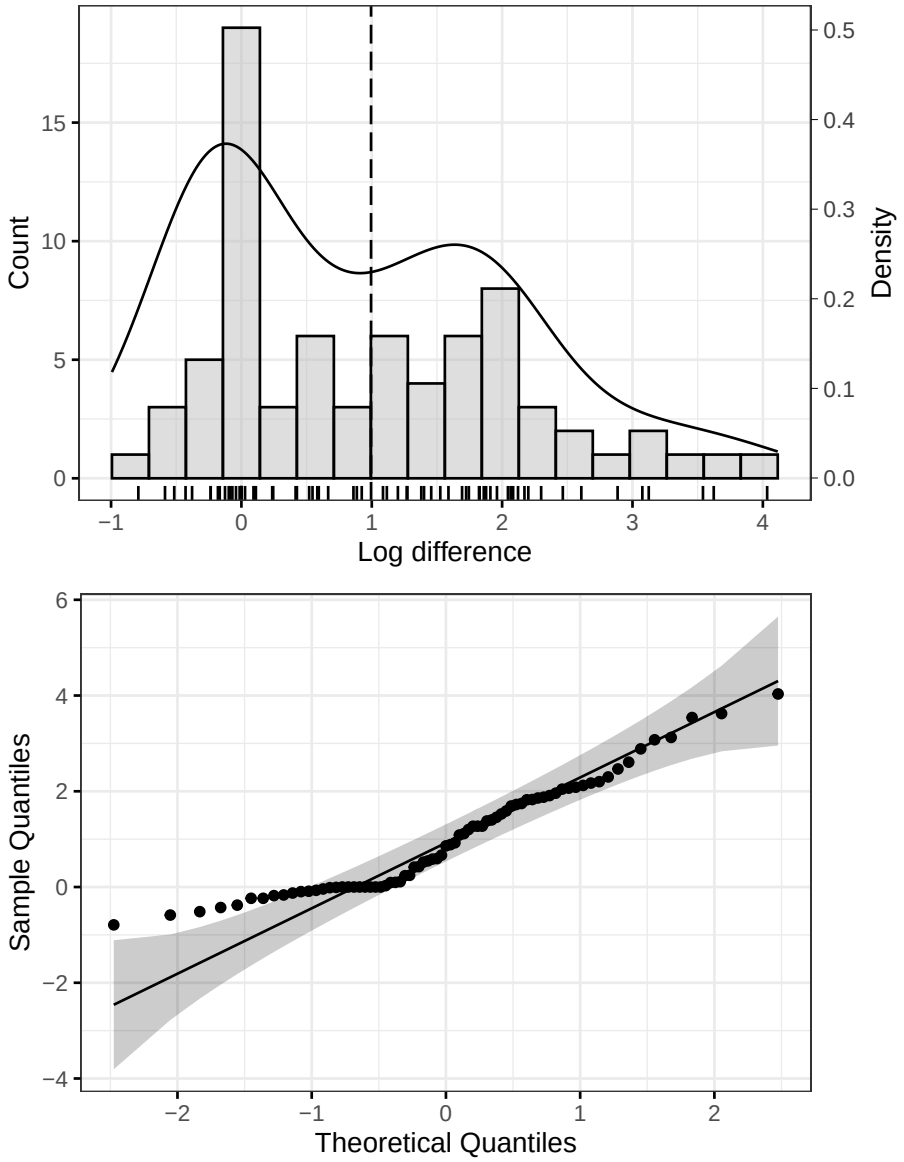

Figure B.1: Histogram with density curve (top) and Q-Q plot (bottom) for the first hypothesis (log difference).

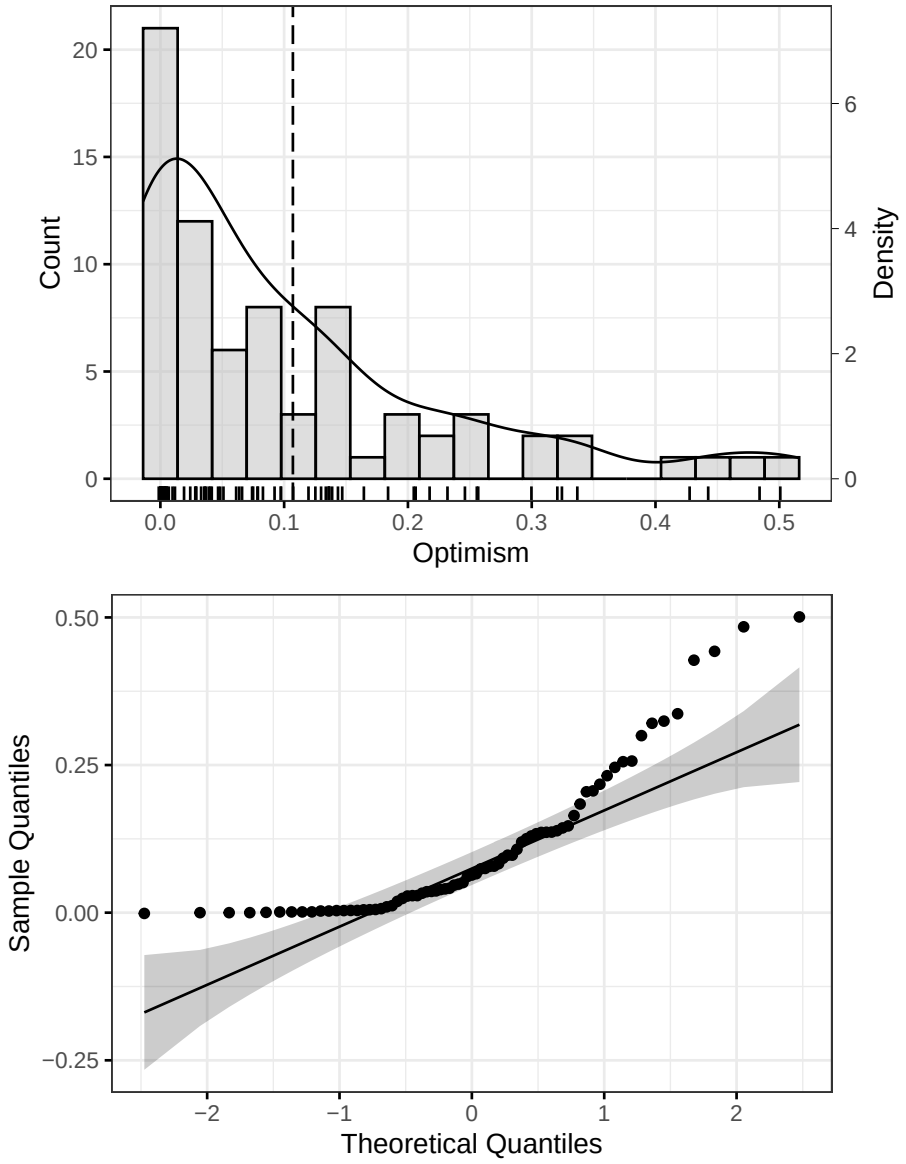

Figure B.2: Histogram with density curve (top) and Q-Q plot (bottom) for the second hypothesis (optimism).

B.2 For the sensitivity analyses

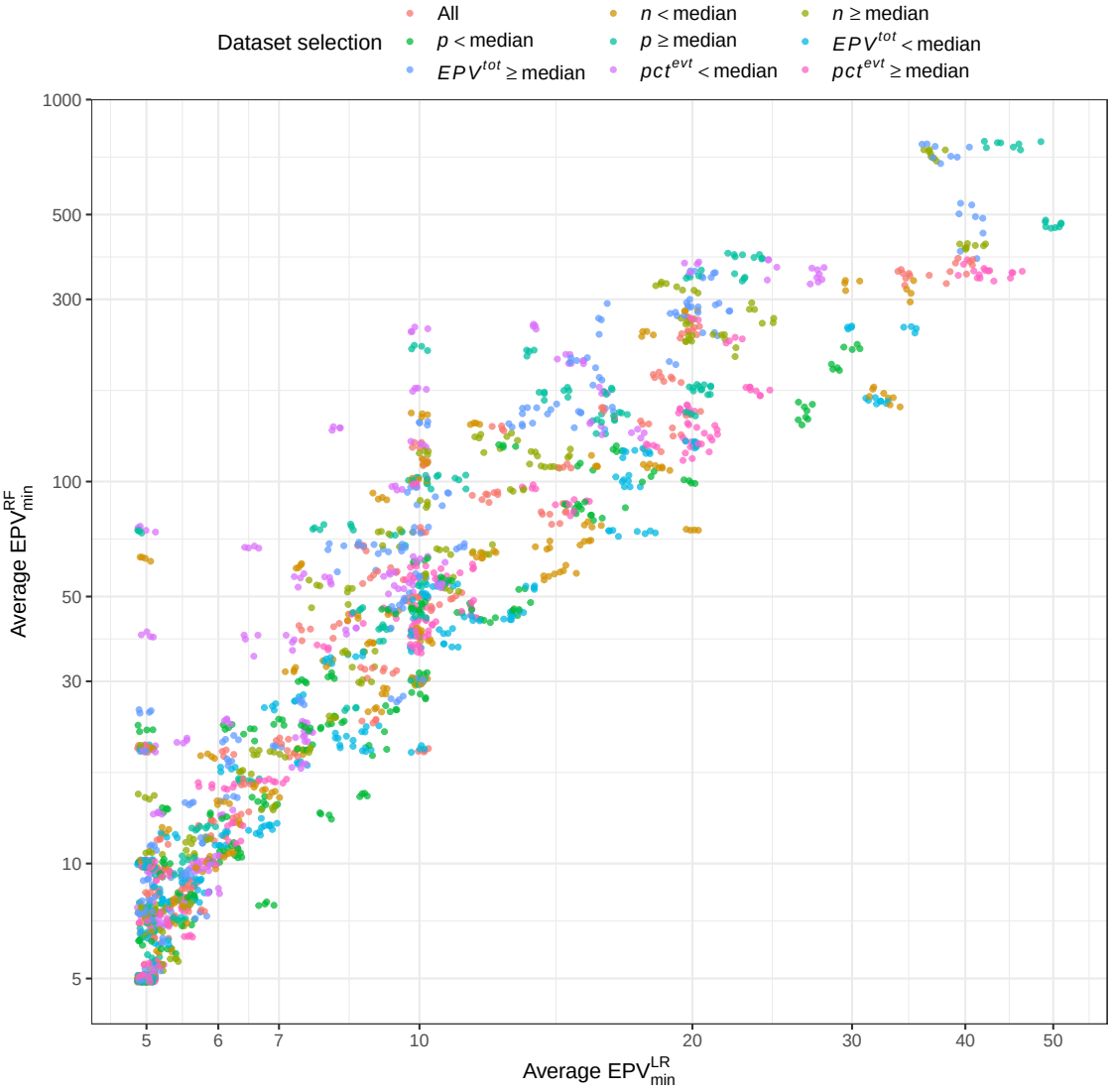

Figure B.3: Scatterplot of the aggregated results of all 2,160 analyses, colored by dataset group. For both axis, a logarithmic scale is used. Slight jitter was added to visually separate overlapping points.

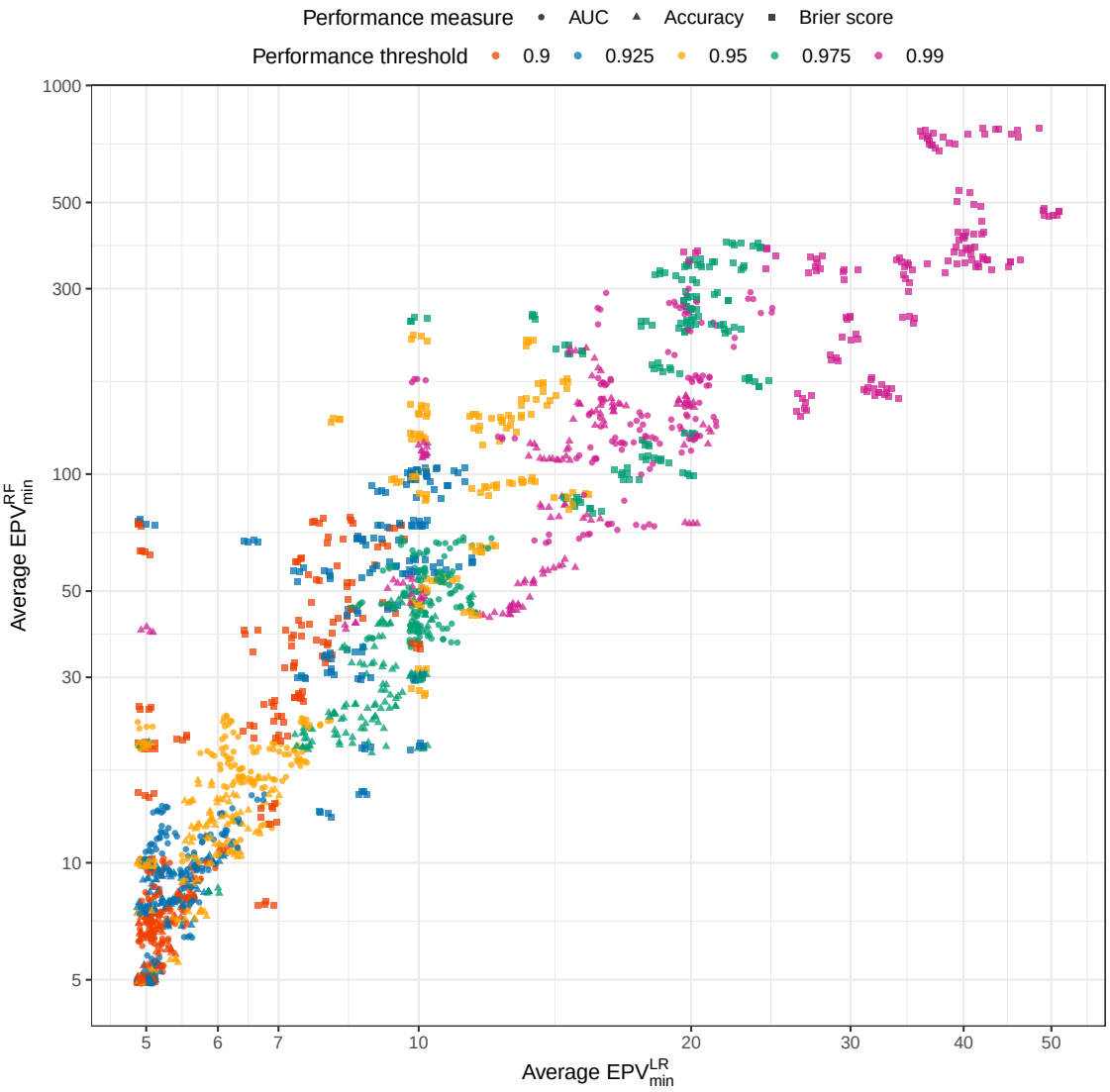

Figure B.4: Scatterplot of the aggregated results of all 2,160 analyses, colored by performance threshold and with shapes reflecting the performance measure. For both axis, a logarithmic scale is used. Slight jitter was added to visually separate overlapping points.

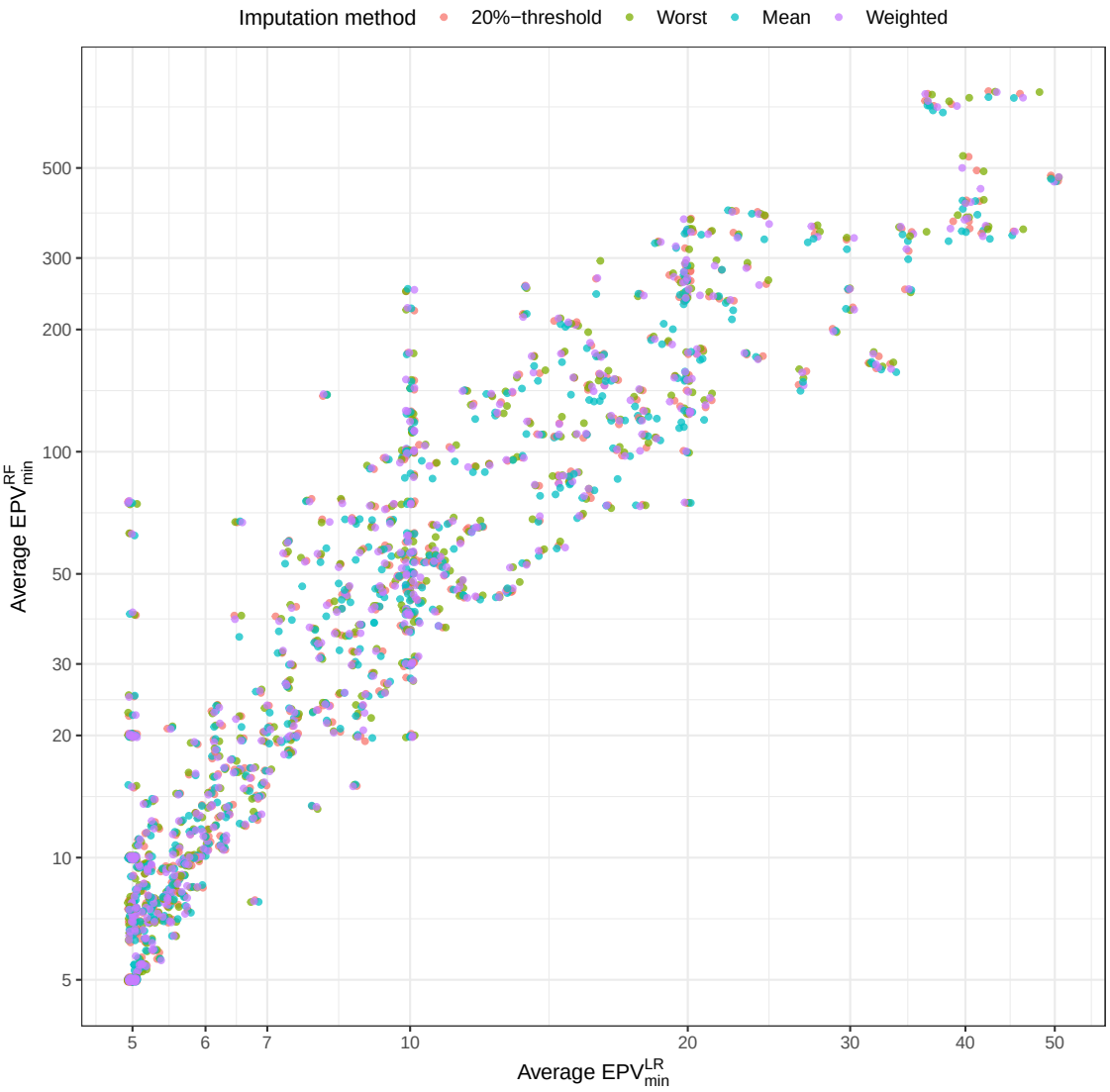

Figure B.5: Scatterplot of the aggregated results of all 2,160 analyses, colored by imputation method. For both axis, a logarithmic scale is used. Slight jitter was added to visually separate overlapping points.

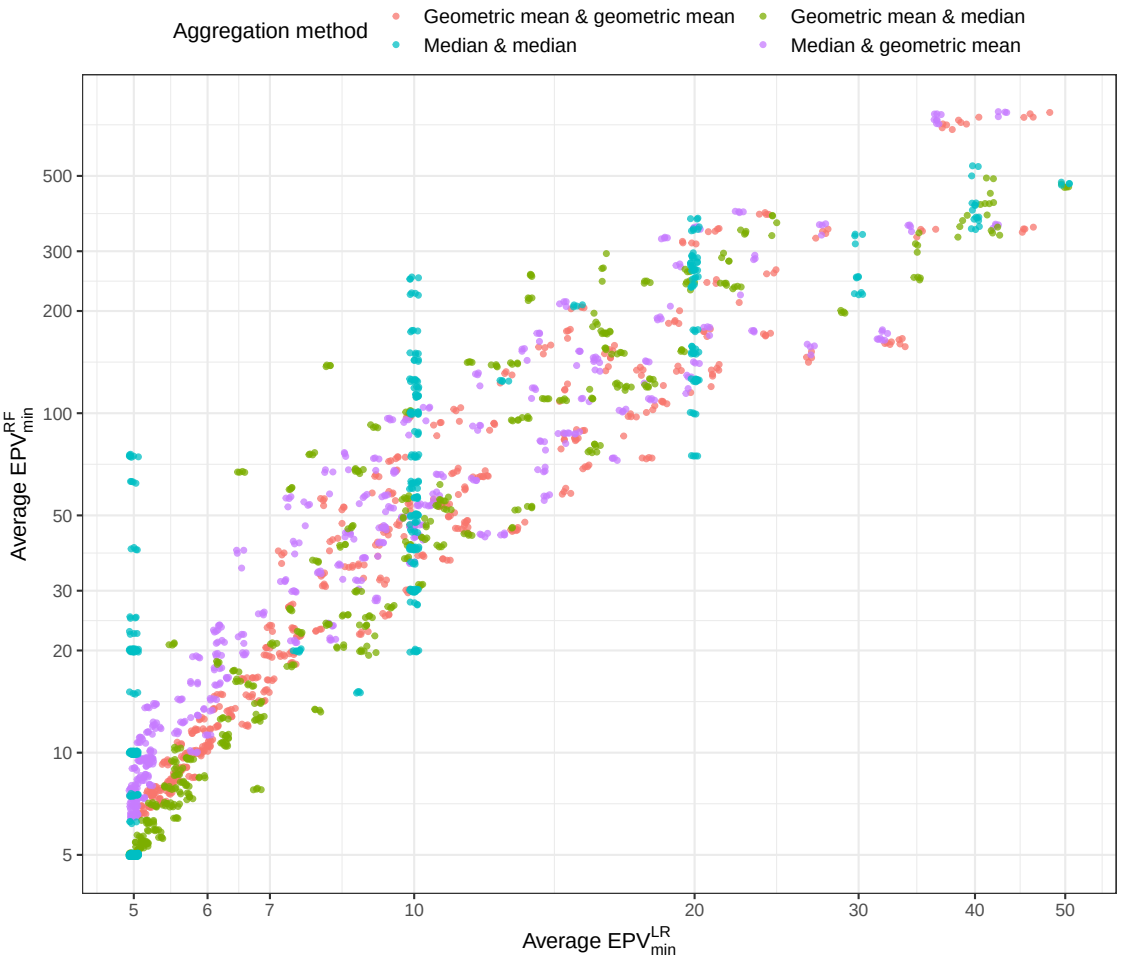

Figure B.6: Scatterplot of the aggregated results of all 2,160 analyses, colored by aggregation method. For both axis, a logarithmic scale is used. Slight jitter was added to visually separate overlapping points.

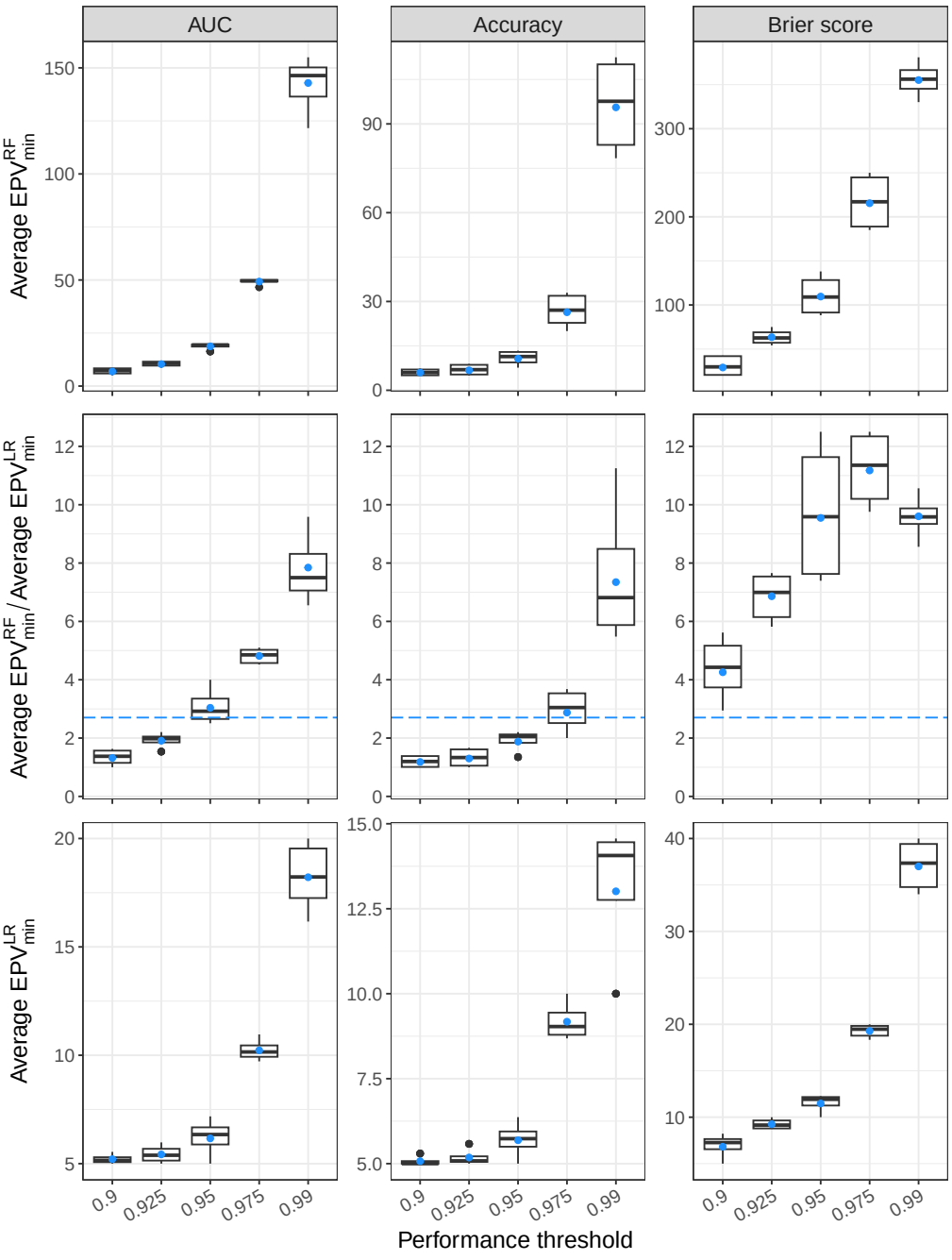

Figure B.7: Boxplots by performance measure and threshold for the aggregated results of the 240 considered analyses, leaving the other analysis options at their defaults (20%-threshold rule imputation and geometric mean aggregation). The blue points show geometric means, and the dashed blue line indicates the geometric mean of the actual confirmatory analysis.

Measure	Threshold	Number of CV iterations		Number of different datasets		
		LR	RF	LR	RF	LR or RF
AUC	90%	0	0	0	0	0
	92.5%	0	0	0	0	0
	95%	0	0	0	0	0
	97.5%	3	172	1	8	8
	99%	31	540	5	23	24
Accuracy	90%	0	0	0	0	0
	92.5%	0	0	0	0	0
	95%	0	1	0	1	1
	97.5%	1	151	1	7	7
	99%	7	459	4	22	23
Brier score	90%	5	435	2	12	13
	92.5%	5	553	2	13	14
	95%	12	665	3	16	17
	97.5%	19	845	3	25	25
	99%	39	1,192	8	40	40

Table B.1: Number of missing values in the evaluation metric by method for the 15 considered performance measure-threshold combinations. The number of missing values in each cell in columns 3 and 4 is out of 3,750 total CV iterations (50 CV iterations $*$ 75 datasets). The numbers of different datasets with missing values in columns 5-7 are always out of 75 total datasets.

C Electronic appendix

The electronic appendix contains an electronic version of this thesis (`MA_Lange.pdf`), one folder (`R Code`) and a `README` document. The folder `R Code` has the four subfolders `Dataset selection`, `Pilot study`, `Main study`, and `Functions`. Also in that folder is the spreadsheet that was used in the dataset selection process to document which datasets met at least one exclusion criterion. More details on the electronic appendix and its contents can be found in the included `README`.

Declaration of authorship

I hereby confirm that I prepared this thesis independently and that the thoughts taken directly or indirectly from other sources are indicated accordingly. The work contained in this thesis has not been previously submitted for examination.

Munich, December 19, 2022

Felix Julian David Lange