

Simon Klau

**Addressing the challenges of uncertainty  
in regression models for high dimensional  
and heterogeneous data from observational  
studies**

Dissertation an der Fakultät für Mathematik, Informatik und Statistik  
der Ludwig-Maximilians-Universität München

Eingereicht am 16.04.2020

Erstgutachterin: Prof. Dr. Anne-Laure Boulesteix  
Zweitgutachterin: Prof. Dr. Manuela Zucknick  
Drittgutachter: PD Dr. Alexander Hapfelmeier

Tag der Disputation: 02.10.2020

## Zusammenfassung

Der Mangel an Replizierbarkeit von wissenschaftlichen Ergebnissen in unterschiedlichen Forschungsdisziplinen hat in der jüngeren Vergangenheit an Aufmerksamkeit gewonnen und umfassende Diskussionen ausgelöst. Eine große Rolle in dieser „replication crisis“ spielen verschiedene Arten von Unsicherheiten, die an unterschiedlichen Punkten der Datenerhebung und statistischen Analyse auftreten. Nichtsdestotrotz werden diesen Unsicherheiten und den daraus resultierenden Folgen auch in der derzeitigen Forschungspraxis oft nur wenig Beachtung geschenkt – mit der Gefahr geringer Zuverlässigkeit und Glaubwürdigkeit von wissenschaftlichen Entdeckungen.

Zur Analyse dieses Umstands und der Entwicklung von Lösungsansätzen für das Problem, werden in dieser Arbeit Messunsicherheit, Datenaufbereitungs-Unsicherheit, Stichproben-Unsicherheit, Methoden-Unsicherheit und Modellunsicherheit definiert, und insbesondere im Kontext von Regressionsmodellen untersucht. Dazu werden Datensätze von Beobachtungsstudien mit einem Fokus auf die Merkmale der Hochdimensionalität und heterogener Variablen herangezogen, die zunehmend an Bedeutung gewinnen. Hochdimensionale Daten, d.h. Daten, die eine größere Anzahl an Variablen als Beobachtungen aufweisen, spielen im Bereich der medizinischen Forschung eine wesentliche Rolle, wo große Mengen an molekularen Daten eines Patienten zunehmend ressourcengünstig erhoben werden können. Liegen verschiedene Arten von molekularen Daten (omics-Daten) vor, ist man außerdem mit dem Umstand der Heterogenität konfrontiert. Darüber hinaus finden sich heterogene Daten in vielen Beobachtungsstudien wieder, in denen Variablen verschiedener Art erhoben werden, oder unterschiedlichen Quellen entstammen.

Die Arbeit besteht im Wesentlichen aus vier Beiträgen, die auf verschiedenen Herangehensweisen an die Thematik basieren und unterschiedliche Schwerpunkte der Untersuchungen setzen.

Der erste Beitrag kann als praktisches Beispiel zur Veranschaulichung von Datenaufbereitungs- und Methoden-Unsicherheit im Kontext der Prädiktion und Variablenselektion anhand sowohl hochdimensionaler als auch heterogener Daten angesehen werden. Zunächst beschäftigt sich dieser Beitrag intensiv mit der Entwicklung der Methode priority-Lasso, einem hierarchischen Verfahren zur Prädiktion mit multiplen omics-Daten. Auf standard Lasso basierend, geht priority-Lasso auf die verschiedenen Datenblöcke ein, indem nach festgelegter Priorität der Blöcke sukzessive Lasso-Modelle auf jedem dieser Blöcke gefittet werden, und der lineare Prädiktor des jeweiligen Fits als Offset im darauffolgenden Fit verwendet wird. In einem zweiten Teil wird diese Methode in einer aktuellen Studie zur akuten myeloischen Leukämie (AML) eingesetzt und einem Vergleich zu standard Lasso unterzogen. Durch verschiedene Möglichkeiten der Variablendefinition und Spezifikation von Einstellungen im Zuge der Methoden-Durchführung werden Datenaufbereitungs- und Methoden-Unsicherheit verdeutlicht, die sich in der Schätzung der Effekte und somit in der Prädiktionsgüte und der Variablenauswahl widerspiegeln.

In einem zweiten Beitrag wird Methoden-Unsicherheit mit Stichproben-Unsicherheit im Kontext der Variablenselektion und des Rankings von „omics“-Biomarkern verglichen.

Dazu wird ein anwenderfreundliches und vielschichtig einsetzbares resampling-basiertes Framework entwickelt. Dieses Framework wird auf Basis hochdimensionaler und teilweise heterogener omics-Datensätze von AML-Patienten angewendet. Dazu werden drei statistische Szenarien beleuchtet: Variablenselektion in multivariabler Regression auf Basis unterschiedlicher Typen von omics-Biomarkern, Ranking von Biomarkern auf Basis ihrer Variablenwichtigkeit in random forests, und Identifikation von Genen auf Basis von Tests auf differentielle Genexpression.

In den Beiträgen 3 und 4 wird zur Veranschaulichung und zum Vergleich von Unsicherheiten unter anderem das „Vibration of Effects“-Framework verwendet, mit dem ursprünglich die Modellunsicherheit in einer umfangreichen epidemiologischen Studie (NHANES) analysiert wurde. Diese Beiträge beschäftigen sich intensiv mit der methodischen Erweiterung des Frameworks auf zusätzliche Unsicherheitstypen.

Zunächst wird diese Erweiterung in Beitrag 3 für Stichproben- und Datenaufbereitungs-Unsicherheit vorgestellt. Zur praktischen Veranschaulichung wird dann ein umfangreicher Datensatz mit heterogener Variablenstruktur aus der psychologischen Forschung herangezogen (SAPA-Project), auf dem Modell-, Stichproben-, und Datenaufbereitungs-Unsicherheit im Kontext von logistischen Regressionsmodellen für unterschiedliche Stichprobengrößen untersucht werden. Neben dem Vergleich dieser einzelnen Unsicherheitstypen wird ein Verfahren vorgestellt, das es erlaubt, die kumulative Modell- und Datenaufbereitungs-Unsicherheit zu quantifizieren und mittels einer Varianzzerlegung ihre relativen Anteile an der Gesamtunsicherheit zu analysieren.

In Beitrag 4 wird das Vibration of Effects Framework zusätzlich auf Messunsicherheit erweitert. Darauf basierend wird auf dem NHANES-Datensatz eine Vergleichsstudie zwischen Modell-, Stichproben- und Messunsicherheit im Kontext der Überlebenszeitanalyse durchgeführt. Dabei wird ein Fokus auf verschiedene Szenarien von Messunsicherheit gelegt, die sich dahingehend unterscheiden, für welche der Variablen im Regressionsmodell ein Messfehler angenommen wird. Mit Hilfe einer umfassenden Simulationsstudie auf Basis der NHANES-Daten wird außerdem das Verhalten verschiedener Unsicherheitsquellen bei steigender Stichprobengröße analysiert.

## Summary

The lack of replicability in research findings from different scientific disciplines has gained wide attention in the last few years and led to extensive discussions. In this ‘replication crisis’, different types of uncertainty play an important role, which occur at different points of data collection and statistical analysis. Nevertheless, the consequences are often ignored in current research practices with the risk of low credibility and reliability of research findings.

For the analysis and the development of solutions to this problem, we define measurement uncertainty, sampling uncertainty, data pre-processing uncertainty, method uncertainty, and model uncertainty, and investigate them in particular in the context of regression analyses. Therefore, we consider data from observational studies with the focus on high dimensionality and heterogeneous variables, which are characteristics of growing importance. High dimensional data, i.e., data with more variables than observations, play an important role in the area of medical research, where large amounts of molecular data (omics data) can be collected with ever decreasing expense and effort. Where several types of omics data are available, we are additionally faced with heterogeneity. Moreover, heterogeneous data can be found in many observational studies, where data originate from different sources, or where variables of different types are collected.

This work comprises four contributions with different approaches to this topic and a different focus of investigation.

Contribution 1 can be considered as a practical example to illustrate data pre-processing and method uncertainty in the context of prediction and variable selection from high dimensional and heterogeneous data. In the first part of this paper, we introduce the development of priority-Lasso, a hierarchical method for prediction using multi-omics data. Priority-Lasso is based on standard Lasso and assumes a pre-specified priority order of blocks of data. The idea is to successively fit Lasso models on these blocks of data and to take the linear predictor from every fit as an offset in the fit of the block with next lowest priority. In the second part, we apply this method in a current study of acute myeloid leukemia (AML) and compare its performance to standard Lasso. We illustrate data pre-processing and method uncertainty, caused by different choices of variable definitions and specifications of settings in the application of the method. These choices result in different effect estimates and thus different prediction performances and selected variables.

In the second contribution, we compare method uncertainty with sampling uncertainty in the context of variable selection and ranking of omics biomarkers. For this purpose, we develop a user-friendly and versatile framework. We apply this framework on data from AML patients with high dimensional and heterogeneous characteristics and explore three different scenarios: First, variable selection in multivariable regression based on multi-omics data, second, variable ranking based on variable importance measures from random forests, and, third, identification of genes based on differential gene expression analysis.

In contributions 3 and 4, we apply the vibration of effects framework, which was initially used to analyze model uncertainty in a large epidemiological study (NHANES), to assess and compare different types of uncertainty. The two contributions intensively address the methodological extension of this framework to different types of uncertainty.

In contribution 3, we describe the extension of the vibration of effects framework to sampling and data pre-processing uncertainty. As a practical illustration, we take a large data set from psychological research with heterogeneous variable structure (SAPA-project), and examine sampling, model and data pre-processing uncertainty in the context of logistic regression for varying sample sizes. Beyond the comparison of single types of uncertainty, we introduce a strategy which allows quantifying cumulative model and data pre-processing uncertainty and analyzing their relative contributions to the total uncertainty with a variance decomposition.

Finally, we extend the vibration of effects framework to measurement uncertainty in contribution 4. In a practical example, we conduct a comparison study between sampling, model and measurement uncertainty on the NHANES data set in the context of survival analysis. We focus on different scenarios of measurement uncertainty which differ in the choice of variables considered to be measured with error. Moreover, we analyze the behavior of different types of uncertainty with increasing sample sizes in a large simulation study.

# Acknowledgments

First of all, I would like to express my deepest gratitude to Anne-Laure: Thank you for giving me the opportunity to write this dissertation and your excellent supervision, for always supporting me and my research with a positive attitude and with great enthusiasm, for your encouragement and for always being available for questions and help.

Moreover, I am utmost grateful to Sabine: Thank you for your help and encouragement, for collaborating with me in so many projects, for inspiring discussions and for sharing your knowledge with me, for proofreading this thesis, and also for your patience and flexibility whenever I spontaneously had to get some food from the cafeteria.

I would also like to express my special thanks to Alethea, who provided valuable language corrections to this thesis and parts of our research projects. Especially your help in converting our manuscripts saved me from desperation.

Furthermore, I would like to thank:

- Prof. Dr. Manuela Zucknick and PD Dr. Alexander Hapfelmeier for reviewing this thesis, and Prof. Dr. Thomas Augustin and Prof. Dr. Christian Heumann for being part of the examination committee.
- All the collaborators of my research projects, including Roman, Vindi, Tobias, Marie-Laure, Felix, Chirag and John.
- My research group and all other colleagues from the IBE for the great time and comfortable atmosphere.
- Bianca and my colleagues from BIPS, who gave me the opportunity to finish my thesis stress-free in a pleasant atmosphere.
- My statistical friends from the Hochschule Magdeburg-Stendal for a great three and a half years (no one will ever forget our historical triumph at the universities soccer league!) and from Ludwig-Maximilians Universität München for a lot of fun and inspiration during my master and beyond.

Finally, and most importantly, I would like to thank my family and my non-statistical friends for backing and support.





# Contents

<b>1</b>	<b>Introduction and motivation</b>	<b>1</b>
<b>2</b>	<b>Topics considered in this work</b>	<b>5</b>
2.1	Types of uncertainty . . . . .	5
2.1.1	Measurement uncertainty . . . . .	5
2.1.2	Sampling uncertainty . . . . .	6
2.1.3	Model uncertainty . . . . .	7
2.1.4	Data pre-processing uncertainty . . . . .	8
2.1.5	Method uncertainty . . . . .	9
2.2	Types of data . . . . .	10
2.2.1	Data from observational studies . . . . .	10
2.2.2	High dimensional data . . . . .	11
2.2.3	Heterogeneous data . . . . .	11
2.2.4	High dimensional and heterogeneous data in our work . . . . .	12
2.3	Types of regression . . . . .	15
2.3.1	Linear regression . . . . .	15
2.3.2	Logistic regression . . . . .	16
2.3.3	Negative binomial regression . . . . .	17
2.3.4	Cox regression . . . . .	17
2.3.5	Lasso regression . . . . .	18
2.3.6	Priority-Lasso . . . . .	19
2.3.7	IPF-Lasso . . . . .	20
2.4	Further issues related to this work . . . . .	23
2.4.1	Some causes and solutions for the replication crisis . . . . .	23
2.4.2	Different ways to deal with uncertainty . . . . .	25
2.4.3	Methods to report the results of different analysis strategies . . . . .	26
2.4.4	Quantifying uncertainties through precise and imprecise probabilities . . . . .	28
2.4.5	A glimpse at Bayesian statistics . . . . .	29
<b>3</b>	<b>Structure and overview of this thesis</b>	<b>31</b>
3.1	Summary of the contributions . . . . .	31
3.2	Further steps . . . . .	33
	<b>Bibliography</b>	<b>35</b>

<b>A</b>	<b>Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data</b>	<b>47</b>
<b>B</b>	<b>Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection</b>	<b>63</b>
<b>C</b>	<b>Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology</b>	<b>83</b>
<b>D</b>	<b>Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework</b>	<b>109</b>

# Chapter 1

## Introduction and motivation

In 2015, Raphael Silberzahn and Eric L. Uhlmann reported an experiment in which they asked 29 teams of researchers with strong statistical backgrounds to answer the same research question of interest with the same data set: ‘Are football (soccer) referees more likely to give red cards to players with dark skin than to players with light skin?’ (Silberzahn and Uhlman, 2015). The results were alarming, but not because they revealed a scandal of racism: In fact, while some teams of researchers found that players with dark skin are significantly more likely to receive red cards, others found that they were less likely to receive them. The reason for this variety of results was the multitude of possible analysis strategies, among which the researchers were free to choose. These researcher degrees of freedom (Simmons et al., 2011) play an important role in the replication crisis in science, which this simple experiment illustrates.

Indeed, the scientific community was confronted with the non-replicability of many studies in the last decade. Although problems of non-replicability were realized in individual cases and appeals for replication were formulated occasionally (Ahlgren, 1969; Smith, 1970; Gliner et al., 2002), it remained a topic of limited influence for a long time. The controversial article of Ioannidis (2005), in which the author argued that most published research findings are false, can be seen as one of the initial sparks provoking attention to the replication crisis. Moreover, the publication of Daryl Bem’s article ‘feeling the future’ (Bem, 2011) led to a wide discussion among scientists beyond the field of parapsychology. Thenceforth, the replication crisis received increasing attention, especially in social science with a focus on psychological research (Open Science Collaboration, 2015; Gelman, 2015), as well as in different fields of biomedical research including epidemiology (Lash, 2017), genetics (Ioannidis et al., 2001), and neuroscience (Gilmore et al., 2017). However, the replication crisis is not limited to these disciplines and present in a variety of scientific fields including chemistry (Gibb, 2014), climatology (Benestad et al., 2016), and economics (Herndon et al., 2013).

The replication crisis can be partly traced back to different types of uncertainty in the process of a statistical analysis, or more generally, in a scientific study. First of all, it is usually impossible to collect data without error, regardless of the type and amount, which confronts researchers with measurement uncertainty. Furthermore, there is sampling uncertainty, since a data set is typically assumed to be obtained from a larger underlying population. This type of uncertainty is also denoted as statistical uncertainty, as standard

concepts in statistical inference account for it. Finally, epistemic uncertainty is related to all choices concerning the analysis strategy and will in this work be denoted as method uncertainty. Here, I will further distinguish between data pre-processing and model uncertainty. While the former addresses the researcher degrees of freedom which are relevant before being able to fit a model to the data, model uncertainty comprises all choices related to the model fit.

These types of uncertainty are expected to affect scientific results differently, depending on a number of issues including the scientific discipline, the research question of interest, the type of study, and characteristics of the data. Different types of uncertainty have gained different amounts of attention in the scientific community so far, which is not necessarily associated with their importance in practice.

In the statistical analysis, regression is an approach used to assess the association between one or more predictors and an outcome of interest. Up to now, there are innumerable types of regression which differ, for example, in the type of the outcome, the specific functional form, or distributional assumptions. Originally, regression methods were developed for data with more observations than variables, i.e., low dimensional data. In contrast, when more variables than observations are available, i.e., the data set is high dimensional, regression methods require more elaborate approaches. In this thesis, regression analysis is chosen as a specific statistical context, which is in itself, however, widely addressed: For low dimensional data, regression analysis for continuous and binary outcomes, count data, and survival outcomes is covered. Moreover, different approaches of regression for high dimensional data are applied in this thesis. In this context, my research group and I also develop a new approach for regression analysis when several types of omics data are available for the same patient (multi-omics data). In general, regression models can be used for prediction or explanation, but we do not focus on one of these points, nor do we restrict ourselves strictly to regression analysis in our work: In some situations, machine learning algorithms will be considered as alternatives to regression analysis (e.g., in the context of variable selection), in other situations, a regression analysis will be the basis for an analysis of variance. Nevertheless, our analyses will always be in the context of supervised learning.

When investigating uncertainties in regression analysis, different types of data can be considered. As the amount of available data increases, it becomes even more important to understand and account for the characteristics of these data correctly (Manzoni et al., 2016). On the one hand, high-dimensional data has gained increasing attention in medical statistics in the past few years. For these data, more variables are available than observations, which makes the application of standard statistical methods impossible. Despite this challenge, it would be incredibly beneficial to be able to use these data as best as possible and to establish some guidelines and standards in their analysis. On the other hand, I define heterogeneous data as data consisting of variables from different sources. From a merely statistical point of view, the inclusion of heterogeneous variables is straightforward. However, variables from different sources often induce further flexibility in data pre-processing or modeling decisions. High dimensional and heterogeneous data can also occur in the same setting, for instance in the form of multi-omics data.

In the first part of this thesis, I will introduce the concepts and methods underlying this work in more detail, starting with different types of uncertainty in section 2.1. In sections 2.2 and 2.3, I will introduce the characteristics of data and types of regression that will be used for the practical applications in our contributing papers. Furthermore, I will address some more general aspects related to this work in section 2.4. In this brief look beyond the horizon, I will discuss some aspects about the replication crisis and how to deal with uncertainties in general. It also comprises short sections about methods of reporting the results of different analysis strategies, uncertainty quantification through precise and imprecise probabilities, and uncertainties from a Bayesian point of view. In chapter 3, I will briefly summarize the papers contributing to this thesis and explain how the challenges introduced in chapter 2 are addressed. The first part ends with some aspects worth investigating based on this work.

The second part of this thesis comprises the four different papers, which have been, or will be published in scientific journals. The first of these papers is an extension of my master thesis and introduces some important methodological basics for some of the other papers, while addressing the challenges of uncertainty only as a subsidiary. For ease of readability, I will denote the papers as contribution 1, contribution 2, contribution 3 and contribution 4 in the following, which refer to the publications Klau et al. (2018), Klau et al. (2020a), Klau et al. (2020b) and Klau et al. (2020c) respectively.



## Chapter 2

### Topics considered in this work

#### 2.1 Types of uncertainty

In the scientific literature, a variety of philosophical concepts to define uncertainty can be found. One of the most prevailing concepts is the distinction between aleatoric and epistemic uncertainty (Der Kiureghian and Ditlevsen, 2009). Epistemic uncertainty refers to uncertainty due to a lack of knowledge, and is in general reducible by gaining further information. In contrast, aleatoric uncertainty occurs in random processes like flipping a coin. In the following, I will introduce measurement uncertainty and sampling uncertainty, which can be classified as aleatoric types of uncertainty. In contrast, model uncertainty, data pre-processing uncertainty and method uncertainty are epistemic types of uncertainty. Although the broad terms of aleatoric and epistemic uncertainty will only be scarcely used in this work, they should serve as rough pre-interpretative guidance at this point.

##### 2.1.1 Measurement uncertainty

With measurement uncertainty, we denote uncertainty that is caused by measurement error in the data. This measurement error occurs before performing a statistical analysis, when collecting the data. Ideally, measurements should have the two properties of accuracy and precision: Accurate measurements are measurements without systematic deviation from the true value (that is unbiasedness), while precise measurements are close to each other (Ulijaszek and Kerr, 1999). Indeed, it is almost impossible to collect data with perfect accuracy and precision – no matter what kind of study is conducted and which type of data is collected. Depending on the research question of interest, the data can be gathered through questionnaires, laboratory measures, measurement devices and experimental protocols, which are all error-prone (Hoffmann et al., 2020). Moreover, the type and strength of measurement error depends on the scientific discipline and its sub-branches. In epidemiology, which this brief overview will focus on, the presence of measurement error is indicated in many cases, nevertheless its impact on scientific findings is often neglected (Brakenhoff et al., 2018). In contrast, few mentions of measurement error in omics data can be found in the scientific literature.

In epidemiology, many different types and characteristics of measurement error can be distinguished and different theoretical concepts exist. For instance, measurement error is assumed to occur either in the exposure or the outcome variable of interest, additive and multiplicative models can be distinguished, and errors have to be modeled differently depending on the type of the variable, e.g., continuous or binary. Here, I refer to Gustafson (2003) for more details on measurement uncertainty in epidemiology, since it is beyond the scope of this work to provide a detailed overview.

A short summary of four types of measurement error in epidemiology is provided by Brakenhoff et al. (2018), for the situation where the association of two continuous variables is studied. For a true variable  $X$ , an observed variable  $Z$ , and measurement error  $U$ , *classical* measurement error can be expressed by adding random values with mean zero and constant variance to the true variable, i.e.,  $Z = X + U$ , with  $U \stackrel{iid}{\sim} N(0, \text{Var}(U))$ . Thus, classical measurement error is independent of the true variable  $X$  and the outcome variable. In a similar situation, measurement error is called *systematic* when the error introduces a bias to the true variable, and *differential* error occurs when the error term depends on the outcome. According to Brakenhoff et al. (2018), the latter can be relevant when the outcome was observed before the covariates, e.g., in case-control studies. Finally, *Berkson* measurement error can be expressed, conversely to classical measurement error, by obtaining the true variable through adding a random component to the observed variable, i.e.,  $X = Z + U$ , where  $U$  is independent of  $Z$ .

There are different methods of accounting for measurement error, for instance regression calibration and simulation extrapolation (see the early references (Rosner et al., 1989) and (Cook and Stefanski, 1994), respectively), as well as approaches from a Bayesian perspective (Richardson and Gilks, 1993). However, these and other approaches are rarely used in practice (Brakenhoff et al., 2018).

In contribution 4 of this thesis, we suggest a framework that provides easy quantification and visualization of measurement uncertainty, and practically apply it on data from health and nutritional epidemiology. For this data set, we assume classical non-differential measurement error and base our practical assessment on general magnitudes of measurement error we found in the literature. In this contribution, we study the effect of a variable of interest on a survival outcome, while taking several adjustment variables into account. To assess measurement error, we focus on three scenarios: In a first scenario, we add measurement error only to the variables of interest, and in a second scenario only to the adjustment variables. In a third scenario, we combine scenario 1 and 2, and assume both the variable of interest and the adjustment variables to be measured with error. Finally, we compare measurement uncertainty to sampling and model uncertainty (see sections 2.1.2 and 2.1.3 for more details on sampling and model uncertainty, respectively) and study the effect of the sample size on all these types of uncertainty on simulated data.

## 2.1.2 Sampling uncertainty

It is usually assumed that a data set which is collected for a scientific study is only a sample from a larger underlying population. Uncertainty that occurs in this context



is referred to as sampling uncertainty in the following, and is clearly the most well-investigated type of uncertainty in statistical modeling. Basic concepts like standard errors address sampling uncertainty, and standard statistical inference widely accounts for it, e.g., through confidence intervals and hypotheses tests (Altman and Bland, 2014). It is well-known that sampling uncertainty decreases with increasing sample size, i.e., confidence intervals become narrower and testing procedures are associated with more power.

Beyond these basic concepts, sampling uncertainty can for instance be assessed through resampling procedures like subsampling or bootstrapping. Several further approaches have been developed for stability investigations in specific situations, e.g., in variable selection (Meinshausen and Bühlmann, 2010) or multivariable regression (Sauerbrei et al., 2011), which can both be applied in potentially high dimensional settings. A specific example of a procedure addressing sampling uncertainty that is used in many statistical applications is cross-validation, where random splits of the data are used for training and validation purposes. Depending on these random subsets, different results can be obtained.

Although sampling uncertainty is well-investigated and well-understood, problems of studies with low power remain present and are often unavoidable, e.g., when a rare disease is analyzed or in high dimensional settings. Furthermore, a tradeoff between small and large samples is often hard to achieve as there are ethical issues that argue for a low sample size. Finally, the relationship between sampling uncertainty and other types of uncertainty is not clear.

In this work, we will directly address sampling uncertainty in contributions 2, 3, and 4. In contribution 2, we will assess sampling uncertainty in the context of different statistical approaches to perform variable selection and ranking in high dimensional settings, and compare it to method uncertainty. For this purpose, we suggest a framework that allows direct comparison of method and sampling uncertainty by splitting the data into two equal-sized parts and applying at least two methods on each of the halves a large number of times. Furthermore, we compare sampling uncertainty to data pre-processing and model uncertainty in contribution 3, and to measurement and model uncertainty in contribution 4. Here, we assess sampling uncertainty by randomly subsampling the data many times. In addition, we are interested in the influence of the sample size on these types of uncertainty in these two contributions.

### **2.1.3 Model uncertainty**

When specifying a probability model, many choices can be made, comprising, for instance, the inclusion and exclusion of covariates, their functional form, or interaction terms. We define the uncertainty that arises through all these choices in the specification of a probability model as model uncertainty. Early work on this topic was provided by Leamer (1983) in the field of econometrics. The so called ‘extreme bounds analysis’ was motivated from Bayesian statistics and aims to determine the most extreme coefficient estimates in a probability model by taking different model choices into account. Extreme bounds analysis was further developed by Breusch (1990) and Granger and Uhlig (1990), who suggest reasonable specifications of the model space in ordinary least squares analysis.

Recent important works focusing on the practical assessment of model uncertainty comprise Patel et al. (2015), Muñoz and Young (2018) and Young (2018). While Patel et al. (2015) investigate model uncertainty in health and nutritional epidemiology through the inclusion and exclusion of covariates, Muñoz and Young (2018) and Young (2018) provide more theoretical contributions which address the question of how to define a model space in practice. Since the latter two papers assume that these definitions of a model space go beyond the choices concerning a probability model, their work could also be mentioned in the context of method uncertainty in section 2.1.5.

Another common way to address model uncertainty is model selection, e.g., by using information criteria like the Akaike information criterion (Akaike, 1974) or the Bayesian information criterion (Schwarz, 1978). However, as there are many different model selection criteria, the choice of a specific one introduces further uncertainty. Moreover, Bayesian model averaging (Hoeting et al., 1999) approaches model uncertainty from a Bayesian perspective, which I will discuss in more detail in section 2.4.5.

In this work, only inclusion and exclusion of variables in a regression model are practically investigated as part of the concept of model uncertainty. In contribution 3, we compare model uncertainty to sampling uncertainty and data pre-processing uncertainty in the context of logistic regressions for varying sample sizes of data from the SAPA project personality test (Condon et al., 2017). In contribution 4, a comparison of model uncertainty to sampling and measurement uncertainty is conducted in the context of survival analysis on data from the National Health and Nutrition Examination Survey (NHANES) and on simulated data of varying sample sizes.

#### **2.1.4 Data pre-processing uncertainty**

Another type of uncertainty occurs due to choices of pre-processing steps concerning the data used for a statistical analysis. For instance, for a sufficiently large data set comprising the demographic characteristics of a cohort, subgroups of observations based on variables like age or sex can be excluded, depending on the research question of interest. Furthermore, variable definitions are often not clear, which concerns both dependent and independent variables in prediction modeling. Other examples of data pre-processing comprise the handling of outliers, data imputation, cleaning, normalization and transformation, but the list could be extended far further.

The choices of pre-processing critically depend on the type of data and the specific statistical application. Wicherts et al. (2016) provide a list of researcher degrees of freedom when conducting a psychological study, which includes some aspects of data pre-processing. The work of Steegen et al. (2016) is a further practical example of data pre-processing uncertainty in psychology. Aside from this, this type of uncertainty has not gained much direct attention in the statistical literature up to now.

However, data pre-processing uncertainty plays an important role in statistics, especially for studies with poor research design or without accurate analysis plans. For these types of studies, the flexibility in pre-processing choices is even greater than in data from other types of studies. In a randomized controlled trial, for instance, the outcome variable is usually clearly defined in advance. In contrast, data sets from retrospective observational studies often consist of several variables that could be suitable for the

specific research purpose, e.g., ‘relationship status’ or ‘marital status’ could both be used to study the characteristics of partnerships in psychological research (an example we are actually confronted with when analyzing data from the SAPA project in contribution 3). Furthermore, this type of uncertainty is important when the data set comprises heterogeneous variables, as overlapping or similar information are expected more often in these situations. For instance, the body mass index could be gathered from a questionnaire, as well as from measurements of height and weight conducted by a second person, and it is not clear which of these variables to choose when they are available in the same data set. Since the choices of pre-processing critically depend on the specific data and the research question of interest, a detailed explanation of these choices can only be done on a case-by-case basis.

In this work, data pre-processing uncertainty is mainly considered in contribution 3, where it is illustrated and quantified on a large data set from personality psychology, the SAPA data set. We extensively explain and discuss the specific choices of pre-processing in this contribution. Moreover, we provide tools to compare data pre-processing uncertainty to sampling and model uncertainty, and to quantify and compare the relative impact of model and data pre-processing uncertainty with sampling uncertainty. This combined uncertainty, which is also referred to as uncertainty due to the analysis strategy in contribution 3, is further explained in section 2.1.5 under the term ‘method uncertainty’.

Another application of data pre-processing uncertainty is provided in contribution 1, where priority-Lasso is applied to data from acute myeloid leukemia (AML) patients. Here, results from two choices of data pre-processing approaches are reported. For the first choice, we consider the European Leukemia Net (ELN) risk score as a first block of data for prediction. For the second choice, we replace this score with all variables that were used for its definition, without further pre-processing of these variables. However, the main focus in this contribution is the method priority-Lasso itself and its prediction performance, and we do not illustrate or quantify data pre-processing uncertainty directly.

### **2.1.5 Method uncertainty**

As a general type of epistemic uncertainty, we define method uncertainty as uncertainty due to the analysis strategy which comprises all operationalization decisions (Simonsohn et al., 2015). Therefore, model choices (section 2.1.3) and data pre-processing choices (section 2.1.4) are considered to be part of the concept of method uncertainty. Furthermore, a researcher may be faced with a large number of additional method choices: For instance, in the simple situation where a continuous variable was measured in two groups that should be compared with a statistical test, a common decision is whether to apply a parametric or nonparametric test. If several statistical tests are performed, the method to adjust for multiple testing is another choice referring to method uncertainty. In prediction modeling using high-dimensional data, a method choice could be whether to run a machine learning algorithm or a penalized regression method. In addition, less crucial choices fall under our definition of method uncertainty, like the number of folds when performing cross-validation for hyperparameter tuning.

Similar to model and data pre-processing uncertainty, various other choices could be listed that refer to method uncertainty, depending on the specific research question of interest. At the same time, it is important to note that clear distinction between these three types of uncertainty is difficult. A simple example of this is the choice of the variable sex in a probability model: Excluding either female or male participants from the study is a data pre-processing choice which also affects the selection of the covariate sex, as it necessarily has to be excluded from the model. A strict definition of these types of uncertainty, however, is outside the scope of this work.

Some practical examples focusing on method uncertainty in the scientific literature are for instance provided by Silberzahn and Uhlman (2015), Simonsohn et al. (2015), Palpacuer et al. (2019), and Chu et al. (2020). In this thesis, we address method uncertainty mainly in contribution 2, where it is compared to sampling uncertainty for three different scenarios in the context of omics biomarker selection and ranking. In the first scenario, method uncertainty is assessed between three different approaches of penalized regression. In the second scenario, we compare the results obtained from running the machine learning algorithm random forest with different parameter settings. In the last scenario, we perform variable selection in the context of differential expression analysis with five different analysis strategies.

Furthermore, contribution 1 provides a practical illustration of method uncertainty, where priority-Lasso and standard Lasso are applied to AML data. Several additional choices for both of the methods could be made and some of these are conducted and reported in our example. For instance, we analyze the performance of priority-Lasso with and without cross-validated offsets. Finally, we compare method and sampling uncertainty in contribution 4 on the SAPA data with the vibration of effects framework. The investigation of method uncertainty is conducted as a joint investigation of model and data pre-processing uncertainty for varying sample sizes.

## 2.2 Types of data

### 2.2.1 Data from observational studies

In biomedical research, there is often a distinction between two major study types, the observational and the experimental study. In an observational study, the researcher does not intervene in the study design, and does not influence the data generating process. In contrast, in an experimental (or interventional) study, the researcher intervenes at some point throughout the study (Thiese, 2014). For both types of studies, there are several design types. Observational studies comprise, for instance, cohort studies, case-control studies and cross-sectional studies (for more information on these types of studies, I refer to McNeil (1996)). These study designs mainly differ in the method of selecting participants and the extend of follow-up over time.

Retrospective studies – a common class of observational studies – are prone to bias (Thiese, 2014), and therefore it could follow that they suffer from low quality data and high measurement uncertainty. For an experimental study, there is often an analysis plan available before the data is collected, and the protocols are inflexible compared

to those of observational studies (Ioannidis, 2008). Thus, we can assume less method uncertainty in experimental studies. As both observational and experimental studies have their specific challenges, we could conclude that the challenges of uncertainty seem to be more prominent for observational studies, on which our work focuses.

### 2.2.2 High dimensional data

In the situation where data is available with many more variables  $p$  than observations  $n$ , i.e.,  $n \ll p$ , the data set is called high dimensional. Similar to big data, high dimensional data underwent a striking boom in recent years, as shown in Figure 2.1. Although high dimensional data is sometimes referred to as big data (Boulesteix et al., 2017b), big data is in most situations characterized only by a large number of observations and hence often encompasses many more observations than variables ( $n \gg p$ ). The computational effort of analyzing big data or high dimensional data is higher than in a conventional situation, but the approach to analyze the two types of data differs remarkably.

The property of high dimensionality is typical for molecular data in medical statistics, where several hundreds or thousands of variables are generated through high throughput experiments. This type of data is often denoted to as ‘omics’ data, as it comprises for instance genomics, proteomics, transcriptomics, or metabolomics, which refer to the investigation of the genome, proteome, transcriptome, or metabolome, respectively. Research questions and goals related to high dimensional data analysis comprise clustering (Steinbach et al., 2004), interaction analysis of variables (Li et al., 2003), differential expression analysis (Anders et al., 2013), variable selection and ranking procedures (Wasserman and Roeder, 2009; Boulesteix and Slawski, 2009), or prediction (Golub et al., 1999), among others. When it comes to the analysis of high dimensional data, a general first decision is whether to reduce the dimension beforehand (e.g., through principal component analysis (Pearson, 1901) or related methods) or to apply methods that account for the high dimensionality directly. In the context of regression analysis, methods are available that address the high dimensionality through shrinkage or internal variable selection procedures.

### 2.2.3 Heterogeneous data

With regard to heterogeneous data, this work refers to data which consists of variables from different sources. Hence, we attribute the property of heterogeneity to data sets comprising several types of omics data for the same patient, e.g., gene expression data, miRNA data, and copy number variations. In this vein, the methods priority-Lasso and IPF-Lasso (section 2.3.6 and 2.3.7) do not only account for high dimensionality, but also for heterogeneity of omics data, as they incorporate a block structure in their algorithm, which is usually defined by variables from different types of data.

Researchers are faced with heterogeneous data in many other situations. In these situations, the data set sometimes comprises variables that measure the same feature. For instance, in an observational study, a patient’s hypertension status could be collected through answering a question in a questionnaire. On the other hand, the same data set might also contain information about the patient’s blood pressure which was measured by

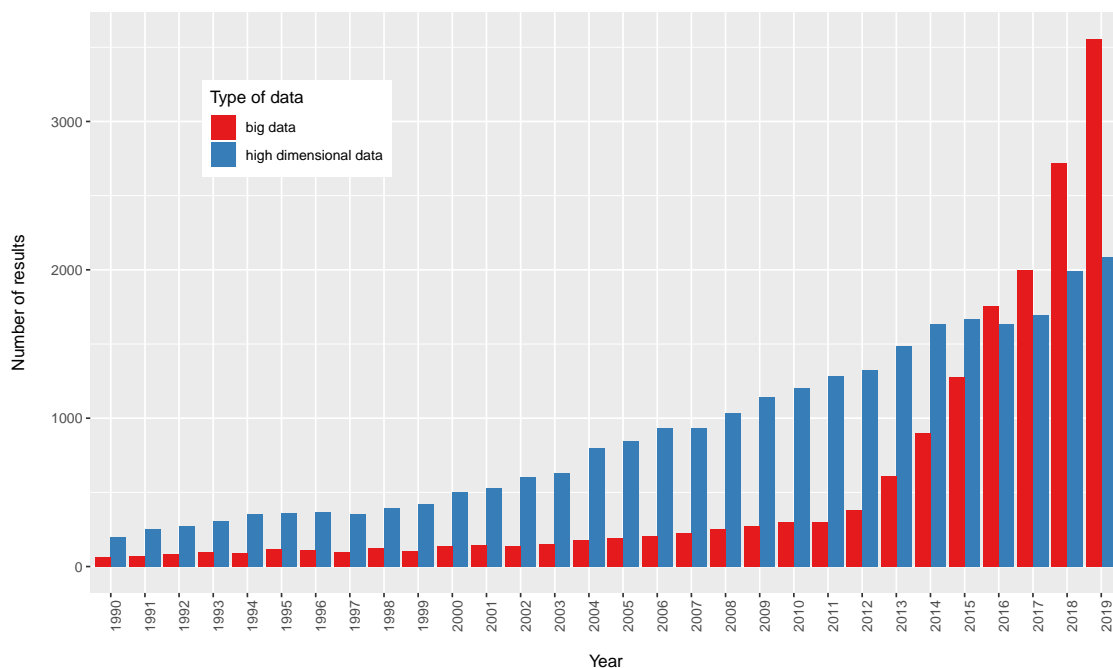


Figure 2.1: Number of results on PubMed when searching for the terms ‘big data’ and ‘high dimensional data’

a clinician. Hence, heterogeneity can introduce further uncertainty, as it allows (or even requires) more data pre-processing and model choices. In contrast to high dimensionality, it is technically not necessary to account for heterogeneity in regression analysis. Usually, variables are made technically comparable, e.g., through standardization of continuous variables, but beyond that, the source of the variables is not directly addressed.

## 2.2.4 High dimensional and heterogeneous data in our work

In contribution 1, we focus on high dimensional and heterogeneous data from AML patients. This data set comprises clinical, gene mutation and gene expression variables, the latter consisting of many more variables than observations (15809 variables versus 447 and 250 observations for the training and test data set, respectively). More specifically, the clinical variables are themselves heterogeneous as they encompass demographic variables like age and sex, but also laboratory measures like white blood cell count, or a performance status assigned after evaluating different health criteria. This heterogeneity, however, will remain disregarded in our application. In this first contribution, we develop the method priority-Lasso which can take high dimensional and heterogeneous data into account. Priority-Lasso addresses the high dimensionality through the Lasso method (section 2.3.5) and the heterogeneity through a hierarchical procedure, where the blocks of data are analyzed successively according to their priority.

In contribution 2, we compare method and sampling uncertainty in the context of omics biomarker selection, where we illustrate our framework in three scenarios. All of these scenarios are exemplified on high dimensional data of AML patients. Two

of these data sets are similar to the gene expression data from contribution 1, with slight differences due to pre-processing. For the second scenario, where we compare the uncertainty between variable rankings obtained by running the random forest method with different tuning parameters and on different subsets of the data, this AML data is gathered by Affymetrix Arrays and consists of 488 patients and 17389 variables. For the third scenario, RNA-Seq data of AML patients is used in the context of differential expression analysis. Furthermore, we consider a data set from the Cancer Genome Atlas (TCGA) for the first scenario, which consists of high dimensional and heterogeneous data. For this data set, clinical data as well as two types of omics data are available (gene expression data and copy number variations), and we analyze method uncertainty between different Lasso-based methods which we apply for the purpose of variable selection.

For the other two contributions, the focus of investigation lies on heterogeneous, but not high dimensional data. In contribution 3, we use data from the SAPA project. Although these data were gathered by the same (or similar) questionnaires, it consists of different types of variables: On the one hand, standard demographic variables are available. On the other hand, variables were gathered that assess the participants personality. In order to obtain valid and interpretable personality scores, these latter variables are analyzed through a factor analysis. Thus, they are accounted for statistically in a different way to the demographic variables.

Another example of heterogeneity is provided by the NHANES data, which explicitly claim to comprise data of different types (Zipf et al., 2013). The sources of data encompass for instance interviews, health examinations, and laboratory tests, but we do not account for these different sources in our application. An overview of all data sets and their basic properties can be found in Table 2.1.

Table 2.1: Overview of data sets from observational studies in this work

Contribution	Number of observations	Number of variables	Characteristics	Cohort	Reference
1	447	15876	high dimensional and heterogeneous (clinical, gene mutation, gene expression (Affymetrix arrays))	acute myeloid leukemia	Herold et al. (2014) Büchner et al. (2016) Büchner et al. (2006)
1	250	15876	high dimensional and heterogeneous (clinical, gene mutation, gene expression (RNA-Seq))	acute myeloid leukemia	Braess et al. (2013) Herold et al. (2017)
2	176	37573	high dimensional and heterogeneous	acute myeloid leukemia	TCGA
2	488	17389	high dimensional (gene expression (Affymetrix arrays))	acute myeloid leukemia	Herold et al. (2014) Büchner et al. (2016) Büchner et al. (2006)
2	241	23369	high dimensional (gene expression (RNA-Seq))	acute myeloid leukemia	Braess et al. (2013) Herold et al. (2017)
3	126884	722	heterogeneous and big	SAPA personality test	Condon et al. (2017)
4	17039	711	heterogeneous	National Health and Nutrition Examination Survey	Patel et al. (2015) Fillenbaum et al. (2009)



## 2.3 Types of regression

In this overview of regression methods, which are used in the four contributions, I will briefly describe the theory, as well as the role these methods play in this work. For the sake of completeness, I will start with linear regression and logistic regression, and will proceed with the negative binomial regression and Cox regression. Finally, I introduce the basic concepts of the three methods for high dimensional data analysis, standard Lasso, priority-Lasso, and IPF-Lasso. More details and further information on all these regression methods can be found in the corresponding references and an overview of their practical use in this thesis is available in Table 2.2.

For consistent notation in this section,  $x_{ij}$  refers to the value of covariate  $j$  for observation  $i$ , where  $j = 1, \dots, p$  and  $i = 1, \dots, n$ . Thus,  $p$  covariates as well as  $n$  observations are available. The case where  $p > 1$  describes multivariable regressions, on which we will focus in this work. For all types of regression, a covariate can take continuous or binary values. Moreover, categorical covariates can be included as several binary covariates by dummy coding. The outcome variable is denoted by  $y_i$ , which can refer to count data, survival times, continuous variables or binary variables. The question of which regression method to apply strongly depends on this property of the outcome. Throughout this section, bold symbols will be used to denote vectors, e.g.,  $\mathbf{y}$  refers to the vector of outcome values.

### 2.3.1 Linear regression

In a linear regression analysis, a continuous outcome is modeled through a linear relationship of one or more predictor variables. For  $i = 1, \dots, n$  observations and  $j = 1, \dots, p$  predictor variables, the linear model is described by

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i,$$

with the assumption of independent and identically distributed errors  $\epsilon_i$ , which satisfy  $E(\epsilon_i) = 0$  and  $\text{Var}(\epsilon_i) = \sigma^2$ . In addition, the errors  $\epsilon_i$  are often assumed to be normally distributed for the purpose of inference. As a result,  $p + 1$  regression coefficients are obtained, encompassing the intercept  $\beta_0$  and slopes  $\beta_j$ , which describe the influence of the variable  $x_j$  on the outcome  $\mathbf{y}$ . The predictor variables can be of different type, e.g., binary, continuous or categorical, the latter being typically dummy coded as several binary variables. More details on a linear regression analysis can be found in Fahrmeir et al. (2007).

In contribution 2, limma-voom (Smyth, 2004; Law et al., 2014; Ritchie et al., 2015), which is applied in scenario 3 as one of the methods for differential expression analysis, is based on linear regression. After transforming the data to obtain normalized log-counts, each variable is modeled through a linear regression. When testing for contrasts, the method uses an empirical Bayes approach borrowing information from other variables, which yields a stabilized variance estimation.

In order to conduct an analysis of variance (ANOVA), a special type of a linear regression can be used, where the continuous outcome is modeled through one or more

categorical variables (factors). In the case of two factors, without considering interactions, the model can be expressed as  $y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$ , with the assumptions that  $\sum_{i=1}^I \alpha_i = 0$ ,  $\sum_{j=1}^J \beta_j = 0$ , and  $\epsilon_{ijk} \stackrel{iid}{\sim} N(0, \sigma^2)$ . This model will be considered in contribution 3, in the analysis of the relative impact of model and data pre-processing vibration. In this application, estimates describing the influence of a covariate on a binary outcome in a logistic regression for different model- and data pre-processing choices are used as the outcome  $y_{ijk}$ , and the two factors indicate the corresponding model and data pre-processing choices.

### 2.3.2 Logistic regression

Similar to a linear regression, the association between an outcome and one or more predictor variables is modeled through a linear relationship in a generalized linear model (GLM). However, the outcome variable considered in a linear regression is usually assumed to be continuous. In contrast, for a generalized linear model, the outcome variable is assumed to follow a distribution from the exponential family (Fahrmeir et al., 2007). Here, I describe the generalized linear model for a binary outcome variable, which has to be modeled with a restriction to an interval between 0 and 1. In this case, the outcome is assumed to be conditionally Bernoulli distributed, i.e.,  $y_i | \mathbf{x}_i \sim \text{Bernoulli}(\pi_i)$ .

To model the probability that the outcome takes the value 1,  $\pi_i = P(y_i = 1)$ , the logit link function can be used:

$$g(\pi_i) = \eta_i = \log \left( \frac{\pi_i}{1 - \pi_i} \right).$$

Thus,  $\pi_i$  can be obtained by

$$\pi_i = P(y_i = 1) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)},$$

with the linear predictor

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}.$$

Accordingly, the odds  $\pi_i / (1 - \pi_i)$  can be formalized through the multiplicative model

$$\frac{\pi_i}{1 - \pi_i} = \frac{P(y_i = 1)}{P(y_i = 0)} = \exp(\beta_0) \cdot \exp(\beta_1 x_{i1}) \cdot \dots \cdot \exp(\beta_p x_{ip}),$$

which can also be logarithmized in order to obtain an additive model and simplify interpretation. The generalized linear model with a logit link is also denoted as ‘logistic regression’, and a detailed explanation is provided by Kleinbaum and Klein (2002), for instance. For outcomes following other distributions from the exponential family, generalized linear models can be applied using other link functions.

In this work, logistic regressions are used in contribution 3, where we assess associations between several binary outcomes of interest and the Big Five personality traits

(John et al., 1999) on the SAPA data in order to explore different types of uncertainty. The binary outcomes comprise the smoking status (smoker vs. non-smoker), levels of education and physical activity (high vs. low), the relationship status (committed vs. non-committed) and obesity (yes or no). In addition to the personality traits, several other adjustment variables are included in the model.

### 2.3.3 Negative binomial regression

Another generalized linear model is the poisson model, which assumes the outcome to be poisson-distributed and is in practice often used to model count data. However, in many situations the restriction that the mean and the variance are equal does not hold, which makes generalizations of the poisson model more appropriate. One of these generalizations is the negative binomial regression, which can be derived and presented with different parameterizations (Hilbe, 2011). The negative binomial distribution can also be expressed in terms of the exponential family. Following the work of Tutz (2011), the model can be derived as a mixture of poisson distributions and expressed by

$$y_i | \mathbf{x}_i \sim \text{NB}(\nu, \mu_i),$$

with the assumptions that  $y_i | \lambda_i \sim \text{Poisson}(\lambda_i)$ ,  $b_i \sim \Gamma(\nu, \nu)$  and  $\lambda_i = b_i \mu_i$ . The density function of the distribution is given by

$$f(y_i | \nu, \mu_i) = \frac{\Gamma(y_i + \nu)}{\Gamma(\nu) + \Gamma(y_i + 1)} \left( \frac{\mu_i}{\mu_i + \nu} \right)^{y_i} \left( \frac{\nu}{\mu_i + \nu} \right)^\nu,$$

with mean  $E(y_i) = \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$  and variance  $\text{Var}(y_i) = \mu_i + \frac{\mu_i^2}{\nu}$ .

In contrast to the poisson model, the negative binomial model is a two-parameter model and can cope with overdispersion. While  $\nu \rightarrow 0$  indicates strong overdispersion, it approximates the poisson distribution for  $\nu \rightarrow \infty$ . Thus, the negative binomial model cannot handle underdispersion.

The negative binomial model is used in some of the methods for differential expression analysis applied in scenario 3 of contribution 2. In this scenario, we consider five different analysis strategies ('methods') to investigate sampling and method uncertainty on RNA-Seq counts of AML patients. Four of these methods are based on negative binomial regression, namely DESeq (Anders and Huber, 2010), DESeq2 (Love et al., 2014), edgeR and glm.edgeR (McCarthy et al., 2012; Robinson et al., 2010). The choice of these methods was motivated by Rigai et al. (2016). For more details on how the negative binomial model is integrated in these methods, I refer to the original work. In the fifth case, limma-voom, the raw data is log-transformed and the analysis can be based on a linear regression (section 2.3.1).

### 2.3.4 Cox regression

When the aim is to model survival data, the Cox proportional hazards model (Cox, 1972) can be used as a regression method. This type of data typically consists of survival times

$T_i$  and censoring times  $C_i$ , where only the minimum of these two times is observed, i.e.,  $T_i^* = \min(T_i, C_i)$ . Furthermore,  $\delta_i = I\{T_i \leq C_i\}$  indicates the status of subject  $i$ . The Cox model is one of the most frequently used models for survival analysis and estimates the influence of covariates  $\mathbf{x}_i$  on the hazard rate  $\lambda(t)$ . With  $f(t)$  and  $F(t)$  as the density function and cumulative distribution function, respectively, this hazard rate can be expressed through  $\lambda(t) = f(t)/S(t)$ , where  $S(t) = 1 - F(t)$  denotes the survival function. In the Cox regression, the hazard rate is modeled as a product of  $\exp(\mathbf{x}_i^T \boldsymbol{\beta})$  and a time-dependent baseline hazard function  $\lambda_0(t)$ :

$$\lambda(t, \mathbf{x}_i) = \lambda_0(t) \exp(\mathbf{x}_i^T \boldsymbol{\beta}).$$

The baseline hazard function corresponds to the hazard in the case where all variables are zero and is estimated non-parametrically – hence, the model is often referred to be semi-parametric. Since the hazard ratio for two subjects 1 and 2,  $\frac{\lambda(t, \mathbf{x}_1)}{\lambda(t, \mathbf{x}_2)} = \exp((\mathbf{x}_1 - \mathbf{x}_2)^T \boldsymbol{\beta})$ , does not depend on time  $t$ , the Cox model satisfies the proportional hazards assumption. In order to estimate survival functions non-parametrically, the Kaplan-Meier estimator can be used, for instance. Moreover, alternatives to the Cox model include parametric models, for which a survival distribution has to be specified. This choice is, however, not straightforward (Bradburn et al., 2003). In comparison to the Cox model, parametric alternatives are only rarely used in practice (Nardi and Schemper, 2003).

The Cox regression will be applied in contribution 4, where we consider the right-censored survival times of participants of the NHANES as an outcome. We use the package `survival` (Therneau, 2015) in order to fit the Cox model in R. Furthermore, the penalized regression methods used in contribution 1 and scenario 1 of contribution 2 are based on a Cox regression, as survival times of AML patients are considered as an outcome in the situation where more variables than observations are available. More details on the regression methods for this high dimensional data setting can be found in the following subsections.

### 2.3.5 Lasso regression

In extension to the types of regression introduced in sections 2.3.1 – 2.3.4, which assume that the data consists of more observations than variables, the least absolute shrinkage and selection operator (Tibshirani, 1996), also denoted as ‘Lasso’, can deal with high dimensional covariate data. In the following, standard regression methods based on this technique are referred to as ‘standard Lasso’, to contrast them clearly from other Lasso-based regression methods, in particular priority-Lasso and IPF-Lasso. Standard Lasso can be used with different types of outcomes, which include for instance continuous and binary variables, as well as survival times. Here, the principles of standard Lasso are briefly explained in terms of a continuous and centered outcome  $\mathbf{y}$ . To obtain the regression coefficients  $\boldsymbol{\beta}$ , the standard Lasso is defined by

$$\sum_{i=1}^n \left( y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \rightarrow \min_{\boldsymbol{\beta}},$$

where  $\lambda > 0$  denotes a penalty parameter, which determines the amount of penalization and is usually chosen through cross-validation. The penalization of the coefficients with

the  $L1$ -norm ensures a shrinkage towards zero, where some coefficients are estimated to be exactly zero. Therefore, standard Lasso provides intrinsic variable selection, in contrast to similar regression methods like ridge regression (Hoerl and Kennard, 1970), which utilizes the  $L2$ -norm.

In the context of survival analysis, Tibshirani (1997) proposed the Cox-Lasso, where the coefficients can be estimated through

$$-l(\boldsymbol{\beta}) + \lambda \sum_{j=1}^p |\beta_j| \rightarrow \min_{\boldsymbol{\beta}},$$

with  $l(\boldsymbol{\beta})$  as the log partial likelihood. Based on these standard Lasso regressions, a large variety of extensions have been proposed for specific purposes, including group Lasso (Yuan and Lin, 2006) and sparse group Lasso (Simon et al., 2013), elastic net regression (Zou and Hastie, 2005), fused Lasso (Tibshirani et al., 2005), and adaptive Lasso (Zou, 2006). In addition, priority-Lasso and IPF-Lasso are based on standard Lasso and can take several blocks of data, such as multi-omics data, into account. The development of priority-Lasso is part of this thesis, and the method is extensively discussed in contribution 1 (Klau et al., 2018) and summarized in section 2.3.6. IPF-Lasso was suggested by Boulesteix et al. (2017a) and its idea will briefly be introduced in section 2.3.7.

Standard Lasso is applied in the context of survival analysis in contribution 1 in comparison to priority-Lasso. Furthermore, method uncertainty in the first scenario of contribution 2 is assessed for the three Lasso-based methods standard Lasso, priority-Lasso and IPF-Lasso. For these applications, we use the implementation from the R-package `glmnet` (Friedman et al., 2010; Simon et al., 2011).

### 2.3.6 Priority-Lasso

As part of this thesis, priority-Lasso is developed as a regression approach based on standard Lasso that can take different blocks of data into account (Klau et al., 2018). Here, I provide a short summary of this method.

Priority-Lasso is motivated by the situation where prior knowledge about these different blocks of data is available, such that they can be ordered according to their priority. In medical research, applications of priority-Lasso can be useful when the data consists of several types of omics markers, but its use is not limited to such multi-omics data. In this and many other examples, the block structure of the data is given implicitly. When the blocks of data are specified, the priority order of these blocks can be defined. This can be done from a practical point of view, for instance according to cost of data collection or whether data are routinely collected in clinical practice, or alternatively, using knowledge about prediction performance, e.g., from earlier research.

The algorithm to obtain the coefficient estimates follows a hierarchical procedure, where every step is based on a standard Lasso regression. With  $m = 1, \dots, M$  blocks of data and a permutation  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)$  of these blocks, the coefficients of the block of data with highest priority  $\pi_1$  can be obtained by

$$\sum_{i=1}^n \left( y_i - \sum_{j=1}^{p_{\pi_1}} x_{ij}^{(\pi_1)} \beta_j^{(\pi_1)} \right)^2 + \lambda^{(\pi_1)} \sum_{j=1}^{p_{\pi_1}} |\beta_j^{(\pi_1)}| \rightarrow \min_{\boldsymbol{\beta}},$$

for a continuous and centered outcome  $\mathbf{y}$ . Similar to standard Lasso,  $\lambda^{(\pi_1)}$  is a tuning parameter which is usually obtained by cross-validation. In the second step, the resulting linear predictor

$$\hat{\eta}_{1,i}(\boldsymbol{\pi}) = \hat{\beta}_1^{(\pi_1)} x_{i1}^{(\pi_1)} + \dots + \hat{\beta}_{p_{\pi_1}}^{(\pi_1)} x_{ip_{\pi_1}}^{(\pi_1)}$$

can be used as an offset in the estimation of the coefficients for the block with next lowest priority. Thus, the second Lasso model, which considers all variables in the block of priority 2,  $\pi_2$ , is estimated through

$$\sum_{i=1}^n \left( y_i - \hat{\eta}_{1,i}(\boldsymbol{\pi}) - \mathbf{x}_i^{(\pi_2)T} \boldsymbol{\beta}^{(\pi_2)} \right)^2 + \lambda^{(\pi_2)} \sum_{j=1}^{p_{\pi_2}} |\beta_j^{(\pi_2)}| \rightarrow \min_{\boldsymbol{\beta}}.$$

Hence, variables from this block are only included in the model, when they provide additional predictive information to that of variables from the block of highest priority. However, since the linear predictor  $\hat{\eta}_{1,i}(\boldsymbol{\pi})$  is overoptimistic with respect to the information contained in block  $\pi_1$ , as we explain in contribution 1, we recommend cross-validating the offsets as well. In the third step, a Lasso regression is fitted to the block of data with third highest priority and the linear predictor from the second fit is used as an offset. This procedure is similarly applied to all blocks of data with lower priority. This ensures that variables from blocks of low priority are only included in the prediction rule when they provide additional information to variables from blocks of higher priority.

Contribution 1 addresses the development of priority-Lasso and compares it to standard Lasso in terms of selected variables and prediction accuracy. Detailed information about the background and motivation are provided as well as a discussion regarding the method and its application. In the practical application, where we aim to predict the survival times of AML patients, we define four blocks of data based on clinical data, gene mutation data and gene expression data, and determine the block order in consultation with a medical expert. Furthermore, priority-Lasso serves as one of the Lasso-based methods used to quantify method uncertainty in scenario 1 of contribution 2. Again, priority-Lasso is applied to AML data in the context of survival analysis. Here, we apply a more objective variant of priority-Lasso, where we run the method for several block orders, and finally choose the best according to cross-validated prediction performance. In both contributions, we use the implementation from the R-package `prioritylasso` (Klau et al., 2019) for the practical applications.

### 2.3.7 IPF-Lasso

Another extension of standard Lasso that can take heterogeneous data into account is IPF-Lasso (Boulesteix et al., 2017a). Similar to priority-Lasso, it was motivated by the situation where different types of omics markers are available for the same patient in a high dimensional setting. The idea is to assign multiplicative factors to the penalty terms (the so-called penalty factors) for each block of data. For a continuous and centered

outcome, the coefficients are estimated by

$$\sum_{i=1}^n \left( y_i - \sum_{m=1}^M \sum_{j=1}^{p_m} x_{ij}^{(m)} \beta_j^{(m)} \right)^2 + \sum_{m=1}^M \lambda_m \sum_{j=1}^{p_m} |\beta_j^{(m)}| \rightarrow \min_{\beta}$$

with  $m = 1, \dots, M$  indicating the block of data, and  $p_m$  the number of variables contained in this block.  $\lambda_m$  refers to the penalty parameters that can be different for each block of data and are chosen in relation to a reference penalty  $\lambda_1$  as  $\lambda_m/\lambda_1$  for block  $m$ . The penalty factors can be specified either by practical considerations or determined through cross-validation procedures out of a number of candidate vectors.

Another option is to use an adaptive variant, where the tuning parameters are determined in two steps based on the data. In the first step, a single candidate vector of penalty factors is determined by using average values of absolute coefficient estimates of a single block of data obtained by standard Lasso or ridge regression. In the second step, the candidate vector is used to calculate  $\lambda_1$  in order to obtain optimal prediction performance. The IPF-Lasso method and its adaptive variant are implemented in the R-package `ipflasso` (Boulesteix et al., 2019), which can deal with continuous, binary and survival outcomes.

In this work, IPF-Lasso is used as one of the methods in scenario 1 of contribution 2, where different methods for penalized Cox regression are compared in a high dimensional and heterogeneous data setting. In this application, the penalty factors are specified through candidate vectors. For instance, in the case of two blocks of data, a candidate vector  $k = (2, 1)^T$  refers to a penalty parameter  $\lambda_2$ , which is two times smaller than the penalty factor for the first block of data ( $\lambda_1$ ). Conversely,  $k = (1, 2)^T$  indicates a  $\lambda_2$  which is twice  $\lambda_1$ , i.e.,  $\lambda_2/\lambda_1 = 2$ . We specified three and seven candidate vectors for two and three blocks of data, respectively, where one block is always penalized twice or half as much as the other blocks. In the end, the best candidate vector is chosen according to cross-validated prediction error.

Table 2.2: Overview of the regression methods used in this work

Contribution	Type of regression	Outcome	Remarks
1	priority-Lasso standard Lasso	survival times survival times	development and application
2	standard Lasso priority-Lasso IPF-Lasso negative binomial linear	survival times survival times survival times RNA-Seq counts RNA-Seq counts	scenario 1 scenario 1 scenario 1 scenario 3, part of the differential expression methods DESeq, DESeq2, edgeR and glm.edgeR scenario 3, after performing a voom-transformation as part of the method limma
3	logistic linear	binary continuous	effect estimates from logistic regression as outcome
4	Cox	survival times	



## 2.4 Further issues related to this work

In this section, I will describe and discuss selected aspects related to this work. As this work is partly motivated by the non-replicability of observational research findings, I will start by discussing some of the causes and solutions related to the replication crisis in section 2.4.1. Section 2.4.2 addresses some general considerations on how to deal with uncertainty. In section 2.4.3, I will introduce some methods which can be used to report the results obtained with different analysis strategies. In section 2.4.4, precise and imprecise probabilities as concepts to quantify uncertainty will be briefly explained and contrasted with our view of uncertainty. Finally, section 2.4.5 goes beyond the frequentist perspective and provides a glimpse at Bayesian statistics.

### 2.4.1 Some causes and solutions for the replication crisis

Although many contributions to the replication crisis in the scientific literature exist, their causes and solutions are extremely broad and complex – hence, this overview will not be exhaustive. Nevertheless, I will try to summarize a few important points.

One of the most obvious explanations for the non-replicability of research findings is scientific misconduct; however, examples of fraud are in fact rare in practice (Ioannidis et al., 2014). Instead, the problems causing the crisis are more intricate and aspects are differently pronounced depending on the scientific field and the type of study.

First of all, measurement error is largely present in epidemiology and medical research (Brakenhoff et al., 2018), and is, according to Loken and Gelman (2017), one of the contributions to the non-replicability of research findings. Secondly, scientific studies are often of low statistical power, as for instance shown by Szucs and Ioannidis (2017) or Button et al. (2013) in psychological research and neuroscience. Maxwell (2004) directly exemplifies the connection to the non-replicability of research results.

As a third point, the multiplicity of analysis strategies plays an important role in the replication crisis (Simmons et al., 2011). Indeed, the choice of the analysis strategy is often not straightforward and can resemble a ‘garden of forking paths’ (Gelman and Loken, 2013). In this context, there are some underlying aspects which further contribute to the multiplicity of analysis strategies: In high dimensional settings, for instance, few comparison studies are conducted (Boulesteix et al., 2017c) compared to the overabundance of methods, the latter being at least partly caused by the demand for new methodological developments (Boulesteix et al., 2018). Together with few guidelines and standards in this area of research (Boulesteix et al., 2017b), this exacerbates the choice of an appropriate analysis strategy. Furthermore, an increasing amount of data that was not initially collected for research purposes, e.g., from twitter accounts (Barberá et al., 2015) or transaction data (Gladstone et al., 2019), leads to a huge number of researcher degrees of freedom. Finally, due to greater computational power, it is possible to run multiple analysis strategies on a single computer (Young and Holsteen, 2017; Young, 2018).

Although these aspects introduce substantial uncertainty, they are in and of themselves nothing evil. In this regard, Gelman and Hennig (2017) underline the importance of the multiplicity of perspectives for the scientific process. However, the question of

how these aspects are handled in practice uncovers some inadequacy. Measurement error, for instance, is often disregarded in medical research (Brakenhoff et al., 2018), and neither reported nor integrated or reduced in such a scientific study (see section 2.4.2 for more details on different strategies to handle uncertainty). When acknowledging the presence of measurement error, there is a tendency for researchers to consider their results as evidence under challenging conditions and to claim even more impressive results in the absence of measurement error. This conclusion is explicitly pointed out as fallacious by Loken and Gelman (2017).

Moreover, p-values are often misinterpreted, and general concern about the concept of significance testing exists (Goodman, 2008; McShane et al., 2019). Beyond that, researchers tend to choose analysis strategies which yield significant p-values, a practice that is called ‘p-hacking’ (Head et al., 2015) or ‘fishing for significance’ (Boulesteix et al., 2017b). Other questionable research practices in psychological research have been collected by John et al. (2012).

These practices are encouraged by two factors: On the one hand, publication bias, i.e., the tendency to publish only research findings with significant results, is widely present in science (Easterbrook et al., 1991; Ferguson and Brannick, 2012). On the other hand, there is a strong need for researchers to publish their work (Fanelli, 2010), a circumstance that is widely known under the slogan ‘publish or perish’. As a consequence, researchers have pressure to obtain significant results, which can be enforced through fishing for significance or other questionable research practices.

Addressing these issues is a problem with multiple dimensions which does not only affect the researcher, but also journals and funding agencies. Here, I will briefly outline a few possible solutions; however, this list is not exhaustive and a full discussion is beyond the scope of this work. The first solution is for journals to encourage researchers to pre-register their papers (Wagenmakers et al., 2012) or to submit registered reports (Chambers, 2013). The idea behind this is to evaluate a study’s methods and analysis, and decide whether it will be published, before the results are known. More generally, the researchers raise their voices for open research and open data (Nosek et al., 2015) in order to increase transparency.

The approach of banning (Amrhein et al., 2019) or abandoning p-values (McShane et al., 2019) is a radical but widely supported solution for the problems related to p-values and significance testing (Goodman, 2008). Moreover, lowering the significance threshold has been suggested, in order to publish only results with strong scientific evidence (Benjamin et al., 2018). More guidelines and standards (Klau et al., 2020a) could shed some light on the darkness that is the multiplicity of analysis strategies, along with comparison or benchmarking studies (Boulesteix et al., 2018). Furthermore, Schooler (2014) claims that metascience and replication studies could counteract the replication crisis. Finally, more radical solutions have been proposed, e.g., to rearrange the whole scientific system involving multiple parties (Knuteson, 2016; Romero, 2019). These solutions also include the creation of new incentives for researchers with the aim of discouraging questionable research practices.

## 2.4.2 Different ways to deal with uncertainty

The following aspects on how to deal with uncertainty are based on the work of Hoffmann et al. (2020). One motivation for these actions is the non-replicability of research findings, which was addressed in section 2.4.1 – nonetheless, there is no direct relation and the following considerations also hold true in general.

First of all, it is important to *acknowledge* that there is uncertainty as part of a scientific study, and that this usually goes beyond sampling uncertainty which most researchers are familiar with. In fact, uncertainties are associated with various other sources which can have important consequences on the results. Acknowledging these uncertainties can be seen as a minimum requirement for statisticians, whose whole work is based on the concept of uncertainty. Moreover, all other aspects discussed in this section will be based on this first and essential point. The acknowledgement of uncertainties is crucial for cautious interpretation of results and can help researchers to avoid overconfidence. In addition, it prevents them from generalizing their conclusions to other data or other methods without a firm basis.

Another aspect in dealing with uncertainties is to *reduce* them as much as possible. The approaches to reduce uncertainty and their feasibility strongly differ depending on the type of uncertainty. For sampling and measurement uncertainty resulting from random errors, a straightforward way is to increase the sample size of the study (Button et al., 2013). Moreover, measurement devices, questionnaires, interview techniques, and data entry processes can be improved to reduce measurement uncertainty (Hoffmann et al., 2020). On the other hand, aspects to reduce epistemic uncertainty comprise, for instance, benchmarking studies (Boulesteix et al., 2017c) and pre-registration practices (Wagenmakers et al., 2012).

Furthermore, uncertainties should be *integrated* into the statistical analysis, as typically done with sampling uncertainty when conducting statistical inference. However, the integration of other types of uncertainty in the analysis should not be neglected. For example, model uncertainty could be addressed by Bayesian model averaging (Hoeting et al., 1999) or multimodel inference (Burnham and Anderson, 2004). To integrate measurement uncertainty in the statistical analysis, regression calibration (Rosner et al., 1989) or simulation extrapolation (Cook and Stefanski, 1994) can be applied. For other types of uncertainty or specific applications, it is not clear how to account for uncertainties, which necessitates the improvement or development of appropriate tools. This is, for instance, relevant in the context of data pre-processing uncertainty, which is not straightforward to integrate in the statistical analysis.

Finally, uncertainties should be systematically *reported*. We believe that researchers should seek visual or quantitative impressions of uncertainties, and share this information with reviewers and readers of their articles. This helps the corresponding researchers to assess their work with regard to uncertainties, and also increases transparency and promotes open research practices. Therefore, there is need for tools to visualize, quantify and compare different types of uncertainty.

Although it is not always feasible to integrate or reduce uncertainty, it is, in summary, important for all members of the statistical community to understand the roles different types of uncertainty play in a scientific study, and to deal with them appropriately.

### 2.4.3 Methods to report the results of different analysis strategies

One of the challenges of dealing with uncertainty, discussed in section 2.4.2, is the reporting of different types of uncertainty. Therefore, straightforward methods to quantify and illustrate uncertainty are needed. In this short overview, I will outline some of the methods suggested to report epistemic uncertainty, including the vibration of effects approach, which is extended and used in contributions 3 and 4. I will also briefly discuss some advantages and disadvantages of these approaches.

**Computational model robustness** Computational model robustness was developed to estimate all possible models from a theoretically informed model space (Muñoz and Young, 2018; Young, 2018). As a model space, the authors do not only consider specific choices in a probability model, but suggest extending the model space for instance to variable definitions or software implementations. Therefore, they address what we denote as method uncertainty in their work. Young (2018) extensively discusses how to define such a model space and presents three approaches. The first approach is motivated by the idea that all models an analyst considered as worth running during the study are worth reporting. An ‘uber log file’ could theoretically save all these models. The second approach, denoted as the ‘task force approach’, combines a wide range of expert opinions, which is close to the crowdsourcing approach of Silberzahn and Uhlman (2015). As a third strategy, the authors suggest combining the uber log file and the task force approach. Taking all the models into account, a ‘modeling distribution’ can be calculated and visualized with kernel density graphs. In such a figure, a favorite model can be indicated.

**Specification Curve** The specification curve analysis was proposed and practically illustrated by Simonsohn et al. (2015) in the field of social science. It considers all operationalization decisions in the data analysis as specifications, and thus addresses what we denote as method uncertainty. Conducting a specification curve analysis can be summarized in three steps: First, all reasonable specifications have to be found, second, all of these specifications have to be calculated, and third, a joint permutation test is performed to test the null hypothesis of no effect. For illustration, Simonsohn et al. (2015) suggest a two-paneled figure, with an upper part showing a ‘curve’ of estimated effects and specification numbers. In this curve, a clear distinction between negative and positive estimates can be made, and significant estimates can be highlighted. In the lower panel, information about the decisions that produce the estimates can be found. A practical application of specification curve analysis was provided by Rohrer et al. (2017) in psychological research, who investigated birth-order effects on personality traits.

**Multiverse analysis** The multiverse analysis was suggested by Steegen et al. (2016) in psychological research with the aim of performing a statistical analysis for different data pre-processing steps. To ensure that the alternative data sets cover reasonable choices, they base their practical application on previously published studies, where

these choices have actually been considered. In order to visualize the results, they suggest showing a histogram of raw p-values. The distribution of p-values obtained by different data pre-processing choices can give information on the robustness of findings due to alternative choices: p-values which are nearly uniformly distributed are not as robust as p-values that indicate increased significance. Furthermore, for a more detailed investigation, results can be reported in grids of p-values, where a p-value can be traced back to the analysis strategy that yielded it. In addition to the practical illustration of Steegen et al. (2016), applications of a multiverse analysis can for instance be found in McBee et al. (2019), Stern et al. (2019), or Credé and Phillips (2017).

**Vibration of effects** The concept of vibration of effects was initially proposed by Ioannidis (2008) and extended by Patel et al. (2015), who used it to practically examine model uncertainty in a large epidemiological study. The developers suggest visualizing results obtained from different analysis strategies with volcano plots. These plots typically show p-values on the y-axis and effect estimates on the x-axis. Moreover, the variability of p-values and effect estimates can be quantified through summary measures. As such, Patel et al. (2015) suggest relative effect estimates and relative p-values, defined as the ratio of the 99th and 1st percentile of effect estimates and the difference between the 99th and 1st percentile of  $-\log_{10}(\text{p-value})$ , respectively. Apart from these primal works, applications of the framework can be found in Palpacuer et al. (2019) and Chu et al. (2020) for different method choices. In our work, we will use and extend the vibration of effects framework in order to assess and compare measurement, sampling, model and data pre-processing uncertainty. Moreover, we will apply it for different types of regression (logistic regression (section 2.3.2) and Cox regression (section 2.3.4)), which results in relative odds ratios and relative hazard ratios as summary measures in order to quantify the variability of effect estimates.

**Discussion of advantages and disadvantages** In contrast to computational model robustness and the vibration of effects, specification curve analysis and multiverse analysis allow easy tracing back of results to the corresponding analytical choices. This, however, results in the disadvantage of there being a limited number of models that can be considered for the visualization. For a large number of analytical choices, this can for instance be accounted for by visualizing only a subset of these decisions. Furthermore, when conducting a multiverse analysis, the focus of visualization can be the histogram rather than the grid of p-values. Similarly, for a specification curve analysis, only the upper panel of the suggested figure can be shown.

With regard to the other approaches, the specification curve analysis implicates a permutation test, which provides a decision over all specifications. However, performing such a test is very computationally demanding and its application has not yet gained acceptance in practice. On the other hand, the vibration of effects framework encompasses relative effect estimates and relative p-values as summary measures of the variability of results. Yet, neither these summary measures nor the permutation test are in principle limited to their specific framework of visualization.

In general, none of the approaches are limited to the type of uncertainty for which they were originally suggested. Using the specification curve only for data pre-processing or model choices is straightforward, and in a multiverse analysis, decisions on model specification can be similarly included to data pre-processing choices. Finally, the vibration of effects framework can be extended to sampling, data pre-processing and measurement uncertainty, as we demonstrate in contributions 3 and 4. In contrast to the other approaches, this framework provides visualization of effect estimates and p-values simultaneously. Moreover, for epistemic uncertainty, it allows the highlighting of points in volcano plots in order to visualize the impact of particular choices. Thus, key choices can be identified.

#### 2.4.4 Quantifying uncertainties through precise and imprecise probabilities

The most established concept to quantify uncertainty, from both a theoretical and a practical perspective, is probability. There are many contributions to probability theory, e.g., by Laplace (Laplace, 1820) and Kolmogorov (Kolmogorov, 1933). The work of these two celebrated scientists on probability is associated with the definitions of classical and axiomatic probabilities, respectively. Moreover, there are objective and subjective perspectives on probability (De Finetti, 1970), from which the frequentist and Bayesian inference can be derived, respectively. Some researchers argue that probability is the only concept needed to quantify uncertainty, while other representations are inadmissible (Lindley, 1982). In contrast, other researchers question whether using a single value to express probability is unrealistically precise (Augustin et al., 2014; Hall et al., 2007).

A generalized version of probability theory that aims to account for this latter issue is imprecise probability. Early contributions on this theory are given by George Boole, who was one of the first to recognize the problem of unrealistic precision in probabilities (Boole, 1854). The basic idea of imprecise probabilities is to provide upper and lower bounds of probability. In particular, when the aim is to quantify the probability of an event  $A$ , the precise probability of  $A$  can be denoted by  $P(A)$ . Upper and lower probabilities of  $A$  can be expressed by  $\underline{P}(A)$  and  $\overline{P}(A)$ , where  $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$ . The extreme case where  $\underline{P}(A) = \overline{P}(A)$  refers to precise probability. In contrast, in the situation where  $\underline{P}(A) = 0$  and  $\overline{P}(A) = 1$ , no information at all can be given on  $A$ .

Based on this idea, the term imprecise probability comprises many theoretical contributions and extensions of formulations of probability. Walley (2000) intended to unify some of these contributions and to provide a general theory of imprecise probabilities. For more details and a comprehensive overview of imprecise probability theory, I refer to Augustin et al. (2014).

Applications of imprecise probability can be mainly found in climatology, where evidence is gathered from a variety of different sources in a range of formats, which are not necessarily numerical (Hall et al., 2007). Kriegler et al. (2009) create subjective probability intervals on the basis of expert elicitation on the occurrence of major changes in the climate system. Hall et al. (2007) combine imprecise probabilities and fuzzy linguistic uncertainties to show upper and lower bounds of global mean temperature.

Ghosh and Mujumdar (2009) downscale multiple general circulation models in order to derive imprecise cumulative distribution functions for monsoon rainfall by applying interval regression. Besides climatology, an important area of application for imprecise probability is engineering (Augustin and Hable, 2010; Oberguggenberger et al., 2009). Especially when it comes to questions of structural safety, imprecise probabilities can be of exceptional importance (Zhang et al., 2017), as wrong decisions can have severe negative consequences.

Together with imprecise probabilities, Aven (2011) lists some other alternative representations of uncertainty, including possibility theory (Dubois and Prade, 1988), evidence-theory by Dempster and Shafer (Dempster, 2008) and fuzzy probabilities (Zadeh, 1968). Ghosh and Mujumdar (2009) consider imprecise probabilities as a generalized form of possibility theory and Dempster-Shafer theory. All these theories are fundamentally different from our practical view on uncertainty, nevertheless it is important to acknowledge them, as they provide important contributions from different perspectives on uncertainty.

## 2.4.5 A glimpse at Bayesian statistics

All aspects of uncertainty addressed in this work are made from a frequentist perspective, which is based on the assumption of the theoretically infinite repeatability of an experiment. Another approach to statistical inference is the Bayesian approach, which allows expression of epistemic uncertainty in terms of probabilities and inclusion of external information in the inference process. The basic concept of Bayesian statistics is to combine prior probabilities, which can be specified according to external information, with a likelihood function with the goal of obtaining posterior probabilities.

The Bayesian approach allows more flexible modeling of uncertainties than the frequentist approach. For instance, an established method to account for model uncertainty is Bayesian model averaging (Hoeting et al., 1999). For a given set of models  $M_1, \dots, M_K$  and data  $D$ , the posterior model probability  $\pi(M_k|D)$  can be obtained by

$$\pi(M_k|D) = \frac{\pi(D|M_k)\pi(M_k)}{\sum_{l=1}^K \pi(D|M_l)\pi(M_l)}.$$

Here,  $\pi(D|M_k)$  describes the integrated likelihood of model  $M_k$  and  $\pi(M_k)$  the prior probabilities which have been specified in advance. On the one hand, the results can be used with a focus on single models, e.g., two models can be compared with Bayes factors. Alternatively, a single model can be selected based on posterior model probabilities. Models can also be averaged and weighted by posterior model probabilities. Hoeting et al. (1999) argue that the average estimates over all models are robust to model choice and show that the average predictive performance is better than any single model. However, Bayesian model averaging is not without challenges in practice, since the choice of models and the specification of prior distributions is not straightforward. For the latter, a simple solution is to assign the same prior probability to each model.

Moreover, the Bayesian framework allows easily integration of measurement uncertainty in the inference process. A detailed overview on Bayesian measurement error

modeling can be found in Gustafson (2003), and practical examples are for instance provided by Richardson and Gilks (1993) and Hoffmann et al. (2017). Data pre-processing choices, however, can address different research questions of interest. As these choices refer to truly different associations, the integration of data pre-processing uncertainty in the Bayesian framework is connected with some difficulties. Apart from this, uncertainties can be flexibly integrated and different types of uncertainty can be combined using Bayesian approaches. The sometimes high computational challenges of these models are usually addressed with Markov chain Monte Carlo algorithms (Carlin and Chib, 1995).

Nevertheless, Bayesian inference is still underrepresented in practical applications in comparison to frequentist approaches. A conceivable explanation could be that the practical implementation is often very complex and the computational challenges are high. Furthermore, this may be a philosophical issue, because many researchers do not feel comfortable with the subjectivity underlying the Bayesian point of view (Gelman, 2008).



# Chapter 3

## Structure and overview of this thesis

### 3.1 Summary of the contributions

In this work, the challenges of uncertainty in regression models for high dimensional and heterogeneous data from observational studies are addressed in four different contributions. I will start the following overview of these contributions with our earliest work (contribution 1), which is an extension of my master thesis and introduces the methodological basics for some of the other contributions. Although there is no direct connection to uncertainties, it can be seen as a practical illustration of data pre-processing and method uncertainty.

In particular, we focus on the development of priority-Lasso (section 2.3.6), a regression method for high-dimensional and heterogeneous data (sections 2.2.2 and 2.2.3) in the first part of this contribution. In the second part, where we practically apply priority-Lasso to AML data, we deal with uncertainties due to data pre-processing and method choices. Here, we illustrate data pre-processing uncertainty (section 2.1.4) through the choice of pre-processing the variables included in the block of priority 1. Furthermore, method uncertainty (section 2.1.5) arises through the choice of whether offsets are cross-validated or not. We compare these different approaches with standard Lasso (section 2.3.5) in terms of included variables and prediction accuracy. In addition, there exist a magnitude of other reasonable data pre-processing and method choices for this research question of interest in general: These choices include decisions on whether to apply standard Lasso or priority-Lasso, how many blocks to specify, which variables to include, how to define the variables, which method settings to use, and many more. Other types of uncertainty, like measurement and sampling uncertainty (sections 2.1.1 and 2.1.2) may also play an important role as they are ubiquitous in biomedical research. However, since the focus of this contribution is the method priority-Lasso itself, we merely acknowledge these further choices and uncertainties without systematically examining them. This contribution was written by me, Vindi Jurinovic, Roman Hornung, Tobias Herold and Anne-Laure Boulesteix, and published as 'Priority-Lasso: A simple hierarchical approach to the prediction of clinical outcome using multi-omics data' in BMC Bioinformatics in 2018 (Klau et al., 2018).

In contribution 2, we develop a framework for the comparison of sampling and method uncertainty (sections 2.1.2 and 2.1.5, respectively). The idea of this framework is, for a given data set, to split this data set in two halves of equal size and run two or more methods on each of these halves. By repeating this procedure with several random splits of the data, sampling and method uncertainty can be assessed by intuitive summary measures. Here, the framework is practically illustrated through three representative examples in the context of variable selection and ranking for high dimensional data (section 2.2.2). In the first example, we select variables from high dimensional and heterogeneous data (section 2.2.3) by using the three penalized regression methods standard Lasso, priority-Lasso and IPF-Lasso (sections 2.3.5 -2.3.7). In the second example, we rank variables from a high dimensional data set through random forest variable importance measures and assess method uncertainty between different tuning parameter settings. Finally, we identify differentially expressed genes from a high dimensional data set with five different approaches of differential expression analysis. These approaches are either based on linear regression (section 2.3.1, limma-voom) or negative binomial regression (section 2.3.3, DESeq, DESeq2, edgeR and glm.edgeR). Beyond these examples, it is straightforward to apply this framework to other statistical scenarios and other types of uncertainty. This contribution was created by me, Marie-Laure Martin-Magniette, Anne-Laure Boulesteix and Sabine Hoffmann, and published with the title ‘Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection’ in *Biometrical Journal* in 2019 (Klau et al., 2020a).

In contribution 3, we focus on sampling, model and data pre-processing uncertainty (sections 2.1.2, 2.1.3 and 2.1.4), and compare these three types of uncertainty on a data set from psychology encompassing heterogeneous variables (section 2.2.3) in a common framework. Therefore, we extend the vibration of effects framework (Ioannidis, 2008; Patel et al., 2015) to sampling and data pre-processing uncertainty and recommend using it as a tool for researchers and readers to systematically examine and report these different types of uncertainty. In our practical application, we study several associations of interests between a binary outcome and one of the Big Five personality traits with logistic regression models (section 2.3.2), where we consider 12 additional variables as adjustment variables in each model. We pre-process the variables in different ways, and run the model with every possible combination of the adjustment variables to study data pre-processing and model uncertainty, respectively. Finally, we subsample the data a large number of times to assess sampling vibration. The large data set from the SAPA personality project with more than 80000 observations allows investigation of these types of uncertainty for varying sample sizes. Furthermore, we assess the relative impact of model and data pre-processing vibration on the total vibration due to the analysis strategy. This contribution was published as a Technical Report on the homepage of the Ludwig-Maximilians-Universität München in 2020 (Klau et al., 2020b). Me, Felix Schönbrodt, Chirag Patel, John Ioannidis, Anne-Laure Boulesteix and Sabine Hoffmann entitled it ‘Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology’.

Similarly to contribution 3, we use the vibration of effects framework to study three types of uncertainty in contribution 4. Here, we develop a strategy to examine measurement uncertainty (section 2.1.1) and compare this type of uncertainty to sampling and model uncertainty (sections 2.1.2 and 2.1.3, respectively) in a practical application with the vibration of effects framework. We perform this application on data from the NHANES, which is a data set from epidemiology consisting of heterogeneous variables (section 2.2.3). The data set comprises variables, for instance, from questionnaires, health examinations, or laboratory tests. In this application, we successively study the association between several variables of interest and mortality in a Cox regression (section 2.3.4). Furthermore, we consider 15 additional adjustment variables for each regression model, which we combine in different ways in order to examine model uncertainty. In our assessment of measurement uncertainty, we vary our strategy to which variables measurement error is added. In addition to the real data analysis, a large simulation study is provided in order to compare these three types of uncertainty for different sample sizes. This contribution was written by me, Sabine Hoffmann, Chirag Patel, John Ioannidis and Anne-Laure Boulesteix. It is available at the homepage of the Ludwig-Maximilians-Universität München (Klau et al., 2020c) and currently under review at the International Journal of Epidemiology.

## 3.2 Further steps

As we show in our contributions, the types of uncertainty we defined in this work strongly depend on the type of analysis and characteristics of the data. Therefore, general conclusions, e.g., of the form ‘data pre-processing uncertainty is always higher than model uncertainty’, or ‘sampling uncertainty is higher than method uncertainty’ cannot be made. Instead, it is essential to study different types of uncertainty in different settings. On the one hand, this should be done by methodological researchers in further comparison studies for the purpose of assessing and understanding uncertainties. This is in line with the need for more comparison studies relative to the vast amount of methodological developments (Boulesteix et al., 2018). On the other hand, as we claim in our contributions, the examination and comparison of different types of uncertainty should be conducted as a standard procedure in practical research.

With our framework to compare sampling and method uncertainty (Klau et al., 2020a) and the extension of the vibration of effects (Klau et al., 2020b,c), we provide such tools that allow easy quantification, visualization and comparison of different types of uncertainty. Furthermore, while computational resources are increasing, researchers often run their analyses with many different analytical choices, like slightly modified data sets or different combinations of covariates. Hence, reporting additional results should cease to be an obstacle from both a methodological (e.g., by using our tools) and practical point of view.

However, these additional analyses are usually not systematically conducted. Therefore, not only the methodological tools should be provided, but the practical implementation to systematically investigate and report different types of uncertainty should be simplified. Hence, there is need for extending our work with a user-friendly software

implementation of this or a similar general framework. Although it is impossible to cover all types of analysis and all choices that can theoretically be made, some basics that are commonly relevant can be implemented. A shiny app could, for instance, be suitable for this purpose. In this implementation, not only a single framework like the vibration of effects can be used for the purpose of illustration. In fact, frameworks like the specification curve analysis or the multiverse analysis have other advantages (see section 2.4.3), and could additionally be utilized to provide more flexibility for the users.

Moreover, future work should focus on methods that not only acknowledge and report, but also integrate and reduce uncertainty. In this regard, it is important to work in an interdisciplinary manner and learn from other fields of research (Rigdon et al., 2020).

The further steps described above are related to this thesis and our view of uncertainty. In addition, it should be the aim of every researcher to contribute to better research and to a solution to the replication crisis by following the suggestions partly introduced in section 2.4.1 as well as possible and good research practices in general.

# Bibliography

- Ahlgren, A. (1969). A modest proposal for encouraging replication. *American Psychologist*, 24(4):471.
- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
- Altman, D. G. and Bland, J. M. (2014). Uncertainty beyond sampling error. *BMJ*, 349:g7065.
- Amrhein, V., Greenland, S., and McShane, B. (2019). Retire statistical significance. *Nature*, 567:305–307.
- Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, 11:R106.
- Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D. (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, 8:1765–1786.
- Augustin, T., Coolen, F. P. A., De Cooman, G., and Troffaes, M. C. M. (2014). *Introduction to imprecise probabilities*. John Wiley & Sons, Chichester, UK.
- Augustin, T. and Hable, R. (2010). On the impact of robust statistics on imprecise probability models: a review. *Structural Safety*, 32(6):358–365.
- Aven, T. (2011). Interpretations of alternative uncertainty representations in a reliability and risk analysis context. *Reliability Engineering & System Safety*, 96(3):353–360.
- Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., and Bonneau, R. (2015). Tweeting from left to right: is online political communication more than an echo chamber? *Psychological Science*, 26(10):1531–1542.
- Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3):407–425.
- Benestad, R. E., Nuccitelli, D., Lewandowsky, S., Hayhoe, K., Hygen, H. O., Van Dorland, R., and Cook, J. (2016). Learning from mistakes in climate research. *Theoretical and Applied Climatology*, 126:699–703.

- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., Bollen, K. A., Brembs, B., Brown, L., Camerer, C., et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2:6–10.
- Boole, G. (1854). *An investigation of the laws of thought, on which are founded the mathematical theories of logic and probabilities*. Macmillan and Co., London.
- Boulesteix, A.-L., Binder, H., Abrahamowicz, M., and Sauerbrei, Willi for the Simulation Panel of the STRATOS Initiative (2018). On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, 60:216–218.
- Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017a). IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*. doi: 10.1155/2017/7691937.
- Boulesteix, A.-L., Fuchs, M., and Schulze, G. (2019). *ipflasso: Integrative Lasso with Penalty Factors*. R package version 1.1.
- Boulesteix, A.-L., Hornung, R., and Sauerbrei, W. (2017b). On fishing for significance and statistician’s degree of freedom in the era of big molecular data. In Pietsch, W., Wernecke, J., and Ott, M., editors, *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data*, pages 155–170. Springer, Wiesbaden.
- Boulesteix, A.-L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, 10(5):556–568.
- Boulesteix, A.-L., Wilson, R., and Hapfelmeier, A. (2017c). 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):1–12.
- Bradburn, M. J., Clark, T. G., Love, S. B., and Altman, D. G. (2003). Survival analysis part II: multivariate data analysis – an introduction to concepts and methods. *British Journal of Cancer*, 89:431–436.
- Braess, J., Kreuzer, K.-A., Spiekermann, K., Lindemann, H. W., Lengfelder, E., Graeven, U., Staib, P., Ludwig, W.-D., Biersack, H., Ko, Y.-D., et al. (2013). High efficacy and significantly shortened neutropenia of dose-dense S-HAM as compared to standard double induction: first results of a prospective randomized trial (AML-CG 2008). *Blood*, 122(21):619.
- Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G. M., Groenwold, R. H. H., and van Smeden, M. (2018). Measurement error is often neglected in medical literature: a systematic review. *Journal of Clinical Epidemiology*, 98:89–97.
- Breusch, T. S. (1990). Simplified extreme bounds. *Modelling Economic Series*, pages 72–82.
- Büchner, T., Berdel, W. E., Schoch, C., Haferlach, T., Serve, H. L., Kienast, J., Schnittger, S., Kern, W., Tchinda, J., Reichle, A., et al. (2006). Double induction containing either two courses or one course of high-dose cytarabine plus mitoxantrone and postremission

- therapy by either autologous stem-cell transplantation or by prolonged maintenance for acute myeloid leukemia. *Journal of Clinical Oncology*, 24(16):2480–2489.
- Büchner, T., Krug, U. O., Gale, R. P., Heinecke, A., Sauerland, M. C., Haferlach, C., Schnittger, S., Haferlach, T., Müller-Tidow, C., Stelljes, M., et al. (2016). Age, not therapy intensity, determines outcomes of adults with acute myeloid leukemia. *Leukemia*, 30(8):1781–1784.
- Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304.
- 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.
- Carlin, B. P. and Chib, S. (1995). Bayesian model choice via markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(3):473–484.
- Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, 49(3):609–610.
- Chu, L., Ioannidis, J. P. A., Egilman, A. C., Vasiliou, V., Ross, J. S., and Wallach, J. D. (2020). Variation of effects in epidemiologic studies of alcohol consumption and breast cancer risk. *International Journal of Epidemiology*. doi: 10.1093/ije/dyz271.
- Condon, D. M., Roney, E., and Revelle, E. (2017). A SAPA project update: on the structure of phrased self-report personality items. *Journal of Open Psychology Data*, 5(1):3.
- Cook, J. R. and Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, 89(428):1314–1328.
- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–220.
- Credé, M. and Phillips, L. A. (2017). Revisiting the power pose effect: how robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological and Personality Science*, 8(5):493–499.
- De Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica, Amsterdam*, 34(2–3):129–145.
- Dempster, A. P. (2008). Upper and lower probabilities induced by a multivalued mapping. In Yager, R. R. and Liu, L., editors, *Classic Works of the Dempster-Shafer Theory of Belief Functions. Studies in Fuzziness and Soft Computing*, volume 219, pages 57–72. Springer, Berlin, Heidelberg.

- Der Kiureghian, A. and Ditlevsen, O. (2009). Aleatory or epistemic? Does it matter? *Structural Safety*, 31(2):105–112.
- Dubois, D. and Prade, H. (1988). *Possibility theory: an approach to computerized processing of uncertainty*. Plenum Press, New York, London.
- Easterbrook, P. J., Gopalan, R., Berlin, J. A., and Matthews, D. R. (1991). Publication bias in clinical research. *The Lancet*, 337(8746):867–872.
- Fahrmeir, L., Kneib, T., and Lang, S. (2007). *Regression*. Springer, Berlin, Heidelberg.
- Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from US states data. *PloS ONE*, 5(4):e10271.
- Ferguson, C. J. and Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, 17(1):120–128.
- Fillenbaum, G. G., Burchett, B. M., and Blazer, D. G. (2009). Identifying a national death index match. *American Journal of Epidemiology*, 170(4):515–518.
- Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.
- Gelman, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, 3(3):445–450.
- Gelman, A. (2015). Working through some issues. *Significance*, 12(3):33–35.
- Gelman, A. and Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(4):967–1033.
- Gelman, A. and Loken, E. (2013). The garden of forking paths: why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University*, pages 1–17.
- Ghosh, S. and Mujumdar, P. P. (2009). Climate change impact assessment: uncertainty modeling with imprecise probability. *Journal of Geophysical Research: Atmospheres*, 114(D18):1–17.
- Gibb, B. C. (2014). Reproducibility. *Nature Chemistry*, 6(8):653–654.
- Gilmore, R. O., Diaz, M. T., Wyble, B. A., and Yarkoni, T. (2017). Progress toward openness, transparency, and reproducibility in cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1396(1):5–18.
- Gladstone, J. J., Matz, S. C., and Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological Science*, 30(7):1087–1096.



- Gliner, J. A., Leech, N. L., and Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST): what do the textbooks say? *The Journal of Experimental Education*, 71(1):83–92.
- Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., et al. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537.
- Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology*, 45(3):135–140.
- Granger, C. W. J. and Uhlig, H. F. (1990). Reasonable extreme-bounds analysis. *Journal of Econometrics*, 44(1-2):159–170.
- Gustafson, P. (2003). *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Chapman and Hall/CRC, London, New York, Washington, D.C.
- Hall, J., Fu, G., and Lawry, J. (2007). Imprecise probabilities of climate change: aggregation of fuzzy scenarios and model uncertainties. *Climatic Change*, 81:265–281.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., and Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3):e1002106.
- Herndon, T., Ash, M., and Pollin, R. (2013). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, 38(2):257–279.
- Herold, T., Jurinovic, V., Batcha, A. M. N., Bamopoulos, S. A., Rothenberg-Thurley, M., Ksienzyk, B., Hartmann, L., Greif, P. A., Phillippou-Massier, J., Krebs, S., et al. (2017). A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *Haematologica*, 103(3):456–465.
- Herold, T., Metzeler, K. H., Vosberg, S., Hartmann, L., Röllig, C., Stölzel, F., Schneider, S., Hubmann, M., Zellmeier, E., Ksienzyk, B., et al. (2014). Isolated trisomy 13 defines a homogeneous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. *Blood*, 124(8):1304–1311.
- Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press, New York.
- Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.
- Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, 14(4):382–401.

- Hoffmann, S., Rage, E., Laurier, D., Laroche, P., Guihenneuc, C., and Ancelet, S. (2017). Accounting for Berkson and classical measurement error in radon exposure using a Bayesian structural approach in the analysis of lung cancer mortality in the French cohort of uranium miners. *Radiation Research*, 187(2):196–209.
- Hoffmann, S., Schönbrodt, F. D., Elsas, R., Wilson, R., Strasser, U., and Boulesteix, A.-L. (2020). The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. *MetaArXiv*, pages 1–39. doi: 10.31222/osf.io/afb9p.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8):696–701.
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5):640–648.
- Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, 18(5):235–241.
- Ioannidis, J. P. A., Ntzani, E. E., Trikalinos, T. A., and Contopoulos-Ioannidis, D. G. (2001). Replication validity of genetic association studies. *Nature Genetics*, 29(3):306–309.
- 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.
- John, O. P., Srivastava, S., et al. (1999). The Big Five trait taxonomy: history, measurement, and theoretical perspectives. In Pervin, L. A. and John, O. P., editors, *Handbook of personality: Theory and research*, pages 102–138. Guilford, New York, London.
- Klau, S., Martin-Magniette, M.-L., Boulesteix, A.-L., and Hoffmann, S. (2020a). Sampling uncertainty versus method uncertainty: a general framework with applications to omics biomarker selection. *Biometrical Journal*, 62(3):670–687. doi: 10.1002/bimj.201800309.
- Klau, S., Schönbrodt, F. D., Patel, C. J., Ioannidis, J. P. A., Boulesteix, A.-L., and Hoffmann, S. (2020b). Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. Technical Report 232, Ludwig-Maximilians-Universität München. doi: 10.5282/ubm/epub.70485.
- Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P. A., and Boulesteix, A.-L. (2020c). Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. Technical Report 233, Ludwig-Maximilians-Universität München. doi: 10.5282/ubm/epub.70493 (accepted at the *International Journal of Epidemiology*).
- Klau, S., Hornung, R., and Bauer, A. (2019). *prioritylasso: Analyzing Multiple Omics Data with an Offset Approach*. R package version 0.2.3.

- Klau, S., Jurinovic, V., Hornung, R., Herold, T., and Boulesteix, A.-L. (2018). Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, 19(322):1–14.
- Kleinbaum, D. G. and Klein, M. (2002). *Logistic regression*. Springer, New York, Berlin, Heidelberg.
- Knuteson, B. (2016). The solution to science’s replication crisis. Available at SSRN, doi: 10.2139/ssrn.2835131.
- Kolmogorov, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Springer, Berlin, Heidelberg, New York.
- Kriegler, E., Hall, J. W., Held, H., Dawson, R., and Schellnhuber, H. J. (2009). Imprecise probability assessment of tipping points in the climate system. *Proceedings of the National Academy of Sciences*, 106(13):5041–5046.
- Laplace, P. S. (1820). *Théorie analytique des probabilités*. Courcier, Paris.
- Lash, T. L. (2017). The harm done to reproducibility by the culture of null hypothesis significance testing. *American Journal of Epidemiology*, 186(6):627–635.
- Law, C. W., Chen, Y., Shi, W., and Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15(R29):1–17.
- Leamer, E. E. (1983). Let’s take the con out of econometrics. *The American Economic Review*, 73(1):31–43.
- Li, J., Liu, H., Downing, J. R., Yeoh, A. E.-J., and Wong, L. (2003). Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. *Bioinformatics*, 19(1):71–78.
- Lindley, D. V. (1982). Scoring rules and the inevitability of probability. *International Statistical Review / Revue Internationale de Statistique*, 50(1):1–11.
- Loken, E. and Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325):584–585.
- Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(550):1–21.
- Manzoni, C., Kia, D. A., Vandrovcova, J., Hardy, J., Wood, N. W., Lewis, P. A., and Ferrari, R. (2016). Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. *Briefings in Bioinformatics*, 19(2):286–302.
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, 9(2):147–163.

- McBee, M. T., Brand, R. J., and Dixon, W. (2019). Challenging the link between early childhood television exposure and later attention problems: a multiverse analysis. *PsyArXiv*, pages 1–43. doi: 10.31234/osf.io/5hd4r.
- McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, 40(10):4288–4297.
- McNeil, D. (1996). *Epidemiological research methods*. John Wiley & Sons.
- McShane, B. B., Gal, D., Gelman, A., Robert, C., and Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(sup1):235–245.
- Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.
- Muñoz, J. and Young, C. (2018). We ran 9 billion regressions: eliminating false positives through computational model robustness. *Sociological Methodology*, 48(1):1–33.
- Nardi, A. and Schemper, M. (2003). Comparing Cox and parametric models in clinical studies. *Statistics in Medicine*, 22(23):3597–3610.
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., et al. (2015). Promoting an open research culture. *Science*, 348(6242):1422–1425.
- Oberguggenberger, M., King, J., and Schmelzer, B. (2009). Classical and imprecise probability methods for sensitivity analysis in engineering: a case study. *International Journal of Approximate Reasoning*, 50(4):680–693.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.
- Palpacuer, C., Hammas, K., Duprez, R., Laviolle, B., Ioannidis, J. P. A., and Naudet, F. (2019). Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Medicine*, 17(174):1–13.
- Patel, C. J., Burford, B., and Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9):1046–1058.
- Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572.
- Richardson, S. and Gilks, W. R. (1993). A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, 138(6):430–442.

- Rigaill, G., Balzergue, S., Brunaud, V., Blondet, E., Rau, A., Rogier, O., Caius, J., Maugis-Rabusseau, C., Soubigou-Taconnat, L., Aubourg, S., et al. (2016). Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis. *Briefings in Bioinformatics*, 19(1):65–76.
- Rigdon, E. E., Sarstedt, M., and Becker, J.-M. (2020). Quantify uncertainty in behavioral research. *Nature Human Behaviour*, pages 1–3. doi: 10.1038/s41562-019-0806-0.
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.
- Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.
- Rohrer, J. M., Egloff, B., and Schmukle, S. C. (2017). Probing birth-order effects on narrow traits using specification-curve analysis. *Psychological Science*, 28(12):1821–1832.
- Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, 14(11):e12633.
- Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8(9):1051–1069.
- Sauerbrei, W., Boulesteix, A.-L., and Binder, H. (2011). Stability investigations of multi-variable regression models derived from low-and high-dimensional data. *Journal of Biopharmaceutical Statistics*, 21(6):1206–1231.
- Schooler, J. W. (2014). Metascience could rescue the ‘replication crisis’. *Nature*, 515(7525):9.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.
- Silberzahn, R. and Uhlman, E. L. (2015). Crowdsourced research: many hands make tight work. *Nature*, 526:189–191.
- Simmons, J. P., Nelson, L. D., 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.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011). Regularization paths for Cox’s proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.
- Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

- Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2015). Specification curve: descriptive and inferential statistics on all reasonable specifications. Available at SSRN, doi: 10.2139/ssrn.2694998.
- Smith, N. C. (1970). Replication studies: a neglected aspect of psychological research. *American Psychologist*, 25(10):970–975.
- Smyth, G. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.
- Stegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5):702–712.
- Steinbach, M., Ertöz, L., and Kumar, V. (2004). The challenges of clustering high dimensional data. In Wille, L. T., editor, *New Directions in Statistical Physics*, pages 273–309. Springer, Berlin, Heidelberg.
- Stern, J., Arslan, R. C., Gerlach, T. M., and Penke, L. (2019). No robust evidence for cycle shifts in preferences for men’s bodies in a multiverse analysis: a response to Gangestad et al. (2019). *PsyArXiv*, pages 1–27. doi: 10.31234/osf.io/pdsuy.
- Szucs, D. and Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3):e2000797.
- Therneau, T. M. (2015). *A Package for Survival Analysis in S*. R package version 2.38.
- Thiese, M. S. (2014). Observational and interventional study design types; an overview. *Biochemia Medica*, 24(2):199–210.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.
- Tibshirani, R. (1997). The Lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.
- Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.
- Tutz, G. (2011). *Regression for Categorical Data*. Cambridge University Press.
- Ulijaszek, S. J. and Kerr, D. A. (1999). Anthropometric measurement error and the assessment of nutritional status. *British Journal of Nutrition*, 82(3):165–177.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., and Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6):632–638.

- Walley, P. (2000). Towards a unified theory of imprecise probability. *International Journal of Approximate Reasoning*, 24(2–3):125–148.
- Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37(5A):2178–2201.
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., and van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Frontiers in Psychology*, 7(1832):1–12.
- Young, C. (2018). Model uncertainty and the crisis in science. *Socius*, 4:1–7.
- Young, C. and Holsteen, K. (2017). Model uncertainty and robustness: a computational framework for multimodel analysis. *Sociological Methods & Research*, 46(1):3–40.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.
- Zadeh, L. A. (1968). Probability measures of fuzzy events. *Journal of Mathematical Analysis and Applications*, 23(2):421–427.
- Zhang, H., Ha, L., Li, Q., and Beer, M. (2017). Imprecise probability analysis of steel structures subject to atmospheric corrosion. *Structural Safety*, 67:62–69.
- Zipf, G., Chiappa, M., Porter, K. S., Ostchega, Y., Lewis, B. G., and Dostal, J. (2013). National health and nutrition examination survey: plan and operations, 1999-2010. *Vital and Health Statistics. Ser. 1, Programs and Collection Procedures*, 1(56):1–37.
- Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.
- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.





## Appendix A

# Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data

### **This chapter is a reprint of:**

Klau, S., Jurinovic, V., Hornung, R., Herold, T., Boulesteix, A.-L. (2018). Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, 19(322):1–14.

### **Copyright:**

BMC Bioinformatics, 2018

### **Author contributions:**

S. Klau developed and implemented priority-Lasso together with A.-L. Boulesteix and performed most of the statistical analyses. The validation of the models was performed by V. Jurinovic. R. Hornung was involved in the implementation of priority-Lasso and initiated the concept of using cross-validated offsets. T. Herold provided the data and was our counterpart for medical questions. All authors were involved in writing the manuscript and read and approved the final version.

### **Acknowledgments:**

The authors thank Jenny Lee for language corrections.

### **Supplementary material available at:**

<https://doi.org/10.1186/s12859-018-2344-6>



METHODOLOGY ARTICLE

Open Access



# Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data

Simon Klau<sup>1\*</sup> , Vindi Jurinovic<sup>1</sup>, Roman Hornung<sup>1</sup>, Tobias Herold<sup>2</sup> and Anne-Laure Boulesteix<sup>1</sup>

## Abstract

**Background:** The inclusion of high-dimensional omics data in prediction models has become a well-studied topic in the last decades. Although most of these methods do not account for possibly different types of variables in the set of covariates available in the same dataset, there are many such scenarios where the variables can be structured in blocks of different types, e.g., clinical, transcriptomic, and methylation data. To date, there exist a few computationally intensive approaches that make use of block structures of this kind.

**Results:** In this paper we present priority-Lasso, an intuitive and practical analysis strategy for building prediction models based on Lasso that takes such block structures into account. It requires the definition of a priority order of blocks of data. Lasso models are calculated successively for every block and the fitted values of every step are included as an offset in the fit of the next step. We apply priority-Lasso in different settings on an acute myeloid leukemia (AML) dataset consisting of clinical variables, cytogenetics, gene mutations and expression variables, and compare its performance on an independent validation dataset to the performance of standard Lasso models.

**Conclusion:** The results show that priority-Lasso is able to keep pace with Lasso in terms of prediction accuracy. Variables of blocks with higher priorities are favored over variables of blocks with lower priority, which results in easily usable and transportable models for clinical practice.

**Keywords:** Cox regression, Lasso, Multi-omics data, Penalized regression, Prediction model, Priority-lasso

## Background

Many cancers are heterogeneous diseases regarding biology, treatment response and outcome. For example, in the context of acute myeloid leukemia (AML), a variety of classifiers and recommendations were published to guide treatment decisions [1]. We and others have recently shown that gene expression markers as well as mutational profiling are able to improve risk prediction based on standard clinical markers [2–5]. Other types of biomarkers such as copy number variation data or methylation data may also be used for this purpose in the future. However, irrespective of the considered specific end point (e.g., overall survival, resistant disease, early death) no model is currently able to precisely predict the outcome

of AML patients. To date, the most powerful prognostic models are based on cytogenetics and gene expression markers [6].

In the present paper, we use the term *omics* to denote molecular biomarkers measured through high-throughput experiments. Beyond the example of AML mentioned above, the integration of multiple types of omics biomarkers with the aim of improved prediction accuracy has been a focus of much attention in the past years, see for example [7] and references therein. While prediction modelling using a single type of omics markers is a well-studied topic, it is not clear how different types of biomarkers should be handled simultaneously when deriving a prediction model.

In addition to the highly important topic of prediction accuracy, encompassing both discrimination ability and calibration, clinical reality requires analysts to take

\*Correspondence: [simonklau@ibe.med.uni-muenchen.de](mailto:simonklau@ibe.med.uni-muenchen.de)

<sup>1</sup>Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Munich, Germany

Full list of author information is available at the end of the article



aspects related to *usability* into account when developing prediction models for clinical practice. Firstly, a model including several hundreds/thousands of variables is much more difficult to implement in clinical practice than a model including only a handful of variables. *Sparcity* is thus an important aspect of the model which contributes to its practical utility in clinical settings. Secondly, a model including variables that are already included in routine diagnostics — such as genetic alterations as recommended by the European LeukemiaNet (ELN) in the case of AML [1], or variables that can be easily assessed such as age or common clinical variables — are more likely to be accepted by physicians than a model including variables measured with new and/or expensive technologies, maybe even at the expense of a slightly lower prediction accuracy. These two points are arguments in favor of models that (preferably) include a small number of variables selected from particular “favorite” sets of variables — as opposed to, say, a large number of variables selected from genome-wide data.

Another aspect related to practical usability is the *transportability* of a prediction model, i.e. the possibility for potential users to apply the prediction model to their own data based on information provided by the model developers [8]. Penalized regression methods yielding sparse models typically yield better transportable models than black-box machine learning algorithms [8, 9]. For example, to apply a Lasso logistic regression model [10] for making predictions for their own patients, users only need the fitted regression coefficients and names of the selected variables to compute the score and, if they want to compute predicted probabilities, the fitted intercept. In contrast, a prediction tool constructed using, for example, the random forest algorithm, can be applied by other researchers or clinicians only if they have access to a software object (such as the output of the R function ‘randomForest’ if the package of the same name is used) or the dataset and the code used to construct it — which may become obsolete after a few years. In this sense, Lasso logistic regression is preferable to random forest as far as transportability and sustainability are concerned. Note that model interpretation is also particularly easy with sparse penalized regression methods.

Finally, coming back to prediction accuracy, we note that medical experts often have some kind of prior knowledge regarding the information content of different sets of variables. For example, they often expect (a particular set of) the clinical variables to have high prediction ability and a large proportion of the gene expression variables to be less relevant. Such prior knowledge should ideally be taken into account while constructing a prediction model.

Motivated by the need, in the context of AML research and other fields, for sparse transportable models selecting preferably variables that are easy to collect or expected to

yield good prediction accuracy, we suggest *priority-Lasso*, a simple Lasso-based approach. Priority-Lasso is a hierarchical regression method which builds prediction rules for patient outcomes (e.g., a time-to-event, a response status or a continuous outcome) from different blocks of variables including high-throughput molecular data while taking clinicians’ preference into account. More precisely, clinicians define “blocks” of variables (which may simply correspond to the type of data, e.g., the block of methylation variables or the block of gene expression variables) and order these blocks according to their level of priority. The prediction model is then fitted in a stepwise manner: In turn, each block of variables is considered as a covariate matrix in Lasso regression, in the sequence of priority specified by the clinician; see the “[Methods](#)” section for more details.

The priority-Lasso procedure is fast and simple. It can cope with all the types of outcome variables accepted by Lasso and, more generally, inherits its properties. The hierarchical principle of priority-Lasso can essentially also be applied to extensions of Lasso, including but not limited to elastic net [11], adaptive Lasso [12] or stability selection [13], but also, more generally, to other prediction methods applicable to high-dimensional covariate data. Last but not least, note that the priority scheme imposed by the clinician merely determines which blocks are prioritized over other blocks with respect to rendering predictive information that is contained in several blocks. Predictive information of blocks with low priority that is not contained in blocks with high priority is still exploited by priority-Lasso (see “[Principles of priority-Lasso](#)” section for details).

The rest of this paper is structured as follows. Section “[Methods](#)” presents the priority-Lasso method and its implementation in detail. In “[Results](#)” section, the method is illustrated with different settings through an application to AML data and compared to standard Lasso in terms of accuracy and included variables. The considered outcome is the survival time and the considered types of data are comprised of clinical data, the mutation status of several genes and gene expression data. Most importantly, prediction models are fitted on a training dataset and subsequently validated on an independent dataset following the recommendations by Royston and Altman [14].

## Methods

We first provide a non-technical introduction into the principles of priority-Lasso in “[Principles of priority-Lasso](#)” section to make these concepts accessible to readers without strong statistical background and to give a succinct overview. We present the method formally in “[Formalization of priority-Lasso](#)” section, treat its implementation in “[R package prioritylasso](#)” section, and describe in “[Validation](#)” section the validation strategy

inspired from Royston and Altman [14] adopted in our illustrative example.

### Principles of priority-Lasso

Priority-Lasso is a method that can construct a prediction model for a clinical outcome of interest (e.g., a time to event or a response status and continuous outcome) based on candidate variables, using an available training dataset. Before running priority-Lasso, the user is required to first specify a block structure for the covariates where each covariate belongs to exactly one of  $M$  blocks and, second, a priority order of these blocks.

A block may be of a particular data type, for example “clinical data”, “gene expression data” or “methylation data”, but the classification of variables into blocks may also be finer. For example, clinical data may be divided into two blocks, e.g., the demographic data (e.g., age or sex) in a first block and clinical data related to the tumor in the second block. Once the blocks of variables are defined, the clinician orders them according to their level of priority. High priority should be given to blocks which are easy and/or inexpensive to collect or are already routinely collected in clinical practice.

After this definition, the prediction model is fitted in a stepwise manner. In the first step, a Lasso model is fitted to the block with highest priority. The goal of this step is simply to explain the largest possible part of the variability in the outcome variable by the covariates from the block with highest priority. In the second step, a Lasso model is fitted to the block with second highest priority using the linear score from the first step as an *offset*, i.e., this linear score is forced into the model with coefficient fixed to 1. In the special case of a metric outcome, this corresponds to fitting a second Lasso model (without the offset) to the residuals from the first Lasso model using the block with second highest priority as covariate matrix. The goal of this second step is thus to use the variables from the second block to explain remaining variability in the outcome variable that could not be explained by covariates from the first block.

In the third step, a Lasso regression is fitted to the block with third highest priority using the linear score from the second step as offset. The special case of a metric outcome is correspondingly equivalent to fitting a Lasso model to the residuals from the second Lasso model using the block with third highest priority. This procedure is iterated until all blocks have been considered in turn. Thus, in the case of a metric outcome, at each step the current block is fitted to the residuals of the previous step. Generalizing to other types of outcome variables, in each step the current block is fitted to the outcome conditional on all blocks with higher priority that were considered in the previous steps. In this way, blocks of variables with low priority enter the model

only if they explain variability that is not explainable by blocks with higher priority. Compared to non-hierarchical approaches, priority-Lasso tends to yield models in which variables from the most prioritized blocks play a more important role.

This procedure was motivated by the fact that there is frequently a strong overlap of predictive information across the considered blocks. For example, some gene expression and gene mutation variables can be associated with the same phenotype, which is why these two different types of omics data may contain similar predictive information. Moreover, clinical covariates and omics covariates often carry similar predictive information. If, in priority-Lasso, a block A is given a higher priority than a block B, this means that the part of the predictive information contained in A and B that is common to both blocks will be obtained from block A. The larger the number of blocks, the lower the information contained in individual blocks, that is not contained in any other block. Thus, in the presence of a large number of blocks there is a high chance that priority-Lasso will exclude variables from blocks of low priority, because the predictive information contained therein may also be contained in the data of blocks of higher priority. Therefore, by providing a priority sequence, the analyst can decide which blocks should be prioritized over others with respect to providing predictive information redundant among blocks. The chosen priority sequence can, however, be expected to have a limited impact on the prediction error for the following reason: If a block A with strong predictive power is attributed a low priority, its predictive power will nevertheless be exploited in the prediction rule. This is because the proportion of the variability of the outcome variable that is only explainable by block A will still be unexplained before block A is considered as a covariate block in the iterative procedure.

### Formalization of priority-Lasso

In the following description, we consider  $M$  blocks of continuous or binary variables that are all to be penalized, and a continuous outcome variable for the sake of simplicity. Extensions to time-to-event and binary outcomes are straightforward using the corresponding variants of Lasso (Cox Lasso and logistic Lasso, respectively, see [15] and [10, 16]). The extension to multicategorical variables is also straightforward using an appropriate coding of the variables.

Let  $x_{ij}$  denote the observed value of the  $j$ th variable ( $j = 1, \dots, p$ ) for the  $i$ th subject ( $i = 1, \dots, n$ ) and  $y_i$  denote the observed outcome of subject  $i$ . For simplicity it is assumed that each variable is centered to have mean zero over the  $n$  observations. The standard Lasso method [10] estimates the regression coefficients  $\beta_1, \dots, \beta_p$  of the  $p$  variables by minimizing the expression

$$\sum_{i=1}^n \left( y_i - \sum_{j=1}^p x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^p |\beta_j|$$

with respect to  $\beta_1, \dots, \beta_p$ , where  $\lambda$  is a so-called penalty parameter. This method performs both regularization (shrinkage of the estimates) and variable selection (i.e., some of the estimates are shrunk to zero, meaning that the variable is excluded from the model). The amount of shrinkage is determined by the parameter  $\lambda$ , which is considered as a tuning parameter of the method and is in practice most often chosen using cross-validation.

We now adapt our notation to the case of variables forming groups that is considered in this paper. From now on, the observations of the  $p_m$  variables from block  $m$  for subject  $i$  are denoted as  $x_{i1}^{(m)}, \dots, x_{ip_m}^{(m)}$ , for  $i = 1, \dots, n$  and  $m = 1, \dots, M$ . The number of blocks  $M$  usually ranges from 2 to, say, 10 in practice, while the number  $p_m$  of variables often varies strongly across the blocks. For example, blocks of clinical variables typically include a very small number of variables, say,  $p_m \approx 10$ , while blocks of molecular variables from high-throughput experiments may include several tens or hundreds of thousands of variables.

Similarly to the definition of  $x_{ij}^{(m)}$ ,  $\beta_j^{(m)}$  denotes the regression coefficient of the  $j$ th variable from block  $m$ , for  $j = 1, \dots, p_m$ , while  $\hat{\beta}_j^{(m)}$  stands for its estimated counterpart.

Let us further denote as  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_M)$  the permutation of  $(1, \dots, M)$  that indicates the priority order:  $\pi_1$  denotes the index of the block with highest priority, while  $\pi_M$  is the index of the block with the lowest priority. For example, if  $M = 4$ ,  $\boldsymbol{\pi} = (3, 1, 4, 2)$  means that the third block has highest priority, the first block has second highest priority, and so on. Conversely, the priority level of a given block is indicated by the position of its index in the vector  $\boldsymbol{\pi}$ .

In the first step of priority-Lasso, the variables from block  $\pi_1$  are used to fit a Lasso regression model. The coefficients  $\beta_1^{(\pi_1)}, \dots, \beta_{p_{\pi_1}}^{(\pi_1)}$  are estimated by minimizing

$$\sum_{i=1}^n \left( y_i - \sum_{j=1}^{p_{\pi_1}} x_{ij}^{(\pi_1)} \beta_j^{(\pi_1)} \right)^2 + \lambda^{(\pi_1)} \sum_{j=1}^{p_{\pi_1}} |\beta_j^{(\pi_1)}|.$$

The linear predictor fitted in step 1 is given as

$$\hat{\eta}_{1,i}(\boldsymbol{\pi}) = \hat{\beta}_1^{(\pi_1)} x_{i1}^{(\pi_1)} + \dots + \hat{\beta}_{p_{\pi_1}}^{(\pi_1)} x_{ip_{\pi_1}}^{(\pi_1)}.$$

In “Principles of priority-Lasso” section we noted that this linear predictor is used as an offset in the second step in which we fit a Lasso model to block  $\pi_2$ . However, the linear score  $\hat{\eta}_{1,i}(\boldsymbol{\pi})$  tends to be over-optimistic with respect to the information usable for predicting  $y_i$  that is contained in block  $\pi_1$ . The reason for the latter is that  $y_i$  was part of the data used for obtaining the estimates

$\hat{\beta}_1^{(\pi_1)}, \dots, \hat{\beta}_{p_{\pi_1}}^{(\pi_1)}$ , which are then used to calculate  $\hat{\eta}_{1,i}(\boldsymbol{\pi})$ . This overoptimism is essentially similar to the well-known overoptimism that results from estimating the prediction error of a prediction rule using the observations in the training dataset. When using this over-optimistic estimate  $\hat{\eta}_{1,i}(\boldsymbol{\pi})$  as an offset in the second step, the influence of block  $\pi_2$  conditional on the influence of block  $\pi_1$  will tend to be underestimated. The reason for this is that by considering the over-optimistic estimate  $\hat{\eta}_{1,i}(\boldsymbol{\pi})$  as an offset, a part of the variability in  $y_i$  is removed that is actually not explainable by block  $\pi_1$  but would possibly be explainable by block  $\pi_2$ . As noted above, this problem results from the fact that  $y_i$  is contained in the training data used for estimating  $\beta_1^{(\pi_1)}, \dots, \beta_{p_{\pi_1}}^{(\pi_1)}$ . As a solution to this problem we suggest estimating the offsets  $\eta_{1,i}(\boldsymbol{\pi})$  using cross-validation in the following way: 1) Split the dataset  $S$  randomly into  $K$  approximately equally sized parts  $S_1, \dots, S_K$ ; 2) For  $k = 1, \dots, K$ : obtain estimates  $\hat{\beta}_{S \setminus S_k, 1}^{(\pi_1)}, \dots, \hat{\beta}_{S \setminus S_k, p_{\pi_1}}^{(\pi_1)}$  of the Lasso coefficients using the training data  $S \setminus S_k$  and for all  $i \in S_k$  ( $k = 1, \dots, K$ ), calculate the cross-validated offsets as

$$\hat{\eta}_{1,i}(\boldsymbol{\pi})_{CV} = \hat{\beta}_{S \setminus S_k, 1}^{(\pi_1)} x_{i1}^{(\pi_1)} + \dots + \hat{\beta}_{S \setminus S_k, p_{\pi_1}}^{(\pi_1)} x_{ip_{\pi_1}}^{(\pi_1)}.$$

In the second step the coefficients of the variables in block  $\pi_2$  are thus estimated by minimizing

$$\sum_{i=1}^n \left( y_i - \hat{\eta}_{1,i}(\boldsymbol{\pi})_{CV} - \sum_{j=1}^{p_{\pi_2}} x_{ij}^{(\pi_2)} \beta_j^{(\pi_2)} \right)^2 + \lambda^{(\pi_2)} \sum_{j=1}^{p_{\pi_2}} |\beta_j^{(\pi_2)}|.$$

Using  $\hat{\eta}_{2,i}(\boldsymbol{\pi}) = \hat{\eta}_{1,i}(\boldsymbol{\pi})_{CV} + \hat{\beta}_1^{(\pi_2)} x_{i1}^{(\pi_2)} + \dots + \hat{\beta}_{p_{\pi_2}}^{(\pi_2)} x_{ip_{\pi_2}}^{(\pi_2)}$  as an offset in the third step in which we fit a Lasso model to block  $\pi_3$  could again lead to underestimating the influence of block  $\pi_3$  conditional on the influences of blocks  $\pi_1$  and  $\pi_2$ . This is because, analogously to the first step, the estimates  $\hat{\beta}_1^{(\pi_2)}, \dots, \hat{\beta}_{p_{\pi_2}}^{(\pi_2)}$  used to calculate  $\hat{\eta}_{2,i}(\boldsymbol{\pi})$  are overly well adapted to the residuals  $y_i - \hat{\eta}_{1,i}(\boldsymbol{\pi})_{CV}$ . Therefore, we again suggest to calculate cross-validated estimates,  $\hat{\eta}_{2,i}(\boldsymbol{\pi})_{CV}$ , of the offsets analogously to the first step.

Priority-Lasso proceeds analogously for the remaining groups until the final ( $M$ th) fit, where the following linear predictor is obtained:

$$\hat{\eta}_{M,i}(\boldsymbol{\pi}) = \sum_{m=1}^M \sum_{j=1}^{p_{\pi_m}} \hat{\beta}_j^{(\pi_m)} x_{ij}^{(\pi_m)}.$$

Note that when the offsets are not estimated by cross-validation but the estimates  $\hat{\eta}_{1,i}(\boldsymbol{\pi}), \dots, \hat{\eta}_{M-1,i}(\boldsymbol{\pi})$  are used, the effects described above of underestimating the conditional influences of the individual blocks accumulate. Thus, the influences of blocks with higher priority are underestimated to a less stronger degree than are blocks with low priority. This could eventually lead to the exclusion of blocks with lower priority that are valuable for



prediction. This is particularly problematic in cases in which low priorities are attributed to blocks with high predictive information. Thus, cross-validated offsets may be used to avoid suboptimal models that may result in cases in which the priority sequence does not attribute high priority to blocks with high predictive power. Note, however, that we are not interested in determining priority sequences that perform optimally from a statistical point of view. Instead, the priority sequence reflects the specific needs of the user, who particularly cares about practicability. Notwithstanding the above mentioned advantages of using cross-validated offsets, we nevertheless also include the version of priority-Lasso without cross-validated offsets in our application study (see “Results” section) for several reasons. Firstly, because the version with cross-validated offsets is more computationally intensive, and thus might not be easily applicable in all situations. Secondly, we aim to illustrate that this version tends to accredit more influence to the blocks with lower priority than does the version without cross-validated offsets. In addition, the suspected tendency of the version without cross-validated offsets to exclude blocks with lower priority might be advantageous in applications in which these blocks contain data types that are expensive to collect or not well established.

### R package *prioritylasso*

The priority-Lasso method (for continuous, binary, and survival outcomes) is implemented in the function ‘prioritylasso’ from our new R package of the same name (version 0.2), which is publicly available from the “Comprehensive R Archive Network” repository. This package uses the implementation of Lasso regression provided by the R package ‘glmnet’ (see [17], and for the special case of Cox-Lasso, see [18]).

The  $M$  penalty parameters  $\lambda^{(\pi_1)}, \dots, \lambda^{(\pi_M)}$  are chosen via cross-validation in the corresponding steps. As in ‘glmnet’, two variants are implemented: The penalty parameter can be chosen either in such a way that the mean cross-validated error is minimal (denoted as ‘lambda.min’), or in such a way that it yields the sparsest model with error within one standard error of the minimum (denoted as ‘lambda.1se’). The latter option yields sparser models. In order to further enforce sparsity at the convenience of the clinician, our package allows to specify a maximum number of non-zero coefficients for each block.

Furthermore, the function ‘prioritylasso’ offers the option to leave the block with highest priority unpenalized (i.e., to set  $\lambda^{(\pi_1)}$  to 0), provided the number of variables  $p_{\pi_1}$  in this group is smaller than the sample size  $n$ . Depending on the outcome, the estimation is then performed via generalized linear regression or via Cox regression [19]. Another variant of the priority-Lasso method is implemented in the function ‘cvm\_prioritylasso’, which makes

it possible to take more than one vector  $\pi$  as the input and choose the best one through minimizing the cross-validation error. This variant is useful in cases where it makes sense to take the group structure into account but the clinician does not feel comfortable assigning clear-cut priorities to each of the groups.

Note that our package solely aims at building prediction models with different types of already prepared omics data available as an  $n \times p$  data matrix. However, generating such multi-omics data matrices from several types of raw data files requires considerable effort. We refer to Bioconductor software packages [20] that allow convenient annotation and organization of multi omics data. As an important example, the ‘MultiAssayExperiment’ data class [21] can be used for data preparation prior to running ‘prioritylasso’.

### Validation

In “Results” section, we apply the priority-Lasso method as well as the classical Lasso to fit prediction models for a time-to-event on a training dataset and subsequently evaluate these models on a validation dataset; see “AML data” section for a description of the data used in this analysis. The present section briefly describes the criteria considered to assess prediction accuracy and the procedures used for validation of the considered models, following the recommendations of Royston and Altman [14]. These authors emphasize in their paper that validation comprises both discrimination and calibration. Hence, we perform both in our analysis and focus on the methods denoted as methods 3, 4, 6, and 7 in their paper.

Firstly, following method 3, we present some measures of discrimination. Instead of Harrell’s C-index, a common measure to quantify the goodness of fit, we show the results of the Uno’s C-index [22], an adapted version of Harrell’s C-index that accounts for censored data and is thus more appropriate in our context. Another useful measure is the integrated Brier score [23] assessing both calibration and discrimination simultaneously, which we calculate over two different time spans: up to two years and up to the time of the last event. To visualize the results, we also show the corresponding prediction error curves obtained using the R package ‘pec’ [24].

Secondly, following method 4 of Royston and Altman [14], we display Kaplan-Meier curves that can be useful for both discrimination and calibration. For each considered prediction model, we define three risk groups, which corresponds to standard practice in the AML context. See for example the newest European Leukemia Net (ELN) genetic risk stratification of AML, which classifies patients into a low-, intermediate-, and a high-risk group [1] and will be referred to as ELN2017 score in the sequel. To build three groups based on a considered score, we choose the two cutpoints that yield the highest logrank statistic in the

training data. We then present the Kaplan-Meier curves of the three risk groups for both training and validation sets. Good separation of the three curves in the validation dataset indicates good discrimination.

These three Kaplan-Meier curves observed for the validation dataset can also be compared to the predicted curves for the three risk groups in the validation dataset (Royston and Altman's method 7). By "predicted curve for a risk group", we mean the average of the individual predicted curves of the patients within this risk group. Good agreement between observed and predicted curves suggests good calibration. Thirdly, as an extension of the graphical check for discrimination, we also examine the hazard ratios across risk groups (Royston and Altman's method 6).

Beyond these methods, we report the AUC, the true positive rate (TPR, also known as sensitivity) and the true negative rate (TNR, also known as specificity) of each score at two years after the diagnosis. This time point was chosen because its ratio of cases to survivors is the closest to 1. The true positive and the true negative rate are calculated with the median of each score as a cutoff for categorizing the scores into two groups. Furthermore, we consider a modified version of Royston and Altman's method 1. They suggest performing a regression with the linear predictor from the model as the only covariate. For a standard Cox model the resulting coefficient is exactly 1 in the training data and should be approximately 1 in the validation data to indicate a good model fit. However, since we perform penalized regression this method is not applicable to our model. Therefore, we modify this criterion in calculating the calibration slopes in both training and validation data. The difference between the slope obtained using the training data and the one obtained using the validation data is a measure for the extent of the overoptimistic assessment of discrimination ability that is obtained using the training data.

## Results

The section starts with a brief description of the AML example dataset ("[AML data](#)" section). Then we present four models fitted using priority-Lasso ("[Results of priority-Lasso](#)" section) and compare them with the current clinical standard model and with two models fitted through standard Lasso (i.e., without taking the block structure into account) in terms of included variables ("[Assessing included variables](#)" section) and performance in the independent validation data ("[Assessing prediction accuracy](#)" section). These models are all fitted with a restricted number of selected variables. The same models without restrictions to the number of variables are presented in Additional file 1 for further comparisons. The complete R code written to perform the analyses is available from Additional file 2.

## AML data

In this study we use two independent datasets, denoted training set and validation set hereafter, including variables belonging to different blocks (see details below). All patients included in the analysis received cytarabine and anthracycline based induction treatment. The training set consists of 447 patients randomized and treated in the multicenter phase III AMLCG-1999 trial (clinicaltrials.gov identifier NCT00266136) between 1999 and 2005 [25, 26]. The patients are part of a previously published gene expression dataset (GSE37642) analyzed with Affymetrix arrays [27]. All patients with a t(15;17) or myelodysplastic syndrome are excluded, as well as patients with missing data.

The validation set consists of all patients with available material treated in the AMLCG-2008 study (NCT01382147) [28], a randomized, multicenter phase III trial ( $n = 210$ ) and additional  $n = 40$  patients that had resistant disease and were treated in the AMLCG-1999 trial. The dataset is publicly available at the Gene Expression Omnibus repository (GSE106291). The detailed inclusion and exclusion criteria were described previously [29]. The patients of the validation set were analyzed by RNAseq. For comparability, all continuous variables are standardized to a mean zero and variance one. All study protocols are in accordance with the Declaration of Helsinki and approved by the institutional review boards of the participating centers. All patients provided written informed consent for inclusion on the clinical trial and genetic analyses.

## Results of priority-Lasso

We apply priority-Lasso on the training dataset ( $n = 447$ , described in "[AML data](#)" section), considering four different scenarios. These scenarios differ in the way the score ELN2017 is included in the analysis and whether or not the offsets are cross-validated (see "[Formalization of priority-Lasso](#)" section). Furthermore, we always apply the 'lambda.min' procedure and 10-fold-cross-validation for the choice of the penalty parameter in each step. However, since prediction performance is not the main concern in our analyses, the 'lambda.1se' approach would also be a reasonable option. In "[Sensitivity analysis](#)" section we show some results with 'lambda.1se' in addition to our main analyses. Furthermore, we allow for a maximum of 10 gene expression variables for each scenario as we want to keep the resulting model as simple as possible and experience has shown that in survival prediction for AML patients only a few gene expression values have a considerable influence on the outcome. Moreover, gene expression values are not easy to implement in clinical routine. We define the following blocks and corresponding priorities:



- Block of priority 1: the score ELN2017 [1]. It can be represented in different ways which are explained in the definition of the scenarios.
- Block of priority 2: 8 clinical variables measured at different scales
- Block of priority 3: 40 binary variables, each of which represents the mutation status for a certain gene
- Block of priority 4: 15809 continuous variables, each of which is the expression value of a certain gene

The order of these blocks have been determined by a physician involved in the project, who has many years of experience in the treatment of patients with AML, as well as experience with AML outcome prediction. These choices are based on practical considerations. However, alternative block orders could be reasonable from other points of view. For example, if the focus is solely on the maximization of prediction performance without any practical constraints, we refer to the function ‘cvm\_prioritylasso’ from our R package ‘prioritylasso’ which chooses the best order of blocks from two or more priority options according to the mean cross-validated performance. In addition to our main analyses that are based on an ordering that takes practical aspects into account as outlined above, we present additional results obtained for other block orders in “[Sensitivity analysis](#)” section.

#### Scenario pl1A

In the first scenario, the block of priority 1 consists of the three-categorical ELN2017 score represented by two dummy variables. We do not penalize this block and do not use cross-validated offsets. In this scenario the selected model includes only 7 variables represented by 8 coefficients: the dummy variables ELN2017\_2 and ELN2017\_3, equaling 1 for the intermediate and the high-risk category, respectively, and 0 otherwise, are selected by definition, because they result from a fit of a standard Cox model without penalization. Moreover, age, the Eastern Cooperative Oncology Group performance status (ECOG) [30], white blood cell count (WBC), lactate dehydrogenase serum level (LDH), hemoglobin level (Hb) and platelet count (PLT) are selected. The selected variables and their coefficients are displayed in the second and third column of Table 1. Variables from blocks with priority 3 (mutation status of 40 genes) and 4 (gene expression) are absent from the model, yielding a particularly sparse model based on variables which are easy to access.

#### Scenario pl1B

This scenario is very similar to pl1A with the difference that the offsets are cross-validated as described in “[Formalization of priority-Lasso](#)” section. Because there are no offsets in the first step of the model fit, the

**Table 1** Variables selected by priority-Lasso in scenarios pl1A and pl1B

Block	Variable	Coef. pl1A	Coef. pl1B
1	ELN2017_2	0.8552	0.8552
	ELN2017_3	1.4324	1.4324
2	Age	0.3540	0.3556
	ECOG (> 1)	0.2794	0.2768
	WBC	0.1029	0.1019
	LDH	0.1744	0.1763
	Hb	0.0529	0.0532
4	PLT	-0.0788	-0.0800
	PHGDH		0.1242
	FAM171B		0.0726
	SH3PXD2B		0.0192
	F12		0.0097
	CD109		0.0599
	FAM92A1		0.0193
	LAPTM4B		0.0079
	FAM24B		0.0378
	DDIT4		0.0424
DOCK1		0.0295	

Column 1: priority of the block the variables are included in. Column 2: variable name. Column 3 and 4: coefficient of the variable in the Cox Lasso model

coefficients of pl1A and pl1B are the same for the block of priority 1 (see Table 1, column 4). For the block of priority 2, the same variables are selected with small differences in their coefficients. While both models do not select variables from the block of priority 3, model pl1B additionally includes 10 gene expression markers—all with only small influence though. Nevertheless, the fact that gene expression markers are included in the model with cross-validated offsets, but not in the model without cross-validated offsets, illustrates the conjecture made in “[Formalization of priority-Lasso](#)” section: When using the priority-Lasso version with cross-validated offsets, more influence tends to be accredited to the blocks with lower priority compared to when using the version without cross-validated offsets.

#### Scenario pl2A

As an alternative approach, considered as sensitivity analysis in the present paper, one may also replace ELN2017 with the 19 variables that are used for its calculation. Because of the far higher number of variables, we penalize this block of priority 1. The results of the scenario without cross-validated offsets (scenario pl2A) are displayed in the third column of Table 2, showing that 14 of these 19 variables are selected. While the selected variables from block 2 are almost the same as in scenario pl1A (except the additional inclusion of sex), now

**Table 2** Variables selected by priority-Lasso in scenarios pl2A and pl2B

Block	Variable	Coef. pl2A	Coef. pl2B	
1	t(8;21)(q22;q22)	-1.0289	-1.0289	
	inv(16)(p13.1q22)	-1.5444	-1.5444	
	NPM1 mut/FLT3-ITD neg or low	-1.0181	-1.0181	
	biCEBPA	-1.2240	-1.2240	
	NPM1 wt/FLT3-ITD pos or low	-0.4358	-0.4358	
	t(9;11)(p21;q23)	0.4635	0.4635	
	Other aberrations	-0.4376	-0.4376	
	KMT2A rearrangements	-0.5440	-0.5440	
	Complex karyotype	0.2970	0.2970	
	Monosomal karyotype	0.0313	0.0313	
	NPM1 wt/FLT3-ITD pos	0.1712	0.1712	
	RUNX1 mutations	0.3065	0.3065	
	ASXL mutations	-0.1224	-0.1224	
	TP53 mutations	0.4306	0.4306	
	2	Age	0.2957	0.2617
		Sex	-0.1011	
ECOG (> 1)		0.3147	0.3206	
WBC		0.0990	0.0589	
LDH		0.1681	0.2371	
Hb		0.0700	0.0671	
PLT		-0.0960	-0.0578	
4	ZBTB37	0.0047	0.0025	
	MF12	0.0090		
	<b>SH3PXD2B</b>	0.0013	0.0418	
	PDK3	-0.0187		
	<b>FAM24B</b>	0.0248		
	SIK3	-0.0063		
	OR7A17	0.0039		
	TBC1D17	-0.0172		
	<b>PHGDH</b>		0.0488	
	<b>FAM171B</b>		0.0134	
	FGD5		0.0359	
	F12		0.0238	
	IRX1		-0.0090	
	FAM92A1		0.0239	
	DDIT4		0.0769	
HSPA2		0.0169		

Column 1: priority of the block the variable is included in. Column 2: variable name. Column 3 and 4: coefficient of the variable in the Cox Lasso model. Variables from the block of priority 4 also appearing in Table 1 are marked in bold

there are 8 gene expression variables selected from the block of priority 4. We can see that these gene expression variables are not necessarily the same as in scenario pl1B.

### Scenario pl2B

Analogously to scenarios pl1A and pl1B, scenario pl2B is the same as pl2A, except that the offsets are calculated with cross-validation. Column 4 of Table 2 contains the results from this model, showing only small differences in the block of priority 2, but again large differences in the selected gene expression markers.

### Assessing included variables

For assessing the fitted models with respect to the selected variables, we consider as a reference two standard Lasso models fitted to the training data using the whole set of variables without taking any block structure into account. The two models differ in the way ELN2017 is treated. In the first Lasso model (variant 'Lasso1') it is considered as the score represented by two dummy variables. In the second Lasso model it is represented by the 19 variables which are used for its definition (variant 'Lasso2'). In order to allow for a fair comparison, we again use the 'lambda.min' procedure and 10-fold-cross-validation to choose the penalty  $\lambda$ . Moreover, we allow the selection of a maximum number of variables equal to the number of all variables in blocks 1-3 for priority-Lasso plus 10. This corresponds to the fact that we did not restrict the number of variables of blocks 1-3 for priority-Lasso, but set the maximum number of gene expression variables to 10. The resulting models (not shown) clearly select more variables than the models obtained with priority-Lasso. Especially the number of gene expression variables is much higher (43 for Lasso1 and 52 for Lasso2), whereas only age for both models and ELN2017\_3 for Lasso1 are selected variables from other types of data. Hence, priority-Lasso favors variables from blocks with high priority compared to standard Lasso and yields models that include considerably less variables.

### Assessing prediction accuracy

In order to compare the different approaches we follow the procedures described in "Validation" section – the results are shown in Table 3. It can be seen that pl1A and pl1B reach the highest sensitivity among the scenarios (0.672), whereas especially the raw ELN2017 score is associated with a far lower value (0.556). In contrast, the specificity is 0.723 for ELN2017, whereas all other scenarios are associated with a specificity between 0.64 and 0.67. However, these results represent only one of many possible time points and cutoffs, so their use is doubtful in our context. The other measures – the AUC, the C-indices, and the integrated Brier score – do not show great differences across the scenarios either. Only ELN2017 is an exception with considerably poorer results. For the AUC, pl1B yields the best result with a value of 0.731, but scenarios pl2B, Lasso1 and Lasso2 are not far worse. For  $C_{Uno}$ , the highest value is 0.664, which is reached by pl2B. The

**Table 3** Validation results for the model scenarios with restrictions to the number of selected variables

	pl1A	pl1B	Lasso1	pl2A	pl2B	Lasso2	ELN2017
TPR	0.672	0.672	0.651	0.640	0.658	0.643	0.556
TNR	0.667	0.658	0.661	0.647	0.664	0.653	0.723
AUC	0.711	0.731	0.726	0.713	0.727	0.725	0.663
$C_{Uno}$	0.653	0.660	0.658	0.658	0.664	0.656	0.619
$IBS_2$	0.175	0.172	0.176	0.175	0.172	0.177	0.181
$IBS_{4.4}$	0.197	0.192	0.191	0.197	0.191	0.193	0.204
Optimism	0.393	0.289	0.920	0.377	0.243	0.984	
$CI_{lower}^L$	0.339	0.304	0.247	0.387	0.327	0.177	0.418
$HR^L$	0.536	0.455	0.363	0.605	0.566	0.286	0.669
$CI_{upper}^L$	0.849	0.652	0.535	0.946	0.981	0.461	1.074
$CI_{lower}^H$	1.175	1.098	0.948	1.515	1.534	0.974	1.314
$HR^H$	1.751	1.651	1.385	2.208	2.199	1.386	1.954
$CI_{upper}^H$	2.612	2.483	2.022	3.216	3.151	1.972	2.907
$p\text{-value}_{LR}$	1.11e-08	1.05e-8	2.22e-10	1.07e-08	1.74e-08	4.99e-11	1.36e-07

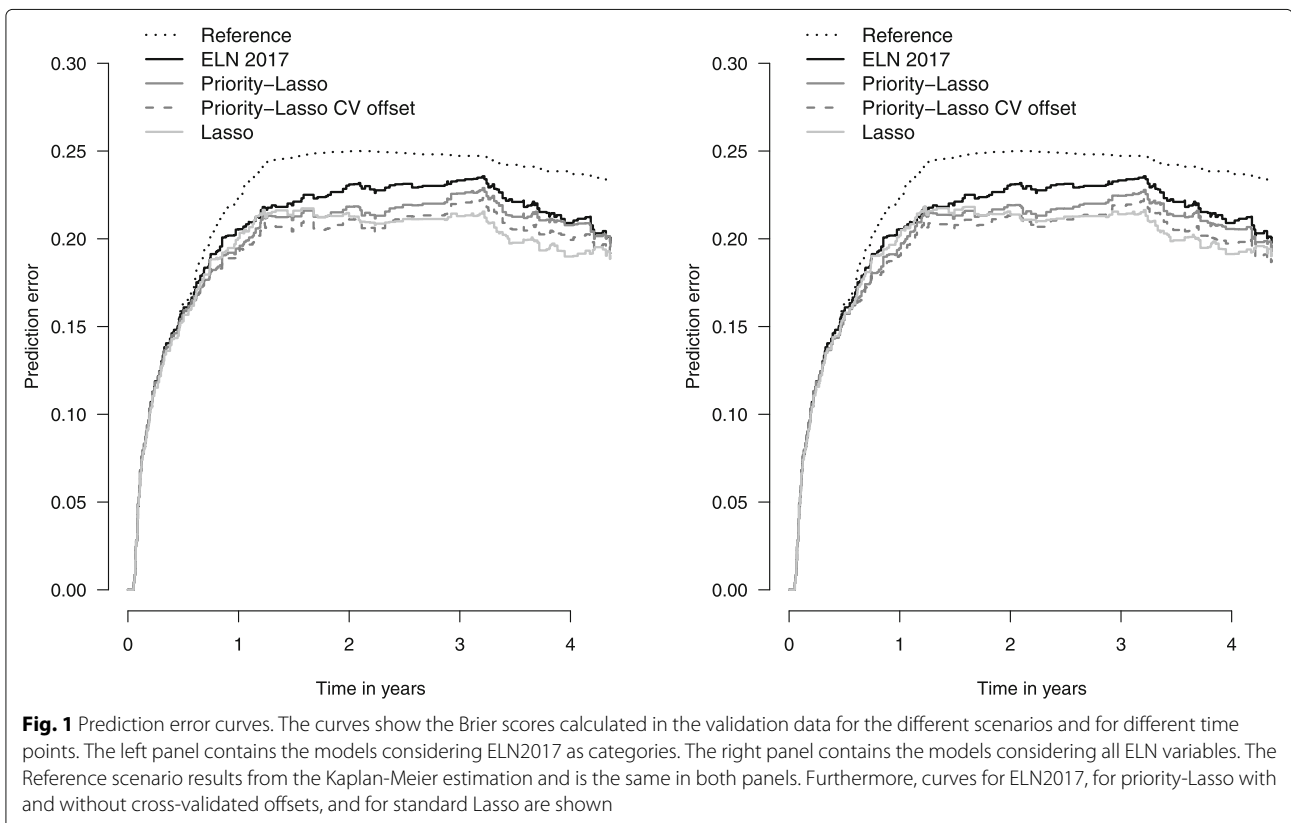
The acronyms in the first column are: TPR: True positive rate; TNR: True negative rate; AUC: Area under the curve,  $C_{Uno}$ : Uno’s C-index,  $IBS_2$ : Integrated Brier score up to 2 years,  $IBS_{4.4}$ : Integrated Brier score up to 4.4 years, Optimism: difference between calibration slopes of training and validation data,  $CI_{lower}^L$ : lower bound of the 95% confidence interval for the hazard ratio of the low risk group,  $HR^L$ : hazard ratio of the low risk group,  $CI_{upper}^L$ : upper bound of the 95% confidence interval for the hazard ratio of the low risk group,  $CI_{lower}^H$ : lower bound of the 95% confidence interval for the hazard ratio of the high risk group,  $HR^H$ : hazard ratio of the high risk group,  $CI_{upper}^H$ : upper bound of the 95% confidence interval for the hazard ratio of the high risk group,  $p\text{-value}$ :  $p$ -value of the likelihood ratio test

integrated Brier score is calculated over two different time spans (up to 2 years and up to 4.4 years, the latter being the time to the last event). After two years, the priority-Lasso fit with cross-validated offsets is better than the other models – no matter how ELN2017 is treated. Over the whole time period, Lasso1 and pl2B give the lowest IBS, followed by Lasso2, indicating a lower prediction error for the Lasso models in the second half of the whole time period. This can also be observed in Fig. 1. Scenarios pl1B and pl2B perform best in the first two years but they are outperformed by Lasso afterwards. As expected, priority-Lasso with cross-validated offsets is always better than without. All fitted models are associated with a much lower prediction error than ELN2017 alone. The results from the prediction error curves do not differ substantially between the two panels of Fig. 1, that is, they are robust with regard to the handling of ELN2017.

The Kaplan-Meier curves for training and validation data are shown in Fig. 2. The discrimination by Lasso is obviously very good in the training data, but worse in the validation data. Especially the difference in survival between intermediate and high risk is not very clear. For both representations of ELN2017, the priority-Lasso models with and without cross-validated offsets feature a similar discrimination, where, however, the results obtained using the version with cross-validated offsets are slightly better. For the scenario with all ELN2017 variables, the priority-Lasso models give the best results in the validation data among all scenarios. In contrast, ELN2017 discriminates less well between the three risk groups. The

results concerning Lasso indicate systematic overfitting in the training data. This is consistent with the results seen in “Assessing included variables” section where Lasso included much more variables than the other methods. It can also be seen from the row ‘optimism’ of Table 3. The difference of the slopes between training and validation data is the largest for the Lasso models, indicating that this method is associated with the highest overoptimism.

A possible way of quantifying the results seen in Fig. 2 is to consider the hazard ratios across risk groups in the validation set as shown in the lower half of Table 3. The intermediate group serves as a baseline here. The result of the likelihood ratio test is significant for all models. The discrimination between low and intermediate group is worst for the ELN2017 score. As already seen in Fig. 2, the discrimination between the low and intermediate group is better for Lasso than priority-Lasso. In contrast, priority-Lasso has a higher hazard ratio for the high risk group, in particular when using all ELN variables. These observations are also consistent with the results shown in Fig. 1, where the prediction was better for priority-Lasso than for Lasso in the earlier years, but worse in the later years. This corresponds to better prediction for shorter survival times and worse prediction for longer survival times, respectively. The fact that ELN2017 is included in the results of priority-Lasso, but not standard Lasso except ELN2017\_3 in Lasso1, also seems to play a role for this issue. Both Fig. 2 and the hazard ratios clearly show that the prediction is better for high risk groups than for low risk groups with the raw ELN2017 score.



Finally, we present the Kaplan-Meier curves for calibration in Fig. 3. For all the scenarios there are groups that reveal some miscalibration. For the Lasso models, especially the high risk groups differ between predicted and observed validation curves. The scenarios pl2A and pl2B show more differences between predictions and observations in the low risk groups than the other scenarios—the same fact applies to pl1A and pl1B in the intermediate risk group.

### Sensitivity analysis

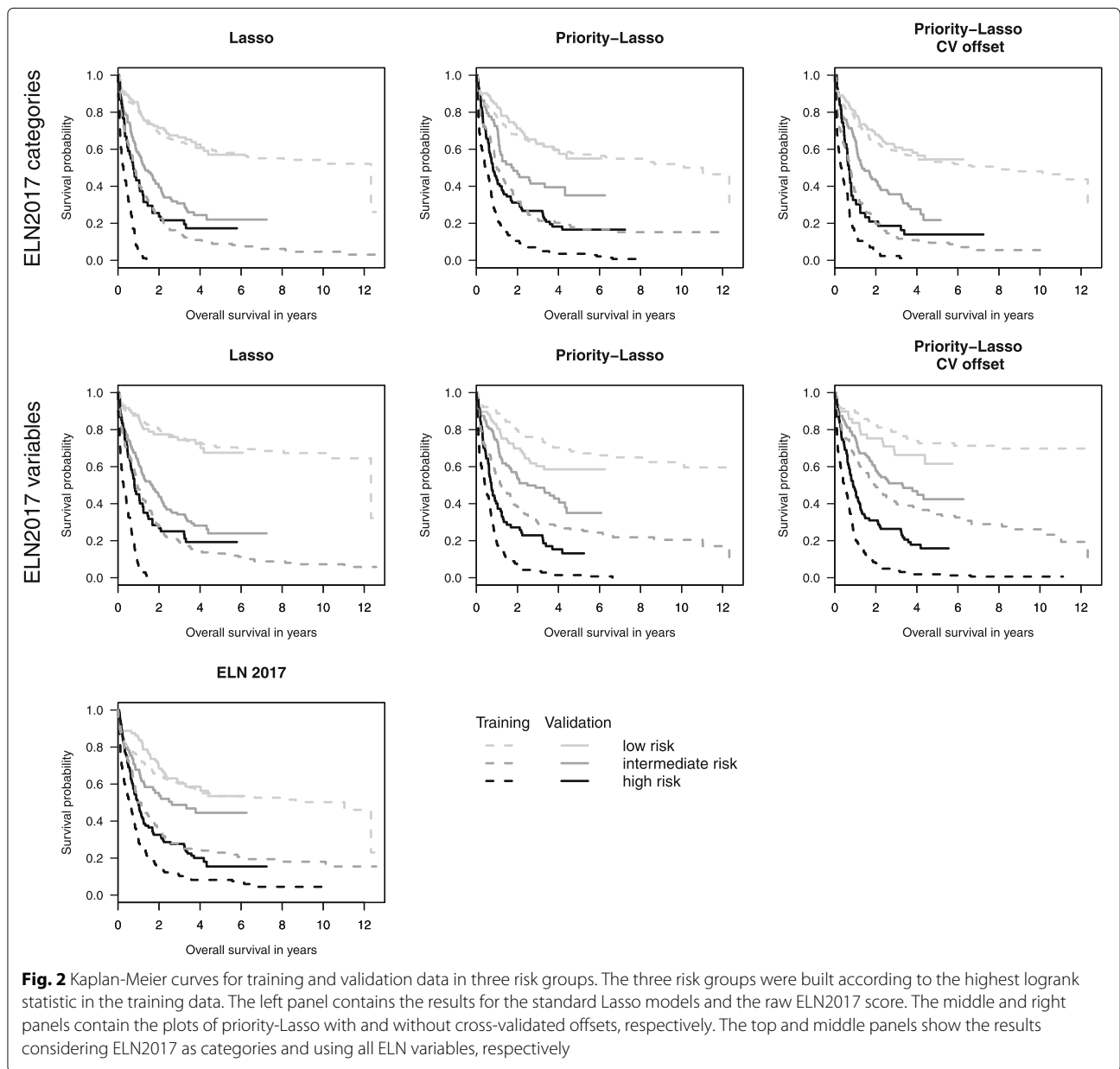
In order to investigate the influence of different block orders on the selected variables, we run the four different scenarios of priority-Lasso with every possible block order (data not shown). The results show that the block order can have substantial influence on the number of selected variables. For the scenarios pl1A and pl1B, sparsest models are obtained with our priority definition, illustrating that priority-Lasso takes advantage of prior knowledge. Higher numbers of variables are obtained for other block orders with maximum values of 45 (pl2A,  $\pi = (4, 3, 1, 2)$  and  $\pi = (4, 3, 2, 1)$ ). Seven of the eight selected variables in pl1A are chosen for almost every scenario of priority-Lasso and block orders, demonstrating their importance even in blocks of low priority. Remarkably, only a small part of them are found in the standard Lasso models (age

in Lasso1 and Lasso2, as well as ELN2017\_3 in Lasso1). It can be further observed that many of the selected gene expression variables are selected for only a small fraction of models.

In additional sensitivity analyses we consider the four scenarios with the 'lambda.1se' setting in order to choose the  $M$  values  $\lambda^{(\pi_1)}, \dots, \lambda^{(\pi_M)}$  as discussed in "R package prioritylasso" section. As expected, the 'lambda.1se' setting leads to a smaller number of selected variables for all scenarios. In total, the number of variables is 4, 10, and 15 for priority-Lasso with ELN categories, priority-Lasso with ELN variables (both with and without cross-validated offsets), and Lasso, respectively. The four different priority-Lasso models solely select variables from blocks 1 and 2. On the other hand, apart from age, Lasso selects only gene expression variables.

### Discussion

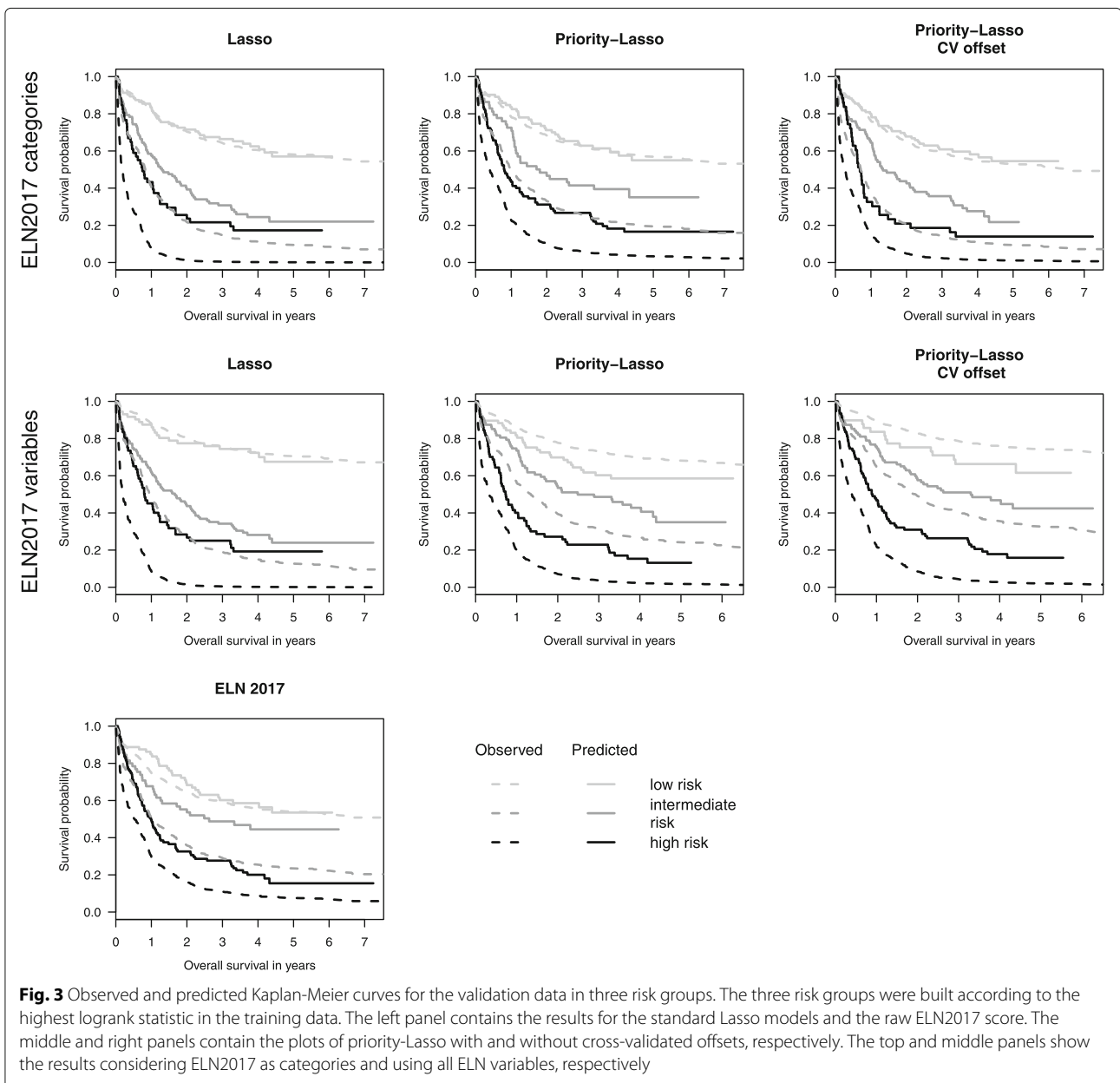
We introduced priority-Lasso, a simple Lasso-based intuitive procedure for patient outcome modelling based on blocks of multiple omics data that incorporates practical constraints and/or prior knowledge on the relevance of the blocks. The procedure essentially inherits most properties of Lasso. Its basic principle is however not limited to Lasso and could be easily adapted to recently developed variants of penalized regression.



An important feature of priority-Lasso is that it directly addresses the problem of redundancies in the predictive information across different blocks: Predictive information contained in the data from specific blocks is incorporated only if it is not contained in data from blocks of higher priority. To date, this idea seems to have been considered only in the TANDEM approach [31], that is, however, restricted to the case of two blocks.

In our illustrative example from leukemia research priority-Lasso was able to reach better prediction accuracy than Lasso. This applies especially to the version of priority-Lasso with cross-validated offsets, however, at

the cost of more computation time and more selected variables than without cross-validated offsets. But even without cross-validated offsets, the models are not substantially worse than Lasso as far as accuracy is concerned. Moreover, they offer considerable advantages in terms of increased sparsity and composition of the models: they include less variables that are currently not included in the recommended diagnostic workup at initial diagnosis, which is an advantage from a practical perspective. Priority-Lasso offers more flexibility than Lasso: it allows the user to define block structures, where for each block a maximum number of selected variables can be specified.



The obtained models can be seen as compromises between “what the data tells us” and what is more realistic and easy to implement in clinical routine. As an extreme variant of priority-Lasso, one could imagine the case of a practitioner fixing the ordering of the variables completely, which amounts to considering blocks of size 1 (each variable forms one block). The other extreme consists of ignoring the block structure and simply fitting a model using Lasso to all variables. The finer the block structure, the less data-driven is the model selection. The number of blocks also influences the maximum possible number of selected variables in the final model. Since a maximum of  $n$  variables can be selected in a Lasso

regression, a selection of  $n$  variables is the maximum for every block in priority-Lasso – hence the maximum possible number of variables selected by priority-Lasso depends on the number of blocks.

Unlike with Bayesian methods, prior knowledge is taken into account only through the definition and ordering of blocks. This feature makes the method less flexible, but also easy to use and interpret for scientists without strong background in statistics. The user does not have to perform any complicated choices in order to apply the method: The first choice to be made is whether or not the offset should be cross-validated – the variant without cross-validation gives more weight to blocks



with high priority, but is prone to overfitting. Moreover, the user may decide to leave the block with highest priority unpenalized in case it satisfies  $p_{\pi_1} < n$ . By default it is treated like the other blocks of data and is thus penalized. As for all penalized regression methods, one can choose the procedure used for optimizing  $\lambda$  (in 'glmnet':  $\lambda_{min}$  or  $\lambda_{1se}$ ), which amounts to deciding between a more complex model with potentially slightly better accuracy and a sparser model. The default is  $\lambda_{min}$ , that is, the  $\lambda$  associated with the minimum cross-validation error in each step. Of course there are additional parameters like the number of folds in the cross-validation procedures that could be modified as well, but are not expected to strongly affect the results.

Note that when working with multi-omics data other, more technical analysis steps are required before building prediction models. The package 'prioritylasso' itself was designed solely to build prediction models and takes the already formatted multi-omics data matrix as input. Fortunately, there are other tools available in Bioconductor that are of great value for the purpose of preparing multi-omics data. For example, the 'MultiAssayExperiment' software package [21] provides useful functions to represent, store, and operate on multi-omics data. It builds a bridge from standard R to Bioconductor and its classes for data representation that cannot be ignored in the context of omics data.

Finally, priority-Lasso offers further practical advantages for clinical practice. Suppose there are (blocks of) variables available only for a subset of patients and missing for the other. A potential approach to efficiently handle such data consists of assigning them a low priority in priority-Lasso. In this way, one can first fit a "basic" model to the blocks that are available for all patients, using all patients. This basic model can then be complemented by variables from the low priority blocks that are missing for a subset of the patients. Importantly, this is also relevant for prediction: Blocks which are not available for all patients in the training data will not be frequently available for new data for the purpose of prediction. In such cases, the basic prediction model can be used to obtain predictions.

## Conclusion

Our results show that priority-Lasso is a flexible and user-friendly prediction method that can reach a similar or even better prediction accuracy compared to standard Lasso. The feature which favors variables of blocks with higher priorities over variables of blocks with lower priority offers a practical advantage and makes the resulting prediction rules easy to use and interpret.

## Additional files

**Additional file 1:** Results of the analyses without restrictions to the maximum number of selected variables. (PDF 215 kb)

**Additional file 2:** R code written to perform the analyses. (ZIP 15 kb)

## Abbreviations

AML: Acute myeloid leukemia; AUC: Area under the curve; C-index: Concordance index; ECOG: Eastern cooperative oncology group; ELN: European leukemiaNet; Hb: Hemoglobin level; IBS: Integrated brier score; LDH: Lactate dehydrogenase serum level; PLT: Platelet count; RNAseq: Ribonucleic acid sequencing; TNR: True negative rate; TPR: True positive rate; WBC: White blood cell count

## Acknowledgements

The authors thank Jenny Lee for language corrections. A small part of this work has been presented orally at the Workshop on Computational Models in Biology and Medicine on the 2nd-3rd March, 2017 at the University of Veterinary Medicine Hannover, and at the 64th Biometrical Colloquium on the 25th-28th March, 2018 at the Goethe University Frankfurt.

## Funding

This project was funded by the Sander Foundation (grant 2014.159.1 to ALB and TH) and by the DFG (grant BO3139/4-2 to ALB). The funding body did not play any role in the design of the study, in collection, analysis, and interpretation of data and in writing the manuscript.

## Availability of data and materials

The datasets used for the analyses are publicly available at the Gene Expression Omnibus (GSE37642 and GSE106291 for the training and validation data, respectively). All R code written to perform the analyses is available from Additional file 2.

## Authors' contributions

SK developed priority-Lasso together with ALB and performed much of the statistical analyses. The validation of the models was performed by VJ. RH was significantly involved in the implementation of priority-Lasso and initiated the concept of using cross-validated offsets. TH provided the data and was our counterpart for medical questions. All authors were involved in writing the manuscript and read and approved the final version.

## Ethics approval and consent to participate

Not applicable.

## Consent for publication

Not applicable.

## Competing interests

The authors declare that they have no competing interests.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Author details

<sup>1</sup>Institute for Medical Information Processing, Biometry and Epidemiology, University of Munich, Munich, Germany. <sup>2</sup>Department of Internal Medicine III, University of Munich, Munich, Germany.

Received: 19 February 2018 Accepted: 29 August 2018

Published online: 12 September 2018

## References

- Döhner H, Estey E, Grimwade D, Amadori S, Appelbaum FR, Büchner T, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*. 2016;129(4):424–47.
- Li Z, Herold T, He C, Valk PJ, Chen P, Jurinovic V, et al. Identification of a 24-Gene Prognostic Signature That Improves the European LeukemiaNet Risk Classification of Acute Myeloid Leukemia: An International Collaborative Study. *J Clin Oncol*. 2013;31(9):1172–81.

3. Ng SW, Mitchell A, Kennedy JA, Chen WC, McLeod J, Ibrahimova N, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*. 2016;540(7633):433–7.
4. Pastore F, Dufour A, Benthous T, Metzeler KH, Maharry KS, Schneider S, et al. Combined Molecular and Clinical Prognostic Index for Relapse and Survival in Cytogenetically Normal Acute Myeloid Leukemia. *J Clin Oncol*. 2014;32(15):1586–94.
5. Walter RB, Othous M, Burnett AK, Löwenberg B, Kantarjian HM, Ossenkoppele GJ, et al. Resistance prediction in AML: analysis of 4601 patients from MRC/NCRI, HOVON/SAKK, SWOG, and MD Anderson Cancer Center. *Leukemia*. 2015;29(2):312–20.
6. Wang M, Lindberg J, Klevebring D, Nilsson C, Mer A, Rantalainen M, et al. Validation of risk stratification models in acute myeloid leukemia using sequencing-based molecular profiling. *Leukemia*. 2017;31(10):2029–36.
7. Boulesteix AL, De Bin R, Jiang X, Fuchs M. IPF-LASSO: Integrative-Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Comput Math Meth Med*. 2017;1–14.
8. Boulesteix AL, Schmid M. Machine learning versus statistical modeling. *Biom J*. 2014;56(4):588–93.
9. Boulesteix AL, Janitza S, Hornung R, Probst P, Busen H, Hapfelmeier A. Making complex prediction rules applicable for readers: Current practice in random forest literature and recommendations. *Biom J*. 2018;1–14.
10. Tibshirani R. Regression shrinkage and selection via the Lasso. *J R Stat Soc Ser B Methodol*. 1996;58:267–88.
11. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B Stat Methodol*. 2005;67(2):301–20.
12. Zou H. The adaptive Lasso and its oracle properties. *J Am Stat Assoc*. 2006;101(476):1418–29.
13. Meinshausen N, Bühlmann P. Stability selection. *J R Stat Soc Ser B Stat Methodol*. 2010;72(4):417–73.
14. Royston P, Altman DG. External validation of a Cox prognostic model: principles and methods. *BMC Med Res Methodol*. 2013;13(1):33.
15. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16(4):385–95.
16. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. *Biostatistics*. 2004;5(3):427–43.
17. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1–22.
18. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *J Stat Softw*. 2011;39(5):1–13.
19. Cox DR. Regression Models and Life-Tables. *J R Stat Soc Ser B Methodol*. 1972;34(2):187–220.
20. Huber W, Carey JV, Gentleman R, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods*. 2015;12(2):115–21.
21. Ramos M, Schiffer L, Re A, Azhar R, Basunia A, Rodriguez C, et al. Software for the Integration of Multiomics Experiments in Bioconductor. *Cancer Res*. 2017;77(21):e39–42.
22. Uno H, Cai T, Pencina MJ, D'Agostino RB, Wei L. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Stat Med*. 2011;30(10):1105–17.
23. Graf E, Schmoor C, Sauerbrei W, Schumacher M. Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*. 1999;18(17-18):2529–45.
24. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. *J Stat Softw*. 2012;50(11):1–23.
25. Büchner T, Krug U, Gale RP, Heinecke A, Sauerland M, Haferlach C, et al. Age, not therapy intensity, determines outcomes of adults with acute myeloid leukemia. *Leukemia*. 2016;30(8):1781–4.
26. Büchner T, Berdel WE, Schoch C, Haferlach T, Serve HL, Kienast J, et al. Double induction containing either two courses or one course of high-dose cytarabine plus mitoxantrone and postremission therapy by either autologous stem-cell transplantation or by prolonged maintenance for acute myeloid leukemia. *J Clin Oncol*. 2006;24(16):2480–9.
27. Herold T, Metzeler KH, Vosberg S, Hartmann L, Röllig C, Stölzel F, et al. Isolated trisomy 13 defines a homogeneous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. *Blood*. 2014;124(8):1304–11.
28. Kreuzer KA, Spiekermann K, Lindemann HW, Lengfelder E, Graeven U, Staib P, et al. High efficacy and significantly shortened neutropenia of dose-dense S-HAM as compared to standard double induction: first results of a prospective randomized trial (AML-CG 2008). *Blood*. 2013;122(21):619.
29. Herold T, Jurinovic V, Batcha AMN, Bamopoulos SA, Rothenberg-Thurley M, Ksienzyk B, et al. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia: *Haematologica*; 2017. <https://doi.org/10.3324/haematol.2017.178442>.
30. Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol*. 1982;5(6):649–55.
31. Aben N, Vis DJ, Michaut M, Wessels LFA. TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types. *Bioinformatics*. 2016;32(17):i413–20.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)





## Appendix B

# Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection

### This chapter is a reprint of:

Klau, S., Martin-Magniette, M.-L., Boulesteix\*, A.-L., and Hoffmann\*, S. (2020). Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. *Biometrical Journal*, 62(3):670–687. doi: 10.1002/bimj.201800309

### Copyright:

Biometrical Journal, 2020

### Author contributions:

S. Klau, A.-L. Boulesteix and S. Hoffmann developed the method. S. Klau designed and conducted the study. M.-L. Martin-Magniette contributed substantially to issues of the differential expression analysis. All authors were involved in writing the manuscript, and read and approved the final version.

\*A.-L. Boulesteix and S. Hoffmann contributed equally to this work.

### Acknowledgments:

This work was funded by the DFG (individual grants BO3139/4-2, BO3139/2-3 and BO3139/6-1 to ALB) and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for its content. The authors thank Jenny Lee and Alethea Charlton for language corrections.



### Supplementary material available at:

<https://doi.org/10.1002/bimj.201800309>



## RESEARCH PAPER

# Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection

Simon Klau<sup>1</sup>  | Marie-Laure Martin-Magniette<sup>2,3,4</sup> | Anne-Laure Boulesteix<sup>1</sup>  | Sabine Hoffmann<sup>1</sup>

<sup>1</sup>Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Munich, Germany

<sup>2</sup>Institute of Plant Sciences Paris Saclay IPS2, CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Orsay, France

<sup>3</sup>Institute of Plant Sciences Paris-Saclay IPS2, Paris Diderot, Sorbonne Paris-Cité, Orsay, France

<sup>4</sup>UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, Paris, France

## Correspondence

Simon Klau, Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Marchioninistraße 15, 81377 Munich, Germany.  
Email: simonklau@ibe.med.uni-muenchen.de

## Funding information

Deutsche Forschungsgemeinschaft, Grant/Award Numbers: BO3139/2-3, BO3139/4-2, BO3139/6-1; Bundesministerium für Bildung und Forschung, Grant/Award Number: 01IS18036A

Anne-Laure Boulesteix and Sabine Hoffmann contributed equally to this work.

## Abstract

Uncertainty is a crucial issue in statistics which can be considered from different points of view. One type of uncertainty, typically referred to as sampling uncertainty, arises through the variability of results obtained when the same analysis strategy is applied to different samples. Another type of uncertainty arises through the variability of results obtained when using the same sample but different analysis strategies addressing the same research question. We denote this latter type of uncertainty as method uncertainty. It results from all the choices to be made for an analysis, for example, decisions related to data preparation, method choice, or model selection. In medical sciences, a large part of omics research is focused on the identification of molecular biomarkers, which can either be performed through ranking or by selection from among a large number of candidates. In this paper, we introduce a general resampling-based framework to quantify and compare sampling and method uncertainty. For illustration, we apply this framework to different scenarios related to the selection and ranking of omics biomarkers in the context of acute myeloid leukemia: variable selection in multivariable regression using different types of omics markers, the ranking of biomarkers according to their predictive performance, and the identification of differentially expressed genes from RNA-seq data. For all three scenarios, our findings suggest highly unstable results when the same analysis strategy is applied to two independent samples, indicating high sampling uncertainty and a comparatively smaller, but non-negligible method uncertainty, which strongly depends on the methods being compared.

## KEYWORDS

high-dimensional data, resampling, stability, variable ranking, variable selection

## 1 | INTRODUCTION

Statistical results are variable. On the one hand, results are affected by *sampling uncertainty*: if we draw different samples from a considered distribution, we obtain different results on each of these samples. Most researchers are familiar with this kind of variability and classical probability models account for it.

On the other hand, one also obtains different results when applying different analysis strategies to address the same research question. These possible analysis strategies result from the combination of every choice that has to be made for an analysis,

for instance, which statistical model to adopt, how to do data preparation and cleaning, how to choose tuning parameters, or how to perform statistical tests. These choices, which have received increasing attention in recent years, are also referred to as researcher degrees of freedom (Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016). As there are usually a great number of possible analysis strategies which can be justified both from a theoretical perspective and from a substantive point of view, these researcher degrees of freedom may give rise to a considerable variability in statistical results. We denote this latter type of variability related to the possibility of different analysis strategies as *method uncertainty*. The term “method” should be understood in a broad sense, including not only the statistical testing or modeling approach but also, for instance, data preparation procedures or parameter values.

It has long been recognized that the results of studies on high-dimensional data can be highly unstable (Michiels, Koscielny, & Hill, 2005; Sauerbrei, Boulesteix, & Binder, 2011). In particular, in the context of variable selection, which is an important task in omics research, there is often very little overlap between biomarkers that are identified by different research groups (Ein-Dor, Kela, Getz, Givol, & Domany, 2005; Ein-Dor, Zuk, & Domany, 2006). On the one hand, this finding can be explained by the fact that in a so-called “ $n \ll p$ ” setting, where there are many more variables than observations, there is a great risk of overfitting (Deroncourt, Hanczar, & Zucker, 2014; Schumacher, Binder, & Gerds, 2007), resulting in very high sampling uncertainty. The problem of sampling uncertainty can be, for instance, addressed by combining selection algorithms with subsampling in stability selection (Meinshausen & Bühlmann, 2010; Shah & Samworth, 2013). On the other hand, method uncertainty can also be expected to be high due to the multitude of methods and the lack of guidance and standards in the omics field (Boulesteix, Hornung, & Sauerbrei, 2017b).

Unlike sampling uncertainty, researchers are often not familiar with method uncertainty and perceive it as disquieting and bothersome. Indeed, method uncertainty may question the validity of the presented results as it is common to base these results on only one among many possible analysis strategies. In order to illustrate this problem, Silberzahn and Uhlman (2015) performed an experiment in giving a data set to 29 independent teams of researchers with strong statistical background with the task of answering the same research question (“are football (soccer) referees more likely to give red cards to players with dark skin than to players with light skin?”). Due to different analysis strategies, the researchers obtained highly varied results, thereby illustrating method uncertainty in this context. This strategy, termed “crowdsourcing,” by Silberzahn and Uhlman (2015) that consists of having the analyses done by several teams can be seen as a possible approach to handle method uncertainty. However, there is no general consent about the best way to do so and several approaches focusing on different types of method uncertainty have been proposed recently. In the multiverse analysis framework (Steenen, Tuerlinckx, Gelman, & Vanpaemel, 2016), for example, the authors consider different ways of preprocessing the data and report the results for all considered choices. Patel, Burford, and Ioannidis (2015), on the other hand, investigate the effect of different combinations of adjustment variables to be included in a multivariable Cox regression and propose simple summary measures to quantify the resulting variability. Finally, Simonsohn, Simmons, and Nelson (2015) propose a specification curve analysis, which does not only allow the visualization of the results obtained with different analysis strategies, but additionally provides a joint permutation-based procedure to test a specific null-hypothesis while accounting for method uncertainty.

In the bioinformatics literature, many studies address the comparison of top lists of biomarkers obtained using different ranking criteria; see, for example, Boulesteix and Slawski (2009) for an early review and Lausser, Müssel, Maucher, and Kestler (2013), Dessì, Pascariello, and Pes (2013) and references therein for illustrations based on high-dimensional gene expression data. Consensus biomarker selection integrating the results of several ranking approaches into a single ranking was proposed in its first variants more than ten years ago (Boulesteix & Slawski, 2009; Dutkowski & Gambin, 2007). It can be seen as the bioinformatics counterpart of the approaches discussed in the previous paragraph as far as the choice of the ranking criterion—a specific type of method uncertainty—is concerned.

While all these approaches are arguably an important step toward making research more transparent, they solely address method uncertainty. Sampling uncertainty is briefly considered along the way in the article by Dessì et al. (2013) mentioned above addressing the comparison between ranking criteria. However, these procedures do not allow (and do not aim at) comparing sampling and method uncertainty. We claim that such a comparison is important in general, particularly in the context of high-dimensional data. Indeed, high-dimensional data analyses are known to be particularly affected both by sampling uncertainty (because of the  $n \ll p$  issue) and method uncertainty (because of the increased researcher degrees of freedom); see Boulesteix et al. (2017b). A better understanding of the sources and extent of uncertainties is desirable both for methodological researchers developing methods (to help them focus their attention on the most critical problems) and for applied scientists applying them in biomedical research projects (to support them in their interpretation of the results).

In this paper, we propose a general resampling-based framework to quantify and compare sampling and method uncertainty and apply it in the context of the ranking and selection of omics markers out of a large number of candidates. The data we consider consist of a phenotype variable of interest as well as the values of  $p$  omics markers collected for  $n$  independent patients

(where  $p$  is typically much larger than  $n$ ). We suggest a simple approach to evaluate the results that allows a straightforward comparison and visualization of these two types of uncertainty. In a nutshell, our approach consists of randomly splitting the given data set in two independent data set halves and applying the same analysis strategy on each of the halves. A comparison of the results obtained on the two data set halves can be used to derive a measure for sampling uncertainty. In the same way, a comparison of the results obtained by two alternative methods on the same data set half can be used to derive a measure for method uncertainty. These steps are repeated a number of times,  $B$ , and the results are analyzed either in terms of the Jaccard index or a rank correlation coefficient for the selection and the ranking of genes, depending on the statistical scenario.

To illustrate our approach and to demonstrate its versatility, we consider three different scenarios when it comes to the selection and ranking of biomarkers. In the first scenario, we focus on the situation where the aim is to perform variable selection in the context of multivariable regression modeling using different types of omics variables. In this context, Lasso-based methods can perform intrinsic variable selection. In the second scenario, we are concerned with the ranking of biomarkers when the aim is the prediction of a binary outcome. In this situation, variable importance measures in random forests can be used to rank biomarkers according to their predictive performance. Finally, in the third scenario, we consider the selection of differentially expressed genes from RNA-seq data using methods for count variables. Following Rigauil et al. (2016), the methods in this third scenario comprise DESeq, DESeq2, edgeR, glm.edgeR, and limma-voom. In each of the three scenarios, the objective is either to obtain a set of selected variables or a ranking of variables (here omics markers). More details on the statistical methods considered in the three scenarios are given in Section 3.1. We introduce the proposed framework in Section 2 and describe the data sets used for application in Section 3.2. Finally, we present the results, consisting of main results, sensitivity analyses, and a simple example of our framework on low dimensional data in Section 4 and end with a discussion in Section 5.

## 2 | FRAMEWORK FOR THE QUANTIFICATION OF SAMPLING AND METHOD UNCERTAINTY

### 2.1 | Principle

We propose a resampling-based framework to assess sampling and method uncertainty in variable selection and ranking, which is based on the comparison of the results obtained for two different data set halves as well as for at least two different method variants or settings. As a consequence, we are not only able to quantify different types of uncertainty, but also to compare them directly in a common framework. For the sake of simplicity, we describe our framework for two method variants or settings—simply denoted as “method settings” from now on. If there are more than two method settings of interest, different pairs of them can be investigated successively.

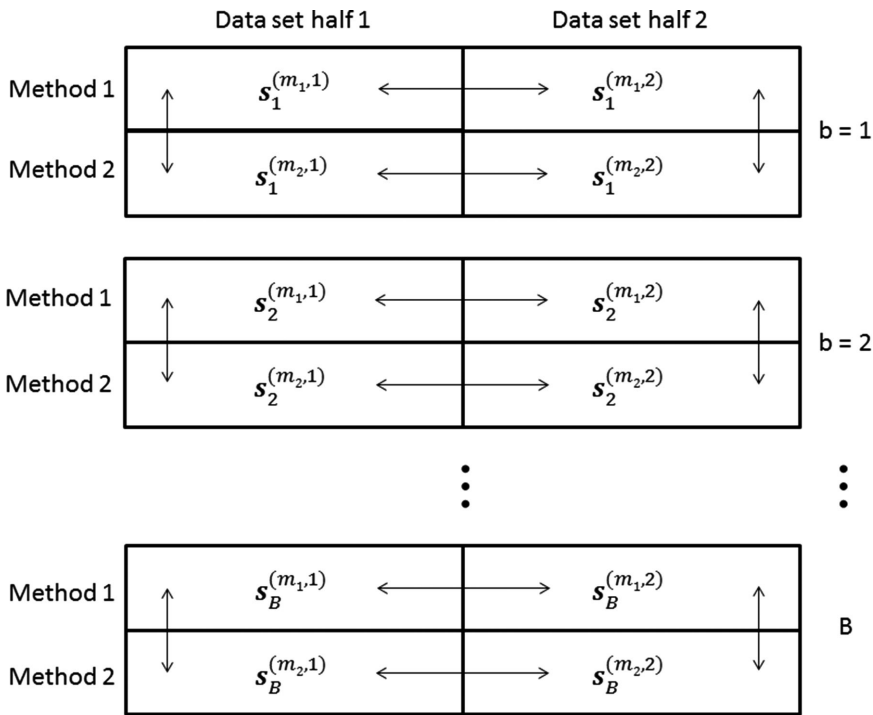
In order to quantify sampling and method uncertainty, the  $n$  patients are first randomly split in two halves for a number  $B$  of times. For each iteration  $b$ ,  $b = 1, \dots, B$ , the two method settings under investigation are applied on each of the data set halves and a subset of selected variables or a ranking of candidate variables is obtained for the  $2 \times 2 = 4$  possible combinations of data set halves and method settings. More formally, for each iteration  $b$ , four vectors  $s_b^{(m,d)}$  of length  $p$  containing the results of the variable selection or ranking are obtained with  $m \in \{m_1, m_2\}$  indicating the method setting and  $d \in \{1, 2\}$  indicating the data set half. In the definition of  $s_b^{(m,d)}$ , we distinguish the two cases of

1. Variable selection:  $s_{j,b}^{(m,d)} = 1$  if variable  $j$  is selected and  $s_{j,b}^{(m,d)} = 0$  otherwise,
2. Variable ranking:  $s_{j,b}^{(m,d)} = r$  if variable  $j$  received rank  $r$ , where  $r = 1$  is the rank of the “best” variable,

for  $j = 1, \dots, p$ . The four vectors  $s_b^{(m,d)}$  are then used to assess and compare the different types of uncertainty as outlined in the sequel of this section. We suggest to assess sampling uncertainty through the comparison of the results  $s_b^{(m,1)}$  and  $s_b^{(m,2)}$  obtained with method settings  $m$  (for  $m \in \{m_1, m_2\}$ ) with the two data set halves. Similarly, we suggest to assess method uncertainty through the comparison of the results  $s_b^{(m_1,1)}$  and  $s_b^{(m_2,1)}$  (resp.  $s_b^{(m_1,2)}$  and  $s_b^{(m_2,2)}$ ) obtained with method settings  $m_1$  and  $m_2$  on the first (resp. the second) data set half.

More precisely, with  $h(\cdot, \cdot)$  being a suitable stability measure, which could for instance be the Jaccard index in the case of variable selection or a type of correlation coefficient in the case of variable ranking (see Section 2.2 for more details on the stability measures we chose in the context of the selection and ranking of biomarkers), we consider

$$h\left(s_b^{(m,1)}, s_b^{(m,2)}\right) \quad (1)$$



**FIGURE 1** Framework for the comparison of sampling uncertainty and method uncertainty

to address sampling uncertainty for method setting  $m$  and the average

$$\frac{1}{2} \left( h(s_b^{(m_1,1)}, s_b^{(m_2,1)}) + h(s_b^{(m_1,2)}, s_b^{(m_2,2)}) \right) \quad (2)$$

over the two data set halves to address method uncertainty, where  $b = 1, \dots, B$ . The lower (1), the higher the sampling uncertainty of method setting  $m$ . The lower (2), the higher the method uncertainty between method setting  $m_1$  and  $m_2$ . Contrasting (1) and (2) provides a convenient comparison of sampling and method uncertainty for a sample size of  $n/2$  where  $n$  is the size of the initial data set. To give a better overview of the procedure, the framework is illustrated in Figure 1.

We calculate stability measures for sampling and method uncertainty for each run  $b = 1, \dots, B$ . To illustrate the results, we show boxplots that allow to visualize their variability over all  $B$  runs, and compute average results over all  $B$  runs. Finally, for additional illustration, dendrograms based on the Ward algorithm for hierarchical clustering, which are based on  $1 - h$  as distance, can be produced.

## 2.2 | Stability measures for variable selection and variable ranking

When the aim is to perform variable selection, we use the Jaccard index as stability measure  $h$  to address sampling and method uncertainty for each iteration  $b$ . For two vectors  $s$  and  $s'$ ,  $h(s, s')$  is defined by

$$h(s, s') = \frac{\sum_{j=1}^p s_j s'_j}{\sum_{j=1}^p (s_j + s'_j - s_j s'_j)}$$

i.e., as the proportion of variables selected in both  $s$  and  $s'$  among variables selected in at least one of them. The Jaccard index can take values between 0 and 1, with higher values indicating more overlap between the sets of selected variables. A Jaccard index of 0 means that completely different variables are selected and hence indicates maximal uncertainty.

To assess the similarity of ranked lists of variables, it is common to use rank correlation coefficients (Boulesteix & Slawski, 2009). Here, we propose to use the Spearman correlation coefficient between the two considered vectors of ranks  $s$  and  $s'$ . Since it makes sense to focus on the best ranked variables, we set all ranks that are larger than  $k$  (where  $k$  is a parameter) to a value of  $(p + k + 1)/2$ , following Critchlow (1985). We will denote this stability measure as the Spearman top- $k$  correlation coefficient in the following. In principle, the Spearman correlation coefficient can take values between  $-1$  and  $1$ , where higher values indicate closer agreement in the two rankings produced on two independent data set halves or for two different method settings. However, the interpretation of this coefficient becomes more difficult when we focus only on the  $k$  best ranked variables, in

particular when  $p$  is much larger than  $k$ . In this situation, the Spearman correlation will be strongly influenced by the number of variables that are included in the  $k$  best ranked variables for the two rankings. We conducted a small simulation study in order to investigate the properties of the Spearman top- $k$  correlation coefficient with  $p = 20,000$ ; we found that this coefficient will only take values greater than  $-0.11$  for  $k = 2000$ ,  $-0.026$  for  $k = 500$ , and  $-0.005$  for  $k = 100$ . Additionally, the interpretation of the Spearman top- $k$  correlation becomes more similar to the interpretation of the Jaccard index in situations where  $p$  is much larger than  $k$ . Nonetheless, the values of the two measures are not directly comparable, because the Spearman top- $k$  correlation will, in general, take larger values for the same number of variables that are included in both rankings.

### 3 | QUANTIFYING UNCERTAINTY IN OMICS BIOMARKER SELECTION AND RANKING

#### 3.1 | The three scenarios

To illustrate our framework for the quantification of sampling and method uncertainty, we apply it in the context of the selection and ranking of molecular biomarkers. In this work, we consider three scenarios: (a) variable selection when predicting a survival outcome based on different types of omics variables, (b) ranking biomarkers based on their performance in predicting a binary outcome and (c) the identification of differentially expressed genes from RNA-seq data yielding a set of selected genes. In each scenario, we use well-known statistical methods that allow the analysis of high-dimensional data. These methods comprise penalized regression via Lasso (scenario 1), random forests (scenario 2), and statistical tests for differential expression analysis from RNA-seq data (scenario 3).

We quantify sampling and method uncertainty for the variable selection procedures performed in scenarios 1 and 3 through the Jaccard index. In scenario 2, where we use random forest variable importance measures to establish a ranking of the considered variables, sampling and method uncertainty is quantified through the Spearman top- $k$  correlation coefficient as described in Section 2.2. We choose  $k = 500$  as a compromise between including a sufficient number of biomarkers and excluding those that may be irrelevant in the ranking. To maintain comparability with the results of scenarios 1 and 3, we also include the results of scenario 2 quantified by using the Jaccard index after a selection of the 20 highest ranking genes in Supporting Information. Note that the purpose of these methods is not necessarily variable selection or variable ranking in the first place. The main aim of the Lasso is to derive prediction rules. In this context, Lasso performs an intrinsic variable selection where the number of selected variables is determined indirectly through the optimization of prediction performance as estimated by cross-validation, i.e., cross-validated Lasso regression automatically produces a set of selected variables. Similar to the Lasso, the main aim of random forests is to derive prediction rules and not to perform variable ranking. In contrast, differential expression analysis is not related to prediction: genes are selected that have significantly different means in the (two) considered groups, i.e., we select all variables considered as significant at the 0.05 level after adjusting the  $p$ -values in order to control the false discovery rate (FDR).

While differential expression analysis merely accounts for univariate associations between the outcome and the predictor variables, both Lasso and random forests consider these associations in a multivariable context, where both the correlation structure and (in the case of random forests) interactions between predictor variables can in principle be accounted for. Note that it is also possible to analyze RNA-seq data with multivariate methods (for instance Lasso-based methods or random forests that are used in scenarios 1 and 2), but here we focus on the case where the aim is to use these data to identify differentially expressed genes. As differential expression analysis considers every gene independently without accounting for the correlation structure between different genes, it can be expected that it will in general identify a larger number of candidate genes than would be selected by a multivariable method. We now present the methods used in the three scenarios in more detail.

##### 3.1.1 | Scenario 1: Variable selection in multivariable regression using different types of omics markers

In scenario 1, the aim is to perform variable selection in a multivariable regression model using different types of omics markers. These analyses can be performed via standard Lasso (Tibshirani, 1996), priority-Lasso (Klau et al., 2018), and IPF-Lasso (Boulesteix, De Bin, Jiang, & Fuchs, 2017a), which is short for “integrative Lasso with penalty factors.” Both priority-Lasso and IPF-Lasso have the advantage of being able to take the different types of variables into account, i.e., different blocks of omics data can be included in an appropriate way. Here, we define each type of data as a block, resulting in three or two blocks, when clinical data are included or excluded, respectively. The other two blocks consist of gene expression and gene deletion data as described in Section 3.2.



In IPF-Lasso, the user has to choose candidate values for the ratios of penalty factors, and the best model is chosen as the one with the smallest cross-validation error. This method is implemented in the R package `ipflasso` (Boulesteix & Fuchs, 2015). Priority-Lasso is a hierarchical approach, where the different blocks of data are considered successively in a Lasso model. The residuals from every fit are used as an offset in the fit of the block with next lowest priority. Therefore, predictive information that is included in several blocks will be obtained from the block with higher priority. The idea is to include prior knowledge in the definition of the order (e.g., by a well-experienced medical researcher) to get a usable model for clinical practice with good prediction accuracy. However, the method can also be used in an impartial way, where the block order corresponding to the best cross-validated error is chosen out of several specifications. Priority-Lasso is implemented in the R package `prioritylasso` (Klau & Hornung, 2017).

Standard Lasso is performed via the R package `glmnet` (see Friedman, Hastie, and Tibshirani (2010), and for the special case of Cox-Lasso Simon, Friedman, Hastie, and Tibshirani (2011)). The different types of data are treated equally and are included together in a Lasso regression. In summary, our analyses concerning variable selection in multivariable regression using different types of omics markers consist of:

#### Scenario 1A: Variable selection without clinical data

1. Priority-Lasso, considering all possible orders of blocks as candidates and choosing the best one by cross-validation
2. IPF-Lasso, considering candidate vectors for penalty factors  $\mathbf{k}^{(1)} = (1, 1)^\top$ ,  $\mathbf{k}^{(2)} = (1, 2)^\top$ , and  $\mathbf{k}^{(3)} = (2, 1)^\top$  and choosing the best one by cross-validation
3. Standard Lasso

#### Scenario 1B: Variable selection with clinical data

4. Priority-Lasso, considering all possible orders of blocks as candidates and choosing the best one by cross-validation
5. IPF-Lasso, considering candidate vectors for penalty factors  $\mathbf{k}^{(1)} = (1, 1, 1)^\top$ ,  $\mathbf{k}^{(2)} = (2, 1, 1)^\top$ ,  $\mathbf{k}^{(3)} = (1, 2, 1)^\top$ ,  $\mathbf{k}^{(4)} = (1, 1, 2)^\top$ ,  $\mathbf{k}^{(5)} = (2, 2, 1)^\top$ ,  $\mathbf{k}^{(6)} = (2, 1, 2)^\top$ , and  $\mathbf{k}^{(7)} = (1, 2, 2)^\top$  and choosing the best one by cross-validation
6. Standard Lasso

We consistently use a five-fold cross-validation procedure and the tuning parameter  $\lambda$  is chosen according to the minimum mean cross-validated error. For ease of computation, priority-Lasso is run without cross-validated offsets (see Klau et al., 2018, for more details). Similarly, it was beyond the scope of this work to consider a broader range of candidate vectors for the penalty factors in IPF-Lasso.

### 3.1.2 | Scenario 2: Ranking biomarkers according to their predictive performance

In scenario 2, we are concerned with the ranking of biomarkers according to their performance when predicting a binary outcome. In this context, we use a variable importance measure based on the Gini index in random forests (Breiman, 2001) to establish a ranking of the candidate biomarkers. In order to quantify method uncertainty in this scenario, we focus on investigations concerning different algorithmic settings rather than considering different methods for classification. Parameters considered for our random forest settings are the number of predictors that are sampled for splitting at each node (called “`mtry`”), the minimal node size (“`min.node.size`”), and the fraction of observations to sample (“`sample.fraction`”). In order to be able to perform variable ranking in a high-dimensional context, we choose to perform the analyses with a number of 10,000 trees for one forest for all our parameter settings. The analysis is done with the R package `ranger` (Wright & Ziegler, 2017). The authors suggest default values for classification of  $mtry = \sqrt{p}$ , where  $p$  is the total number of variables, a `min.node.size` of 1, and a `sample.fraction` of 1 since we sample with replacement. These default values provide our first setting. In settings 2 and 3, we change `mtry` to  $p/10$  and  $p/5$ , respectively, which results in more appropriate values for the high-dimensional context (Goldstein, Polley, & Briggs, 2011). In setting 4, we tune `mtry`, `min.node.size`, and `sample.fraction` with respect to the Brier score. We perform 100 iterations, from which we take 5% of the iterations with the lowest Brier scores and obtain the final tuning parameters as an average of these results. The tuning is performed with the R package `tuneRanger` (Probst, 2018) and the underlying procedure is described in more detail in Probst, Wright, and Boulesteix (2018). In summary, the four settings for the random forests applied in this scenario are:

1. `mtry` =  $\sqrt{p} = 132$ , `min.node.size` = 1, `sample.fraction` = 1.00
2. `mtry` =  $p/10 = 1739$ , `min.node.size` = 1, `sample.fraction` = 1.00



3.  $mtry = p/5 = 3478$ ,  $min.node.size = 1$ ,  $sample.fraction = 1.00$
4.  $mtry = 2235$ ,  $min.node.size = 9$ ,  $sample.fraction = 0.85$

### 3.1.3 | Scenario 3: Identifying differentially expressed genes from RNA-seq data

In the context of differential expression analysis with RNA-seq data, we aim to find differentially expressed genes. Among the many possible methods that are available in this context, we focus on those used in Rigai et al. (2016) where the authors compared methods for differential expression analysis in the presence of few biological replicates with synthetic data sets. We thus consider DESeq (Anders & Huber, 2010), DESeq2 (Love, Huber, & Anders, 2014), edgeR and glm.edgeR (McCarthy, Chen, & Smyth, 2012; Robinson, McCarthy, & Smyth, 2010), and limma-voom (Law, Chen, Shi, & Smyth, 2014; Ritchie et al., 2015). These methods differ primarily in the choice of the probability distribution, the strategy used to estimate the mean and variance parameters, and the prefiltering approach. Following Rigai et al. (2016), we apply DESeq without the filtering of genes with low counts beforehand. All other methods are consistently used with the filtering strategy recommended by their authors. Apart from limma-voom, all methods are based on the negative binomial distribution. The analysis with limma-voom, on the other hand, is based on a normal distribution after conducting a voom-transformation of the data. For details concerning the specific estimation and modeling of the mean–variance relationship and the specific test procedures, we refer to the explanations of the authors of the methods. The genes are finally selected if their corresponding  $p$ -values are below a predefined threshold chosen consistently as 0.05 after adjusting for multiple testing with the Benjamini–Hochberg procedure to control the FDR.

## 3.2 | Application to acute myeloid leukemia data sets

For all our analyses, we use data from acute myeloid leukemia (AML) patients. Since different situations are addressed in the three scenarios, different AML data sets are used for each of them.

### 3.2.1 | AML data from TCGA

In scenario 1 (see Section 3.1), our aim is to build a prediction model based on different types of omics data. Hence, we choose a data set including different types of variables (blocks of data) for this scenario. We consider an AML data set from The Cancer Genome Atlas (TCGA, <https://portal.gdc.cancer.gov/>) and work with gene expression variables, copy number variations (CNVs), as well as clinical variables as predictors. The censored outcome variable of interest is overall survival. In order to allow for a fair comparison of variables in our models, we standardize all continuous variables to mean zero and variance one. The 15 clinical variables, which are only used for part of the investigations, include sex, age at initial pathologic diagnosis, leukocyte level, hemoglobin level, platelet count, percent value of blast cells in bone marrow, as well as eight binary variables, representing the mutation status for certain genes. Furthermore, we include the three-categorical variable cytogenetic risk. We exclude patients with missing values on at least one of the variables, resulting in a total of 176 patients of which 125 experienced the event of interest for the analyses without clinical data. For the analyses including clinical variables, there are 149 patients (of which 89 died), as there are more missing values. The gene expression and CNV data consist of 19,204 and 18,354 genes, respectively.

### 3.2.2 | AMLCG-1999 trial data

In order to perform the analyses for the random forests scenario (scenario 2), we consider a data set consisting of patients who were randomized and treated in the multicenter phase III AMLCG-1999 trial (clinicaltrials.gov identifier NCT00266136) between 1999 and 2005 (Büchner et al., 2016, 2006). It consists of gene expression data analyzed with Affymetrix arrays (Herold et al., 2014). As outcome variable, we consider the resistance to induction treatment. The data set consists of 488 patients of which 119 are resistant and 17,389 gene expression variables. It is publicly available from the Gene Expression Omnibus repository (GSE37642).

### 3.2.3 | AMLCG-2008 study

For the differential expression analysis scenario (scenario 3), we use a data set of 218 AML patients treated in the AMLCG-2008 study (NCT01382147) (Kreuzer et al., 2013), a randomized, multicenter phase III trial. Additionally, 40 patients who had resistant disease and were treated in the AMLCG-1999 trial are included. The final data set consists of  $n = 241$  patients after excluding those with missing values. They are described by their transcriptome comprising  $p = 23,368$  genes measured by the

**TABLE 1** Stability measures for the main results averaged over 100 runs. The values for sampling uncertainty, obtained by comparing two different data set halves, are shown on the diagonal. The values for method uncertainty, obtained by averaging over the results obtained with data set half 1 and data set half 2, are shown on the off-diagonal. In scenarios 1 and 3, uncertainty is quantified by Jaccard indices and in scenario 2 by the Spearman top- $k$  correlation with  $k = 500$ , as described in Section 2.2. The acronyms “Lasso,” “PL,” and “IPF” represent standard Lasso, priority-Lasso, and the IPF-Lasso, respectively. “sqrt( $p$ ),” “ $p/5$ ,” “ $p/10$ ” indicate the different values for the number of covariates to possibly split at each node (mtry), with  $p$  as the number of covariates. “tuning” refers to the scenario in which mtry, min.node.size, and sample.fraction were tuned

Scenario 1A	Lasso	PL	IPF		
Lasso	0.00	0.50	0.35		
PL		0.00	0.35		
IPF			0.00		
Scenario 1B	clin: Lasso	clin: PL	clin: IPF		
clin: Lasso	0.00	0.19	0.17		
clin: PL		0.10	0.27		
clin: IPF			0.12		
Scenario 2	sqrt( $p$ )	$p/10$	$p/5$	tuning	
sqrt( $p$ )	0.06	0.78	0.77	0.78	
$p/10$		0.05	0.87	0.84	
$p/5$			0.05	0.84	
tuning				0.05	
Scenario 3	DESeq	DESeq2	edgeR	glm.edgeR	limma-voom
DESeq	0.00	0.01	0.01	0.01	0.01
DESeq2		0.01	0.53	0.55	0.17
edgeR			0.02	0.95	0.17
glm.edgeR				0.02	0.17
limma-voom					0.02

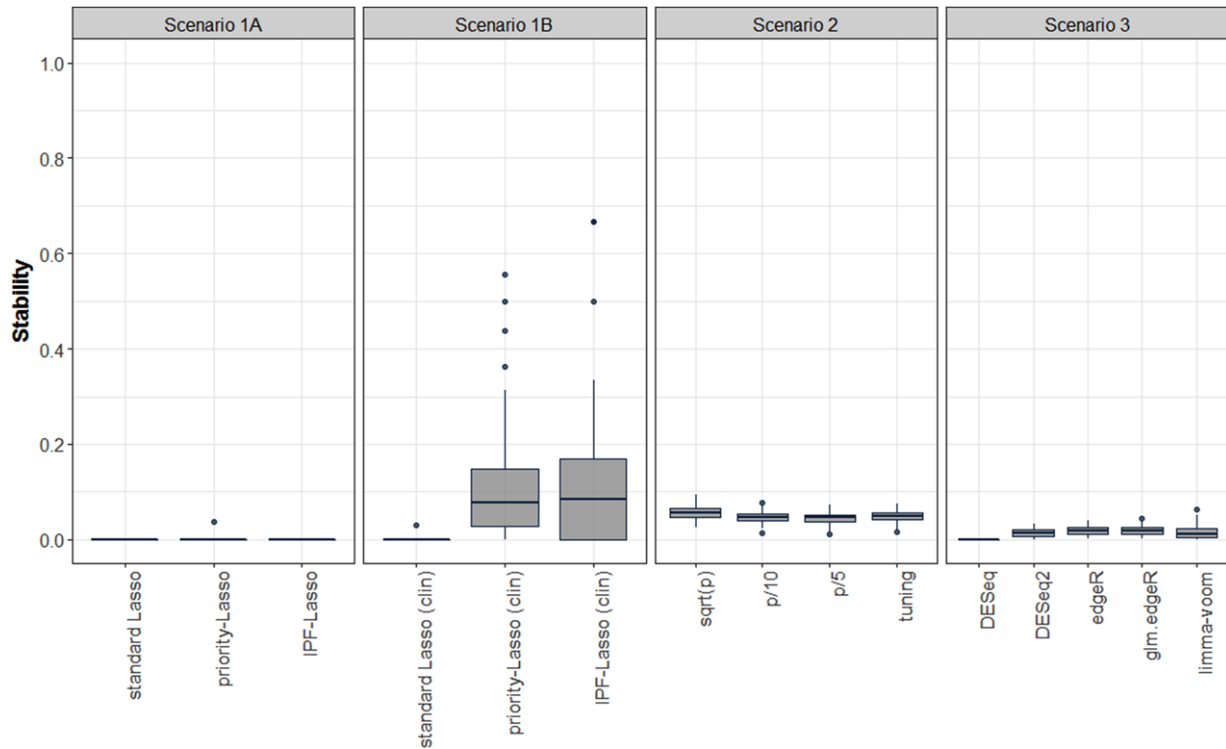
RNA-seq technology. For more detailed information, we refer to Herold et al. (2017), where the data set was used to predict the resistance to induction treatment. In our study, the resistance to induction treatment serves as a binary condition for differential expression analysis: 73 patients are resistant and 168 non-resistant.

## 4 | RESULTS

### 4.1 | Main analyses

All methods and scenarios that we described in Section 3.1 are conducted with a number of  $B = 100$  runs, i.e., the data set is split 100 times in two halves and each method setting is applied on both data set halves. For each run, stability measures are calculated to quantify both sampling and method uncertainty, as described in Section 2. The mean stability measures for the three scenarios, averaged over the 100 repetitions, are shown in Table 1. The boxplots, visualizing the distribution of the results over the 100 runs, are shown in Figures 2 and 3.

First of all, the most obvious results are the remarkably low stability measures, i.e., Jaccard indices and rank correlation coefficients, when comparing the same method or method setting on two data set halves, indicating high sampling uncertainty in general. In scenario 3, where we were concerned with the identification of differentially expressed genes from RNA-seq data, we observe only a small fraction of runs where selected genes from the two data set halves overlap, resulting in very low Jaccard indices between 0.01 and 0.02 for most methods and even in a Jaccard index of 0.00 for DESeq. The ranking of biomarkers according to their predictive performance in random forests in scenario 2 results in values for the Spearman top- $k$  correlation around 0.05, indicating high sampling uncertainty in this situation. The results quantifying the uncertainty for scenario 2 with the Jaccard index are shown in Supporting Information. They confirm the high sampling uncertainty with values comparable to those obtained in scenario 3. Similarly, for scenario 1A, where variable selection in multivariable regression using different types of omics markers is performed without clinical data, we observe low values for the Jaccard index measuring sampling uncertainty. In scenario 1B, however, where clinical data is included in the variable selection process, priority-Lasso and IPF-Lasso identify

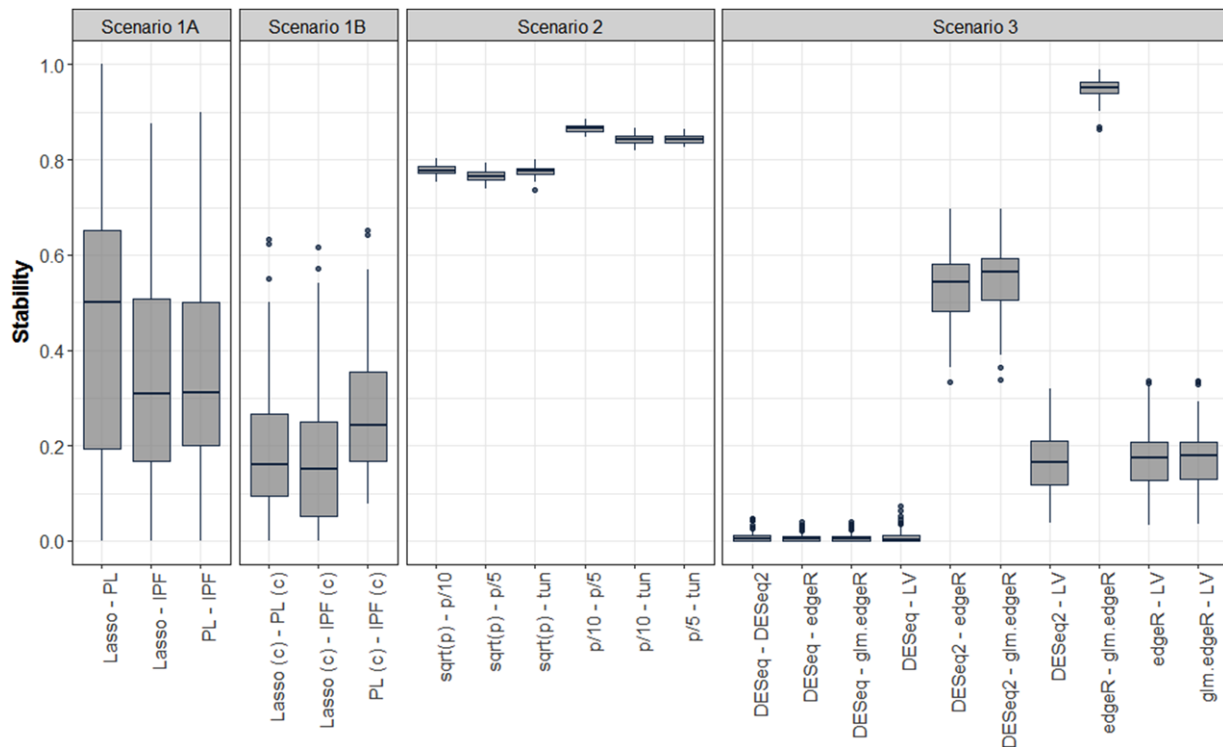


**FIGURE 2** Boxplots visualizing the distribution of the stability measures (Jaccard indices for scenarios 1 and 3, and Spearman top- $k$  correlations with  $k = 500$ , as described in Section 2.2, for scenario 2) measuring sampling uncertainty for 100 runs. For each run, the data set was split in two halves and the corresponding method was applied on both halves. Boxplots for random forest method settings (scenario 2) contain the abbreviations “sqrt(p),” “p/10,” “p/5,” and “tuning.” The first three indicate the different values for the number of covariates to possibly split at each node (mtry), with  $p$  as the number of covariates. “tuning” refers to the scenario in which mtry, min.node.size, and sample.fraction were tuned

more stable sets of selected variables and therefore show lower sampling uncertainty. This finding can be explained by the fact that these two methods account for the block structure and therefore have a higher chance of including important clinical variables in the final model. However, mean Jaccard indices of 0.10 and 0.12 for priority-Lasso and IPF-Lasso, respectively, still indicate unstable sets of selected variables and therefore high sampling uncertainty.

The comparison of two alternative methods on the same data set half results in more stable results than the comparison of the same method on two data set halves, i.e., method uncertainty was lower than sampling uncertainty in all three scenarios. In scenario 1, where the aim is to perform variable selection in a multivariable regression using different types of omics markers, we observe Jaccard indices that indicate moderate to high method uncertainty. When comparing the values of scenarios 1A and 1B, however, it can be noticed that method uncertainty is actually higher in scenario 1B when clinical data is taken into consideration. In this situation, the difference between the set of variables selected by Lasso on the one side and priority-Lasso and IPF-Lasso on the other side might again be explained by the fact that the two latter methods are able to select important clinical variables in the final model as they account for the block structure in the data. However, while one might expect high overlap between priority-Lasso and IPF-Lasso in scenario 1B, there is a slight increase in method uncertainty between these two methods when clinical data is included in the variable selection process. This finding might be explained by the fact that IPF-Lasso is more conservative than priority-Lasso in our analyses. Indeed, as can be seen in Table S1, the number of selected variables is higher for priority-Lasso than for IPF-Lasso, in particular for scenario 1B when clinical data is included in the analyses. The boxplots presented in Figures 2 and 3 show that the variability in the 100 Jaccard indices for scenario 1 is very high, especially without clinical data, which might be related to the small number of selected variables, which is particularly low for scenario 1A (see Table S1).

In scenario 2, where we performed a ranking of biomarkers according to their predictive performance, the Spearman top- $k$  correlation coefficients comparing the results of different method settings are high and show little variability, indicating low method uncertainty in this situation (see Figure 3). The low method uncertainty between the settings is even more noticeable for comparisons including the settings 2, 3, and 4 with larger values of mtry. The corresponding Jaccard indices are similar but due to the properties explained in Section 2.2 slightly lower than the correlations (see Table S2). The cluster dendrogram visualizing

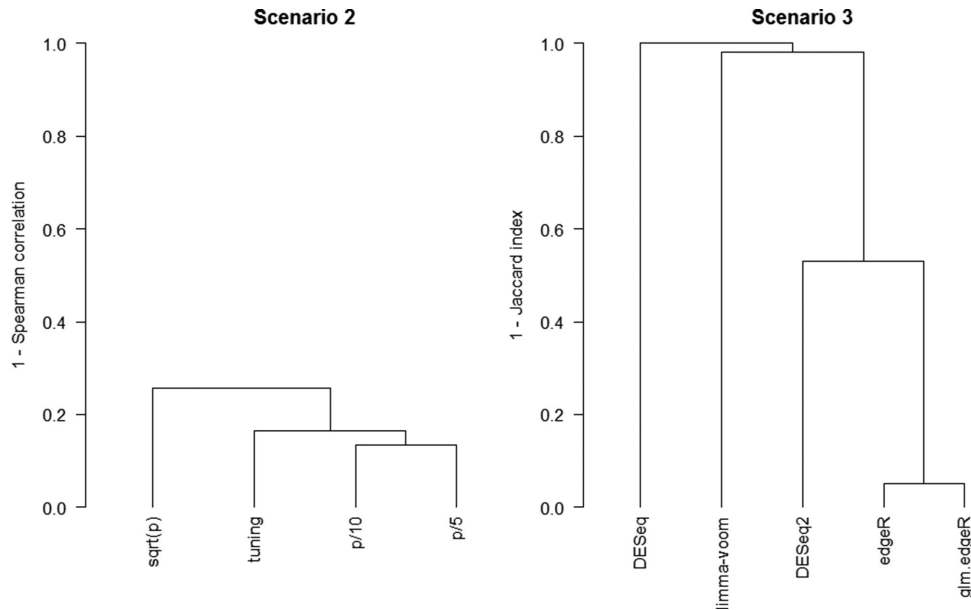


**FIGURE 3** Boxplots visualizing the distribution of the stability measures (Jaccard indices for scenarios 1 and 3, and Spearman top- $k$  correlations with  $k = 500$ , as described in Section 2.2, for scenario 2) measuring method uncertainty for 100 runs. For each run, the different methods or method settings were applied on both data set halves and results are analyzed in terms of the stability measures. The abbreviations “Lasso,” “PL,” and “IPF” represent standard Lasso, priority-Lasso and IPF-Lasso, respectively. “sqrt( $p$ ),” “ $p/10$ ,” and “ $p/5$ ” indicate the different values for the number of covariates to possibly split at each node ( $mtry$ ), with  $p$  as the number of covariates. “tun” refers to the scenario in which  $mtry$ ,  $min.node.size$ , and  $sample.fraction$  were tuned. Furthermore, “LV” stands for the limma-voom method

the results concerning method uncertainty in scenario 2 is shown in the left panel of Figure 4. It confirms the results observed in Table 1 and Figure 3.

However, the findings concerning method uncertainty in scenarios 1 and 2 have to be interpreted with great caution. Indeed, both the Lasso-based methods and the random forest method are based on resampling, i.e., there is already an inherent variability: the results are dependent on the chosen seed values. When conducting the same analysis for two subsequent runs on the same data set half with different seed values and comparing the results, we also observe average values for the Jaccard index between 0.43 and 0.52 in scenario 1A. For scenario 2, we observe values of the Spearman top- $k$  correlation between 0.77 and 0.86 which is very close to the values obtained by comparing different method settings. This finding suggests that the stability measures observed concerning method uncertainty in scenario 2 do not reflect the variability due to the parameter settings, but rather the randomness which is inherent in the random forest method itself.

Finally, in scenario 3, which was concerned with the identification of differentially expressed genes from RNA-seq data, the results quantifying method uncertainty are highly dependent on the methods that are being compared. In particular, when DESeq is compared to the other methods, high uncertainty can be observed with mean Jaccard indices of 0.01. This finding might be related to the fact that DESeq identifies only a very small number of differentially expressed genes. While all other methods identify more than 192 variables as differentially expressed, DESeq only selects 2.5 variables on average as can be seen in Table 2. The methods edgeR and glm.edgeR, on the other hand, share similar properties, resulting in low method uncertainty when comparing these methods as indicated by a Jaccard index of 0.95. The comparisons between the latter two methods and DESeq2, which is also based on negative binomial counts, leads to slightly less stable sets of identified variables with Jaccard indices of 0.53 and 0.55, indicating moderate method uncertainty. Finally, comparing limma-voom to edgeR, glm.edgeR, and DESeq2 yields Jaccard indices of 0.17, indicating moderate to high method uncertainty. While this method is less conservative than DESeq and therefore results in a comparable number of identified variables as DESeq2, edgeR, and glm.edgeR, it is the only method based on a normal distribution, thereby leading to less overlap and lower Jaccard indices than are observed in the comparisons involving only DESeq2, edgeR, and glm.edgeR. In the same manner as for the random forest scenario, the method



**FIGURE 4** Dendrograms for scenarios 2 and 3. “sqrt(p),” “p/10,” and “p/5” indicate the different values for the number of covariates to possibly split at each node (mtry), with  $p$  as the number of covariates

**TABLE 2** Numbers of variables selected in scenario 3 and the corresponding sensitivity analyses. The results for two data set halves are averaged across these halves. All results, except those obtained with the full data set, contain the results averaged across 100 runs. A balanced split means that a constant resistant/non-resistant ratio is maintained

Full data ( $n = 241$ )	DESeq	DESeq2	edgeR	glm.edgeR	limma-voom
Unfiltered	107	1147	2313	1526	1512
Filtered		1431	1144	1181	1423
<b>Data set halves (<math>n = 120</math>)</b>					
Unfiltered	2.5	252.2	749.6	319.7	206.8
Filtered		322.1	280.4	285.6	192.2
<b>Balanced split (<math>n = 146</math>)</b>					
Unfiltered	41.2	611.6	1383.4	758.7	740.0
Filtered		765.7	586.3	596.3	636.9
<b>Equal library sizes (<math>n = 53</math>)</b>					
Unfiltered	0.4	72.2	108.6	83.2	20.4
Filtered		98.8	76.7	77.0	28.4

uncertainty in the differential expression analysis scenario is visualized with a cluster dendrogram in the right panel of Figure 4. Unsurprisingly, DESeq and limma-voom form single clusters. In contrast, edgeR and glm.edgeR group together very early and melt with DESeq2 at a later point.

## 4.2 | Sensitivity analyses

In a second step, we perform several sensitivity analyses to investigate possible factors that could have an influence on the results of our main analyses presented in Section 4.1. In scenario 1, we explore the correlation structure of the variables selected on the two different data set halves in order to assess whether the small overlap between the selected subsets of variables can be explained by the fact that these subsets of variables carry the same information. In scenario 2, we investigate the impact of different values for  $n$  on sampling and method uncertainty and use two data sets with higher signal-to-noise ratio. Moreover, we study the sensitivity of results when changing the value  $k$  for the Spearman top- $k$  correlation in this scenario. Finally, we study the impact of filtering and a limitation to a balanced sample and patients with equal library sizes in the differential expression analysis (scenario 3). All analyses of this section were conducted with  $B = 100$ .

### 4.2.1 | Exploring the correlation structure of the selected variables

One possible explanation for the small overlap in the genes that are selected by Lasso-based methods on two data set halves observed in scenario 1A could be that there are highly correlated groups of biomarkers that carry the same information (De Bin & Sauerbrei, 2018). In this vein, one may imagine that, while there is not an exact overlap between the biomarkers selected on the data set halves, the selected biomarkers are at least highly correlated. However, further examination of the correlation between the variables (not shown), selected by a method from the two independent data sets in scenario 1, does not back up the presumption that genes selected from one data set half might be highly correlated to the genes selected from the other data set half. Indeed, the correlation is consistently low or at most medium. Correspondingly, when using elastic net regression (Zou & Hastie, 2005), which is especially recommended to address the problem of groups of variables with high pairwise correlation, sampling uncertainty does not decrease (data not shown). Here, it yields similar results to standard Lasso even when trying different mixing parameters.

### 4.2.2 | Exploring the choice of $k$ for the Spearman top- $k$ correlation

For scenario 2, where we perform a ranking of biomarkers according to their predictive performance and analyze the results with the Spearman top- $k$  correlation, the choice of  $k = 500$  is somewhat arbitrary. In order to gain additional insight concerning the impact of this choice, we perform our analyses with  $k \in \{100, 200, 1000, 2000\}$ . While  $k$  only modestly impacts the results, we observe slightly lower sampling uncertainty, expressed through higher correlations for higher values of  $k$ . Conversely, the correlations quantifying method uncertainty slightly decrease for higher  $k$ , indicating higher uncertainty in this context.

### 4.2.3 | Investigations with other data sets

In a third sensitivity analysis, we investigate the influence of the signal-to-noise ratio and the number of observations  $n$  by considering two additional high-dimensional data sets with other outcome variables: one that merely has a higher signal-to-noise ratio and a second one that has both a higher signal-to-noise ratio and a larger sample size. First, a data set of AML patients from the Gene Expression Omnibus repository (GSE61804) is considered. The data set consists of 54,675 gene expression variables and 323 patients. When the outcome to be predicted is gender, we observe an AUC of 0.99 for a random forest, specified with 10,000 trees and default values for the other parameters, i.e., gender is an outcome that can be predicted with high accuracy. For comparison, the prediction of resistant patients with the data initially used for the random forest scenario (scenario 2) yields a considerably lower AUC of 0.64. Unsurprisingly, we find higher correlations between two rankings of genes selected on two data set halves compared to the results of our main analyses, i.e., sampling uncertainty is lower for this data set with higher signal-to-noise ratio. However, correlation coefficients between 0.11 and 0.17 still indicate moderate to high uncertainty. The second data set we consider in our sensitivity analyses is a gene expression data set consisting of different cancer types with 1,545 patients in total and 10,936 variables that is available on the OpenML platform (Vanschoren, van Rijn, Bischl, & Torgo, 2013) (data set ID 1128). We predict whether the tissue is a breast tissue or another tissue type and obtain an AUC of 0.98 for a standard random forest with 10,000 trees. Conducting our analyses with  $B = 100$  runs yields a mean correlation coefficient of 0.53 for method uncertainty and values between 0.40 and 0.77 for sampling uncertainty, depending on the method setting. The relatively high number of observations also allows us to perform the same analysis on subsets of the data. Here we consider subsets of sizes  $n \in \{100, 200, 500, 800, 1200\}$ . The results, which are visualized in Figure S2, reveal on the one hand a decreasing sampling uncertainty for higher sample sizes, indicated by higher correlations. On the other hand, the correlations quantifying method uncertainty are less affected by  $n$  and we observe even a slight increase in method uncertainty for higher sample sizes. In conclusion, the sensitivity analyses performed on other data sets support the hypothesis that the high sampling uncertainty, observed in our main analyses, is at least partly due to the relatively low signal-to-noise ratio and the relatively small sample size.

### 4.2.4 | Sensitivity analyses for the differential expression scenario

In order to investigate how method uncertainty is influenced by filtering in the differential expression scenario, we rerun the methods where filtering of genes with low counts was part of the analysis, i.e., all methods except DESeq, without this filtering step. As shown in Table 2, when averaging the  $B = 100$  runs, more genes are detected as differentially expressed by most of the methods. Here, DESeq2 is the only exception with a decrease from 322.1 to 252.2 detected genes when the data is unfiltered. Correspondingly, method uncertainty for the comparisons involving DESeq remains very high as DESeq only identifies a small number of differentially expressed genes. For method comparisons in which DESeq is not involved, the Jaccard indices are slightly lower for the filtered analyses.

Furthermore, we slightly modify our data set in order to obtain a balanced ratio of resistant and non-resistant patients. Since we are only interested in method uncertainty here, we do not split the data set in two halves. Instead, we randomly draw samples



of non-resistant patients with a sample size that is equal to the number of resistant patients a number of 100 times. We perform the analyses on each of the data sets that are obtained by combining the random subset with the set of resistant patients. More genes are selected than in the original analysis (see Table 2), and method uncertainty is lower for all comparisons. Here, on average, 41.2 genes are considered as differentially expressed by the DESeq method. Nevertheless, the general structure of the results does not change.

Finally, Law et al. (2014) show that DESeq can have low power to detect differentially expressed genes when library sizes are unequal. Therefore, we conduct a further sensitivity analysis where we consider only patients with a total number of mapped reads between 10 and 15 million. This reduced our original data set to 106 patients. In contrast to the observations made by Law et al. (2014), that higher power can be obtained with equal library sizes, our results do not markedly change. Instead, the number of selected genes and the Jaccard indices quantifying method uncertainty decrease—not only for DESeq but for all of the methods. Altogether, the additional analyses presented in this paragraph provide an indication that the results are not substantially affected by changes in our data set composition. Instead, they seem to be affected by the sample size as well as characteristics whose detailed comprehension is beyond the scope of this paper, but should be further investigated in future analyses.

### 4.3 | Example on low dimensional data

In order to illustrate the versatility of our framework, we complement our analyses concerning the variable selection and ranking on high-dimensional data by a simple example where we apply the framework to low-dimensional data. In this example, we quantify method and sampling uncertainty when predicting the percentage of body fat in the body fat data set (data set ID 560,  $n = 252$ ) that is available on the OpenML platform (Vanschoren et al., 2013). The different methods we consider here consist of (a) a linear regression model with age and height as predictors (model 1) and (b) a linear regression model with age, height, and weight as predictors (model 2). We use the correlation between the outcome predictions on independent test data ( $1/3$  of the original data set, i.e., 84 observations) as stability measure. We run our framework with  $B = 100$  and observe average correlations of 0.75 and 0.97 when fitting models 1 and 2 to two different data set halves, respectively, indicating that contrary to the results on high-dimensional data, sampling uncertainty is very low in this example. On the other hand, we observe an average correlation of 0.20 when comparing the predictions of the two models on the same data set half. However, similar to the main results of our work, it has to be noted that method uncertainty strongly depends on the methods being compared, i.e., we would expect it to be lower when the two models differ with respect to the inclusion of a less influential predictor than weight.

## 5 | DISCUSSION

### 5.1 | Summary

In this paper, we proposed a general resampling-based framework to quantify sampling and method uncertainty which makes it possible to compare these two types of uncertainty in a simple and comprehensive way. In this framework, the random splitting of a data set allows quantification of sampling uncertainty by comparing the results when performing the same analysis strategy on each of the data set halves. A comparison of the results obtained when applying two alternative analysis strategies on the same data set half, on the other hand, can be used to derive a measure of method uncertainty. We applied our framework in the selection and ranking of omics markers in the context of AML and considered three different scenarios: variable selection in multivariable regression using different types of omics markers (scenario 1), the ranking of biomarkers according to their predictive performance as estimated by random forest variable importance measures (scenario 2), and the identification of differentially expressed genes from RNA-seq data (scenario 3).

In scenario 2, where sampling uncertainty is quantified through the Spearman top- $k$  correlation coefficient with  $k = 500$ , we observe high sampling uncertainty with values around 0.05. For the differential expression analyses in scenario 3, we observed very high sampling uncertainty with Jaccard indices that imply almost no overlap between the biomarkers selected on the two data set halves. In contrast, in scenario 1B, where variable selection in multivariable regression using different types of omics markers and clinical data was performed using Lasso-based methods, sampling uncertainty was slightly lower for priority-Lasso and IPF-Lasso that take the block structure in the data set into account. However, this is also related to the special composition of the data set itself. Indeed, when clinical data was incorporated in the analyses, these two methods, which can properly account for this external information, selected more stable sets of variables. Accordingly, Binder and Schumacher (2008, 2009) show that the proper incorporation of clinical and pathway information, respectively, can both yield more convincing results and improve predictive performance when analyzing high-dimensional microarray data.

As might be expected, method uncertainty was strongly affected by the methods being compared. For example, in scenario 1, high method uncertainty was observed for a comparison of standard Lasso and IPF-Lasso with clinical data. In contrast, in scenario 2, we found low method uncertainty for a comparison of random forest settings 2 and 3 where only  $m$ try was changed from  $p/10$  to  $p/5$ . Extremely heterogeneous results concerning method uncertainty could be observed in the differential expression scenario (scenario 3). Here, comparisons where DESeq was involved led to very high method uncertainty. In contrast, for a comparison of edgeR and glm.edgeR, a very high Jaccard index, indicating low method uncertainty, could be observed. We performed a number of sensitivity analyses where we studied the impact of filtering and a limitation to patients with equal library sizes and to a balanced sample, which essentially led to the same results.

Rather than being mock examples with extraordinary low sample size or signal-to-noise ratio, the data sets considered here have been used in actual research projects for AML (Herold et al., 2017; Klau et al., 2018) and can be seen as typical for data sets that are used for the identification of omics markers. The extremely high sampling uncertainty we observed is therefore alarming as this finding suggests that the omics markers identified in typical studies are very sensitive to random variations in the data set, raising concerns about the replicability of the results of these studies. While this observation is in accordance with many previous studies that were concerned with the stability of variable selection procedures based on high-dimensional data (see for instance Michiels et al. (2005) and Ein-Dor et al. (2005, 2006)), we found such an extremely poor overlap between markers selected on two data set halves disconcerting on samples including several hundreds of patients.

## 5.2 | Strengths and limitations

The aim of the framework we proposed in this work is to quantify sampling and method uncertainty and to allow direct comparisons between these two sources of uncertainty. A deeper understanding of the different sources of uncertainty is not only useful for methodological researchers developing new methods to help them focus their attention on the most critical problems, but also for applied scientists in the interpretation of their results. In particular, the framework can be useful as a complement of standard sensitivity analysis where the latter typically allows consideration of the results of only a very limited number of alternative analysis strategies while the former can give a more general idea on the variability of results which can arise from these strategies. As the variability resulting from a single source of uncertainty is typically very difficult to interpret, it is important in this context to be able to quantify sampling and method uncertainty on a common scale in order to allow a direct comparison between the two. Finally, the resampling-based framework we propose here can be seen as a flexible tool to give an idea on how likely it is that a research finding will replicate with an independent sample or with the same sample when an alternative analysis strategy is chosen. Considering the very high values of sampling uncertainty we observed in the three scenarios, it is unlikely that a selection or a ranking of biomarkers on an independent sample of AML patients would yield the same results. In this situation, it might be sensible to conclude that more patients have to be recruited before valid conclusions can be drawn. If we had observed higher values for method uncertainty, we might on the other hand have concluded that further efforts should be directed to the selection of the most suitable method, for instance by comparing the results of several methods in a simulation study that specifically reflects the properties of our data. In this vein, sampling uncertainty and method uncertainty as quantified by our framework can be used to judge the stability of research findings. Achieving values of sampling and method uncertainty that indicate a certain stability may be seen as a minimum requirement that has to be fulfilled in order to assure the credibility and reproducibility of results.

In the three scenarios considered in this work, we observed high sampling uncertainty and a comparatively smaller, but non-negligible method uncertainty. However, when comparing the values we observed for method uncertainty between the different scenarios, one has to keep in mind that the stability measures that we used in these scenarios are not directly comparable. Additionally, the concept of method uncertainty is somewhat elusive. Indeed, method uncertainty strongly depends on the methods being compared and it is in general neither feasible nor reasonable to consider all possible methods and method settings, but only those that are justified both from a theoretical and a substantive point of view. In this vein, we restricted the possible method settings in this work. For instance, we only used a set of possible candidate vectors for the penalty factors in IPF-Lasso in scenario 1. Moreover, another possibility would have been to include the clinical covariates as mandatory in this scenario. We investigated how this choice would have changed the results in an additional analysis (results not shown) and found in this case that no other covariates were chosen by standard Lasso, which results in very low sampling uncertainty, but that the non-clinical variables that were selected by priority-Lasso are highly unstable among different data set halves. Similarly, we based all analyses in scenario 3 on a threshold for the  $p$ -value of 0.05, although alternative thresholds could lead to other results. In a second additional analysis (results not shown), we investigated how the results would have changed with an alternative threshold of 0.157 (Sauerbrei et al., 2011) in scenario 3. In this additional analysis, more variables were chosen and slight changes could be observed in the values of the stability measures.



In addition to the limited number of analytical decisions, it has to be noted that the resampling-based framework we propose does not allow to quantify sampling and method uncertainty for the total sample size  $n$  of the original data set, but merely for  $n/2$ . It can be argued that this decreased sample size leads to a loss in statistical power and therefore to an overestimation of sampling and method uncertainty for the original data set. Using bootstrap resampling (i.e., drawing  $n$  observations with replacement) instead of subsampling could be an alternative solution, but there is evidence that the bootstrap leads to an inflation of type 1 error and to a high inclusion frequency of noise variables (De Bin, Janitza, Sauerbrei, & Boulesteix, 2016; Janitza, Binder, & Boulesteix, 2016). Moreover, outliers may have a devastating effect in  $n \ll p$  settings when drawn several times in the same bootstrap sample (De Bin et al., 2016). Most importantly, bootstrap samples partially overlap with each other, making a direct comparison of sampling and method uncertainty (as performed in our framework) impossible. For these reasons, the use of bootstrap sampling in place of  $n/2$ -subsampling does not appear to be recommended in our context. The decrease in sample size in the latter resampling approach is arguably the price one has to pay in order to avoid any overlap between the two data halves, thereby allowing the derivation of an unbiased stability measure for  $n/2$ . Moreover, when the ultimate aim is to translate research findings into clinical practice, it is important to not only validate results on the considered data set but also to use external validation in order to show that the results can be generalized to other samples and even to samples drawn under different conditions (Gerds, Cai, & Schumacher, 2008). The stability of selected genes is likely to be higher on the considered data set than on an independent data set. Therefore, it may be preferable to choose a pessimistic measure of stability performance as this measure will probably give a more accurate picture of the stability that can be expected when one aims at generalizing the results.

### 5.3 | Extensions

In this work, we focused on sampling and method uncertainty, where method uncertainty was either defined as the variability in results when applying different methods (scenario 1), different parameter settings (scenario 2), or different analysis strategies (scenario 3). However, the framework can be applied to many other issues beyond the selection of omics biomarkers and the different types of uncertainty addressed in this paper are of course not the only sources of variability in statistical results. A further aspect worth mentioning is, for instance, related to imprecise methods of data collection, which could be denoted as “measurement uncertainty.” Similarly, data preprocessing uncertainty is not particularly addressed in our application, but, in contrast to measurement uncertainty, it could be easily incorporated in our framework. Finally, it is straightforward to extend future analyses to stability measures other than the Jaccard index and the Spearman top- $k$  correlation coefficient to quantify different sources of uncertainty in the analysis of high- and low-dimensional data, as we have already demonstrated with a simple example in Section 4.3.

### 5.4 | Outlook

The importance of investigations concerning the stability of variable selection procedures and the role of resampling-based approaches in these investigations have long been recognized (Baty, Jaeger, Preiswerk, Schumacher, & Brutsche, 2008; Sauerbrei & Schumacher, 1992; Sauerbrei et al., 2011). However, even today, evaluations of the stability of research findings are rarely carried out in a systematic way. While the instability of variable selection procedures is an important topic in general, it becomes even more fundamental in the analysis of high-dimensional molecular data. Indeed, the high-dimensionality of this data leads to high sampling uncertainty as analyses are prone to overfitting and may be sensitive to the inclusion or exclusion of a few patients. Additionally, the complexity of high-dimensional molecular data leads to higher method uncertainty than might be expected in low-dimensional data as there are a great number of researcher degrees of freedom (Boulesteix et al., 2017b).

Moreover, in biometrical research, there are strong incentives for presenting work that entails new methods (Boulesteix, Binder, Abrahamowicz, & Sauerbrei, 2018). As a consequence, it is impossible for a researcher to keep pace with the multitude of methods being published every month in statistical journals (Sauerbrei et al., 2014). In contrast, the evaluation and the comparison of alternative methods is usually only given very little attention (Boulesteix, Wilson, & Hapfelmeier, 2017c). Given the multitude of methods and the lack of guidance based on evidence from neutral comparison studies, method uncertainty in the omics field is already very large and likely to further increase in the future. It can therefore be argued that we have to raise awareness of method uncertainty, which, in contrast to sampling uncertainty, has so far received only very limited attention from the statistical community. First of all, it seems to be essential to establish guidance based on neutral simulation studies, which may ultimately lead to a reduction in method uncertainty. As it is difficult to establish high-dimensional simulations that realistically reflect real data situations, simulation studies using real data sets as a basis (Rigaiil et al., 2016) as well as illustrations with real data are other important options for developing guidance.

Furthermore, it is important to develop and to promote approaches that handle the multiplicity of results from possible analysis strategies to enable researchers to systematically report method uncertainty in a given study. In this context, one can distinguish approaches that merely acknowledge method uncertainty, i.e., that measure the variability of results obtained when applying alternative analysis strategies, and approaches that can also account for method uncertainty: To acknowledge method uncertainty is a first step, but it is difficult to interpret results that merely acknowledge method uncertainty when the aim is to answer the initial question of interest. The next step is therefore to account for method uncertainty, i.e., to a method that allows the answering of the initial research question while considering the results of a set of reasonable analysis strategies. Consensus clustering, for instance, accounts for method uncertainty in clustering by combining the results of different clustering techniques. Additionally, it can be used to assess the stability of the discovered cluster solution. Similarly, model averaging approaches can be used to summarize the results of alternative models with weighting approaches that range from the use of posterior model probabilities in the context of Bayesian model averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999) to more pragmatic approaches where weights can be derived through bootstrap resampling (Augustin, Sauerbrei, & Schumacher, 2005; Holländer, Augustin, & Sauerbrei, 2006).

While these approaches might be somewhat difficult to implement and more resource intensive than classical approaches, ignoring method uncertainty and researcher degrees of freedom can lead to inflated effect sizes and type 1 error probabilities (Simmons et al., 2011). In contrast, approaches that adequately account for method uncertainty are more likely to produce findings that can be replicated in independent studies. In a “Post-Truth Era” (Mann, 2018; Vernon, 2017), where the research community and the general public have recently been rocked with the recognition that many research findings do not replicate on independent data sets (Begley & Ellis, 2012; Ioannidis et al., 2009; Open Science Collaboration, 2015), acknowledging and accounting for method uncertainty thus seem like important steps to reestablish the credibility of biomedical research, and of scientific evidence in general.

## ACKNOWLEDGEMENTS

This work was funded by the DFG (individual grants BO3139/4-2, BO3139/2-3 and BO3139/6-1 to ALB) and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors of this work take full responsibilities for the content. The authors thank Jenny Lee and Alethea Charlton for language corrections.

## CONFLICT OF INTEREST

The authors declare no conflict of interest.

## ORCID

Simon Klau  <https://orcid.org/0000-0002-7857-1263>

Anne-Laure Boulesteix  <https://orcid.org/0000-0002-2729-0947>

## REFERENCES

- Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*, R106.
- Augustin, N., Sauerbrei, W., & Schumacher, M. (2005). The practical utility of incorporating model selection uncertainty into prognostic models for survival data. *Statistical Modelling*, *5*, 95–118.
- Baty, F., Jaeger, D., Preiswerk, F., Schumacher, M. M., & Brutsche, M. H. (2008). Stability of gene contributions, & identification of outliers in multivariate analysis of microarray data. *BMC Bioinformatics*, *9*, 289.
- Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531–533.
- Binder, H., & Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, *9*, 14.
- Binder, H., & Schumacher, M. (2009). Incorporating pathway information into boosting estimation of high-dimensional risk prediction models. *BMC Bioinformatics*, *10*, 18.
- Boulesteix, A.-L., Binder, H., Abrahamowicz, M., & Sauerbrei, W. (2018). On the necessity and design of studies comparing statistical methods. *Biometrical Journal*, *60*, 216–218.
- Boulesteix, A.-L., De Bin, R., Jiang, X., & Fuchs, M. (2017a). IPF-LASSO: Integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*, *14*.
- Boulesteix, A.-L., & Fuchs, M. (2015). *ipflasso: Integrative Lasso with penalty factors*. R package version 0.1.

- Boulesteix, A.-L., Hornung, R., & Sauerbrei, W. (2017b). On fishing for significance and statistician's degree of freedom in the era of big molecular data. In W. Pietsch, J. Wernecke, & M. Ott (Eds.), *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data*, (pp. 155–170), Wiesbaden, Germany: Springer.
- Boulesteix, A.-L., & Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in Bioinformatics*, *10*, 556–568.
- Boulesteix, A.-L., Wilson, R., & Hapfelmeier, A. (2017c). 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.
- Braess, J., Kreuzer, K.-A., Spiekermann, K., Lindemann, H. W., Lengfelder, E., Graeven, U., ... Hiddemann, W. (2013). High efficacy and significantly shortened neutropenia of dose-dense S-HAM as compared to standard double induction: First results of a prospective randomized trial (AML-CG 2008). *Blood*, *122*, 619.
- Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.
- Büchner, T., Berdel, W. E., Schoch, C., Haferlach, T., Serve, H. L., & Kienast, J., ... Hiddemann, W. (2006). Double induction containing either two courses or one course of high-dose cytarabine plus mitoxantrone and postremission therapy by either autologous stem-cell transplantation or by prolonged maintenance for acute myeloid leukemia. *Journal of Clinical Oncology*, *24*, 2480–2489.
- Büchner, T., Krug, U., Gale, R. P., Heinecke, A., Sauerland, M., Haferlach, C., ... Hiddemann, W. (2016). Age, not therapy intensity, determines outcomes of adults with acute myeloid leukemia. *Leukemia*, *30*, 1781–1784.
- Critchlow, D. E. (1985). *Metric Methods for Analyzing Partially Ranked Data*. New York: Springer.
- De Bin, R., Janitz, S., Sauerbrei, W., & Boulesteix, A.-L. (2016). Subsampling versus bootstrapping in resampling-based model selection for multi-variable regression. *Biometrics*, *72*, 272–280.
- De Bin, R., & Sauerbrei, W. (2018). Handling co-dependence issues in resampling-based variable selection procedures: a simulation study. *Journal of Statistical Computation and Simulation*, *88*, 28–55.
- Dernoncourt, D., Hanczar, B., & Zucker, J.-D. (2014). Analysis of feature selection stability on high dimension and small sample data. *Computational Statistics & Data Analysis*, *71*, 681–693.
- Dessi, N., Pascariello, E., & Pes, B. (2013). A comparative analysis of biomarker selection techniques. *BioMed Research International*, *2013*, 1–10.
- Dutkowski, J., & Gambin, A. (2007). On consensus biomarker selection. *BMC Bioinformatics*, *8*, S5.
- Ein-Dor, L., Kela, I., Getz, G., Givol, D., & Domany, E. (2005). Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, *21*, 171–178.
- Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *National Academy of Sciences*, *103*, 5923–5928.
- Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*, 1–22.
- Gerds, T. A., Cai, T., & Schumacher, M. (2008). The performance of risk prediction models. *Biometrical Journal*, *50*, 457–479.
- Goldstein, B. A., Polley, E. C., & Briggs, F. B. (2011). Random forests for genetic association studies. *Statistical Applications in Genetics and Molecular Biology*, *10*, 32.
- Herold, T., Jurinovic, V., Batcha, A. M. N., Bamopoulos, S. A., Rothenberg-Thurley, M., Ksienzyk, B., ... Spiekermann K. (2017). A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. *Haematologica*, *103*, 456–465.
- Herold, T., Metzeler, K. H., Vosberg, S., Hartmann, L., Röllig, C., Stölzel, F., ... Greif, P. A. (2014). Isolated trisomy 13 defines a homogeneous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. *Blood*, *124*, 1304–1311.
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: a tutorial. *Statistical Science*, *14*, 382–401.
- Holländer, N., Augustin, N., & Sauerbrei, W. (2006). Investigation on the improvement of prediction by bootstrap model averaging. *Methods of Information in Medicine*, *45*, 44–50.
- Ioannidis, J. P., Allison, D. B., Ball, C. A., Coulibaly, I., Cui, X., Culhane, A. C., ... van Noort, V. (2009). Repeatability of published microarray gene expression analyses. *Nature Genetics*, *41*, 149–155.
- Janitz, S., Binder, H., & Boulesteix, A.-L. (2016). Pitfalls of hypothesis tests and model selection on bootstrap samples: Causes and consequences in biometrical applications. *Biometrical Journal*, *58*, 447–473.
- Klau, S., & Hornung, R. (2017). prioritylasso: Analyzing multiple omics data with an Offset approach. R package version 0.2.1.
- Klau, S., Jurinovic, V., Hornung, R., Herold, T., & Boulesteix, A.-L. (2018). Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, *19*, 322.
- Lausser, L., Müssel, C., Maucher, M., & Kestler, H. A. (2013). Measuring and visualizing the stability of biomarker selection techniques. *Computational Statistics*, *28*, 51–65.
- Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*, R29.
- Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*, 550.
- Mann, D. L. (2018). Fake news, alternative facts, and things that just are not true: Can science survive the post-truth era? *JACC: Basic to Translational Science*, *3*, 573–574.
- McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*, 4288–4297.
- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*, 417–473.

- Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: a multiple random validation strategy. *The Lancet*, *365*, 488–492.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.
- Patel, C. J., Burford, B., & Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*, 1046–1058.
- Probst, P. (2018). tuneRanger: Tune random forest of the ranger package. R package version 0.2.
- Probst, P., Wright, M., & Boulesteix, A.-L. (2018). Hyperparameters and tuning strategies for random forest. arXiv:1804.03515.
- Rigaill, G., Balzergue, S., Brunaud, V., Blondet, E., Rau, A., Rogier, O., ... Delannoy, E. (2016). Synthetic data sets for the identification of key ingredients for RNA-seq differential analysis. *Briefings in Bioinformatics*, *19*, 65–76.
- Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*, e47.
- Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*, 139–140.
- Sauerbrei, W., Abrahamowicz, M., Altman, D. G., le Cessie, S., Carpenter, J., & STRATOS initiative (2014). STRENGTHENING Analytical Thinking for Observational Studies: The STRATOS initiative. *Statistics in Medicine*, *33*, 5413–5432.
- Sauerbrei, W., Boulesteix, A.-L., & Binder, H. (2011). Stability investigations of multivariable regression models derived from low- and high-dimensional data. *Journal of Biopharmaceutical Statistics*, *21*, 1206–1231.
- Sauerbrei, W., & Schumacher, M. (1992). A bootstrap resampling procedure for model building: Application to the cox regression model. *Statistics in Medicine*, *11*, 2093–2109.
- Schumacher, M., Binder, H., & Gerds, T. (2007). Assessment of survival prediction models based on microarray data. *Bioinformatics*, *23*, 1768–1774.
- Shah, R. D., & Samworth, R. J. (2013). Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *75*, 55–80.
- Silberzahn, R., & Uhlman, E. L. (2015). Crowdsourced research: Many hands make tight work. *Nature*, *526*, 189–191.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.
- Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*, 1–13.
- Simonsohn, U., Simmons, J., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. <https://doi.org/10.2139/ssrn.2694998>.
- Steege, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*, 702–712.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*, 267–288.
- Vanschoren, J., van Rijn, J. N., Bischl, B., & Torgo, L. (2013). OpenML: networked science in machine learning. *SIGKDD Explorations*, *15*, 49–60.
- Vernon, J. L. (2017). Science in the post-truth era. *American Scientist*, *105*, 2.
- Wicherts, J. M., Veldkamp, C. L., Augusteyn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832.
- Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, *77*, 1–17.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *67*, 301–320.

## SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

**How to cite this article:** Klau S, Martin-Magniette M-L, Boulesteix A-L, Hoffmann S. Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. *Biometrical Journal*. 2019;1–18. <https://doi.org/10.1002/bimj.201800309>

## Appendix C

# Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology

### **This chapter is a reprint of:**

Klau, S., Schönbrodt, F. D., Patel, C. J., Ioannidis, J. P. A., Boulesteix, A.-L., and Hoffmann, S. (2020). Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. Technical Report 232, Ludwig-Maximilians-Universität München. doi: 10.5282/ubm/epub.70485

### **Copyright:**

### **Author contributions:**

S. Klau, S. Hoffmann and A.-L. Boulesteix developed the study concept. S. Hoffmann and S. Klau conducted the study and wrote the manuscript. S. Klau performed the statistical analysis. F. Schönbrodt provided some professional insights into psychological research. C. Patel, J. Ioannidis, F. Schönbrodt and A.-L. Boulesteix substantially contributed to the manuscript. All authors approved the final version.

### **Acknowledgments:**

This work was funded by the DFG (individual grant BO3139/4-3). The authors of this work take full responsibilities for its content. The authors thank Alethea Charlton for language corrections.

### **Supplementary material available at:**

<https://osf.io/59gf4/>





LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

INSTITUT FÜR STATISTIK



Simon Klau, Felix Schönbrodt, Chirag Patel, John Ioannidis,  
Anne-Laure Boulesteix, Sabine Hoffmann

# Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology

Technical Report Number 232, 2020  
Department of Statistics  
University of Munich

<http://www.statistik.uni-muenchen.de>



Comparing the vibration of effects due to model, data  
pre-processing and sampling uncertainty on a large data set in  
personality psychology

Simon Klau<sup>\*1</sup>, Felix D. Schönbrodt<sup>2,3</sup>, Chirag J. Patel<sup>4</sup>, John P.A. Ioannidis<sup>5,6,7,8</sup>,  
Anne-Laure Boulesteix<sup>1,3</sup>, and Sabine Hoffmann<sup>1,3</sup>

<sup>1</sup>Institute for Medical Information Processing, Biometry, and Epidemiology,  
Ludwig-Maximilians-Universität München, Munich, Germany

<sup>2</sup>Department of Psychology, Ludwig-Maximilians-Universität München, Munich, Germany

<sup>3</sup>LMU Open Science Center, Ludwig-Maximilians-Universität München, Munich, Germany

<sup>4</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

<sup>5</sup>Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA

<sup>6</sup>Department of Epidemiology and Population Health, Stanford University, Stanford, CA, USA

<sup>7</sup>Department of Biomedical Data Science, Stanford University, Stanford, CA, USA

<sup>8</sup>Department of Statistics, Stanford University, Stanford, CA, USA

February 5, 2020

---

\*Corresponding author: e-mail: [simon.klau@yahoo.de](mailto:simon.klau@yahoo.de), Department of Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Marchioninstr. 15, D-81377, Munich, Germany



## Abstract

Researchers have great flexibility in the analysis of observational data. If combined with selective reporting and pressure to publish, this flexibility can have devastating consequences on the validity of research findings. We extend the recently proposed vibration of effects approach to provide a framework comparing three main sources of uncertainty which lead to instability in observational associations, namely data pre-processing, model and sampling uncertainty. We analyze their behavior for varying sample sizes for two associations in personality psychology. While all types of vibration show a decrease for increasing sample sizes, data pre-processing and model vibration remain non-negligible, even for a sample of over 80000 participants. The increasing availability of large data sets that are not initially recorded for research purposes can make data pre-processing and model choices very influential. We therefore recommend the framework as a tool for the transparent reporting of the stability of research findings.

*Keywords*— metascience, researcher degrees of freedom, stability, replicability, Big Five

## 1 Introduction

In recent years, a series of attempts to replicate results of published research findings on independent data have shown that these replications tend to produce much weaker evidence than the original study (Open Science Collaboration, 2015), leading to what has been referred to as a ‘replication crisis’. While there have been a number of widely publicized examples of fraud and scientific misconduct (Ince, 2011; van der Zee, Anaya, & Brown, 2017), many researchers agree that this is not the major problem causing the crisis (Gelman & Loken, 2014; Ioannidis, Munafo, Fusar-Poli, Nosek, & David, 2014). Instead, the problems seem to be more subtle and partly due to the multiplicity of possible analysis strategies (Goodman, Fanelli, & Ioannidis, 2016; Open Science Collaboration, 2015). In this vein, there is evidence that the instability of observational associations can partly be explained by the fact that researchers tend to run several analysis strategies on a given data set, but to report only one of them selected post-hoc (Simmons, Nelson, & Simonsohn, 2011).

Indeed, there are a great number of implicit and explicit choices that have to be made when analyzing observational data. It is necessary to make various decisions when specifying a probability model to study the association between possible predictor variables and an outcome of interest. In addition to possible choices involved in the specification of a probability model, denoted as ‘model uncertainty’ in the following, there are numerous judgments and decisions that are required even before being able to fit the model to the data. When pre-processing the data, there are many possibilities regarding, not only

the definition of predictor and outcome variables, but also data inclusion and exclusion criteria, and the treatment of outliers (Wicherts et al., 2016). We denote this type of uncertainty as ‘data pre-processing uncertainty’.

Apart from the problems arising through the multiplicity of possible analysis strategies, there seem to be more fundamental issues in the analysis of observational data that originate from the low statistical power which characterizes many psychological studies (Maxwell, 2004; Szucs & Ioannidis, 2017). In psychology, effect sizes tend to be small and sample sizes are typically small to moderate. This combination leads to studies with low statistical power and therefore high sampling uncertainty when the same analysis strategies are applied to different samples with the aim of answering the same research question. High sampling uncertainty increases the false positive rate while simultaneously decreasing the chances of being able to replicate the results of studies that detect a true effect.

In recent years, a plethora of solutions to the replication crisis have been proposed in different disciplines. There are several approaches that allow the reporting of the results of a large number of possible analysis strategies (Muñoz & Young, 2018; Simonsohn, Simmons, & Nelson, 2015; Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016; Young, 2018), including the vibration of effects, proposed by Ioannidis (2008) and further developed by Patel, Burford, and Ioannidis (2015), and Palpacuer et al. (2019). Alternatively, the flexibility in the choice of analysis strategies can be reduced before analyzing the data through pre-registration and registered reports (Chambers, 2013; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). Similarly, the instability of observational findings arising from sampling uncertainty can be assessed through resampling (Meinshausen & Bühlmann, 2010; Sauerbrei, Boulesteix, & Binder, 2011) or sampling uncertainty can be reduced by increasing the sample size (Button et al., 2013; Maxwell, 2004; Schönbrodt & Perugini, 2013). While the solutions which have been proposed so far address important pieces of the problem by either focusing on the multiplicity of analysis strategies or on sampling uncertainty, it is important to be able to investigate sampling, model and data pre-processing uncertainty in a common framework to understand the full picture. Klau, Martin-Magniette, Boulesteix, and Hoffmann (2019) rely on a resampling procedure to compare method and sampling uncertainty, but focus their application on the selection and ranking of molecular biomarkers.

In this work, we use the vibration of effects approach (Ioannidis, 2008) to assess model, data pre-processing and sampling uncertainty in order to provide a tool for applied researchers to quantify and compare the instability of research findings arising from all three sources of uncertainty. We study this instability for varying sample sizes for two associations in personality psychology, namely between neuroticism and relationship status, and extraversion and physical activity, by analyzing a large and publicly available data set.

## 2 Methods

### 2.1 The data and research questions of interest

We use a large data set from the SAPA project personality test (Condon, Roney, & Revelle, 2017) which is publicly available at the Dataverse repository (<https://dataverse.harvard.edu/dataverse/SAPA-Project>). The sample consists of 126884 participants who were invited to complete an online survey between 2013 and 2017 in order to evaluate the structure of personality traits. The data set comprises information about a large pool of 696 personality items which were completed by the participants on a 6-point scale ranging from 1 (*very inaccurate*) to 6 (*very accurate*) and a set of additional variables including gender, age, country, job status, educational attainment level, physical activity, smoking status, relationship status and body mass index (BMI) of participants.

In this work, we use these data to assess the extent to which observational associations between the Big Five (agreeableness, conscientiousness, extraversion, neuroticism, openness to experience) and the variables physical activity, educational achievement, relationship status, smoking habits and obesity are influenced by data pre-processing, model and sampling uncertainty. In order to investigate the behavior of the three types of vibration with increasing sample size, we consider different subsets of the original data with subset sizes  $n \in \{500, 5000, 15000, 50000, 84543\}$ , where 84543 is the size of the complete data set after excluding participants with missing observations. Lower sample sizes than the original sample size were obtained by generating random subsamples from the original data set, without replacement. In the application of our framework, we consider six associations of interest, comprising five for which we found empirical evidence in the psychological literature. In the presentation of our results, we focus on the association between neuroticism and relationship status (Malouff, Thorsteinsson, Schutte, Bhullar, & Rooke, 2010) and between extraversion and physical activity (Rhodes & Smith, 2006). Additional results on the association between agreeableness and smoking (Malouff, Thorsteinsson, & Schutte, 2006), neuroticism and obesity (Gerlach, Herpertz, & Loeber, 2015), and conscientiousness and education (Sorić, Penezić, & Burić, 2017) can be found in the Supplementary Material, together with results on openness and physical activity, for which no evidence for an association could be found (Rhodes & Smith, 2006).

### 2.2 Quantifying the instability of observational associations due to different sources of uncertainty through the vibration of effects framework

We describe each association of interest through a logistic regression model in which we estimate the effect of the predictor of interest (e.g., neuroticism or extraversion) on the binary outcome of interest (e.g., relationship status or physical activity) to obtain odds ratios (OR) and corresponding p-values, while controlling for the effect of several covariates. As potential control variables, we consider all variables introduced in section 2.1 that are not part of the association of interest. For instance, the association between neuroticism and relationship status comprises the control variables age, gender, continent, job

status, BMI, smoking, education, physical activity, conscientiousness, agreeableness, extraversion and openness. For the association between physical activity and extraversion, we replace these two variables in the list of potential control variables with neuroticism and relationship status. This results in a total number of 12 control variables for each associations of interest.

We quantify the instability of these observational associations through the vibration of effects framework proposed by Ioannidis (2008). In the application of the framework by Patel et al. (2015), the authors consider the association between a predictor of interest and a survival outcome, and assess the vibration by defining a large number of models, resulting from the inclusion or exclusion of a number of potential control variables. To quantify the variability in these results, they calculate two summary measures, namely relative hazard ratios and relative p-values (RP). These summary measures are defined as the ratio of the 99th and 1st percentile of hazard ratios and the difference between the 99th and 1st percentile of  $-\log_{10}(\text{p-value})$ , respectively. Furthermore, the authors propose visualizing  $-\log_{10}(\text{p-values})$  and hazard ratios with volcano plots. These plots allow easy detection of a Janus effect, which is characterized by significant results in both positive and negative directions.

In this work, we will refer to the type of vibration investigated by Patel et al. (2015) as ‘model vibration’ and extend the framework to subsamples of the data and data pre-processing choices in order to compare model vibration to ‘sampling vibration’ and ‘data pre-processing vibration’, which we will introduce in more detail in sections 2.2.2 and 2.2.3. Following the proposal of Patel et al. (2015), we define the relative odds ratio (ROR) as the ratio of the 99th percentile and 1st percentile of the OR. The ROR provides a more robust and intuitive measure of variability than the variance. The minimal possible value of ROR is 1, indicating no vibration of effects at all, while larger ROR values indicate larger vibration.

### **2.2.1 Model vibration**

In order to assess model vibration for a given association of interest, we will consider a logistic regression model for which we take any possible combination of control variables into account. Following Patel et al. (2015), we will consider age and gender as baseline variables which are included in every model, resulting in a total number of  $2^{10} = 1024$  possible models for a given association of interest.

### **2.2.2 Sampling vibration**

To quantify sampling vibration, we use a resampling-based framework where we draw a large number of random subsets from our data set and fit the same logistic regression model on each of these subsets. In particular, we draw 1000 subsets of size  $0.5n$ , with  $n$  as the number of observations from the data sets defined in section 2.2, which comprise different numbers of observations themselves. Although each subset is drawn without replacement, the observations of subsets overlap between repetitions.

### 2.2.3 Data pre-processing vibration

The data pre-processing choices we are considering include the handling of outliers, eligibility criteria and the definition of predictor and outcome variables. These data pre-processing choices are based on studies found in the literature. For a given association of interest, we fit a logistic regression model for each data pre-processing strategy.

**Eligibility criteria** The eligibility criteria are based on the variables age, gender and the country of participants. For age, either the full group of participants is included in the analyses (definition 1) or a subgroup is defined by excluding participants who are younger than 18 (definition 2), which can be justified by their inability to legally provide consent (Barchard & Williams, 2008). Furthermore, studies about associations involving the Big Five personality traits are often carried out on subgroups of gender or countries, as was for instance shown by Malouff et al. (2006) and Malouff et al. (2010) for the variables smoking and physical activity. Therefore, with regards to the gender of participants, we either perform the analyses with all participants (definition 1), only with male participants (definition 2), or only with female participants (definition 3). Finally, we distinguish two alternative study populations based on the participants' country. Either all participants are included in the analyses and continent is considered as a categorical control variable (definition 1), or we include only participants from the United States, which presents the single largest country in the data set. In this case (definition 2), we exclude the control variable specifying the continent from the analyses. In total, this results in  $3 \cdot 2 \cdot 2 = 12$  possible combinations for the definition of eligibility criteria.

**Handling of outliers** A further data pre-processing choice is the handling of outliers. A variety of different outlier definitions can be found in the literature. Bakker and Wicherts (2014), for instance, provide a large range of z-values that are used to define outliers. Furthermore, it is either possible to remove or winsorize outlier values (Osborne & Overbay, 2004). Here, we focus on three different choices concerning all continuous covariates, comprising the five personality dimensions, as well as age and BMI: Firstly, we perform no further pre-processing with these covariates (definition 1). As a second option, we delete observations with absolute z-values greater than 2.5 (definition 2). Finally, we perform winsorization to achieve absolute z-values less than or equal 2.5 (definition 3). Thereby we replace values with  $z > 2.5$  by 2.5, and values with  $z < -2.5$  by  $-2.5$ .

**Dichotomization of outcome and covariates** In the definition of the outcome and covariates, we only consider the influence of different pre-processing choices for the three variables smoking, physical activity and education. All three variables are recorded with a certain number of categories (nine categories for smoking, six categories for physical activity and seven categories for education) and have to be dichotomized in order to be able to model them as a binary outcome in a logistic regression model. For all three variables, literature search revealed a lack of common definitions. For smoking and physical

activity for instance, summaries of these definitions are provided by Malouff et al. (2006) and Rhodes and Smith (2006), respectively. Similarly, the term education is very ambiguous, and even the more specific phrase of academic achievement exhibits a large variety of definitions (Fan & Chen, 2001). Therefore, we aim at reasonable dichotomizations of our given categories. For smoking, we either consider a definition based on never smokers vs. all other categories of smoking (definition 1) or based on non-smokers (never smokers and study participants who did not smoke the previous year) versus all other study participants (definition 2). For physical activity, we either assume a definition based on the two categories ‘less than once per week’ versus ‘once per week or more’ (definition 1) or, alternatively, ‘less than once per month’ versus ‘less than once per week or more’ (definition 2). Finally, in the definition of education we distinguish between study participants with a high level of education and study participants with a low level of education. In this distinction, we either assign current university students to the group with a high level of education (definition 1), because they will soon obtain a university degree or to the group with a low level of education (definition 2), as they have not obtained a degree yet. All other variables (job status, relationship status, BMI) are included in the analyses without considering alternative pre-processing choices. Therefore, we should acknowledge that the vibration of effects due to pre-processing choices can be larger than what is illustrated here. For more details on the variables which were collected in the SAPA project, we refer to Condon et al. (2017).

**Personality scores** The definitions of the five personality dimensions, i.e., openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism, are based on the corresponding personality items. There are a large number of different strategies to combine several items to a scale value. Indeed, the SAPA data set contains almost 700 items that were designed to assess personality, but each participant only completed a subset of these items. In order to determine a score on each of the personality dimensions, a correlation matrix, which is based on pairwise complete cases can be analyzed through factor analysis. As the Big Five personality traits were initially constructed as orthogonal factors (Saucier, 2002), we consider orthogonal rotation techniques as a first option (definition 1) for the factor analysis. However, Saucier (2002) argues that the scales used to measure the Big Five are not orthogonal in practice. In fact, a more common option in factor analysis of the personality traits is the use of oblique rotation techniques (definition 2). The assignment of items to the five personality dimensions can be realized by determining a minimal factor loading that has to be achieved to assign an item to a factor. Here, we either choose a minimal factor loading of 0.3 (definition 1) or of 0.4 (definition 2). The score of a participant can then be calculated by taking the mean score of all items that were assigned to a given factor. This strategy might lead to missing values for some participants on the personality dimensions as it is only reasonable to calculate such a score if there is a minimum number of completed items. Here, we use a required minimum value of 5 completed items.

While there are numerous analysis strategies to determine the personality score of a participant, it is not

in the scope of this study to consider all possible analysis strategies. Therefore, we limit the number of possible data pre-processing strategies by only considering the two choices: orthogonal vs. oblique rotation, and mean scores on items assigned to a factor with loadings greater than 0.3 or 0.4. While these variable definitions are based on the raw data set with all observations, the other data pre-processing choices are subsequently implemented on the data sets of different sizes.

The combination of the definition of personality scores with all other data pre-processing choices results in 1152 different data pre-processing strategies in total. These represent only a subset of a larger number of choices that may be made, in theory. However, in practical terms, they represent the main choices that are likely to be considered.

### **2.3 Comparing the vibration of effects due to different types of uncertainty**

For each association of interest, we quantify and compare model, data pre-processing and sampling uncertainty through the vibration of effects framework for varying sample sizes. In order to assess the variability in effect estimates and p-values for one type of vibration, the other types of vibration have to be fixed to a ‘favorite’ specification. For instance, if we focus on sampling vibration only, we need to decide on a favorite model as well as a favorite data pre-processing choice. As favorite data pre-processing choice, we consider data pre-processing without any subgroup analysis, without special handling of outliers, and with variable definition 1 for education, smoking and physical activity. Additionally, the favorite definition of the personality traits is performed with the oblique rotation technique and factor loadings greater than 0.3. Our favorite model choice simply consists in the model that contains all potential control variables. Furthermore, if the aim is to assess data pre-processing vibration or model vibration, we define the full data set as our favorite sample.

### **2.4 Comparing the vibration due to the choice of the analysis strategy with sampling vibration**

In addition to the investigation of individual types of vibration, we aim at quantifying the joint impact of model and data pre-processing choices on the variability of results. For simplicity, we will refer to the combination of a model and all necessary data pre-processing choices as analysis strategy. Correspondingly, this combination of choices results in  $1024 \times 1152 = 1179648$  analysis strategies. However, not every possible combination yields useful and valid results. For instance, when we consider the data pre-processing choice where the association of interest is only explored for participants from the US, the model including continent as a control variable is not valid. Thus, the total amount of feasible analysis strategies falls to 884736.

In the joint investigation of model and data pre-processing choices, the calculation of ROR is straightforward and can give an estimate for the absolute amount of vibration caused by the analysis strategy.

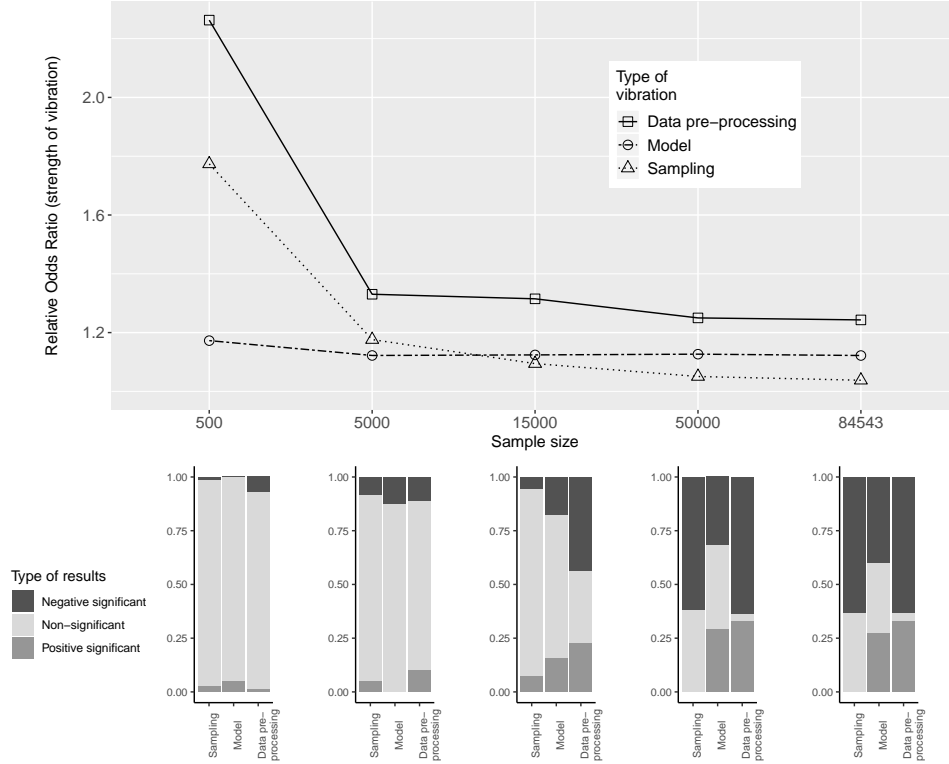


Figure 1: Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between neuroticism and relationship status.

Additionally, we quantify the relative impact of data pre-processing and model choices on the vibration that is caused by the choice of the analysis strategy. This is done by modelling  $\log(\text{OR})$ , corresponding to the regression coefficient of the predictor of interest, with two categorical covariates, indicating data pre-processing and model choices, in a linear model and by performing a variance decomposition through an analysis of variance (ANOVA).

### 3 Results

#### 3.1 The variability in effect estimates for one type of vibration

For more stable results, we repeat the analyses of all types of vibration for sample sizes of 500, 5000 and 15000 ten times and average the results across the obtained RORs. For the visualization of vibration patterns, however, we choose one representative plot out of the total number of ten. For a sample size of 50000, we consider the variability between RORs as negligible and run the analyses on only one sampled data set.

For the association between neuroticism and relationship status and the association between extraversion and physical activity, results of measures quantifying the variability in effect estimates for one type of vibration are visualized in Figures 1 and 2, respectively.



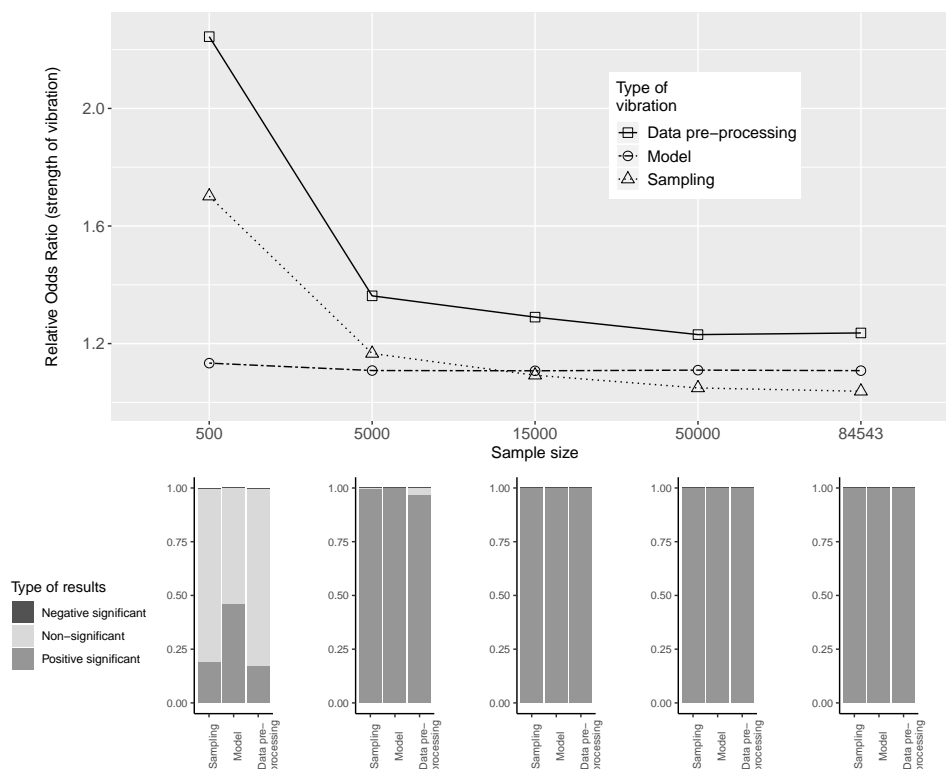


Figure 2: Data pre-processing, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association between extraversion and physical activity.

Corresponding figures for the other associations are provided in the Supplementary Material. In the upper panels, RORs are displayed against the sample size  $n$  for the three types of vibration (data pre-processing, model and sampling). For all investigated sample sizes, data pre-processing vibration is higher than model and sampling vibration for both associations of interest. For the lowest sample size of 500, high RORs can be observed, for instance of 2.26 for the association between neuroticism and relationship status. For larger sample sizes, data pre-processing vibration decreases and tends to a value of 1.24 for both associations of interest. A similar behavior can be observed for sampling vibration, but with consistently lower RORs. These tend to 1 for large sample sizes, in contrast to those RORs quantifying data pre-processing vibration. Therefore, the influence of a specific sample can be expected to be negligible for sufficiently large sample sizes. Compared to data pre-processing and sampling vibration, model vibration is less influenced by the sample size. Although we observe a slight decrease for RORs quantifying model vibration for increasing sample sizes, it is lower than sampling and data pre-processing vibration for small sample sizes and does not tend to a value of 1 for larger sample sizes.

In the lower panels of Figures 1 and 2, bar plots provide information about the percentage of significant results for each sample size and each type of vibration for the three categories: "negative significant", "non-significant", and "positive significant". For all three types of vibration, most results are not significant for a sample size of 500 while for the larger sample sizes the results are mostly significant: Here,

the association between neuroticism and relationship status shows a Janus effect with both negative and positive significant results for model and data pre-processing vibration. For sampling vibration, on the other hand, only negative-significant or non-significant effects can be observed for large sample sizes. For the association between extraversion and physical activity, all types of vibration yield positive significant effects for sample sizes larger than 5000, which is in accordance with the results from the literature (Rhodes & Smith, 2006). Hence, a Janus effect cannot be observed for this association.

These results are underlined by volcano plots (Figures 3 and 4), which contain exact patterns of  $-\log_{10}(\text{p-value})$  and ORs for three different sample sizes. Here, irregular patterns in data pre-processing vibration can be detected, which contrasts with the more consistent patterns of sampling and model vibration. A closer look at the data pre-processing vibration reveals that three clusters can be clearly distinguished, resulting from the pre-processing choice for the control variable gender. For male participants, neuroticism is associated with a committed relationship. On the other hand, there are two clusters with the opposite sign: Both female participants as well as the full data without subgroups for the variable gender are associated with a predominantly negative association between neuroticism and relationship status. The larger the sample size, the more clearly the clusters can be distinguished.

### 3.2 The relative impact of model and data pre-processing choices and the cumulative impact of both

Results for the total amount of vibration caused by model- and data pre-processing choices are visualized in Figure 5 for the association between neuroticism and relationship status, and Figure 6 for the association between extraversion and physical activity. In these figures, the top panels allow for a comparison of this joint vibration, also referred to as vibration due to the analysis strategy, and sampling vibration. The vibration caused by the analysis strategy is higher than sampling vibration for both associations. For a low sample size of  $n = 500$ , it is considerably higher than for larger sample sizes with RORs of 2.02 and 2.00 for the association between neuroticism and relationship status, and extraversion and physical activity, respectively. For a sample size of 5000, ROR values of 1.39 and 1.36 can be observed for these associations, which indicate lower vibration. For larger sample sizes, however, RORs do not show any further obvious decrease. For sample sizes greater than 500, the vibration remains in the range of ROR values of 1.34 and 1.40 for the association between neuroticism and relationship status. Similarly, the RORs for sample sizes greater than 500 are in the range of 1.27 and 1.36 for the association between extraversion and physical activity.

Pie charts in the bottom panels illustrate the relative impact of model and data pre-processing choices on the total vibration caused by the choice of the analysis strategy. Due to the high computational burden of the variance decomposition, we randomly select three of the ten data sets for low sample sizes of 500, 5000 and 15000 to estimate the relative impact of data pre-processing and model choices and average the results over the three selected data sets. For both associations, the relative impact of data pre-processing

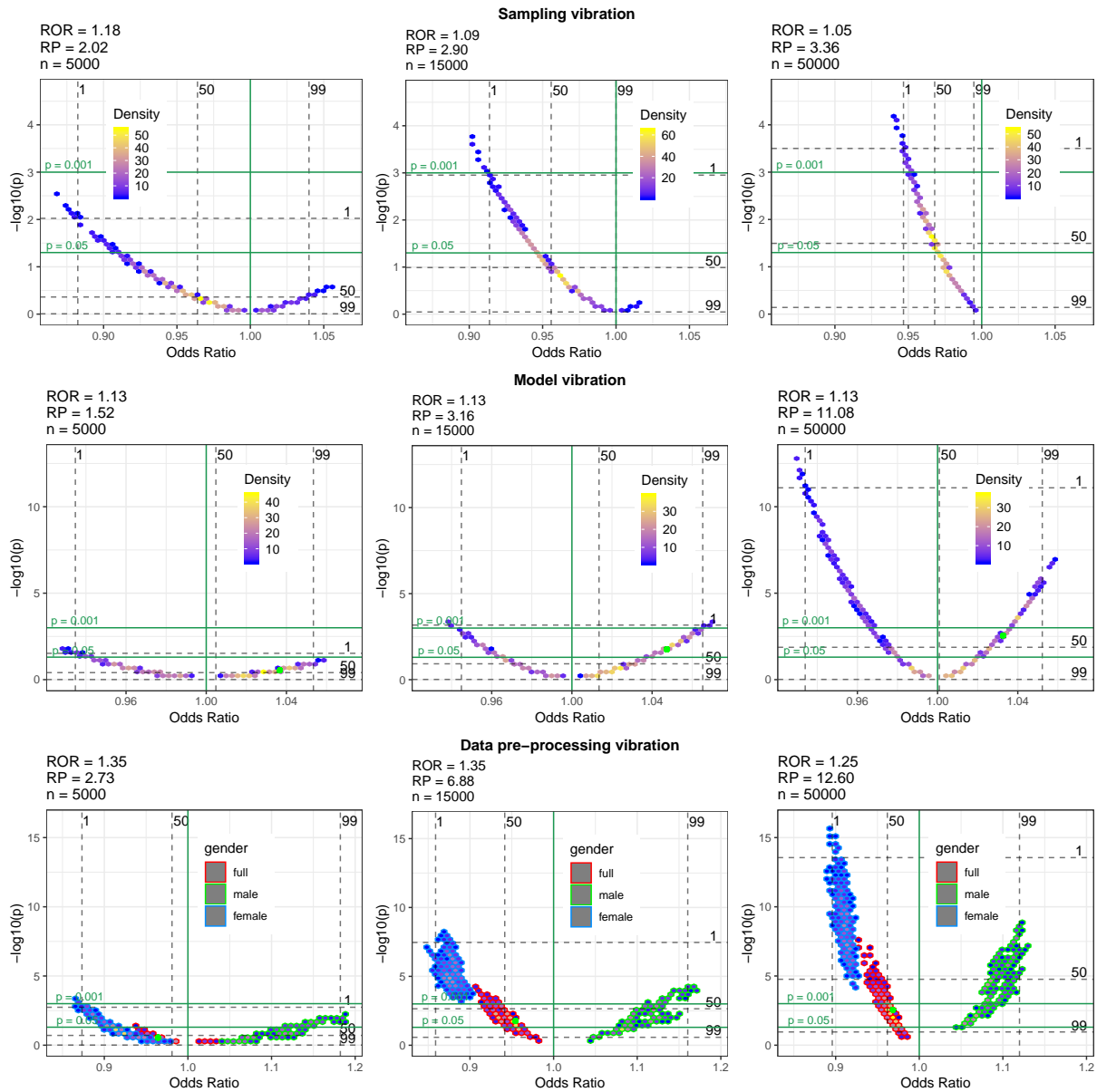


Figure 3: Volcano plots for different types of vibration and different sample sizes ( $n$ ) for the association between neuroticism and relationship status. The summary measures ROR and RP indicate relative odds ratios and relative p-values, respectively. Green dots indicate results obtained with favorite model choices (middle row) and favorite data pre-processing choices (bottom row).

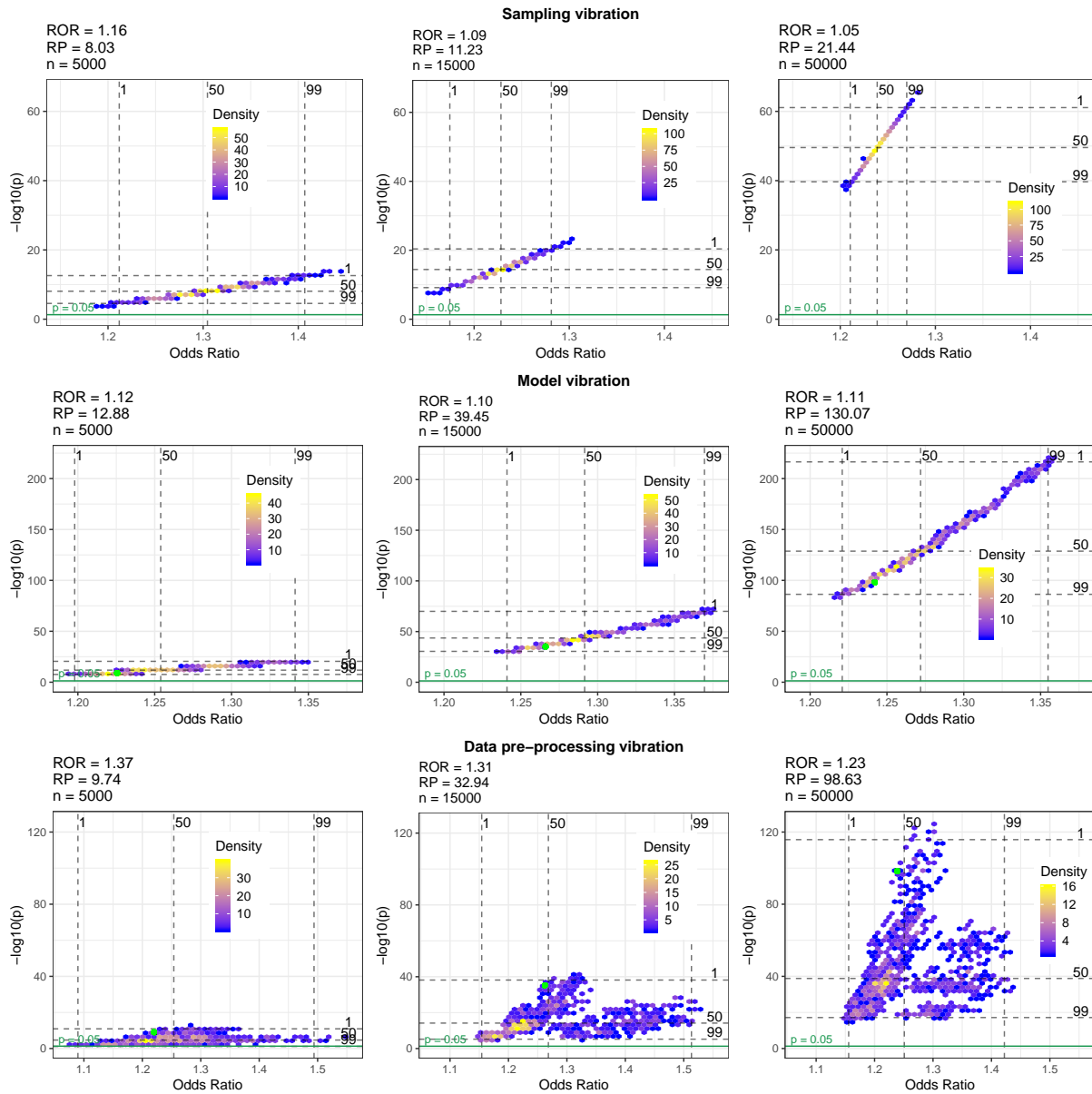


Figure 4: Volcano plots for different types of vibration and different sample sizes ( $n$ ) for the association between extraversion and physical activity. The summary measures ROR and RP indicate relative odds ratios and relative p-values, respectively. Green dots indicate results obtained with favorite model choices (middle row) and favorite data pre-processing choices (bottom row).

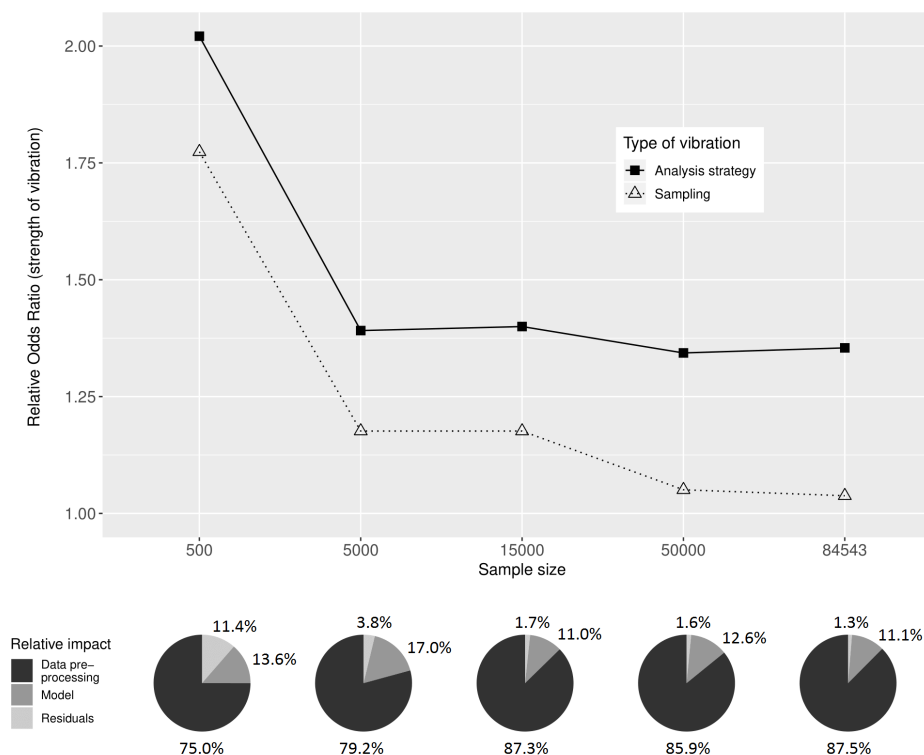


Figure 5: Cumulative model and data pre-processing vibration (‘analysis strategy’) compared to sampling vibration (top panel), and relative impact of model and data pre-processing vibration for different sample sizes (bottom panel) for the association between neuroticism and relationship status.

choices by far exceeds the impact of model vibration. Indeed, at most 22.2% of the total vibration due to the analysis strategy can be explained by model choices for the association between extraversion and physical activity. For the association between neuroticism and relationship status, the relative model impact is even lower, with a maximum value of 17%.

A more detailed investigation of data pre-processing vibration as part of the total vibration shows that the variable gender has the largest impact of the data pre-processing choices on the vibration of effects for the both associations of interest. Indeed, for the association between neuroticism and relationship status, 86% of data pre-processing vibration can be explained by the impact of gender for the largest sample size, which is in accordance with Figure 5. For the association between extraversion and physical activity, the relative impact of gender on data pre-processing vibration is 59.2% for the full data set.

## 4 Discussion

### 4.1 Summary

Researchers have great flexibility in the analysis of observational data. If this flexibility is combined with selective reporting and pressure to publish significant results, it can have devastating consequences on the replicability of research findings. In this work, we extended the vibration of effects approach, proposed

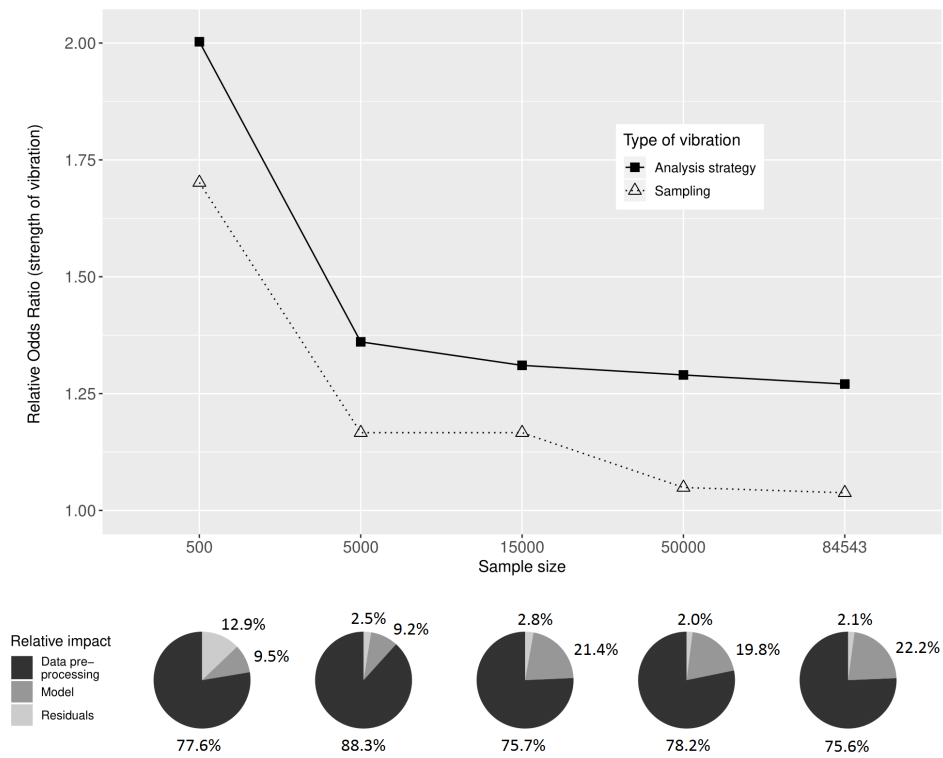


Figure 6: Cumulative model and data pre-processing vibration (‘analysis strategy’) compared to sampling vibration (top panel), and relative impact of model and data pre-processing vibration for different sample sizes (bottom panel) for the association between extraversion and physical activity.

by Ioannidis (2008), to quantify and compare the impact of model and data pre-processing choices on the stability of observational associations. Through this extension, the vibration of effects framework allows assessment of the extent to which the instability of research findings in observational studies can be explained by the choice of analysis strategy and enables comparison of the relative impact of different choices with sampling uncertainty.

We illustrated three different types of vibration on the SAPA data set, considering reasonable data pre-processing choices and modeling strategies based on a logistic regression model, focusing on two associations of interest in personality psychology. In addition, we quantified sampling vibration by considering the results obtained from random subsets of the data set in use. We found that data pre-processing vibration was higher than model and sampling vibration for all sample sizes considered in our analyses. For high sample sizes, sampling vibration decreased and became negligible, while model and data pre-processing vibration showed an initial decrease with increasing sample size and then remained constantly non-negligible. When considering all possible combinations of model and data pre-processing choices to compare the relative impact of each source of uncertainty, we found that data pre-processing choices explained by far more variability in results than model choices.

## 4.2 Limitations

When interpreting our results, it is important to keep in mind that both model vibration and data pre-processing vibration are in reality rather elusive concepts as they critically depend on the number and the type of analysis strategies under consideration. In theory, there are an infinite number of models and an infinite number of possible data pre-processing strategies, so any attempt to quantify the variability in an effect estimate resulting from every possible analysis strategy is doomed to fail. As it is futile to quantify the vibration in results arising from every possible strategy, we decided to focus on reasonable analysis strategies, i.e., those that could have been selected in an actual research project. Following Patel et al. (2015), we merely focused on a special type of model vibration, namely the vibration of effects that is due to the inclusion or exclusion of all potential control variables. Vibration of effects may be larger in situations where very complex models are involved, encompassing a very large number of control variables. Conversely, it may have less of an impact in data-poor studies with few variables measured and considered. Furthermore, we only considered linear effects and did not examine interaction terms, which may be essential in some settings.

Finally, we considered a number of possible data pre-processing strategies that is comparable to the number of models in order to allow a fair comparison of data pre-processing uncertainty and model uncertainty. As the combination of model and data pre-processing choices was in the order of magnitude of one million, it would not have been feasible from a computational point of view to consider a larger combination of models and data pre-processing strategies. As a consequence, we have to be careful when generalizing the findings of our study to other data sets and applications. While there is a firm theoretical

basis to predict sampling vibration, the behavior of model and data pre-processing vibration critically depends on the particular data set and the number of possible choices under consideration. Efforts to standardize analytical options are underway in some scientific fields building consensus among investigators and these efforts may result in diminishing the space for potential vibration of effects.

While the number of analysis strategies we considered in this work was limited by the computational feasibility of our analyses, it has to be noted that this number might in principle be reduced by only selecting those models that show a reasonable fit to the data. In the vibration of effects framework, the results of all possible models are reported, regardless of the fit of these models. In this respect, the vibration of effects framework differs from other approaches like Bayesian Model Averaging (Hoeting, Madigan, Raftery, & Volinsky, 1999), where a single summary measure is obtained that accounts for model uncertainty by weighting every model under consideration by its probability of being the true model. On the other hand, the vibration of effects framework shows greater flexibility than Bayesian Model Averaging as it can report the results of not only different models, but also different data pre-processing choices. Contrary to the choice of a model, data pre-processing choices may often be based on untestable assumptions, concerning for instance the nature of outlying observations, or they may arise because scientific theories are generally not precise enough to allow for a one-to-one mapping to statistical hypotheses (Steege et al., 2016). Contrary to model uncertainty, data pre-processing uncertainty therefore cannot be reduced by comparing the fit of different data pre-processing strategies to the data, but only through conceptual rigor (Schaller, 2016) and the standardization of experimental conditions (Elson, Mohseni, Breuer, Scharkow, & Quandt, 2014).

### 4.3 Conclusion and Outlook

When analyzing observational data, it is necessary to make model and data pre-processing choices which rely on many explicit and implicit assumptions. The vibration of effects framework provides investigators with a tool to quantify the impact of these choices on the stability of observational associations, helping them focus their attention on the choices that have the most influence and are therefore worth further investigation or discussion. To establish it as a tool, we recommend visualizing data pre-processing, model and sampling vibration with volcano plots as we have demonstrated in the Supplementary Material for the association between neuroticism and relationship status. The corresponding analysis took 21.6 minutes on a 64-bit Debian GNU/Linux 10 system with Intel Xeon CPU E5-2640. Moreover, the systematic reporting of RORs and p-value characteristics for these types of vibration is a simple but informative guideline for quantifying the stability of published results. The framework can also be useful for readers in the interpretation of these results: When used as a tool to report the robustness of observational associations, it helps readers (including reviewers) to interpret these results in the context of all the possible results that could have been obtained with alternative, equally justified analysis strategies. When the research data of a publication are made publicly available, which is more and more common to



enhance transparency, a reader can use the vibration of effects framework to assess the extent to which the originally reported results are fragile or incredible because they depend on very specific analytical decisions. In this vein, it is possible to specify a number of model and data pre-processing choices and to apply the framework to assess the variability in effect estimates arising from these possible analysis strategies. In our application of the framework in personality psychology, we observed many cases in which both significant and non-significant results could be obtained, depending on the choice of the analysis strategy. In extreme cases, it was even possible to obtain both positive and negative significant associations and this phenomenon persisted for a very large sample size of over 80000 participants.

The number of decisions which have to be made in the analysis of observational data becomes even more important when analyzing data that are not initially recorded for research purposes. While the increasing availability of large data sets, for instance in the form of Twitter accounts (Barberá, Jost, Nagler, Tucker, & Bonneau, 2015) or transaction data (Gladstone, Matz, & Lemaire, 2019), offer unprecedented opportunities to study complex phenomena of interest, they also increase the number of untestable assumptions which must be made in the data pre-processing and choice of model used to describe the data. In light of our results, we suggest using the vibration of effects framework as a tool to assess the robustness of conclusions from observational data.

## Acknowledgements

This work was funded by the DFG (individual grant BO3139/4-3). The authors of this work take full responsibilities for its content. The authors thank Alethea Charlton for language corrections.

## Bibliography

- Bakker, M., & Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type I error rate in independent samples t tests: The power of alternatives and recommendations. *Psychological Methods, 19*(3), 409–427. doi: 10.1037/met0000014
- Barberá, P., Jost, J. T., Nagler, J., Tucker, J. A., & Bonneau, R. (2015). Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological Science, 26*(10), 1531–1542. doi: 10.1177/0956797615594620
- Barchard, K. A., & Williams, J. (2008). Practical advice for conducting ethical online experiments and questionnaires for United States psychologists. *Behavior Research Methods, 40*(4), 1111–1128. doi: 10.3758/BRM.40.4.1111
- Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability

- of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. doi: 10.1038/nrn3475
- Chambers, C. D. (2013). Registered reports: A new publishing initiative at cortex. *Cortex*, *49*(3), 609–610. doi: 10.1016/j.cortex.2012.12.016
- Condon, D., Roney, E., & Revelle, E. (2017). A SAPA project update: On the structure of phrased self-report personality items. *Journal of Open Psychology Data*, *5*(1), 3. doi: 10.5334/jopd.32
- Elson, M., Mohseni, M. R., Breuer, J., Scharnow, M., & Quandt, T. (2014). Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research. *Psychological Assessment*, *26*(2), 419–432. doi: 10.1037/a0035569
- Fan, X., & Chen, M. (2001). Parental involvement and students' academic achievement: A meta-analysis. *Educational Psychology Review*, *13*(1), 1–22. doi: 10.1023/A:1009048817385
- Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*(6), 460–465.
- Gerlach, G., Herpertz, S., & Loeber, S. (2015). Personality traits and obesity: A systematic review. *Obesity Reviews*, *16*(1), 32–63. doi: 10.1111/obr.12235
- Gladstone, J. J., Matz, S. C., & Lemaire, A. (2019). Can psychological traits be inferred from spending? Evidence from transaction data. *Psychological Science*, *30*(7), 1087–1096. doi: 10.1177/0956797619849435
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, *8*(341), 341ps12–341ps12. doi: 10.1126/scitranslmed.aaf5027
- Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, *14*(4), 382–401.
- Ince, D. (2011). The duke university scandal – what can be done? *Significance*, *8*(3), 113–115. doi: 10.1111/j.1740-9713.2011.00505.x
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. doi: 10.1097/EDE.0b013e31818131e7
- Ioannidis, J. P. A., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. doi: 10.1016/j.tics.2014.02.010
- Klau, S., Martin-Magniette, M.-L., Boulesteix, A.-L., & Hoffmann, S. (2019). Sampling uncertainty versus method uncertainty: A general framework with applications to omics

- biomarker selection. *Biometrical Journal*, 1–18. doi: 10.1002/bimj.201800309
- Malouff, J. M., Thorsteinsson, E. B., & Schutte, N. S. (2006). The five-factor model of personality and smoking: A meta-analysis. *Journal of Drug Education*, 36(1), 47–58. doi: 10.2190/9EP8-17P8-EKG7-66AD
- Malouff, J. M., Thorsteinsson, E. B., Schutte, N. S., Bhullar, N., & Rooke, S. E. (2010). The five-factor model of personality and relationship satisfaction of intimate partners: A meta-analysis. *Journal of Research in Personality*, 44(1), 124–127. doi: 10.1016/j.jrp.2009.09.004
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2), 147–163. doi: 10.1037/1082-989X.9.2.147
- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 417–473. doi: 10.1111/j.1467-9868.2010.00740.x
- Muñoz, J., & Young, C. (2018). We ran 9 billion regressions: Eliminating false positives through computational model robustness. *Sociological Methodology*, 48(1), 1–33. doi: 10.1177/0081175018777988
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. doi: 10.1126/science.aac4716
- Osborne, J. W., & Overbay, A. (2004). The power of outliers (and why researchers should always check for them). *Practical Assessment, Research & Evaluation*, 9(6), 1–8.
- Palpacuer, C., Hammas, K., Duprez, R., Laviolle, B., Ioannidis, J. P. A., & Naudet, F. (2019). Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. *BMC Medicine*, 17(174), 1–13. doi: 10.1186/s12916-019-1409-3
- Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, 68(9), 1046–1058. doi: 10.1016/j.jclinepi.2015.05.029
- Rhodes, R. E., & Smith, N. E. I. (2006). Personality correlates of physical activity: A review and meta-analysis. *British Journal of Sports Medicine*, 40(12), 958–965. doi: 10.1136/bjism.2006.028860
- Saucier, G. (2002). Orthogonal markers for orthogonal factors: The case of the Big Five. *Journal*

- of Research in Personality*, 36(1), 1–31. doi: 10.1006/jrpe.2001.2335
- Sauerbrei, W., Boulesteix, A.-L., & Binder, H. (2011). Stability investigations of multivariable regression models derived from low-and high-dimensional data. *Journal of Biopharmaceutical Statistics*, 21(6), 1206–1231. doi: 10.1080/10543406.2011.629890
- Schaller, M. (2016). The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). *Journal of Experimental Social Psychology*, 66, 107–115. doi: 10.1016/j.jesp.2015.09.006
- Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, 47(5), 609–612. doi: 10.1016/j.jrp.2013.05.009
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. doi: 10.1177/0956797611417632
- Simonsohn, U., Simmons, J., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications.  
doi: 10.2139/ssrn.2694998
- Sorić, I., Penezić, Z., & Burić, I. (2017). The Big Five personality traits, goal orientations, and academic achievement. *Learning and Individual Differences*, 54, 126–134. doi: 10.1016/j.lindif.2017.01.024
- Stegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, 11(5), 702–712. doi: 10.1177/1745691616658637
- Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLOS Biology*, 15(3), 1–18. doi: 10.1371/journal.pbio.2000797
- van der Zee, T., Anaya, J., & Brown, N. J. (2017). Statistical heartburn: An attempt to digest four pizza publications from the cornell food and brand lab. *BMC Nutrition*, 3(54), 1–15. doi: 10.1186/s40795-017-0167-x
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7(6), 632–638. doi: 10.1177/1745691612463078
- Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., &

- van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, 7(1832), 1–12. doi: 10.3389/fpsyg.2016.01832
- Young, C. (2018). Model uncertainty and the crisis in science. *Socius*, 4, 1–7. doi: 10.1177/2378023117737206



## Appendix D

# Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework

### This chapter is a reprint of:

Klau\*, S., Hoffmann\*, S., Patel, C. J., Ioannidis, J. P. A. and Boulesteix, A.-L. (2020). Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. Technical Report 233, Ludwig-Maximilians-Universität München. doi: 10.5282/ubm/epub.70493 (accepted at the *International Journal of Epidemiology*)

### Copyright:

### Author contributions:

S. Klau, S. Hoffmann and A.-L. Boulesteix developed the study concept. S. Hoffmann and S. Klau conducted the study and wrote the manuscript. S. Klau performed the statistical analysis. C. Patel, J. Ioannidis, and A.-L. Boulesteix substantially contributed to the manuscript. All authors approved the final version.

\*S. Klau and S. Hoffmann contributed equally to this work.

### Acknowledgments:

This work was funded by the DFG (individual grant BO3139/4-3) and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors thank Alethea Charlton for language corrections.

### Supplementary material available at:

[https://github.com/simonsimon01/VoE\\_nhanes](https://github.com/simonsimon01/VoE_nhanes)







LUDWIG-  
MAXIMILIANS-  
UNIVERSITÄT  
MÜNCHEN

INSTITUT FÜR STATISTIK



Simon Klau, Sabine Hoffmann, Chirag Patel, John Ioannidis,  
Anne-Laure Boulesteix

# Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework

Technical Report Number 233, 2020  
Department of Statistics  
University of Munich

<http://www.statistik.uni-muenchen.de>



Examining the robustness of observational associations  
to model, measurement and sampling uncertainty with the  
vibration of effects framework

Simon Klau<sup>\*†1</sup>, Sabine Hoffmann<sup>†1,2</sup>, Chirag J. Patel<sup>3</sup>, John P.A. Ioannidis<sup>4,5,6,7,8</sup>, and  
Anne-Laure Boulesteix<sup>1,2</sup>

<sup>1</sup>Institute for Medical Information Processing, Biometry, and Epidemiology,  
Ludwig-Maximilians-Universität München, Munich, Germany

<sup>2</sup>LMU Open Science Center, Ludwig-Maximilians-Universität München, Munich, Germany

<sup>3</sup>Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA

<sup>4</sup>Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, CA, USA

<sup>5</sup>Department of Epidemiology and Population Health, Stanford University School of Medicine,  
Stanford, CA, USA

<sup>6</sup>Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA

<sup>7</sup>Department of Statistics, Stanford University School of Humanities and Sciences, Stanford, CA, USA

<sup>8</sup>Department of Medicine, Stanford University School of Medicine, Stanford, CA, USA

February 5, 2020

---

\*Corresponding author: e-mail: [simon.klau@yahoo.de](mailto:simon.klau@yahoo.de), Department of Medical Information Processing, Biometry and Epidemiology, Ludwig-Maximilians-Universität München, Marchioninstr. 15, D-81377, Munich, Germany

<sup>†</sup>These authors contributed equally to this work.

## Abstract

*Background:* The results of studies on observational associations may vary depending on the study design and analysis choices as well as due to measurement error. It is important to understand the relative contribution of different factors towards generating variable results, including low sample sizes, researchers' flexibility in model choices, and measurement error in variables of interest and adjustment variables.

*Methods:* We define sampling, model and measurement uncertainty, and extend the concept of vibration of effects in order to study these three types of uncertainty in a common framework. In a practical application, we examine these types of uncertainty in a Cox model using data from the National Health and Nutrition Examination Survey. In addition, we analyze the behavior of sampling, model and measurement uncertainty for varying sample sizes in a simulation study.

*Results:* All types of uncertainty are associated with a potentially large variability in effect estimates. Measurement error in the variable of interest attenuates the true effect in most cases, but can occasionally lead to overestimation. When we consider measurement error in both the variable of interest and adjustment variables, the vibration of effects are even less predictable as both systematic under- and overestimation of the true effect can be observed. The results on simulated data show that measurement and model vibration remain non-negligible even for large sample sizes.

*Conclusion:* Sampling, model and measurement uncertainty can have important consequences on the stability of observational associations. We recommend systematically studying and reporting these types of uncertainty, and comparing them in a common framework.

**Keywords**— measurement error, metascience, observational study, replicability, researcher degrees of freedom, stability

# 1 Introduction

Observational associations in epidemiology can be unstable and occasionally difficult to replicate in subsequent studies [1, 2, 3, 4]. The instability sometimes leads to contradictory findings from similar epidemiological studies, raising challenges to the interpretation and credibility of epidemiological evidence [5].

There are many factors which contribute to this instability: Small sample sizes may lead to high instability in the estimates of the magnitude of an association and its statistical significance. Another key factor which may play an important role in the instability of research findings in epidemiology includes diverse model specification choices, such as which variables are adjusted for. As we have shown in earlier research, the inclusion and exclusion of potential adjustment variables, can cause a large variability in results when estimating the association between an exposure and an outcome variable of interest using a given data set [6]. Finally, measurement error in exposure and outcome variables may further exacerbate the instability of observational associations.

While sampling uncertainty is classically accounted for when deriving p-values and confidence intervals to report the results of epidemiological studies, methods to account for model and measurement uncertainty are not commonly used when analyzing observational data. Instead, results are usually presented as if the chosen model were the only possible model, even though different authors may consider very different sets of adjustment variables when analyzing the same research question of interest [6, 7]. The large majority of observational analyses are not pre-registered and do not have explicitly pre-specified analysis plans [8]. Concerning measurement error, there is a widespread and persistent belief that the effects of exposure measurement error and exposure misclassification are relatively benign, as they will merely result in a bias in parameter estimates towards the null and loss in statistical power [9, 10, 11]. However, these presumed consequences of exposure measurement error and exposure misclassification, which are sometimes mentioned in the discussion of epidemiological findings to argue that an observed association may potentially have been underestimated, only hold in cases where the variable of interest is the only co-variate in the model which is measured with error. If the included adjustment variables are also subject to measurement error, which is almost always the case in epidemiological studies, it is more difficult to predict whether measurement error will attenuate or inflate risk estimates [12, 13, 14, 15].

Due to the multiplicity of possible analysis strategies, the relatively small sample sizes of many epidemiological studies and the ubiquity of measurement error, model, sampling and measurement uncertainty all appear to play important roles in the instability of observational associations and may contribute to the non-replicability of research findings. It would be interesting to quantify and compare these different sources of uncertainty in a common framework.

The aim of this work is to extend the vibration of effects approach [7], which we previously used to assess model and sampling uncertainty [6, 16], to measurement uncertainty in order to provide a tool to investigate the robustness of observational associations to these three types of uncertainty.

We will illustrate this approach with data from the National Health and Nutrition Examination Survey (NHANES) and consider three different scenarios for measurement vibration. In the first scenario, we introduce measurement error only in the exposure of interest. This type of error is expected to reduce the strength of the association. Secondly, we introduce measurement error only in the adjustment variables. This second scenario occurs in practice when there are special efforts being made to reduce measurement error to a minimum for the exposure of interest or if a method for measurement error correction has been applied to account for measurement error in this variable. Finally, we consider a more realistic scenario for measurement error where error is present both in the variable of interest and in the adjustment variables. Additionally, we compare measurement vibration with model vibration and sampling vibration. We complement the analyses on real data with results on simulated data to investigate the behavior of the three types of vibration for increasing sample sizes.

## 2 Methods

### 2.1 Model and sampling vibration

We previously introduced the concept of vibration of effects to quantify the variability in results when studying an association of interest under a broad range of model specifications [7]. The idea of this approach is to quantify the variability of results through a vibration ratio, which we defined as the ratio of the largest versus smallest effect estimate for the same association of interest under different analysis choices. Moreover, we applied this framework to assess the vibration of effects arising through the specification of the probability model to data from the NHANES [6]. We showed that this type of vibration, which we obtained through the inclusion or exclusion of all potential adjustment variables, can have important consequences on the estimation of the effect of the variable of interest on all-cause mortality in a Cox regression. The vibration ratios used were the relative hazard ratio (RHR) and the relative p-value (RP). In this second study, these vibration ratios describe the ratio of the 99<sup>th</sup> and 1<sup>st</sup> percentile of hazard ratios and the difference between the 99<sup>th</sup> and 1<sup>st</sup> percentile of  $-\log_{10}(\text{p-value})$ , respectively. In addition, we suggested showing volcano plots with p-values at the y-axis and effect estimates at the x-axis. These volcano plots allow easy detection of patterns like the Janus pattern, which is characterized by significant estimates in both a positive and negative direction.

Furthermore, we previously applied the vibration of effects framework when fitting the same model on different subsamples of the data [16], and compared this type of vibration, denoted as ‘sampling vibration’ in the following, with ‘model vibration’ as assessed in [6]. When studying sampling vibration, a favorite model has to be chosen from all models considered in the assessment of model vibration. For this model, we suggested drawing a large number of  $B$  random subsets of the data and fitting the same statistical model on each of these subsets [16]. Similar to model vibration, vibration ratios and volcano

plots can be used to illustrate sampling vibration.

## 2.2 Measurement vibration

In this work, we suggest further extending the vibration of effects framework to illustrate measurement uncertainty. For continuous variables, we focus on an additive classical non-differential measurement error model  $Z = X + U$ , where  $Z$  is the observed exposure,  $X$  is the true exposure, and  $U$  is a measurement error term, which is independent of the true exposure  $X$ . Measurement error for a continuous variable can be assessed by quantifying the correlation  $\rho_{XZ}$  between true exposure  $X$  and observed exposure  $Z$  in a calibration sample. For binary variables, the magnitude of misclassification can be quantified through sensitivity and specificity values.

Following [11], we consider the observed values for a given variable to be the true exposure values  $X$ . We can generate virtual error-prone observed values  $Z$  for continuous variables based on a given correlation  $\rho_{XZ}$  as follows. As shown in the Supplement, we first calculate the variance of observed exposure  $Z$  as the variance of true exposure  $X$  divided by  $\rho_{XZ}^2$ . We can then determine the measurement error variance by subtracting the variance of  $X$  from the variance of  $Z$ :

$$\text{Var}(U) = \frac{\text{Var}(X)}{\rho_{XZ}^2} - \text{Var}(X). \quad (1)$$

As a final step, to obtain observed exposure  $Z$ , measurement error values  $U$  can be generated from a normal distribution with mean zero and variance  $\text{Var}(U)$ , and added to the true exposure  $X$ .

Furthermore, we suggest adding exposure misclassification to binary variables by using values for sensitivity and specificity. In particular, for a binary variable with observed values 0 or 1, all values of 1 can be replaced by random values from a binomial distribution with a probability of success that is equal to the sensitivity. Similarly, all values of 0 can be replaced by random values from a binomial distribution with a probability of success equal to  $1 - \text{specificity}$ . As shown in the Supplement, for ordinal variables, we follow a strategy which is similar to the simulation strategy for continuous variables by assuming latent variables which follow a normal distribution.

Similarly to sampling vibration, we have to choose a favorite model among the models that are considered in the assessment of model vibration. For this model, we repeat the procedure of adding random measurement error a number of  $B$  times. With  $B$  different results obtained by adding measurement error to the variables, the vibration of effects framework can be used. To quantify the results, we suggest using the 99<sup>th</sup> and 1<sup>st</sup> percentiles of effect estimates and p-values as vibration ratios to define relative effect estimates and relative p-values, similar to model and sampling vibration. Moreover, these results can be visualized with volcano plots.

## 2.3 The National Health and Nutrition Examination Survey cohort data

### 2.3.1 Data set description

We analyze cohorts from the NHANES, modeling all-cause mortality with a variable of interest and 15 adjustment variables (for more details on data collection and pre-processing see [6]). For this work, we run the analyses successively with 30 variables of interest, which were chosen from a pool of 417 variables. We selected these 30 variables due to a small amount of missing values ( $< 15\%$ ), and, for ease of interpretation, ensured that they were either binary or continuous. For illustrative purposes, out of these thirty variables, we will limit the presentation of results to two continuous variables of interest (thigh circumference and HDL-cholesterol), as well as the two binary variables diabetes (defined as self-reported doctor diagnosed diabetes and fasting glucose  $> 125$  mg/dl) and heart disease (defined as self-reported doctor diagnosed heart attack or coronary disease). Results for the other 26 variables of interest can be found in the Supplementary Material. The 15 adjustment variables used were selected in line with our recent work [6]. They comprise variables of continuous, binary and ordinal type.

### 2.3.2 Assessing model and sampling vibration

In order to assess model vibration for the NHANES data, we follow [6], where we focused on the particular type of model vibration that is due to the inclusion or exclusion of all potential adjustment variables in a Cox regression. Furthermore, we include the variables age and sex as baseline variables in every model. The combination of the 13 remaining adjustment variables yields  $2^{13} = 8192$  different models. For the investigation of sampling vibration, we consider  $B = 1000$  subsets of size  $0.5n$ , where  $n$  is the number of observations. Moreover, we use the model with all 15 adjustment variables as favorite model.

### 2.3.3 Assessing measurement vibration

In order to assess the vibration of effects due to measurement uncertainty in the NHANES data, we first have to get an idea of the magnitude of measurement error that we can expect in this study. In the absence of a calibration sample specific to the NHANES data, which would allow quantification of the exact magnitude of measurement error in this study, we decided to search in the literature for information on the precision with which the variables of interest and adjustment variables used in our analyses are typically measured. To obtain a representative range of measurement error, we aimed to collect high and low values of sensitivity, specificity and correlations for each variable. As we found only scarce information for most variables, we decided to calculate average values for sensitivity, specificity and correlations for high and low measurement error to obtain representative values which we applied to all error-prone variables. For more detailed information for the different variables and references see Supplementary Tables 2 and 3. Using the average values for high and low measurement error as limits in a uniform distribution, we randomly draw a correlation and values for sensitivity and specificity for

each iteration  $b = 1, \dots, B$ . In the case of continuous variables, this strategy resulted in correlation coefficients between observed exposure and true exposure uniformly distributed between 0.73 to 0.9. For binary variables, we draw values for sensitivity and specificity from a uniform distribution between 0.56 and 0.85, and between 0.73 and 0.98, respectively. Finally, we generate measurement error for different types of variables following the procedure described in section 2.2. Similar to the assessment of sampling vibration, we use the model with all 15 adjustment variables as a favorite model and repeat the procedure  $B = 1000$  times. In accordance with [11], we assume the variables age and sex to be without measurement error, and the same is assumed to apply to race/ethnicity.

### **2.3.4 Comparing different scenarios of measurement vibration with sampling and model vibration**

In the assessment of measurement vibration for the NHANES data, we distinguish between three different scenarios: 1) We add measurement error to the variable of interest but not to the adjustment variables, or, conversely, 2) we add measurement error to all adjustment variables except age, sex, and race/ethnicity, and consider the variable of interest to be measured without error, and 3) we add measurement error to both the variable of interest and the adjustment variables (except age, sex and race/ethnicity). For all scenarios, we assume that information on the outcome has no measurement error, an assumption that is justifiable given the completeness and accuracy of NHANES data on death ascertainment. Finally, we compare these three scenarios, which illustrate measurement vibration, with model and sampling vibration, and focus in the interpretation of results on relative hazard ratios and volcano plots. In these volcano plots, we consider a p-value  $< 0.05$  as significant. For all analyses on the NHANES data, we use the `coxph` function from the R-package `survival`. Due to the complex sampling structure of the NHANES data, we account for participant weights, as well as for the clusters pseudostrata and pseudosampling units by using a robust sandwich variance estimator. For all types of vibration, we standardize the continuous variables of interest to ensure comparability.

## **2.4 Simulation study**

In addition to the analyses on real data, we conduct a simulation study with the aim of comparing measurement, sampling and model vibration for sample sizes that can both be smaller and larger than the initial sample size of the NHANES data. In this simulation study, we generate data with sample sizes  $n \in \{500, 1000, 5000, 10\ 000, 50\ 000, 100\ 000, 200\ 000\}$ . The simulated data is based on the NHANES data in the sense that we adopt the correlation structure as well as the effect sizes of the variables on the real data. More details about the data generation are described in the Supplement. Finally, we assess the three types of vibration in the same way as introduced in section 2.3. For measurement vibration, we consider only the scenario with measurement error in both the variable of interest and the adjustment variables.



## 3 Results

### 3.1 Results on the NHANES data

Figures 1 – 4 show volcano plots of model, sampling and measurement vibration for the three different scenarios of measurement error introduced in section 2.3.4 for the four selected variables of interest, i.e., diabetes, heart disease, thigh circumference and HDL-cholesterol. In these figures, we provide additional quantitative information about RHRs and RPs.

In the most realistic scenario for measurement error, i.e., when there is measurement error in the variable of interest and the adjustment variables, both significant and non-significant results can be observed for all variables of interest. Measurement vibration in this scenario is higher than model and sampling vibration in terms of RHRs for three of four variables of interest (diabetes, heart disease and HDL-cholesterol). In the assessment of sampling vibration, both significant and non-significant results are obtained for all variables of interest and sampling vibration is higher than model vibration for diabetes, heart disease and thigh circumference. In contrast to measurement and sampling uncertainty, model uncertainty does not change the significance of results for diabetes, heart disease and thigh circumference, where all results are significant. Only for HDL-cholesterol does model uncertainty change the significance of results. While we observe a Janus pattern for HDL-cholesterol in the case of sampling vibration, we can clearly distinguish two clusters for thigh circumference in the case of model vibration. These clusters result from the choice of whether BMI was included or excluded as an adjustment variable.

Despite a general tendency of measurement error to lead to an attenuation in effect estimates and loss of statistical power when present only in the variable of interest, we can also observe cases where measurement error leads to an inflated effect estimate and a smaller p-value compared to the results without measurement error in this scenario. This tendency is particularly evident for HDL-cholesterol and diabetes and can also be observed for the large majority of the variables of interest illustrated in the Supplement. When measurement error is only present in the adjustment variables, we can observe a clear bias towards the null for thigh circumference, while there is a substantial bias away from the null for diabetes and HDL-cholesterol. Finally, in the more realistic scenario when measurement error is present both in the variable of interest and in the adjustment variables, the effects of measurement error are more difficult to summarize as they seem to combine the effects of a general attenuation towards the null, which occurs due to the measurement error in the variable of interest, and the effect attenuation or inflation which occurs due to measurement error in the adjustment variables.

### 3.2 Results on simulated data

Figures 5 – 8 provide RHRs quantifying the variability in effect estimates for simulated data of varying sample sizes. In the lower panels of these figures, bar plots visualize the percentage of significant results for each sample size and each type of vibration for the three categories: negative significant, non-significant,

and positive significant.

For all variables of interest, RHRs decrease with increasing sample size. This is most obvious for sampling vibration, which is larger than model and measurement vibration for small sample sizes and tends to 1 with increasing sample size. Model and measurement vibration, on the other hand, remain non-negligible even for a sample size of 200 000: For diabetes, heart disease and thigh circumference RHRs  $> 1.1$  can be observed. For HDL-cholesterol, model vibration decreases to 1.1 and measurement vibration to 1.02 for the largest sample size. In the comparison of model and measurement vibration, measurement vibration is lower for thigh circumference and higher for diabetes and heart disease for all sample sizes. For HDL-cholesterol, measurement vibration is higher than model vibration for small sample sizes, and lower for large sample sizes.

When focusing on the results with regard to the type of significance, both significant and non-significant results are present for small sample sizes and all types of vibration for the three variables diabetes, heart disease and thigh circumference. For large sample sizes, the results indicate significance with either only positive sign or only negative sign (without showing a Janus pattern). For HDL-cholesterol, in contrast, a Janus pattern can be observed for measurement and model vibration for both small and large sample sizes. For sampling vibration, most of the results are significant with positive sign for the largest sample size, but non-significant results occur as well. As shown in the Supplement, eight of the other 26 variables of interest can be associated with a Janus pattern for at least one type of vibration for the largest sample size.

## 4 Discussion

In this work, the vibration of effects approach [7], which we previously used to assess the variability in observational associations for different model specifications [6], and applied to different subsamples of the data [16], was extended to exposure measurement uncertainty. Through this extension, it is possible to quantify and compare model, sampling and measurement uncertainty in a common framework when investigating the stability of research findings in observational studies. We studied these three sources of uncertainty on real data for different scenarios of measurement vibration and on simulated data for varying sample sizes. In accordance with [17] and in contrast to what is commonly assumed in the literature [9, 10], we found in our analyses on the NHANES data set that even in the simple situation where there is only measurement error in the variable of interest, measurement error can lead to occasional overestimations of parameter estimates. This phenomenon is well-illustrated in [17] and especially occurs in the situation of low sample sizes. Yet, even for larger sample sizes, the additional variance in the estimator, which is introduced by measurement error, can induce overestimations of parameter estimates.

For the more realistic scenario of measurement error, where both the variable of interest and the adjustment variables were assumed to be prone to measurement error, measurement vibration was even

less predictable as both bias towards the null and systematic inflations of effect estimates occurred in this situation. For this latter scenario, measurement vibration, as quantified through RHRs, exceeded model vibration and sampling vibration for 27 and 12 of the 30 associations of interest that we studied, respectively. In our simulation study we found that, while all types of uncertainty decreased for increasing sample sizes, model and measurement vibration persisted non-negligibly for large sample sizes in contrast to sampling vibration.

For most probability models, there are theoretical results on the behavior of sampling uncertainty. In contrast, the consequences of model and measurement uncertainty on parameter estimates in observational studies in epidemiology are very difficult to predict. Model uncertainty is, in principle, reducible by considering the fit of the different candidate models to the data (note, however, that there are different possible ways to do that, implying some sort of method uncertainty). In contrast, a reduction in sampling uncertainty and measurement uncertainty requires more effort as it can only be achieved by increasing the sample size or by using more precise measurement tools, respectively. Finally, in the comparison between the different types of vibration, one must keep in mind that measurement uncertainty does not only lead to a variability in effect estimates, but also to bias.

Measurement error may also be a prominent feature for outcomes assessed in observational studies. This was not an issue for the mortality outcome that we used in the NHANES analyses, but measurement error in the outcome may be as large as or even larger than measurement error in the exposure and adjustment variables in many other circumstances. In these cases, a similar approach can be used to investigate the vibration of effects due to outcome measurement error.

Currently, statistical inference that is commonly applied to analyze epidemiological studies only accounts for sampling uncertainty. Neglecting model and measurement uncertainty can lead to an underestimation of uncertainty and overconfidence in results, and therefore to contradictory findings when studying the same association of interest in different epidemiological studies. To improve the replicability and credibility of epidemiological findings, it is therefore vital to either pre-emptively reduce these sources of uncertainty during the planning of epidemiological studies, to integrate them when deriving statistical results, or to systematically report their consequences on parameter estimation. While there are a number of methods to account for model and measurement uncertainty in epidemiological studies, including Bayesian model averaging [18], multimodel inference [19], simulation extrapolation, regression calibration [20] and Bayesian hierarchical approaches [21], these methods are only rarely applied in practice. Moreover, to our knowledge there are currently no methods which can simultaneously account for measurement error in the exposure of interest and all adjustment variables, or methods that can account for sampling, model and measurement uncertainty in a common framework. In cases where we can neither reduce nor integrate the uncertainty when deriving statistical results, it is important to study the robustness of results by systematically assessing the impact of all three types of uncertainty on parameter estimation. Some caveats need to be discussed regarding our vibration of effects approach. Firstly, there may be a lack

of consensus among experts about which variables can legitimately be considered adjustment variables in a model, and which combinations of adjustments are acceptable and most plausible. The plausible set may be a reduced subset of the full set of all theoretical combinations. However, even experts will often have difficulties agreeing which variables are indispensable. Empirical studies suggest that most observational studies do not include the majority of those variables for which there is a theoretical consensus that they should be considered as adjustment variables [22]. Other empirical work shows that, even within the same publication, estimates of reported associations for the same exposure-outcome pair under different analyses and models can yield large differences in effect estimates [23]. Therefore, we argue that considering a substantial number of variables and all their combinations is a legitimate exercise. Secondly, data on the extent of measurement error for exposures, outcomes, and adjustment variables may be missing entirely, or existing data from other datasets may not be representative of the respective measurement errors in a new dataset. In these cases, investigators should meticulously record what is known and what is unknown about these measurement errors. Using the proposed vibration of effects framework will allow them to show what influence different sizes of measurement error could have on the stability of the results.

Acknowledging these caveats, the vibration of effects approach provides a flexible tool to systematically assess and compare sampling, model and measurement uncertainty in a common framework. Finally, encouraging the wider use of the vibration of effects concept for understanding model, sampling, and measurement uncertainty may further sensitize researchers to the need to think more carefully about these sources of instability. For example, studies rarely report the extent of measurement error for the exposures of interest, and don't make a systematic effort to summarize the existing evidence about these measurement errors. It is possible that, in many studies, such evidence does not even exist. Similarly, consideration of confounding and choice of adjustment variables is often sketchy and not well-documented [24, 25]. In the current illustrative simulations we used a broad range of possible error, but in specific future studies investigators may be able to have a better sense, even at the design phase, of what magnitude of errors need to be anticipated. Moreover, the set of candidate adjustment variables would best be pre-emptively defined. Regardless, the vibration of effects estimations may help place the instability or robustness of study results into better context.

## **Acknowledgements**

This work was funded by the DFG (individual grant BO3139/4-3) and by the German Federal Ministry of Education and Research (BMBF) under Grant No. 01IS18036A. The authors thank Alethea Charlton for language corrections.

## Supplementary Material

Supplementary Material related to this article can be found at [https://github.com/simonsimon01/VoE\\_nhanes/blob/master/Supplement.pdf](https://github.com/simonsimon01/VoE_nhanes/blob/master/Supplement.pdf).

## Bibliography

- [1] Taubes G, Mann CC. Epidemiology faces its limits. *Science*. 1995;269:164–169.
- [2] Ioannidis JPA, Tarone RE, McLaughlin JK. The false-positive to false-negative ratio in epidemiologic studies. *Epidemiology*. 2011;22:450–456.
- [3] Schoenfeld JD, Ioannidis JPA. Is everything we eat associated with cancer? A systematic cookbook review. *The American Journal of Clinical Nutrition*. 2012;97:127–134.
- [4] Lash TL. The harm done to reproducibility by the culture of null hypothesis significance testing. *American Journal of Epidemiology*. 2017;186:627–635.
- [5] Boffetta P, McLaughlin JK, La Vecchia C, Tarone RE, Lipworth L, Blot WJ. False-positive results in cancer epidemiology: a plea for epistemological modesty. *JNCI: Journal of the National Cancer Institute*. 2008;100:988–995.
- [6] Patel CJ, Burford B, Ioannidis JPA. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*. 2015;68:1046–1058.
- [7] Ioannidis JPA. Why most discovered true associations are inflated. *Epidemiology*. 2008;19:640–648.
- [8] Boccia S, Rothman KJ, Panic N, et al. Registration practices for observational studies on ClinicalTrials.gov indicated low adherence. *Journal of Clinical Epidemiology*. 2016;70:176–182.
- [9] Thomas D, Stram D, Dwyer J. Exposure measurement error: influence on exposure-disease relationships and methods of correction. *Annual Review of Public Health*. 1993;14:69–93.
- [10] Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occupational and Environmental Medicine*. 1998;55:651–656.
- [11] Brakenhoff TB, Mitroiu M, Keogh RH, Moons KG, Groenwold RH, van Smeden M. Measurement error is often neglected in medical literature: a systematic review. *Journal of Clinical Epidemiology*. 2018;98:89–97.
- [12] Michels KB, Bingham SA, Luben R, Welch AA, Day NE. The effect of correlated measurement error in multivariate models of diet. *American Journal of Epidemiology*. 2004;160:59–67.
- [13] Day NE, Wong MY, Bingham S, et al. Correlated measurement error - implications for nutritional epidemiology. *International Journal of Epidemiology*. 2004;33:1373–1381.
- [14] Kipnis V, Freedman LS. Impact of exposure measurement error in nutritional epidemiology. *JNCI: Journal of the National Cancer Institute*. 2008;100:1658–1659.

- [15] Brakenhoff TB, Van Smeden M, Visseren FL, Groenwold RH. Random measurement error: Why worry? An example of cardiovascular risk factors. *PLOS ONE*. 2018;13:e0192298.
- [16] Klau S, Schönbrodt FD, Patel CJ, Ioannidis JPA, Boulesteix AL, Hoffmann S. Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology; 2020. Submitted to *Psychological Science*.
- [17] Loken E, Gelman A. Measurement error and the replication crisis. *Science*. 2017;355:584–585.
- [18] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: a tutorial. *Statistical Science*. 1999;14:382–417.
- [19] Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*. 2004;33:261–304.
- [20] Bennett DA, Landry D, Little J, Minelli C. Systematic review of statistical approaches to quantify, or correct for, measurement error in a continuous exposure in nutritional epidemiology. *BMC Medical Research Methodology*. 2017;17:1–22.
- [21] Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*. 1993;138:430–442.
- [22] Serghiou S, Patel CJ, Tan YY, Koay P, Ioannidis JPA. Field-wide meta-analyses of observational associations can map selective availability of risk factors and the impact of model specifications. *Journal of Clinical Epidemiology*. 2016;71:58–67.
- [23] Chu L, Ioannidis JPA, Egilman AC, Vasiliou V, Ross JS, Wallach JD. Vibration of effects in epidemiologic studies of alcohol consumption and breast cancer risk. *International Journal of Epidemiology*. 2020;.
- [24] Munkholm K, Faurholt-Jepsen M, Ioannidis JPA, Hemkens LG. Consideration of confounding was suboptimal in the reporting of observational studies in psychiatry: a meta-epidemiological study. *Journal of Clinical Epidemiology*. 2020;119:75–84.
- [25] Hemkens LG, Ewald H, Naudet F, et al. Interpretation of epidemiologic studies very often lacked adequate consideration of confounding. *Journal of Clinical Epidemiology*. 2018;93:94–102.

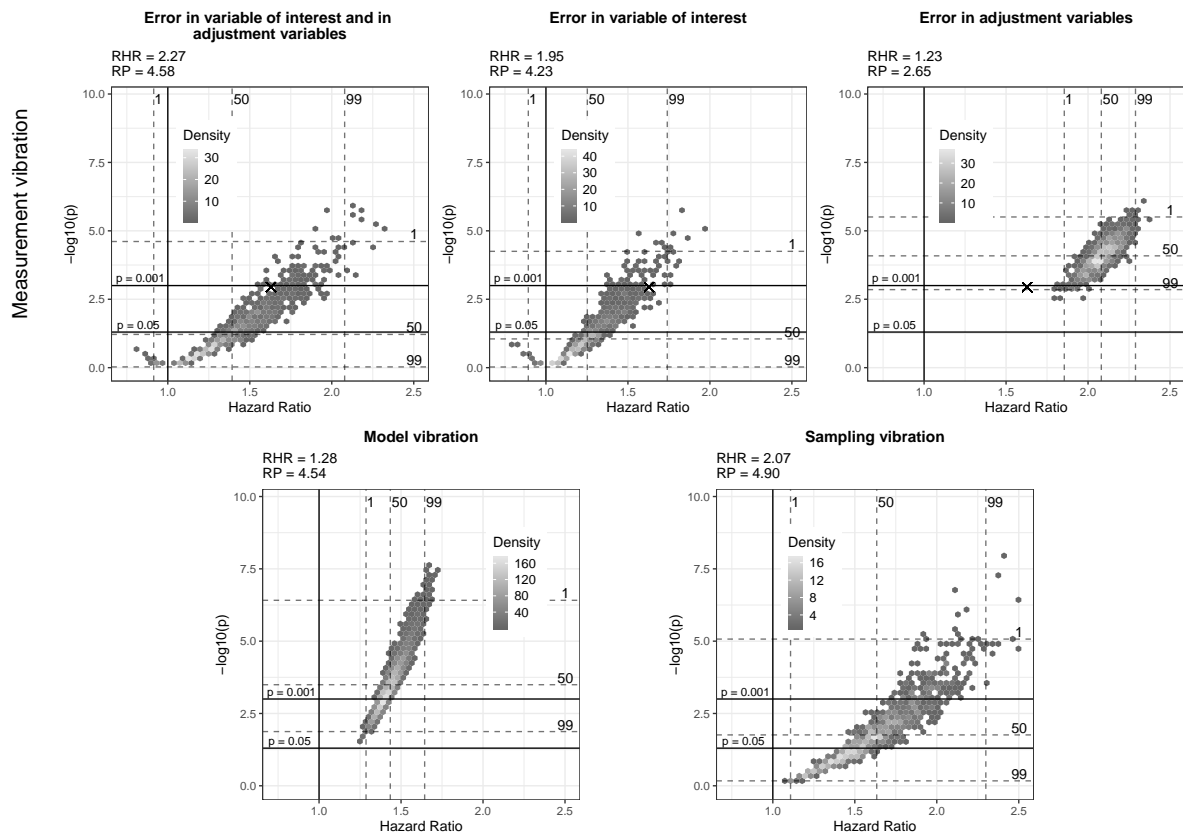


Figure 1: Volcano plots for different types of vibration and different scenarios of measurement vibration when **diabetes** is the variable of interest. The summary measures RHR and RP indicate relative hazard ratios and relative p-values, respectively. The black cross in the top panel indicates the model without measurement error.



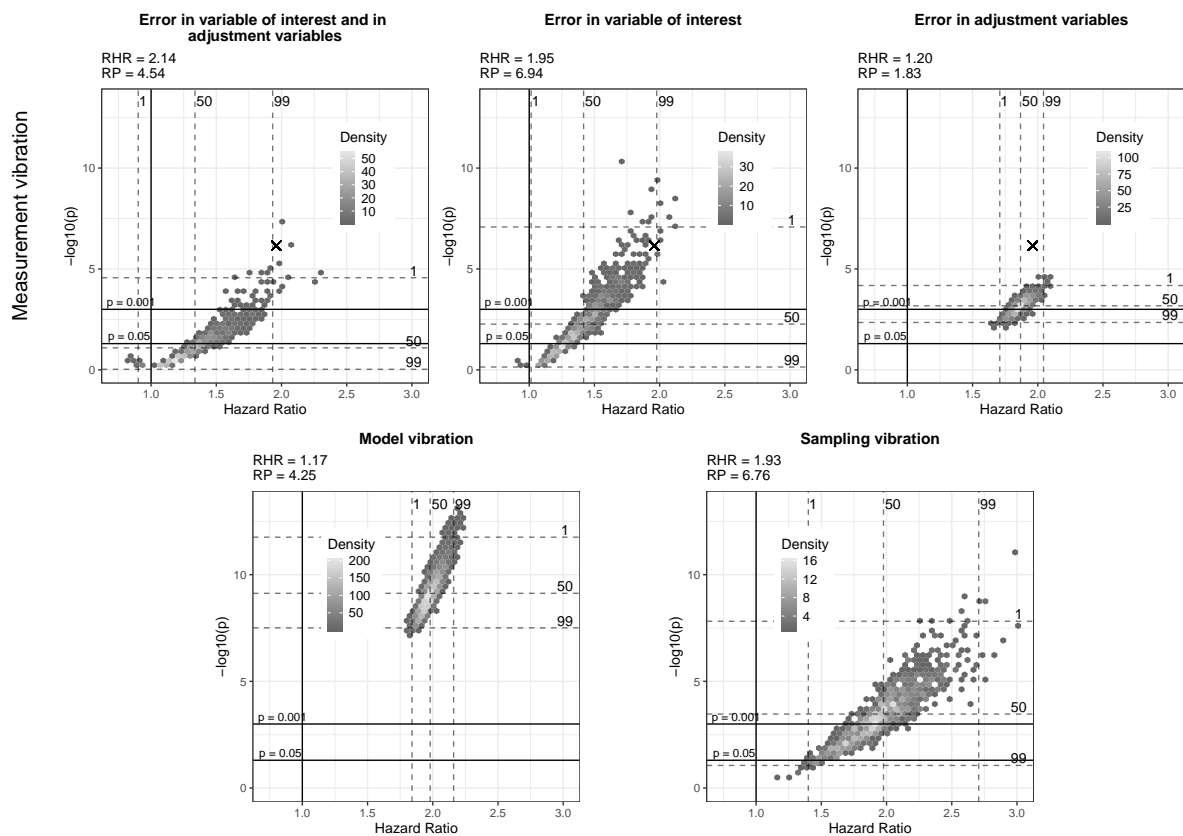


Figure 2: Volcano plots for different types of vibration and different scenarios of measurement vibration when **heart disease** is the variable of interest. The summary measures RHR and RP indicate relative hazard ratios and relative p-values, respectively. The black cross in the top panel indicates the model without measurement error.

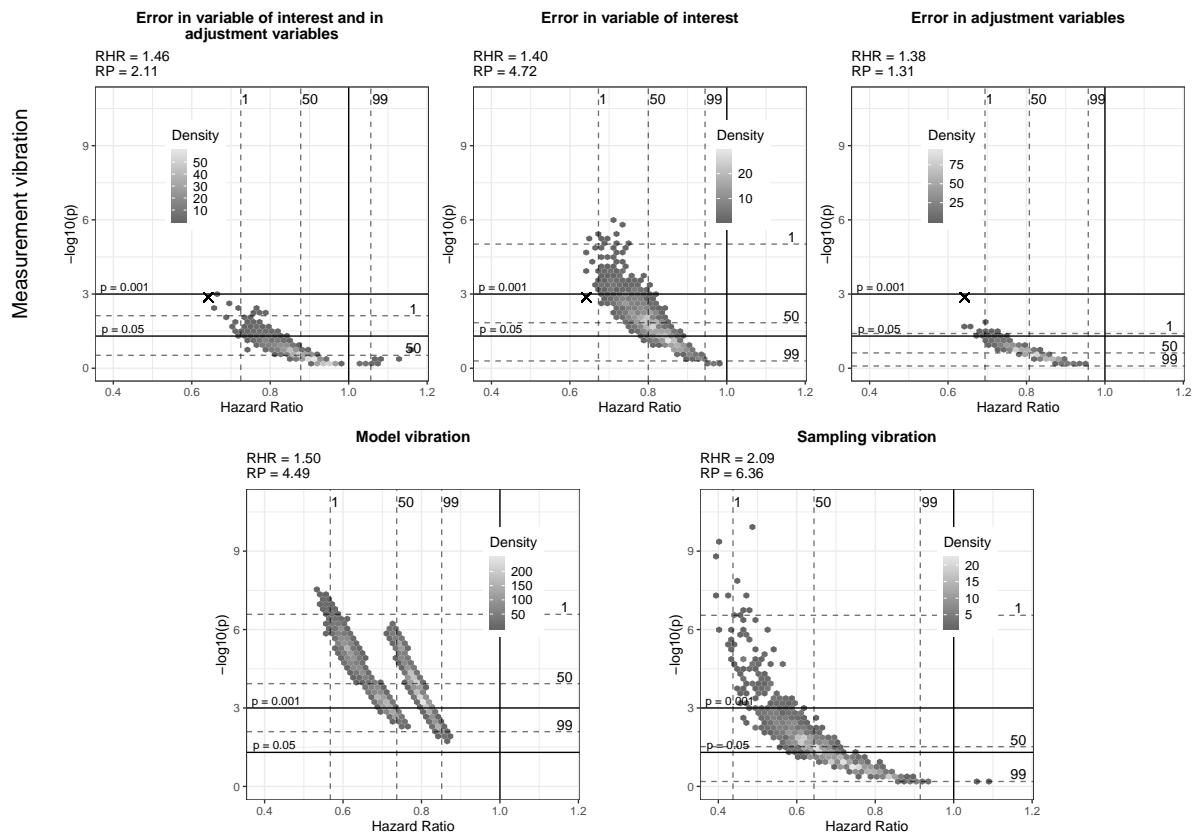


Figure 3: Volcano plots for different types of vibration and different scenarios of measurement vibration when **thigh circumference** is the variable of interest. The summary measures RHR and RP indicate relative hazard ratios and relative p-values, respectively. The black cross in the top panel indicates the model without measurement error.

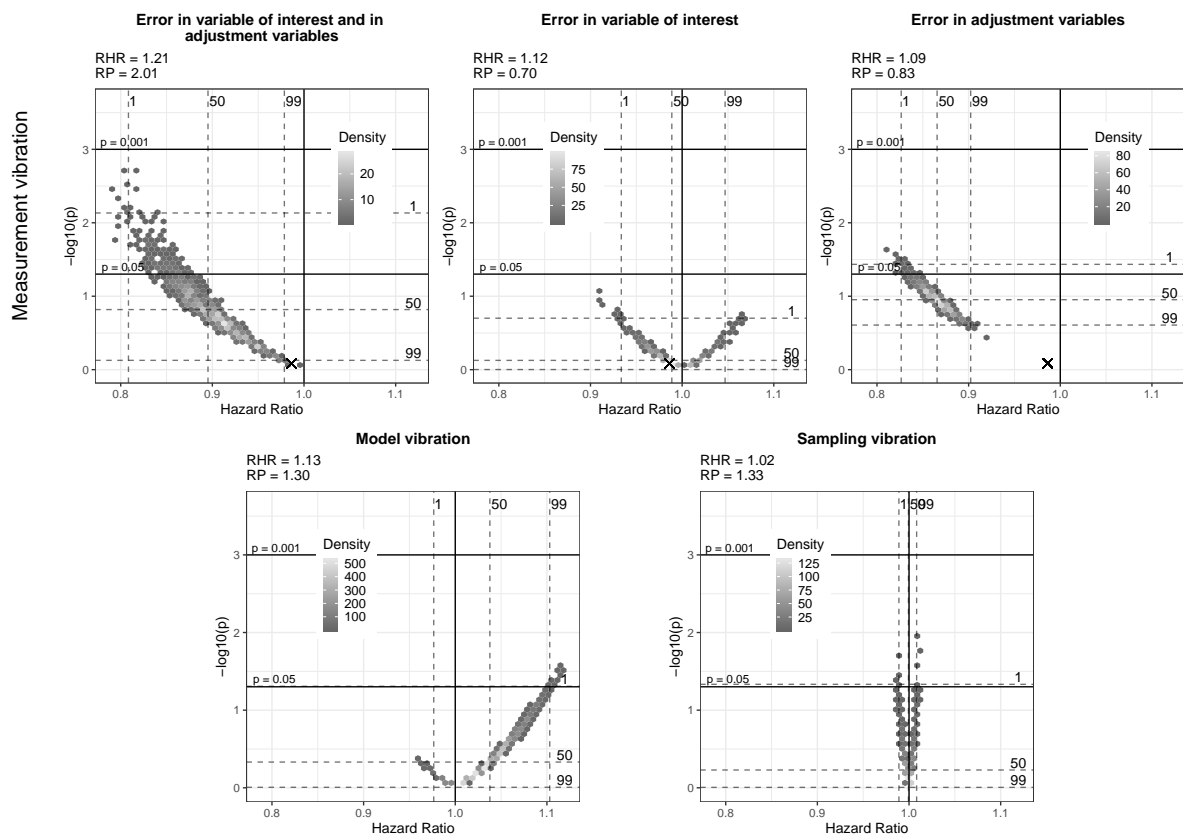


Figure 4: Volcano plots for different types of vibration and different scenarios of measurement vibration when **HDL-cholesterol** is the variable of interest. The summary measures RHR and RP indicate relative hazard ratios and relative p-values, respectively. The black cross in the top panel indicates the model without measurement error.

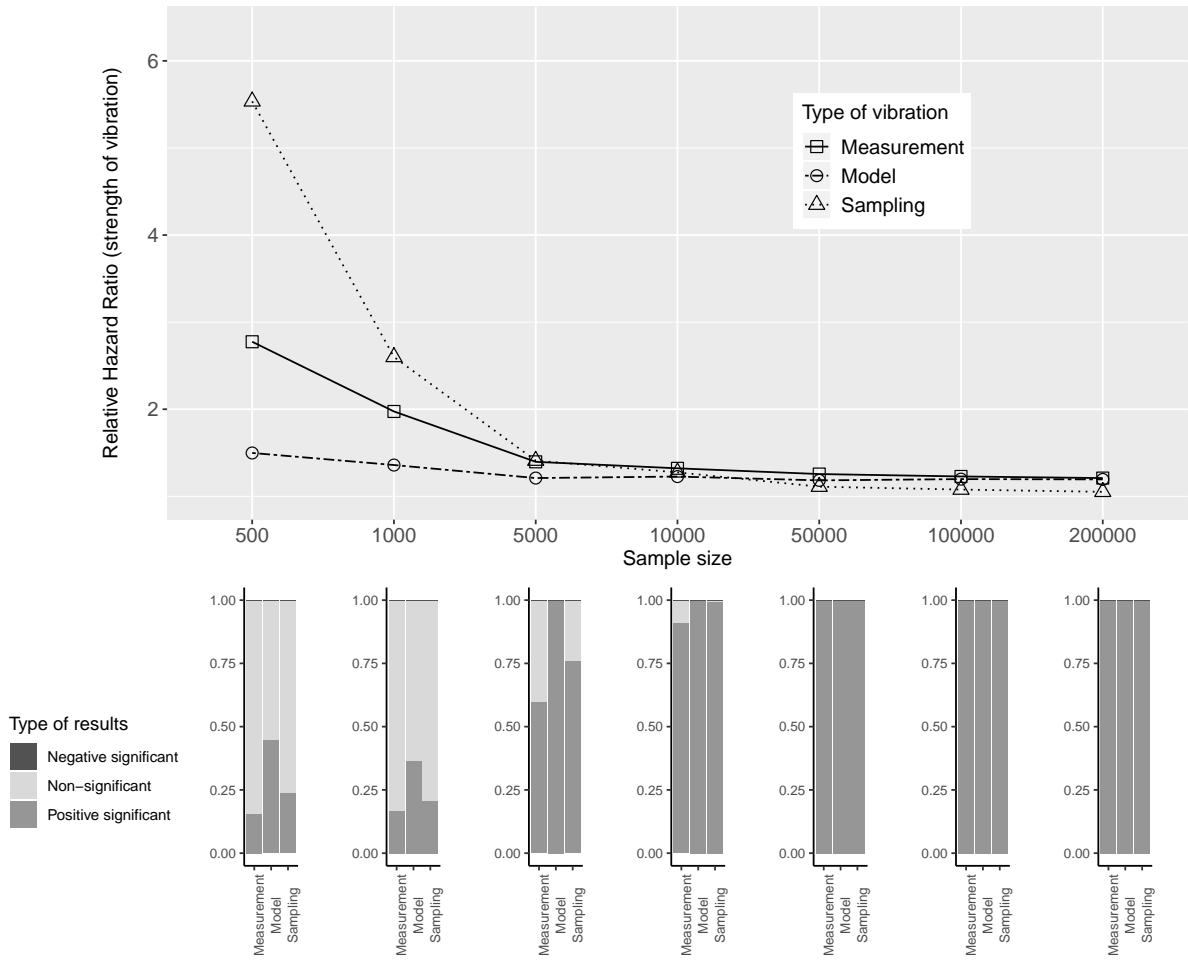


Figure 5: Measurement, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association of **diabetes** with mortality.

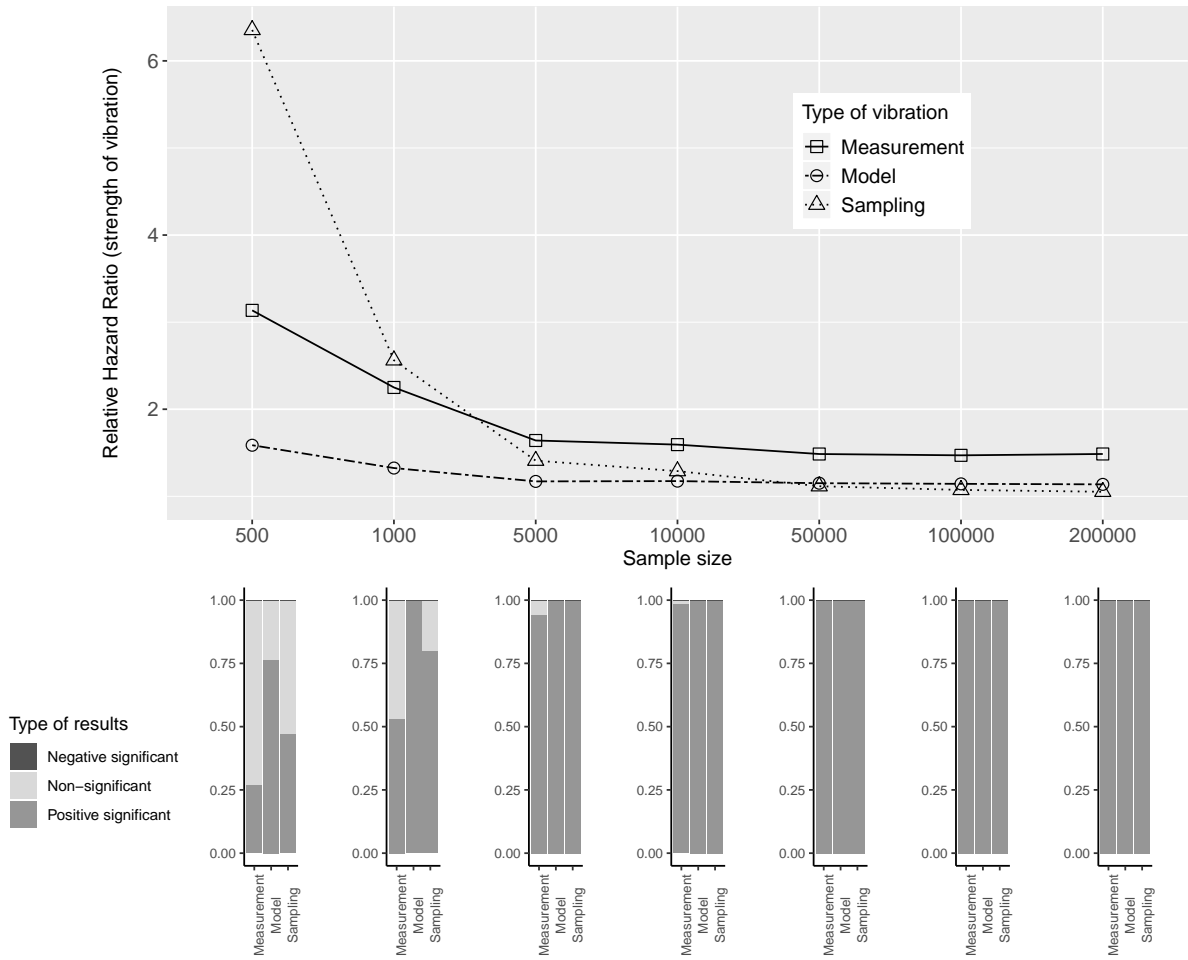


Figure 6: Measurement, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association of **heart disease** with mortality.

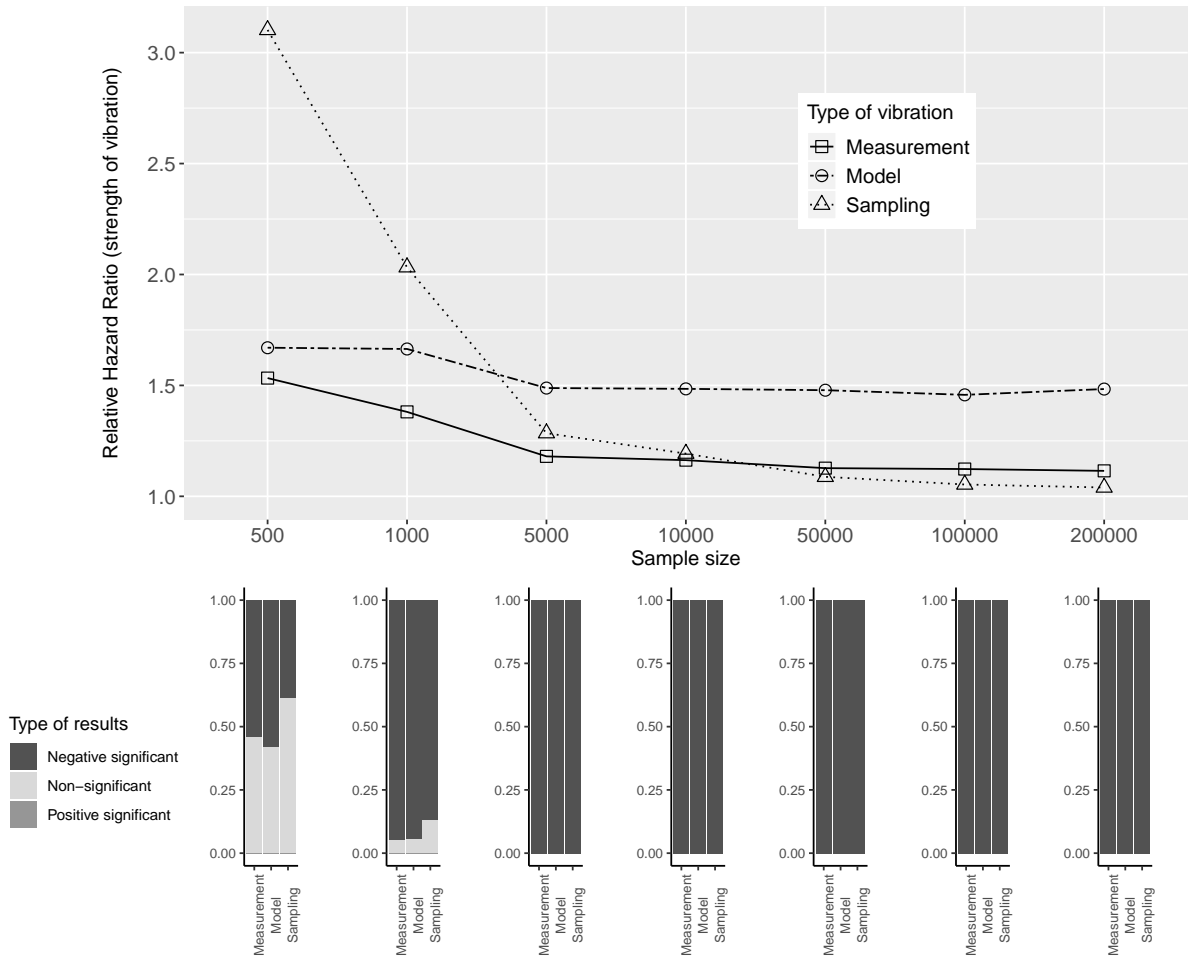


Figure 7: Measurement, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association of **high circumference** with mortality.

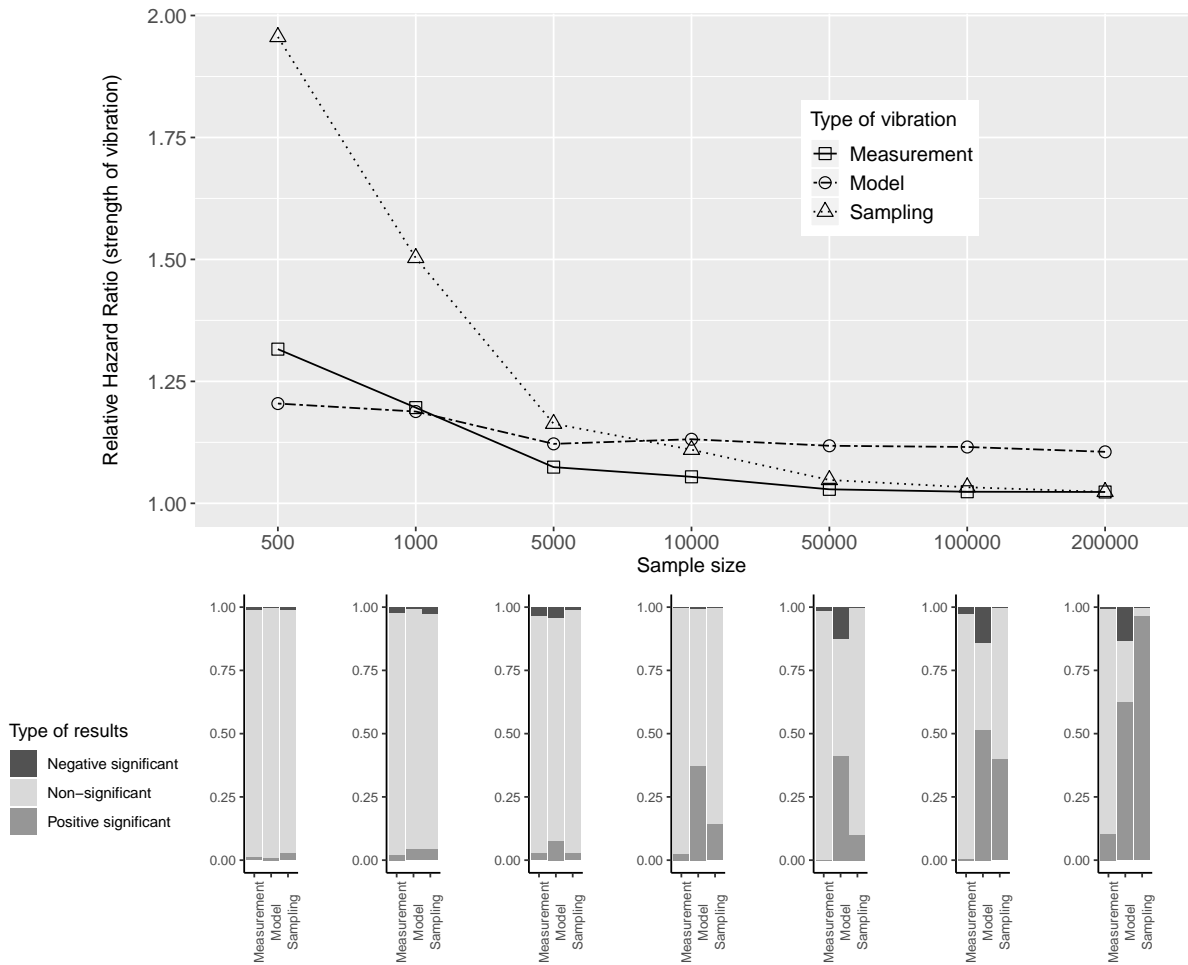


Figure 8: Measurement, model, and sampling vibration for different sample sizes (top panel), and bar plots visualizing the type of results in terms of significance of estimated effects (bottom panel) for the association of **HDL-cholesterol** with mortality.

# **Eidesstattliche Versicherung**

(Siehe Promotionsordnung vom 12. Juli 2011, § 8 Abs. 2 Pkt. 5)

Hiermit erkläre ich an Eides statt, dass die Dissertation von mir selbstständig, ohne unerlaubte Beihilfe angefertigt ist.

Bremen, den 13.10.2020

---

Simon Klau