Addressing researcher degrees of freedom in applications, methodological research, and teaching

Dissertation

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

eingereicht von
Maximilian Michael Mandl
22.05.2025

1. Gutachterin: Prof. Dr. Anne-Laure Boulesteix
2. Gutachter: Prof. Dr. Andreas Mayr
3. Gutachterin: Prof. Dr. Inke König

Tag der mündlichen Prüfung: 31.07.2025

»Es irrt der Mensch, solang er strebt.«

— J. W. v. Goethe, *Faust I*

ACKNOWLEDGEMENTS

First of all, this doctoral thesis would have not been possible without the great support of my supervisor Prof. Dr. Anne-Laure Boulesteix. Thank you for your great guidance, patience, and all the freedom you gave me during the past five years. I also want to thank Prof. Dr. Inke König and Prof. Dr. Andreas Mayr for being my reviewers, and Prof. Dr. Christian Heumann and Prof. Dr. Thomas Augustin for being part of my examination panel. I am also very grateful to all my co-authors and colleagues at the IBE, the Department of Statistics, the University of Cambridge, and the Imperial College London—thank you for all the great work and our weekly meetings with fruitful discussions in the last years.

Throughout my academic career I met a lot of wonderful people from different fields and I am thankful for all the opportunities they gave me. Thank you, Prof. Dr. Monika Schnitzer, Prof. Dr. Christoph Trebesch, and Prof. Dr. Hans Otto Seitschek for inspiring me outside the world of statistics. Furthermore, this path would not have been possible without the scholarships from the *Studienstiftung* and the *Max-Weber-Programm*, who funded me throughout my Bachelor's and Master's degrees and enabled me to study and research in England and Germany.

Outside of the academic world, my friends made my life so much more beautiful—and besides, had a hard time listening to all my worries and complaints in stressful times. You are not only wonderful friends, you have become a part of my family in the last decade. Moreover, I am grateful for the guidance, encouragement, and great discussions in Feldafing in the past years—even though the both of you came late into my life, you are a significant part of it. As well, I'm deeply grateful to the wonderful family from Holzkirchen—thank you for always welcoming me and making me feel at home. I also want to thank my parents—thank you for making me the person I am today. Furthermore, I will always be indebted to my sisters—without your unconditional support, love, and obviously the paved path in terms of your academic achievements, this thesis would not have been possible.

Finally, I want to thank my wife—my rock in the storm! My words cannot express how I feel and how grateful I am for having you in my life—but maybe Isolde's final song can:

> *»In dem wogenden Schwall,*
> *in dem tönenden Schall,*
> *in des Welt-Atems wehendem All,*
> *—ertrinken,*
> *—versinken,*
> *—unbewusst,*
> *—höchste Lust!«*
>
> — R. Wagner, *Tristan und Isolde*

SUMMARY

This thesis investigates the concept of *researcher degrees of freedom* (RDF), which refers to the analytical flexibility available to researchers performing data analysis tasks. This flexibility can lead to different and potentially conflicting results, even when analyzing the same research question with the same dataset. When combined with selective reporting, RDF can inflate effect sizes, increase false positive rates, and produce over-optimistic results. This work is centered around the concept of RDF in three domains of statistical research: applied, methodological, and didactic.

At first, RDF are conceptualized as a multiple testing problem, recognizing that different analytical strategies often result in highly dependent hypothesis tests. Traditional methods like the Bonferroni correction are overly conservative in this context, resulting in a substantial loss of power. As a solution, we propose using the minP method in the context of RDF—a permutation-based approach that offers better power while controlling the family-wise error rate. This method is applied in a study investigating the relationship between perioperative paO_2 levels and post-operative complications in a neurosurgical context. The results demonstrate that the minP approach provides valid p-value adjustments, enabling selective reporting without inflating type I error rates.

The second manuscript focuses on Mendelian Randomisation (MR), a causal inference method using genetic variants as instruments to infer causal effects of risk factors on an outcome. A method to detect and account for pleiotropic single-nucleotide polymorphisms (SNPs) in both uni- and multivariable MR analyses is introduced. An inflation factor is applied to the general heterogeneity statistic to correct for overdispersion. This study provides the basis for one of the use cases presented in the next manuscript.

The third manuscript shifts the focus to RDF in methodological research. Generally, the flexibility in benchmarking new statistical methods—such as selecting favorable datasets or comparison methods—can make new/preferred methods appear artificially superior. This manuscript deals with the interpretation of real data examples and introduces the so-called *storytelling fallacy*: the selective interpretation of real-data examples to support the superiority of a new/preferred method. Researchers often develop domain-specific narratives that favor their method while ignoring alternative, equally plausible interpretations. This practice can result in misleading conclusions and an over-optimistic representation of new methods. The concept is illustrated by examples related to pleiotropy detection in MR analyses and COVID–19 prediction modeling.

Finally, the fourth manuscript addresses RDF in the context of statistical education. Students may develop unrealistic expectations about data analysis, believing that there is a single correct way to address research questions and find significant results. In reality, empirical research involves many valid analytic decisions, which may lead to different and even conflicting outcomes. A seminar course for advanced undergraduate and early graduate students was developed to address this issue. It combines theoretical and practical modules to raise awareness of RDF and train students in open science principles.

ZUSAMMENFASSUNG

Diese Arbeit befasst sich mit dem Konzept der sogenannten *researcher degrees of freedom* (RDF), also der analytischen Flexibilität, die Wissenschaftler:innen im Rahmen statistischer Analysen haben. Diese Flexibilität kann dazu führen, dass selbst bei identischer Forschungsfrage und gleicher Datengrundlage unterschiedliche—mitunter sogar widersprüchliche—Ergebnisse entstehen. In Kombination mit selektivem Reporting kann dies dazu führen, dass Effekte überschätzt, falsch-positiv-Raten erhöht und Ergebnisse insgesamt zu optimistisch dargestellt werden. RDF werden in dieser Arbeit aus drei verschiedenen Perspektiven beleuchtet: der angewandten, der methodologischen und der didaktischen Forschung.

Im ersten Manuskript werden RDF im Kontext multipler Testprobleme formalisiert. Dabei liegt der Fokus insbesondere auf der Korrelation zwischen Hypothesen, die sich aus unterschiedlichen Analysestrategien ergeben. Klassische Korrekturverfahren wie die Bonferroni-Methode erweisen sich in diesem Zusammenhang als zu konservativ und gehen mit einem deutlichen Verlust an Power einher. Als Alternative wird die sogenannte minP-Methode vorgestellt—ein Permutationsverfahren, das bei schwacher Kontrolle der family-wise error rate eine höhere Power ermöglicht. Angewendet wird diese Methode in einer Studie zum Zusammenhang zwischen perioperativen paO_2-Werten und postoperativen Komplikationen in der Neurochirurgie. Die Ergebnisse belegen die Validität der minP-Methode unter analytischer Unsicherheit.

Das zweite Manuskript führt eine Methode zur Identifikation pleiotroper Single-Nukleotid-Polymorphismen (SNPs) in Mendelian-Randomisation-Analysen ein—ein Verfahren der kausalen Inferenz, das genetische Varianten als Instrumentalvariablen nutzt. Die vorgestellte Methode erkennt pleiotrope SNPs sowohl im uni- als auch im multivariaten Kontext, indem sie einen Inflationsfaktor zur Korrektur von Überdispersion in der allgemeinen Heterogenitätsstatistik schätzt. Diese Arbeit bildet zugleich die Grundlage für einen der Anwendungsfälle, die im darauffolgenden Manuskript diskutiert werden.

Auch in der methodologischen Forschung spielen RDF eine Rolle, wenngleich sie dort seltener thematisiert werden. Die Flexibilität bei der Entwicklung neuer statistischer Verfahren—etwa durch die gezielte Auswahl günstiger Datensätze oder Vergleichsmethoden—kann zu einer überoptimistischen Darstellung der Performance der jeweiligen Methode führen. Im dritten Manuskript definieren wir die *storytelling-fallacy* als den selektiven Einsatz von anekdotischem Expertenwissen, um die vermeintliche Überlegenheit einer bevorzugten Methode durch reale Datenbeispiele zu stützen. Dabei werden häufig fachspezifische Interpretationen herangezogen, während ebenso plausible alternative Erklärungen unberücksichtigt bleiben. Dies kann insbesondere in Benchmarking-Studien zu invaliden Ergebnissen führen. Veranschaulicht wird dieses Problem anhand zweier Beispiele: bei der Erkennung von Pleiotropie in MR-Analysen und bei der Prädiktion von COVID-19-Infektionen.

Abschließend befasst sich das vierte Manuskript mit RDF im Kontext der Hochschullehre. Studierende haben häufig die Vorstellung, es gebe einen „richtigen“ Analyseansatz, der zwangsläufig zu signifikanten oder interessanten Ergebnissen führt. Diese Erwartung steht im Widerspruch zu den vielen möglichen Analyseansätzen, die zu verschiedenen Ergebnissen führen können. Vor diesem Hintergrund wurde ein Seminarkonzept für fortgeschrittene Bachelor- und Masterstudierende erarbeitet, das mithilfe von theoretischen Modulen und praktischen Übungen das Bewusstsein für RDF schärfen und die zentralen Prinzipien der Open-Science-Bewegung vermitteln soll.

Contents

1 Introduction

> *»[A]s Kempthorne (1976) suggests, statistics may be seen as "applied philosophy of science"«.*
>
> — Mayo (1980)

1.1 Foundations

Contrary to the proposition in the opening quote of this section, the perception of statistics within society and broader audiences of applied statistical research differs substantially: Empirical research is often seen as clear-cut and unambiguous. However, this understanding may be deceptive. For this reason, the general objective of this thesis is to create awareness for and deal with different sources of uncertainty and the consequent *researcher degrees of freedom* (RDF) in empirical research in an applied and methodological context. In addition, it highlights the role of academic education as one of the potential remedies to counteract the claimed crisis of empirical sciences in the 21$^{\text{st}}$ century.

Uncertainty is an inherent feature of empirical analyses. Statistical reasoning is usually based on a general logic of induction, i.e., we use past or present observations to form future expectations and make predictions about the world. These conclusions are evidently subject to considerable uncertainty, standing in contrast to the certainty typically associated with deductive reasoning.

For instance, if all the swans we have seen are white, we may conclude that all swans are white—until we encounter a black swan. While such an observation is highly unlikely, it is not impossible. This type of reasoning is known as inductive inference, as it involves concluding the unobserved based on what has been observed (Henderson, 2024). The problem becomes clearer when we compare the rules of deduction and induction. A generally valid deductive rule, for example, might be the following:

$$\begin{aligned}
&\forall x \in D : (A(x) \rightarrow B(x)) \\
&A(y), \text{with}\, y \in D \\
&\Rightarrow B(y)
\end{aligned} \tag{1}$$

In simple terms: The rule depends on the premises that, e.g., *all men are mortal* and *Ronald Fisher is a man.* Thus, we can conclude that *Ronald Fisher is mortal.* Rules of this nature are valid because a situation in which the premises are true and the conclusion is false is logically impossible (Harman and Kulkarni, 2006). In contrast, inductive rules might take the following form:

$$\begin{aligned}
&\exists x \in D : (A(x) \rightarrow B(x)) \\
&A(y), \text{with}\, y \in D \\
&\Rightarrow B(y)
\end{aligned} \tag{2}$$

The rule above is not valid in the same way as a deductive rule. The premises could be true, even if the conclusion is false (Harman and Kulkarni, 2006). Let's take the example mentioned above: Suppose all the swans we know are white. In this case, the premise would be true. However, if we later observe a black swan, the conclusion that all swans are white would be proven false. Nevertheless, this is the underlying logic of empirical research, where conclusions are based on observed evidence, but may change with new findings.[1]

Empirical results are still often seen as unequivocal. In recent years, however, the scientific community has tried to address the causes and implications of the so-called *replication crisis* (Baker, 2016; Loken and Gelman, 2017), i.e., the crisis therefore lies in the differing conclusions drawn from prior studies, which may not align with the findings of replication studies based on independent data. A high analytical variability became evident in this context. Even for very simple settings, there is a multitude of different possible analysis strategies to address a particular research question—with potentially different and even contradicting results (Gelman and Loken, 2014; Silberzahn et al., 2018). See, e.g., the conflicting results and subsequent discussions of Fields et al. (2019), Childers and Maggard-Gibbons (2019), Turner et al. (2019), and Childers and Maggard-Gibbons (2021) on studies examining the association of infectious complications and the use of retrieval bags for patients who had undergone laparoscopic appendectomies. However, contradictory results are not always a stimulating basis for scientific debate. Instead, they lead to a severe loss of trust in science among policymakers and the general public—as the COVID–19 pandemic showed with regard to the biomedical sciences (Caplan, 2023).

1.2 Uncertainty and RDF

Empirical research is subject to various sources of uncertainty. Hoffmann et al. (2021) categorise these into data preprocessing, parameter, model, method, measurement, and

[1] Note, that there is also the concept of *abduction* (Douven, 2021) which is outside the scope of this short introduction.

sampling uncertainty as depicted in Table 1. While measurement and sampling uncertainty may be classified as aleatoric uncertainty, which is generally unavoidable due to inherent variability in a process, the other forms of uncertainty are of epistemic nature, i.e., resulting from a lack of knowledge. See, e.g., Gruber et al. (2023) and Hüllermeier and Waegeman (2021) for an extensive introduction in the field of machine learning. Consequently, the potential for RDF arises from the epistemic sources of uncertainty. When combined with aleatoric uncertainty and selective reporting—which means reporting only the most favorable or "noteworthy" result—the outcome is often a lack of replicability (Hoffmann et al., 2021). Note that there are many (and sometimes confusing) definitions of the term *replication*. This work adopts a common, pragmatic definition, i.e., re-evaluating prior research findings using new data but the same or similar methods. Nosek and Errington (2020) provide a broad starting point for exploring the discussion around the term *replication*.

	Aleatoric Uncertainties
Measurement	Randomness arising from the inherent imprecision of measurements (see, e.g, Brakenhoff et al. (2018)).
Sampling	Randomness due to the selection of a subset from a population (see, e.g., Klau et al. (2020)).
	Epistemic Uncertaintes
Data preprocessing	Uncertainty introduced during the selection of relevant data, feature engineering, cleaning, and transformation processes (see, e.g. Klau et al. (2023)).
Parameter	Uncertainty due to the specification and choice of input parameters (see, e.g., Baio and Dawid (2015) and Probst et al. (2019)).
Model	Uncertainty arising from the specification of the model (see, e.g. Chatfield (1995)).
Method	Uncertainty resulting from the specific implementation and computational methods used for estimation (see, e.g., Sauerbrei et al. (2014)).

Table 1: Description of the sources of uncertainties (Hoffmann et al., 2021).

The RDF—resulting from the epistemic uncertainties—are defined as the free choices within the statistical analysis pipeline, representing the flexibility a researcher has regarding all decisions made during the analysis (Boulesteix et al., 2017; Simmons et al., 2011; Wicherts et al., 2016). The concept of RDF is closely related to other topics covered by the meta-science community, some of them referred to as *questionable research practices* (QRPs) (Andrade, 2021): *P-hacking*, or *fishing for significance* (Boulesteix et al., 2017; Gelman and Loken, 2013; Head et al., 2015), refers to the process of trying different analysis strategies to find a significant result. As Andrade (2021) states: *»the* *purpose is not to test a hypothesis but to obtain a significant result«*. *HARKing* stands for "Hypothesizing After the Results are Known" and refers to presenting a hypothesis developed after seeing the results (post-hoc) as if it had been proposed beforehand (a priori) (Kerr, 1998). These practices are sometimes linked to the incentive structure in academia. *Publication bias* (Marín-Franch, 2018; Turner et al., 2012) means that studies with positive or significant results are more likely to be published. However, it is important to emphasize that we do not imply that scientists always intentionally engage in the above-mentioned QRPs. In fact, selective reporting often occurs subconsciously, without any malicious intent, or by pure accident. Hodges et al. (2023) present a prime example—RDF may display a wide variety, and even using different software frameworks, such as R, Stata, or SPSS, may change results due to algorithmic variations or computational errors.

1.3 Meta-science and how to address the replication crisis

According to Schooler (2014), meta-science is defined as

> *»the science of science, [that] uses rigorous methods to examine how scientific practices influence the validity of scientific conclusions. It has its roots in the philosophy of science and the study of scientific methods, but is distinguished from the former by a reliance on quantitative analysis, and from the latter by a broad focus on the general factors that contribute to the limitations and successes of research.«*

Originating in philosophy, modern meta-science seeks to understand the causes and implications of the replication crisis in the empirical sciences and to offer tools for addressing it. To overcome the problems regarding RDF mentioned in the previous section, Hoffmann et al. (2021) suggest distinguishing between different ways to deal with the underlying sources of uncertainty: *reduce*, *report*, *accept*, and *integrate* uncertainties. In the following, some influential approaches and ideas are introduced.

One way to *report* uncertainty is the assessment of an association of interest through the so-called *vibration of effects* (VoE) framework (Ioannidis, 2008; Klau et al., 2021, 2023; Patel et al., 2015). Figure 1 depicts an exemplary volcano plot from Contribution 4. Its x-axis displays the estimated effect sizes of the variable of interest for different model specifications within a regression setting, while the y-axis shows the $-log_{10}$ transformed p-values for each model. Usually, the VoE framework can be used to assess the impact of different sources of uncertainty and to check the robustness and flexibility of results. Simple summary measures may be applied, such as relative effect estimates or p-values, and additionally volcano plots, like the one in Figure 1, can be used to identify, e.g., *Janus effects*—inspired by the two-headed Roman god—where results are significant for both positive and negative effects (Patel et al., 2015). This allows us to see how strongly the results depend on the chosen analysis strategy.

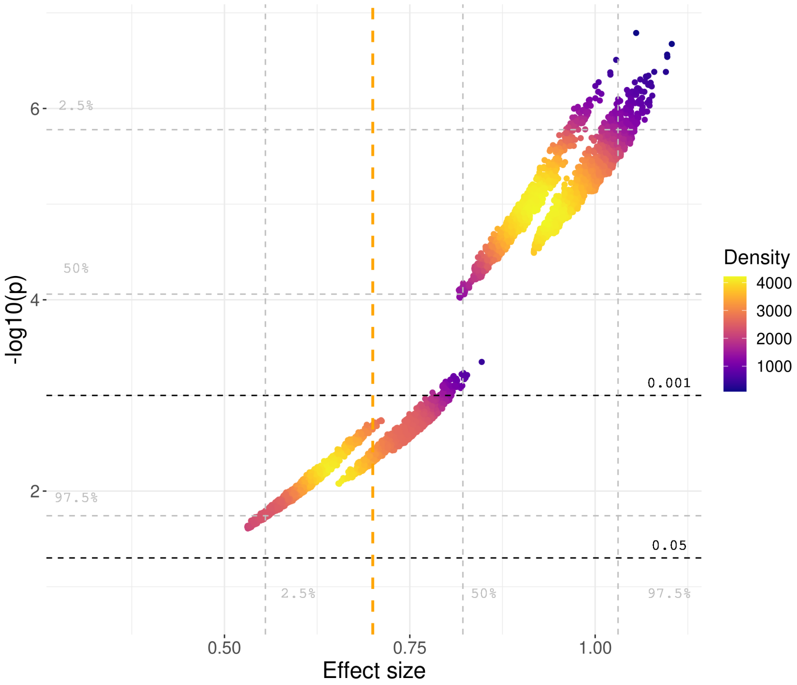

Figure 1: Exemplary VoE plot from Contribution 4 (Mandl et al., 2024b). Different estimates (x-axis) come from different model specifications of a simple regression problem with the respective $-log_{10}$ transformed p-values. The respective quantiles (2.5%, 50%, 97.5%) are represented by the grey dashed lines for both axes. The black dashed lines additionally mark different significance levels (0.001 and 0.05). The orange dashed line depicts the true underlying effect.

Figure 1 shows effect estimates for different model specifications from the didactic experiment conducted in Contribution 4. Students were asked to analyse simulated data and report the effect estimate of the variable of interest. The vertical dashed orange line shows the true underlying effect. Using specific model specifications makes it easy to obtain inflated and significant estimates—note that even for this simple RDF, 2^p models may be chosen, with p being the number of covariates. Criticism has been raised that the framework often incorporates implausible and unrealistic analytical strategies. As a result, the plot interpretation may not accurately reflect the set of specifications researchers would typically select.

Furthermore, Rohrer et al. (2017) and Simonsohn et al. (2020) introduced a permutation-based procedure called *specification curve.* The specification curve only visualises a set of reasonable analytical strategies and evaluates the joint distribution of the estimated effects using a null hypothesis of the median effect across all the specifications being zero. This can be seen as some kind of *multiverse analysis* carried out by a single researcher or team. Within a typical *multiverse analysis* (Olsson-Collentine et al., 2023; Steegen et al., 2016) researchers or different teams perform a whole set of alternative analysis approaches that arise from RDF to examine the robustness of results.

Approaches to *integrating* sources of uncertainty include using Bayesian models, e.g., hierarchical modeling or model averaging. In this context, Rehms et al. (2024) provide a flexible Bayesian framework that incorporates various sources of uncertainty in the modeling of infectious diseases. While these approaches offer a great deal of flexibility, they naturally come with certain costs—such as requiring specific assumptions about prior distributions. This should not be taken as a criticism of Bayesian methods per se. However, it does highlight the methodological flexibility they entail, which in turn allows for multiple valid analysis strategies. Due to the commonly held criticism directed at Bayesian models, reference should be made to Gelman and Hennig (2017) and their discussion on the concepts of subjectivity and objectivity in statistics. What is interesting here is that, due to the multiplicity of analysis strategies, many subjective decisions also have to be made in frequentist statistics—a problem that is more often attributed to Bayesian approaches.

Furthermore, aleatoric uncertainties may also be *reduced* through an increase of sample size (Button et al., 2013) or standardised experimental conditions (Jarvis and Williams, 2016).

In many cases, however, there is little we can do but to *accept* the inherent uncertainty in statistical findings. This can help to prevent us from taking single studies and findings too seriously—whether exploratory or confirmatory—and reminds us that generalisations are *built upon the shoulders of giants*[2], relying on cumulative knowledge and efforts.

1.4 Related Topics

Even though meta-research has gained significant popularity in recent years, it still faces considerable criticism from outside and even within its own community—a great overview can be found in Rubin (2023). Mayo and Hand (2022) *»[...] [argue] that the central criticisms arise from misunderstanding and misusing the statistical tools, and that in fact the purported remedies themselves risk damaging science«*. These ongoing debates arise from different approaches to understand the underlying reasons for the replication crisis. This thesis focuses on rather pragmatic solutions regarding the non-replicability of results from a statistician's point of view, while acknowledging that other more conceptual or philosophical problems might exist: For example, some

[2] This phrase refers to the famous aphorism by Isaac Newton. Its origin is discussed in the book by Merton (1993).

research focuses on the incentive structure of science in general, e.g., Smaldino and McElreath (2016) argue that the use of poor methods leads to high false discovery rates. The current incentive system, in which publication output is key to a scientist's career, rewards researchers using these methods. As a result, these practices tend to spread to students and colleagues, much like traits favored by natural selection in biology. Additional criticism is raised concerning the fundamental concepts of statistics: Typical discussions involve the role of p-values (Wasserstein and Lazar, 2016; Wasserstein et al., 2019), the concept of significance testing (Mayo and Hand, 2022; Schneider, 2015), subjectivism vs. objectivism (Gelman and Hennig, 2017; Gelman and Shalizi, 2013), or the discussion on exploratory vs. confirmatory research (Fife and Rodgers, 2022). We recognize that these topics also contribute to explaining the replication crisis. Nonetheless, this thesis does not adopt a specific stance—we neither defend nor promote the use of specific methods, analytical strategies, or philosophies. Our approach is based on a pragmatic understanding of the current situation and how to (at least partly) fix it.

Returning to the opening quote of this dissertation: Why should statistics be viewed as an applied philosophy of science? Statistics, and generally the empirical sciences, are concerned with drawing inferences from data in the presence of uncertainty. Because ground truths are typically inaccessible, scientists formulate hypotheses, test them against evidence, and revise or reject them accordingly—and this fundamental process of reasoning under uncertainty is precisely what motivates this thesis.

The remainder of this work is organized as follows: Section 2 introduces important methodological concepts relevant to the contributions of this thesis. Section 3 gives short summaries of the contributions with regard to the three main categories addressed in this thesis: Section 3.1 presents RDF in an applied statistical context and outlines a methodology for adjusting for RDF (Contribution 1). Section 3.2 is divided into two parts: First, we introduce a method to detect pleiotropic SNPs in the context of Mendelian Randomisation (Contribution 2); second, this preceding example (alongside with another case from Woehrle et al. (2024)) is used to discuss RDF in the context of methodological research (Contribution 3). Section 3.3 focuses on academic teaching and highlights the importance of familiarising students with meta-scientific concepts like RDF (Contribution 4). Finally, Section 4 offers concluding remarks and an outlook. The four manuscripts included in this cumulative dissertation can be found in the appendix.

2 Methods

> *»There are two cultures in the use of statistical modeling to reach conclusions from data.«*
>
> — Breiman (2001b)

As Breiman (2001b) describes, there are two (main) approaches or philosophies for drawing conclusions from data. This thesis does not take a position on which culture is preferable as it employs a broad range of different philosophies/cultures, which are introduced in the following. Section 2.1 addresses the multiple testing problem which is central to Contribution 1 (Mandl et al., 2024a). Section 2.2 introduces the concept of Mendelian Randomisation, discussed in Contributions 2 and 3 (Mandl et al., 2025b, 2025a). Subsequently, Section 2.3 covers some introductory concepts of supervised machine learning, as applied in Contributions 1 and 3 (Mandl et al., 2024a, 2025b).

2.1 Multiple Testing

In the following, let H_0 and H_1 be the null and alternative hypothesis in a general, unspecified statistical testing scenario. Table 2 summarizes the used notations for type I and type II errors along with their complementary probabilities based on the truth of the null hypothesis and whether it is rejected or not.

Decision	**H_0 true**	**H_1 true**
Reject H_0	Type I error: α	Power $= 1 - \beta$
Fail to reject H_0	$1 - \alpha$	Type II error: β

Table 2: The four possible situations in a hypothesis test setting depending on the null hypothesis being true/false and the respective test decision.

Now, assume that we are testing m hypotheses H_0^i for $i = 1, ..., m$. Usually, we are interested in controlling the number of type I errors V, see Table 3.

Decision	**H_0 true**	**H_1 true**
Reject H_0	Type I errors: V	True Positives: S
Fail to reject H_0	True Negatives: U	Type II errors: T

Table 3: Testing m hypotheses: V, U, S, and T are random variables that count each outcome for the m tests. These are unobservable, as the true underlying state of the hypotheses is unknown.

Note, that Contribution 1 and this section follows the definitions and mathematical notations of Dudoit et al. (2003). More specifically, we want to control $P(V \geq 1)$, i.e.,

$P(\text{Reject at least one true } H_0^i)$ which means making at least one type I error when testing H_0^i for $i = 1, ..., m$—exactly the definition of the family–wise error rate (FWER). For error rates like the FWER, we distinguish between a *strong* and a *weak* control. The difference depends on the specific composition of the subset $\Lambda_0 \subseteq \{1, ..., m\}$ of true and false H_0's for the underlying unknown data-generating mechanism. A multiple testing correction is said to achieve strong control of the FWER if it controls $P(\text{Reject at least one true } H_0^i)$ under any combination of true and false H_0's, i.e., any subsets of $\Lambda_0 \subseteq \{1, ..., m\}$ of true and false H_0^i's. Alternatively, a procedure is said to achieve a weak control of the FWER if all H_0^i for $i = 1, ..., m$ are true, i.e., we want to control $P(\text{Reject at least one true } H_0^i | \cap_{i=1}^m H_0^i)$. In the context of this thesis, we are primarily concerned with avoiding false positives, particulary in the case when all H_0^i for $i = 1, ..., m$ are true—i.e., the weak control is satisfactory. Note that in the following, and according to Contribution 1, we can formalize the problem of RDF as a multiple testing scenario, with m being the number analytical specifications.

2.1.1 Bonferroni correction

The Bonferroni procedure is arguably the most well-known and straightforward method—it provides strong control of the FWER, meaning it controls $P(\text{reject at least one true } H_0^i)$ under any configuration of true and false null hypotheses. Bonferroni-adjusted significance levels $\tilde{\alpha}$ and p-values $\tilde{p}_i$ are equivalently defined as

$$\tilde{\alpha} = \frac{\alpha}{m} \quad \text{and} \quad \tilde{p}_i = \min(mp_i, 1) \tag{3}$$

with m being the number of tests or analytical specifications, α the global significance level, and p_i the unadjusted p-values. However, the Bonferroni adjustment yields low power in rejecting false null hypotheses, particularly in the presence of strong dependencies among the tests (Bland and Altman, 1995). To address this issue, the minP procedure is introduced in the following section.

2.1.2 The minP approach

The single-step minP adjustment (Westfall et al., 1993; Westfall and Young, 1993) is a procedure that corrects for multiple testing, while indirectly accounting for the dependencies between tests and thus leading to an increased power compared to the Bonferroni approach. Note, that it controls the FWER only weakly, which is sufficient in the context of RDF. Let P_ℓ denote the random variable corresponding to the unadjusted p-value for H_0^ℓ (Dudoit et al., 2003). The adjusted p-values, denoted as $\tilde{p}_i$ for $i = 1, \ldots, m$, are derived from the distribution of the smallest p-value among

$p_1, \ldots, p_m$. The *single-step minP* adjusted p-values are thus defined as:

$$\tilde{p}_i = P\left(\min_{1\leq \ell \leq m} P_\ell \leq p_i \mid \cap_{i=1}^{m} H_0^i\right) \tag{4}$$

To apply the minP adjustment method, two key components are required: (1) the original unadjusted p-values p_i—e.g., derived from the different analytical specifications, and (2) p-values obtained from the analysis of permuted datasets, as we approximate the probability in Equation (4) using permutations of the original data that mimic the global null-hypothesis $\cap_{i=1}^{m} H_0^i$. Adjusting the minimal p-value involves counting how often the minimal p-values from the permuted datasets are smaller than the original minimal p-value. This proportion constitutes the adjusted p-value.

For illustrative purposes, consider a researcher comparing variable X between two independent groups, using either a two-sample t-test or a nonparametric Wilcoxon test. This results in two analytical specifications ($m = 2$). In practise, the number of specifications, m, can be considerably higher, and the degree of adjustment depends on both the total number of specifications and the interdependence of the hypotheses. The researcher conducts these two tests, resulting in two unadjusted p-values $\tilde{p}_i$. To implement the minP method, the procedure involves performing B iterations, in which the group assignments are randomly shuffled without replacement for each iteration. Hypothesis tests are then applied to each permuted dataset, resulting in a matrix of results with dimensions $[m, B]$, where m is the number of specifications and B represents the number of permutations. Each iteration provides a set of p-values corresponding to the two specifications, leading to two p-values per iteration. To adjust the initial minimal p-value, the number of instances in which the permuted minimal p-values are smaller than the original p-value is counted. This count yields the min-P adjusted p-value $\tilde{p}_i$. For a more formal description of the approach, see Procedure 1.

Procedure 1: minP adjustment (Westfall et al., 1993; Westfall and Young, 1993) in the context of RDF.

```
Input: Original dataset $D$ including exposure $X$ and outcome $Y$, analytical
  specifications $i = 1, ..., m$
1. Apply each analytical specification $i = 1, ..., m$ to $D$ to calculate the
  unadjusted p-values $p_1, \dots, p_m$
2. Randomly shuffle $Y$ without replacement and generate $B$ permuted datasets
  $\tilde{D}_j$ with $j = 1, ..., B$
3. for $\tilde{D}_j$ with $j = 1, ..., B$ do
    3.1 Apply each analytical specification $i = 1, ..., m$
    3.2 Return $\min_{1 \leq \ell \leq m} p_{\ell j}$
end
4. for $p_i$ with $i = 1, ..., m$ do
    4.1 Calculate minP adjusted p-values: $\tilde{p}_i = \mathbb{1}_{\{\min_j p_{\ell j} \leq p_i\}}/B$
end
```

2.2 Mendelian Randomisation—an instrumental variable approach

The use of instrumental variables originates in econometrics (Angrist and Krueger, 2001). The key idea is to isolate the variation in a risk factor that is caused by an exogenous instrumental variable (IV), and use this variation to estimate the risk factor's causal effect on an outcome. Applying the IV approach to epidemiology in order to estimate causal effects from observational data was a logical next step as alleles are assigned randomly during meiosis before birth—and thus, independently of the potential confounding. The development of the Mendelian Randomisation (MR) approach can be traced in Katan (1986) and Davey Smith and Ebrahim (2003).

The key idea of (multivariable) MR is to utilise genetic variants G_i as IVs in order to infer causal effects θ_j of exposures X_j on an outcome Y for $i \in 1, ..., n$ and $j = 1, ..., d$ (Burgess et al., 2013), as illustrated in Figure 2.

Multivariable MR incorporates multiple risk factors into a single model, accounting for measured pleiotropy. Pleiotropy is a violation to the exclusion restriction assumption, which refers to the effect of any genetic variant G_i on the outcome Y through any pathway other than the risk factors X_j included in the MR model—as depicted by the red dashed lines in Figure 2. To define a *valid* IV, the genetic variants in the multivariable MR analysis must satisfy the following assumptions for each genetic variant G_i and each risk factor, where $i = 1, ..., n$ and $j = 1, ..., d$ (Burgess et al., 2013; Zuber et al., 2020):

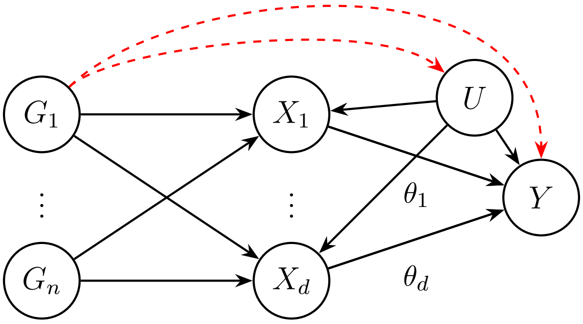

Figure 2: Causal directed acyclic graph for the multivariable MR scenario in Mandl et al. (2025a). Genetic variants G_i ($i \in 1...n$), a set of confounders U and causal effects of the risk factors X_j ($j \in 1...d$) on the outcome Y being θ_j. The red dashed lines represent an effect caused by pleiotropy.

IV1: Relevance—each genetic variant G_i is associated with at least one of the risk factors X_j.

IV2: Exchangeability—each genetic variant G_i is not associated with any confounder of the risk factor-outcome associations.

IV3: Exclusion restriction—each genetic variant G_i is independent of the outcome Y conditional on the risk factors X_j and confounders U.

RF1: Relevance—each risk factor X_j needs to be strongly instrumented by at least one genetic variant G_i.

RF2: No multi-collinearity—each risk factor X_j considered in the analysis cannot be linearly explained by the genetic associations of any other risk factor X_j or by the combined genetic associations of several other risk factors included in the analysis.

If IV1 to RF2, linearity, and homogeneity assumptions hold, the consistent estimates of the direct causal effects θ_j can be obtained from individual-level data via a *two-stage least squares* (2SLS) approach or through the multivariable two-sample summary-level *inverse variance weighted* (IVW) method, with weights $se(\hat{\beta}_{Y_i})^{-2}$ being the inverse of the estimated variance for genetic variant i and $\hat{\beta}_{X_{ij}}$, and $\hat{\beta}_{Y_i}$ being the genetic effects of G_i on X_{ij} and Y_i for variant i and risk factor j (Burgess and Thompson, 2015):

$$\hat{\beta}_{Y_i} = \sum_{j=1}^{d} \theta_j \hat{\beta}_{X_{ij}} + \varepsilon_i. \tag{5}$$

As with many methods, the underlying assumptions can never be fully verified. In the context of detecting pleiotropic effects, tests for heterogeneity may be used to check if

IV3 holds. These tests are commonly based on *Cochran's Q*, i.e., the test statistic is based on the weighted sum of squared residuals—with variations in the choice of the weighting factors (Del Greco et al., 2015). Cochran's Q is especially well-suited in two–sample summary–level MR which can be interpreted as a meta-analysis across genetic variants. Sanderson et al. (2019) introduced a generalized version of the Q-statistic for multivariable MR, which is defined as

$$Q = \sum_{i=1}^{n} \left(\frac{1}{\omega_i}\right) \left(\hat{\beta}_{Y_i} - \sum_{j}^{d} \hat{\theta}_j \hat{\beta}_{X_{ij}}\right)^2 \sim \chi^2_{(n-d)}, \tag{6}$$

with SNP index i, risk factor index j, and ω_i the weight approximated either using 1^{st} or 2^{nd} order weights. Cochran's Q follows a χ^2_{n-d} distribution with $n-d$ degrees of freedom under H_0. $\omega_i = \sigma^2_{Y_i}$ are 1^{st} order weights, with σ_{Y_i} being the standard error of $\hat{\beta}_{Y_i}$, and are known to lead to overdispersion in Q, which inflates the type I error rate (Bowden et al., 2018). Accordingly, Contribution 2 introduces the GC-Q method based on the work by Devlin and Roeder (1999) which corrects for overdispersion in the heterogeneity statistic.

2.3 Supervised Machine Learning

2.3.1 General principle

Following the notation of Bischl et al. (2023), supervised machine learning (ML) is concerned with fitting a model $f : \mathcal{X} \to \mathbb{R}^g$ with $g \in \mathbb{N}$ based on training data $\mathcal{D}$ with observations $(\mathbf{x}^i, y^i) \in \mathcal{X} \times \mathcal{Y}$ with $i \in 1, ...n$ in order to make predictions on unseen data drawn from the same underlying data-generating mechanism. $\mathcal{D}$ is sampled i.i.d. from an unknown distribution $\mathbb{P}_{XY}$. For a classification problem[3], $|\mathcal{Y}| = g$ with g being the number of classes. The goal is to fit a model f that generalises well to unseen data from $\mathbb{P}_{XY}$. $\hat{f}$—the approximation of f—is induced by a learner or inducer algorithm $\mathcal{I} : \mathcal{D} \times \Lambda \to \mathcal{H}$, with the inducer $\mathcal{I}$ being configured by hyperparameters $\lambda \in \Lambda$ and $\mathcal{H}$ being the function space of the model or hypothesis space. Let $L : \mathcal{Y} \times \mathbb{R}^g \to \mathbb{R}^+_0$ be a loss function that evaluates the deviation of the prediction from the actual label, then the true generalisation error is defined as $\mathcal{R}(f) := \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}[L(y, f(\mathbf{x}))]$. Many inducer algorithms rely on empirical risk minimisation to learn $\hat{f}$, i.e. they minimise $\mathcal{R}_{emp}(\hat{f}) = \mathbb{E}_{(\mathbf{x},y)\sim\mathbb{P}_{XY}}[L(y, \hat{f}(\mathbf{x}))]$. To prevent overfitting, the learned model is usually assessed on independent test data. Resampling methods like cross-validation can be applied to improve the stability of results. Frameworks like `mlr3` (Lang et al., 2019) make it quite easy to try different analytical strategies and use a variety of different

[3] Note: The classification setting can be easily transferred to the regression setting.

learners (or inducers) $\mathcal{I}$ to approximate the underlying relationship f. Here, we focus on *Random Forests* (RFs) and *Support Vector Machines* (SVMs).
The RF algorithm was first introduced by Breiman (2001a). Unlike standard bagging trees, the RF reduces correlation among individual trees by introducing randomness in the selection of the features. At each split within a decision tree, a random subset of features is selected from the full set as potential split candidates. The increased diversity of the ensemble improves the overall robustness and accuracy of the model (Hastie et al., 2009; James et al., 2013).
SVMs generalise the basic idea of the Support Vector Classifier that aims to find optimal separation between classes. This is done by maximising the margin between the so-called support vectors which are the closest data points to the decision boundary. For non-linear separation problems, SVMs employ the kernel trick to map features into higher–dimensional spaces and thus enable linear separation of the transformed data. The use of regularization parameters and soft margins enables SVMs to handle noise and reduce the risk of overfitting (Cortes and Vapnik, 1995; Hastie et al., 2009).

2.3.2 Permutation Feature Importance

The Permutation Feature Importance (PFI) metric for random forests was introduced by Breiman (2001a). Later, a generalised model-agnostic version was introduced by Fisher et al. (2019). The procedure is straight-forward: First, we estimate the original model error $e = L(y, \hat{f}(\mathbf{x}))$, with the fitted model $\hat{f}$, the features $\mathbf{x}$, the label y, and the loss function L (usually on the test set). Then for each feature x_j with $j = 1, ..., p$, we permute x_j and generate a new feature matrix $\mathbf{x}^{\text{perm,j}}$. Then, we estimate the new model error based on $\mathbf{x}^{\text{perm,j}}$: $e^{\text{perm,j}} = L(y, \hat{f}(\mathbf{x}^{\text{perm,j}}))$ and subsequently calculate the PFI$_j$ as either quotient or difference of $e^{\text{perm,j}}$ and e (Fisher et al., 2019; Molnar, 2025). The core concept of the PFI is to measure how much the prediction error increases when the values of a feature are randomly shuffled, thereby breaking its original relationship with the outcome. If the feature is important for the model's predictions, this disruption will significantly reduce performance. In contrast, permuting irrelevant features will have little or no impact on the model's accuracy (Molnar, 2025).

2.3.3 Applications

Supervised ML methods were used in both Contribution 1 and Contribution 3. Contribution 1 builds on the work of Becker-Pennrich et al. (2022) and uses a RF regressor ($\mathcal{I}$) to impute perioperative paO_2 values (y) for patients undergoing neurosurgery using non–invasive features ($\mathbf{x}$). These paO_2-levels were subsequently aggregated for each patient and used to test their association to post-operative complications.

Besides the GC-Q method from Contribution 2, Contribution 3 addresses a second use case that is based on the work of Woehrle et al. (2024) and includes a supervised ML classification task: The objective was to evaluate a diagnostic tool called *E-Nose*, which may be used to distinguish between individuals infected with SARS-CoV-2 and those who are not. Generally, the tool is able to detect volatile organic compounds in human breath. The analysis is conducted using an array of ten metal oxide semiconductor sensors. These produce specific signal patterns, as shown in Figure 3. In Contribution 3, two models were used to classify SARS-CoV-2 infected patients—a RF and a SVM classifier. Since the raw sensor data is complex, specific global time series features were extracted for each sensor (Hyndman et al., 2024): *Stability* is defined as the variance of means of each time series based on k non-overlapping windows: $\mathrm{S} = \mathrm{Var}(\mu_1, \mu_2, \ldots, \mu_k)$, with μ_l being the mean of the l–th window. *Lumpiness* is defined as the variance of variances of these k non-overlapping windows: $\mathrm{L} = \mathrm{Var}(\mathrm{Var}_1, \mathrm{Var}_2, \ldots, \mathrm{Var}_k)$, with Var_l denoting the variance of the l–th window. Furthermore, other features such as the *standard deviation*, the *minimum* and *maximum* were extracted. In total, this resulted in a feature matrix $\mathbf{x}$ with dimension 126×120—i.e., 120 features and 126 patients—and a binary label y indicating the COVID–19 status. After applying a nested resampling procedure (Varma and Simon, 2006) for hyperparameter tuning and performance evaluation, PFI was interpreted for both learners. The PFI measures were used in Contribution 3 to illustrate the *storytelling fallacy*, which is based on a newly introduced RDF in methodological research.

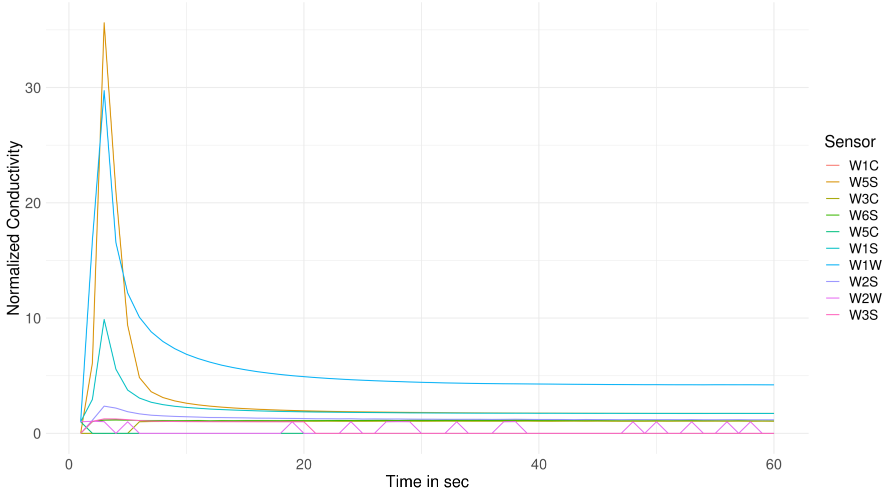

Figure 3: Raw data generated from the E-Nose analysis (Woehrle et al., 2024) for the upper respiratory tract of a single patient. The curves show temporal changes in conductivity for 10 different sensors over a period of 60 seconds, with each line representing one sensor. Conductivity values are normalized with baseline values. For each of these ten time series, global features were extracted.

3 Summary of contributions

The next section presents a summary of the four contributions of this dissertation, structured according to the three focus areas: RDF in applied, methodological, and didactic research.

3.1 Contribution 1: Mandl et al. (2024a) RDF in applied research—a multiple testing problem

»One statistical analysis must not rule them all.«

— Wagenmakers et al. (2022)

As Hoffmann et al. (2021) state, routinely collected clinical data pose even greater challenges than other types of data, as it is not initially generated for scientific purposes. Because such data are often messy and the data-generating process is typically unknown, issues related to measurement, data preprocessing, and model/method uncertainty play a central role in their analysis. Now, let us start with a common but loosely defined research hypothesis along the lines of

> *»partial arterial pressure of oxygen (paO_2) during neurosurgery has an impact on post-operative complications of the patients«*

that was investigated in Contribution 1 (Mandl et al., 2024a). The project was based on a study by Becker-Pennrich et al. (2022) that had used routinely collected data from a Munich-based university hospital to impute missing paO_2 values during neurosurgery using machine learning methods—as the continuous measurement of these is usually too invasive. Its motivation stems from the observation that, although the harmful consequences of hypoxia have been well established, the clinical effects of hyperoxia remain insufficiently understood and are subject to continued debate.

As mentioned in Section 1, many analysis strategies exist to address the research question at hand. Some potential choices include: the handling of missing values, the model choice to predict unobserved paO_2-levels, the choice of (hyper-)parameters, the creation of the risk and outcome variables, and the testing method. In this project, we came up with a total number of $n = 48$ reasonable analysis strategies, i.e., including all steps performed before the final statistical test and the choice of test itself. The problem arises when selectively reporting a single analysis strategy, which in turn leads to over-optimistic and non-replicable results.

When different analytical approaches—resulting in some kind of statistical test—are applied to the same research question, it effectively results in a *multiple testing problem*. Let m be the number of analysis strategies a researcher considers and H_0^i, $i = 1, \ldots, m$ the corresponding null hypotheses tested for each of the m analyses. Our vaguely defined research hypothesis may be formalized differently, as shown in Table 4.

Option 1	**Option 2**
$\mathbf{H_0}$ The mean paO_2 level is equal in patients with and without post-operative complications. **vs.** $\mathbf{H_1}$ The mean paO_2 level is not equal in these two groups.	$\mathbf{H_0}$ The rate of post-operative complications is equal for patients with $paO_2 < 200\,\text{mmHg}$ and those with $paO_2 \geq 200\,\text{mmHg}$. **vs.** $\mathbf{H_1}$ The rate of post-operative complications is not equal for patients with $paO_2 < 200\,\text{mmHg}$ and those with $paO_2 \geq 200\,\text{mmHg}$.

Table 4: Two different formalizations of the vaguely defined research hypothesis as introduced in Mandl et al. (2024a).

If the existence of various analysis strategies is now defined as a multiple testing problem, one can draw on established methods from this field of research. Our main goal is to control the family-wise error rate (FWER), i.e., the probability to make at least one type I error when testing H_0^i with $i = 1, ..., m$: $P(\text{reject at least one true } H_0^i)$. Therefore, we propose using the single-step minP adjustment method (Westfall et al., 1993; Westfall and Young, 1993) as this approach stands out due to its relatively intuitive principle and higher power compared to, e.g., the Bonferroni correction, while still ensuring weak control of the FWER.

The central idea of Contribution 1 is the formalisation of different analytical strategies as multiple testing problem. The proposed minP approach is subsequently tested in a (real data based) simulation study to compare the type I error rates to an unadjusted and Bonferroni adjusted strategy. Furthermore, we compare these approaches in the above-mentioned study on the impact of paO_2 on post-operative complications after neurosurgery.

We conclude that methods like the minP correction may help to choose the strategy that provides the strongest evidence, while controlling the type I error rate and thereby reducing the risk of non-replicable results. Further potential developments are discussed in Section 4.

3.2 Contributions 2–3: Mandl et al. (2025a) and Mandl et al. (2025b) RDF in methodological research—the *storytelling fallacy*

> *»[A]ccept that there is no universally best method […].«*
>
> — Strobl and Leisch (2024)

The concept of RDF has mainly been discussed in fields of applications of statistics, such as epidemiology and psychology. However, we can assess methodological research analogously. Methodological research in statistics focuses on developing and evaluating new methods. Researchers typically face similar decision problems when designing methodological comparison studies. The use of specific datasets, preprocessing procedures, and method variants that are subsequently selectively reported may lead to over-optimistic results of the preferred/newly introduced method. Illustrations regarding RDF within methodological studies can be found in Jelizarow et al. (2010), Nießl et al. (2022, 2024), Ullmann et al. (2023), and Pawel et al. (2024).
In this context, Contribution 3 introduces a new kind of RDF which focuses on real data examples that are often part of methodological studies. Authors might tell a domain-specific "story" supporting the results of their preferred method(s). Nonetheless, there is a multitude of equally plausible stories that may lead to different conclusions. In this sense, we define the so-called *storytelling fallacy* as the selective use of anecdotal domain-specific knowledge to support the superiority of the preferred method in real data examples. This practice may lead to biased and non-replicable results. The concept is illustrated using two examples of our own research (Mandl et al., 2025a; Woehrle et al., 2024). In this section, we focus on the example dealing with the detection of pleiotropic SNPs in MR analyses, as introduced in Contribution 2:

The GC-Q method is based on the work of Devlin and Roeder (1999) and uses a correction factor in the standard heterogeneity statistic to correct for overdispersion. In applied analyses using the two–sample summary–level MR setting, excess heterogeneity may arise from several additional sources, including widespread but negligible pleiotropic effects, or small discrepancies in allele frequencies between the samples used for exposure and outcome associations. In Contribution 2, we propose to correct for overdispersion by using an estimated inflation factor to identify and remove outlying instruments that may be invalid due to pleiotropy. This correction approach is inspired by the *Genomic Control* method, which was originally developed in the context of genome-wide association studies (GWAS) (Devlin and Roeder, 1999). Motivated from a Bayesian standpoint, the distribution of the local q-statistic—considering the presence of pleiotropic SNPs—can be represented by a mixture of two χ^2 distributions. A straightforward frequentist estimate of the inflation parameter can be computed directly from the data. Subsequently, the GC-Q method is evaluated against a variety of other approaches for detecting outlying SNPs, using both an extensive simulation study and real data examples.

With this new methodology in mind, we illustrate the so-called *storytelling fallacy*. Usually, methodological evaluations include real data applications. In the following, the causal effect of circulating vitamin D levels (exposure: X) on multiple sclerosis (MS, outcome: Y) using an Mendelian Randomisation (MR) approach is investigated. Table 5 shows the results of two pleiotropy detection methods: Method A (Standard) and the newly introduced Method B (GC-Q). While Method A identifies one pleiotropic SNP—*rs4944958*—Method B detects no pleiotropic SNP.

Method	pleiotropic SNP
Method A (Standard)	rs4944958
Method B (GC-Q)	—

Table 5: Real data analysis from Mandl et al. (2025a): pleiotropic SNPs for the analysis of vitamin D on MS in the univariable MR analysis.

Imagine playing devil's advocate for each approach. We can easily develop two domain-specific stories to support each method's superiority, as shown in Contribution 3. Assume that we want to justify the use of Method A with a plausible domain-specific Story A:

> »*Genetic variants linked to vitamin D fall into three main categories: those involved in the direct vitamin D pathway, U/V absorption, and cholesterol metabolism (Manousaki et al., 2020). The SNP rs4944958, flagged by method A, is an intron of the NADSYN1 gene, which influences cholesterol precursor synthesis. Given cholesterol metabolism's potential role as a confounder in multiple sclerosis (MS) (Murali et al., 2020), this suggests that rs4944958 may exert horizontal pleiotropy. This interpretation aligns with the results of method A.*«

However, the development of a plausible Story B is equally applicable to Method B:

> »*The SNP rs4944958, identified by method A, is considered a perfect proxy for rs12785878 (Mokry et al., 2015), a genetic variant consistently associated with serum vitamin D levels in multiple studies. For this reason,*

> *rs4944958 has been explicitly included in related MR studies, e.g., Mokry et al. (2015). Although it also plays a role in cholesterol metabolism, the relevance of this pathway in mediating effects on multiple sclerosis remains unresolved (Lorincz et al., 2022). Consequently, excluding rs4944958 based on presumed pleiotropy appears unjustified, and its removal by method A likely represents a false positive finding. These considerations are consistent with the findings from method B.*«

These two contrasting interpretations (or "stories") that might be constructed to support the superiority of both Method A and B, demonstrate that the interpretation of real data examples is inherently uncertain—and highlights the need for caution when concluding from real data analyses in methodological studies. This raises the question of whether a newly introduced method must always outperform its competitors in every analytical setting. We agree with Strobl and Leisch (2024) and think that we have to »*accept that there is no universally best method*«. Furthermore, from a general scientific perspective, employing multiple analytical strategies may yield valuable insights—therefore, it is advisable to consider and potentially combine several approaches rather than relying on a single one.

3.3 Contribution 4: Mandl et al. (2024b) RDF in a teaching context

> *»[I]t seems to me that statistics is often sold as a sort of alchemy that transmutes randomness into certainty [...].«*
>
> — Gelman (2016)

In addition to examining RDF in applied and methodological contexts, an essential perspective on the problem is missing: university education. As Gelman (2016) notes in the opening quote of this section, statistics is often portrayed as a tool for transforming randomness into certainty—and not only to those outside the field, but also to those within. Statistics and data analysis education often emphasises learning specific techniques sequentially and isolatedly. For example, students might first take a course focused on linear models, without addressing realistic analytical challenges such as handling missing data and outliers. These issues are typically introduced in more advanced courses—however, without adopting a holistic analytical approach. In the classroom, students usually work with clean, well-structured data and are guided through (toy) examples with clear, straightforward solutions. These exercises often include significant results, which may condition students to expect similar outcomes in their future real-world tasks.

While this approach mis pedagogically convenient, it might give students the misleading impression that every data analysis task has a single correct solution and will naturally produce significant results. In practise, analytical challenges often arise simultaneously and interact in complex ways. These misconceptions are further reinforced by academic publications, in which researchers typically present only one analytical strategy and report noteworthy and significant results, without detailing the many alternative strategies that may have been explored. This lack of transparency can mislead readers into believing that data analysis is a straightforward, deterministic process, rather than an iterative and interpretive one.

We argue that exposing students only to clean data sets and one single, supposedly correct analysis strategy that reliably produces significant or interesting results leaves them poorly prepared for the complexities of real-world data analysis. More critically, they may struggle to recognize and guard against the risks of selective reporting. While many courses focus on the correct application of statistical models, it is equally important to foster students' awareness of aleatoric and epistemic uncertainties and RDF. Creating this awareness is essential for promoting responsible and transparent empirical research.

To address these challenges, Contribution 4 presents a seminar course designed for advanced undergraduate and early graduate students in fields such as statistics or data science. The course aims to raise awareness of the multiplicity of analysis strategies and to equip students with theoretical frameworks and practical tools to deal with the problem of RDF through a combination of conceptual instructions and hands-on sessions.

As shown in Table 6, the concept is structured around four fundamental elements: The first element (**I**) consists of the key concepts of reproducibility and replicability, including an introduction to version control frameworks and R-Markdown. The second fundamental part (**II.1/2**) of the course focuses on replicability: Each student performs two data analysis tasks with specific pre-defined instructions. Between these tasks (**III**), they are introduced to a theoretical module that addresses relevant meta–scientific topics, such as different sources of uncertainties, selective reporting, and potential strategies for addressing these challenges. While the practical task evaluates the extent of selective reporting, the second theoretical module acts as an intervention to prevent students from selectively reporting only the most promising results. The effect of the intervention might be assessed by comparing the results of the two practical sessions, since the second task follows the (meta-scientific) intervention and involves re-analysing data generated from the same data-generating mechanism as in the first task. The students' experiences with the practical sessions, the intervention effect, and their key takeaways (e.g., over-optimistic results) are discussed in the debriefing sessions (**IV**).

	Topic	Details
I	Reproducibility	Introduction to frameworks like Git and R-Markdown.
II.1	1st practical session	First assignment.
III	Lecture on meta-scientific concepts	Uncertain choices in the analysis of empirical data, selective reporting, and ways to address these issues.
II.2	2nd practical session	Second assignment.
IV	Debriefing:	Review and discussion of results.

Table 6: Sample structure of the course as described in Mandl et al. (2024b).

Given that a lack of awareness of RDF has contributed to the replication crisis, addressing these issues in academic teaching is crucial. In particular, we must raise awareness of the negative consequences of selective reporting in the education of future (empirical) scientists. The growing availability of large, complex datasets introduces greater analytical challenges and increases the uncertainty of analytical choices even further. Therefore, students must be adequately equipped to address these challenges.

4 Conclusion and Outlook

> *»[A] solution to the current crisis is to acknowledge the inherent uncertainty in scientific findings.«*
>
> — Hoffmann et al. (2021)

As mentioned in Section 1, this thesis is motivated from a philosophical standpoint. Statistics draws inferences while dealing with different sources of uncertainty. It tries to unveil causal relationships, unravel complex associations, and test hypotheses against evidence. Some of these techniques come at a certain price: Causal inference is usually strongly assumption-driven, while opening the "black box" of certain ML models seems to be nearly impossible. If one takes a step further towards the underlying principle of the field itself, it even gets more complex—e.g., what does the underlying truth look like? Is there a true parameter value, or is the value itself uncertain? This is where Heisenberg's uncertainty principle—which could open a whole new discussion here—comes to mind. That said, this thesis is part of a broader range of pragmatic approaches aimed at addressing some of the problems of the replication crisis. Emphasising both applied and methodological research is one way to see the problem holistically. We do not believe that only the improper application of methods is the problem. Rather, the incentive structure within academia further exacerbates the implications of this crisis. Furthermore, we must address these problems from the very beginning in academic teaching to end this self-reinforcing loop. Meta-science is on the right track to do so—yet it is essential to remain constructive rather than merely critical or destructive. In the following, we present the main conclusions drawn from this thesis' contributions and offer an outlook on future work.

4.1 Conclusion

RDF come in many shapes and forms—be it in an applied or methodological context. Furthermore, teaching meta-scientific concepts must not be overlooked, as doing so allows us to foster awareness from the very beginning. Some RDF are quite apparent, whereas others are not that obvious initially.

In applied statistical research, effective methods exist for addressing RDF. Moreover, we can draw on existing tools and methods if the multiplicity of analysis strategies is understood as a multiple testing problem. One of the goals of this thesis is to build bridges between different communities that have each developed solutions to known problems. In our case, these are the meta-science and the multiple testing community. Furthermore, *methodological studies* can be biased in favor of the authors' preferred method(s) through selective reporting. Numerous recommendations—often grounded in principles of scientific integrity and open science—have been developed to ensure that such studies are conducted as neutrally and objectively as possible. Nonetheless, the *storytelling fallacy* shows that some RDF might be harder to spot than others. Thus, careful interpretation is not only required for simulation studies but also for real data examples.

Finally, with the growing volume of data and new methodologies, it is crucial to prepare future empirical scientists for the issues discussed in this thesis. Therefore, *academic teaching* must raise awareness of the implications of uncertainty and RDF to ensure the validity of statistical results.

So, where do we go from here to address these problems? First and foremost, we must acknowledge the inherent uncertainty in scientific findings. Subsequently, we must revisit and clarify our research goals. One way to achieve this for the case of methodological research is through the various *»phases of methodological research«* (Heinze et al., 2024). The underlying idea is to conceptualize methodological research in analogy to clinical trials: Early-phase studies might introduce new methodological ideas based on theoretical considerations while late-phase studies aim to generate robust evidence across diverse contexts. This also aligns with the idea to abandon the *»one beats them all philosophy«* (Strobl and Leisch, 2024)—i.e., no method should be expected to perform better in every possible situation.

Although the expectation that new methods must outperform existing ones in every scenario and that research findings must always be “positive”, are often only implicit in the research community and among journal decision-makers, challenging this view could improve the underlying incentive structure. Ultimately, this may encourage a more balanced and less biased presentation of results in both applied and methodological research.

4.2 Outlook

As a final point, some topics are beyond the scope of this thesis, but should nonetheless be briefly addressed. Contribution 1 introduced a way to deal with the multiplicity of analysis strategies. However, in multivariate regression settings with multiple confounders, defining a valid permutation scheme may be challenging. Permuting the exposure breaks its link to confounders, while permuting the outcome disrupts exposure–outcome and confounder–outcome associations. More advanced permutation methods may be preferable (Berrett et al., 2020; Girardi et al., 2024). Alternatively, asymptotic approaches, such as the one by Ristl et al. (2020), offer an efficient way to

adjust p-values and confidence intervals for the multiplicity of analysis strategies.
Nevertheless, these approaches do not represent a universal solution, and despite their flexibility in many contexts, they rely on assumptions—such as parametric test structures and specific input data requirements—that do not always hold in complex settings.
Furthermore, validation-based strategies (Daumer et al., 2008)—as used in the ML literature—involve sample-splitting. In this context, multiple analytical strategies might be run on a "training" set first. The preferred analysis strategy is validated on a holdout set. Like the minP method in Contribution 1, this reduces power. Still, the concept is compelling, as a key advantage is that it controls the type I error rate even when researchers are unaware of the risks linked to RDF or unable/unwilling to clearly report the number of analytical approaches conducted during their analyses.
Some theoretical considerations for using the validation approach and related power considerations have already been discussed in Labonne and Fafchamps (2017).
So, which approach is superior? Choosing the right approach obviously depends on the context, e.g., power considerations might result in favoring a particular strategy.
Moreover, these approaches might be combined.
Finally, sometimes researchers unintentionally "fool" themselves (Nuzzo, 2015) which makes the control of the type I error rate impossible. Study registration involves a (publicly available) pre-specified analysis plan (Hardwicke and Wagenmakers, 2023; Munafò et al., 2017; Nosek et al., 2018) and might help to address this issue. We encourage the use of this practise to improve openness and trustworthiness in empirical research (Naudet et al., 2024). However, even in clinical trials, there is a debate about whether analysis plans are detailed enough to prevent potential selective reporting (Greenberg et al., 2018). Ultimately, even in the case of meta-scientific practices—such as the approaches discussed in this thesis—we simply may have to accept that there is no universally best method.

References

Andrade, C. (2021). HARKing, Cherry-Picking, P-Hacking, Fishing Expeditions, and Data Dredging and Mining as Questionable Research Practices. *The Journal of Clinical Psychiatry*, *82*(1), 25941. https://doi.org/10.4088/JCP.20f13804

Angrist, J. D., and Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69–85. https://doi.org/10.1257/jep.15.4.69

Baio, G., and Dawid, A. P. (2015). Probabilistic sensitivity analysis in health economics. *Statistical Methods in Medical Research*, *24*(6), 615–634. https://doi.org/10.1177/0962280211419832

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454. https://doi.org/10.1038/533452a

Becker-Pennrich, A. S., Mandl, M. M., Rieder, C., Hoechter, D. J., Dietz, K., Geisler, B. P., Boulesteix, A.-L., Tomasi, R., and Hinske, L. C. (2022). Comparing supervised machine learning algorithms for the prediction of partial arterial pressure of oxygen during craniotomy. *medRxiv*. https://doi.org/10.1101/2022.06.07.22275483

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 175–197. https://doi.org/10.1111/rssb.12340

Bischl, B., Binder, M., Lang, M., Pielok, T., Richter, J., Coors, S., Thomas, J., Ullmann, T., Becker, M., Boulesteix, A.-L., Deng, D., and Lindauer, M. (2023). Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery*, *13*(2), e1484. https://doi.org/10.1002/widm.1484

Bland, J. M., and Altman, D. G. (1995). Multiple significance tests: The Bonferroni method. *BMJ*, *310*(6973), 170. https://doi.org/10.1136/bmj.310.6973.170

Boulesteix, A.-L., Hornung, R., and Sauerbrei, W. (2017). On fishing for significance and statistician's degree of freedom in the era of big molecular data. In W. Pietsch, J. Wernecke and M. Ott (Eds.), *Berechenbarkeit der Welt? Philosophie und Wissenschaft im Zeitalter von Big Data* (pp. 155–170). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-12153-2_7

Bowden, J., Del Greco M, F., Minelli, C., Zhao, Q., Lawlor, D. A., Sheehan, N. A., Thompson, J., and Davey Smith, G. (2018). Improving the accuracy of two-sample summary-data mendelian randomization: Moving beyond the nome assumption. *International Journal of Epidemiology*, *48*(3), 728–742. https://doi.org/10.1093/ije/dyy258

Brakenhoff, T. B., Mitroiu, M., Keogh, R. H., Moons, K. G., Groenwold, R. H., and van Smeden, M. (2018). Measurement error is often neglected in medical literature: A systematic review. *Journal of Clinical Epidemiology*, *98*, 89–97. https://doi.org/10.1016/j.jclinepi.2018.02.023

Breiman, L. (2001a). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Burgess, S., Butterworth, A., and Thompson, S. G. (2013). Mendelian Randomization analysis with multiple genetic variants using summarized data. *Genetic Epidemiology*, *37*(7), 658–665. https://doi.org/10.1002/gepi.21758

Burgess, S., and Thompson, S. G. (2015). Multivariable Mendelian Randomization: The use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, *181*(4), 251–260. https://doi.org/10.1093/aje/kwu283

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., and Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Caplan, A. L. (2023). Regaining Trust in Public Health and Biomedical Science following Covid: The Role of Scientists. *Hastings Center Report*, *53*(5), 105–109. https://doi.org/10.1002/hast.1531

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *158*(3), 419–444. https://doi.org/10.2307/2983440

Childers, C. P., and Maggard-Gibbons, M. (2019). Re: Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection? *Surgery*, *166*(1), 127–128. https://doi.org/10.1016/j.surg.2019.01.019

Childers, C. P., and Maggard-Gibbons, M. (2021). Same data, opposite results?: A call to improve surgical database research. *JAMA Surgery*, *156*(3), 219–220. https://doi.org/10.1001/jamasurg.2020.4991

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

Daumer, M., Held, U., Ickstadt, K., Heinz, M., Schach, S., and Ebers, G. (2008). Reducing the probability of false positive research findings by pre-publication validation–Experience with a large multiple sclerosis database. *BMC Medical Research Methodology*, *8*(18). https://doi.org/10.1186/1471-2288-8-18

Davey Smith, G., and Ebrahim, S. (2003). Mendelian randomization: Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, *32*(1), 1–22. https://doi.org/10.1093/ije/dyg070

Del Greco, F., Minelli, C., Sheehan, N. A., and Thompson, J. R. (2015). Detecting pleiotropy in Mendelian Randomisation studies with summary data and a continuous outcome. *Statistics in Medicine*, *34*(21), 2926–2940. https://doi.org/10.1002/sim.6522

Devlin, B., and Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997–1004. https://doi.org/10.1111/j.0006-341X.1999.00997.x

Douven, I. (2021). Abduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2021/entries/abduction/

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, *18*(1), 71–103. https://doi.org/10.1214/ss/1056397487

Fields, A. C., Lu, P., Palenzuela, D. L., Bleday, R., Goldberg, J. E., Irani, J., Davids, J. S., and Melnitchouk, N. (2019). Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection? *Surgery*, *165*(5), 953–957. https://doi.org/10.1016/j.surg.2018.11.012.

Fife, D. A., and Rodgers, J. L. (2022). Understanding the exploratory/confirmatory data analysis continuum: Moving beyond the "replication crisis". *American Psychologist*, *77*(3), 453–466. http://dx.doi.org/10.1037/amp0000886

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, *20*(177), 1–81. https://www.jmlr.org/papers/v20/18-760.html

Gelman, A. (2016). The problems with p-values are not just with p-values. In: Ronald L Wasserstein and Nicole A Lazar, The ASA Statement on p-values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://dx.doi.org/10.1080/00031305.2016.1154108

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. https://doi.org/10.1111/rssa.12276

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*. https://sites.stat.columbia.edu/gelman/research/unpublished/p_hacking.pdf

Gelman, A., and Loken, E. (2014). The statistical crisis in science: Data-dependent analysis–a "garden of forking path"–explains why many statistically significant comparisons don't hold up. *American Scientist*, *102*(6), 460–466. https://doi.org/10.1511/2014.111.460

Gelman, A., and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

Girardi, P., Vesely, A., Lakens, D., Altoè, G., Pastore, M., Calcagnı, A., and Finos, L. (2024). Post-selection inference in multiverse analysis (pima): An inferential framework based on the sign flipping score test. *Psychometrika*, *89*(2), 1–27. https://doi.org/10.1007/s11336-024-09973-6

Greenberg, L., Jairath, V., Pearse, R., and Kahan, B. C. (2018). Pre-specification of statistical analysis approaches in published clinical trial protocols was inadequate. *Journal of Clinical Epidemiology*, *101*, 53–60. https://doi.org/10.1016/j.jclinepi.2018.05.023

Gruber, C., Schenk, P. O., Schierholz, M., Kreuter, F., and Kauermann, G. (2023). Sources of uncertainty in machine learning–a statisticians' view. *arXiv preprint: 2305.16703*. https://doi.org/10.48550/arXiv.2305.16703

Hardwicke, T. E., and Wagenmakers, E.-J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, *7*(1), 15–26. https://doi.org/10.1038/s41562-022-01497-2

Harman, G., and Kulkarni, S. R. (2006). The problem of induction. *Philosophy and Phenomenological Research*, *72*(3), 559–575. https://www.jstor.org/stable/40040950

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

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. https://doi.org/10.1371/journal.pbio.1002106

Heinze, G., Boulesteix, A.-L., Kammer, M., Morris, T. P., White, I. R., and the Simulation Panel of the STRATOS initiative. (2024). Phases of methodological research in biostatistics—Building the evidence base for new methods. *Biometrical Journal*, *66*(1), 2200222. https://doi.org/10.1002/bimj.202200222

Henderson, L. (2024). The Problem of Induction. In E. N. Zalta and U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2024/entries/induction-problem/

Hodges, C. B., Stone, B. M., Johnson, P. K., Carter III, J. H., Sawyers, C. K., Roby, P. R., and Lindsey, H. M. (2023). Researcher degrees of freedom in statistical software contribute to unreliable results: A comparison of nonparametric analyses conducted in SPSS, SAS, Stata, and R. *Behavior Research Methods*, *55*(6), 2813–2837. https://doi.org/10.3758/s13428-022-01932-2

Hoffmann, S., Schönbrodt, F., Elsas, R., Wilson, R., Strasser, U., and Boulesteix, A.-L. (2021). The multiplicity of analysis strategies jeopardizes replicability: Lessons learned across disciplines. *Royal Society Open Science*, *8*(4), 201925. https://doi.org/10.1098/rsos.201925

Hüllermeier, E., and Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, *110*(3), 457–506. https://doi.org/10.1007/s10994-021-05946-3

Hyndman, R., Kang, Y., Montero-Manso, P., O'Hara-Wild, M., Talagala, T., Wang, E., and Yang, Y. (2024). *Tsfeatures: Time series feature extraction* [R package version 1.1.1.9000]. https://pkg.robjhyndman.com/tsfeatures/

Ioannidis, J. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning* (Vol. 112). Springer Cham. https://doi.org/10.1007/978-3-031-38747-0

Jarvis, M. F., and Williams, M. (2016). Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends in Pharmacological Sciences*, *37*(4), 290–302. https://doi.org/10.1016/j.tips.2015.12.001

Jelizarow, M., Guillemot, V., Tenenhaus, A., Strimmer, K., and Boulesteix, A.-L. (2010). Over-optimism in bioinformatics: An illustration. *Bioinformatics*, *26*(16), 1990–1998. https://doi.org/10.1093/bioinformatics/btq323

Katan, M. (1986). Apoupoprotein e isoforms, serum cholesterol, and cancer. *The Lancet*, *327*(8479), 507–508. https://doi.org/10.1016/S0140-6736(86)92972-7

Kempthorne, O. (1976). Statistics and the philosophers. In W. L. Harper and C. A. Hooker (Eds.), *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science: Proceedings of an International Research Colloquium held at the University of Western Ontario, London, Canada, 10–13 May 1973, Volume II, Foundations and Philosophy of Statistical Inference* (pp. 273–314). Springer Netherlands. https://doi.org/10.1007/978-94-010-1436-6_8

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klau, S., Hoffmann, S., Patel, C. J., Ioannidis, J. P., and Boulesteix, A.-L. (2021). Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. *International Journal of Epidemiology*, *50*(1), 266–278. https://doi.org/10.1093/ije/dyaa164

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

Klau, S., Schönbrodt, F., Patel, C. J., Ioannidis, J. P., Boulesteix, A.-L., and Hoffmann, S. (2023). Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. *Meta-Psychology*, *7*. https://doi.org/10.15626/MP.2020.2556

Labonne, J., and Fafchamps, M. (2017). Using split samples to improve inference on causal effects. *Political Analysis*, *25*(4), 465–482. http://dx.doi.org/10.1017/pan.2017.22

Lang, M., Binder, M., Richter, J., Schratz, P., Pfisterer, F., Coors, S., Au, Q., Casalicchio, G., Kotthoff, L., and Bischl, B. (2019). mlr3: A modern object-oriented machine learning framework in R. *Journal of Open Source Software*, *4*(44), 1903. https://doi.org/10.21105/joss.01903

Loken, E., and Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*(6325), 584–585. https://www.science.org/doi/abs/10.1126/science.aal3618

Lorincz, B., Jury, E. C., Vrablik, M., Ramanathan, M., and Uher, T. (2022). The role of cholesterol metabolism in multiple sclerosis: From molecular pathophysiology to radiological and clinical disease activity. *Autoimmunity Reviews*, *21*(6), 103088. https://doi.org/10.1016/j.autrev.2022.103088

Mandl, M. M., Becker-Pennrich, A. S., Hinske, L. C., Hoffmann, S., and Boulesteix, A.-L. (2024a). Addressing researcher degrees of freedom through minp adjustment. *BMC Medical Research Methodology*, *24*(1), 152. https://doi.org/10.1186/s12874-024-02279-2

Mandl, M. M., Boulesteix, A.-L., Burgess, S., and Zuber, V. (2025a). Outlier detection in mendelian randomization. *Statistics in Medicine*, *44*(15-17), e70143. https://doi.org/10.1002/sim.70143

Mandl, M. M., Hoffmann, S., Bieringer, S., Jacob, A. E., Kraft, M., Lemster, S., and Boulesteix, A.-L. (2024b). Raising awareness of uncertain choices in empirical data analysis: A teaching concept toward replicable research practices. *PLOS Computational Biology*, *20*(3), 1–10. https://doi.org/10.1371/journal.pcbi.1011936

Mandl, M. M., Weber, F., Wöhrle, T., and Boulesteix, A.-L. (2025b). The impact of the storytelling fallacy on real data examples in methodological research. *arXiv preprint: 2503.03484*. https://doi.org/10.48550/arXiv.2503.03484

Manousaki, D., Mitchell, R., Dudding, T., Haworth, S., Harroud, A., Forgetta, V., Shah, R. L., Luan, J., Langenberg, C., Timpson, N. J., et al. (2020). Genome-wide association study for vitamin d levels reveals 69 independent loci. *The American Journal of Human Genetics*, *106*(3), 327–337. https://doi.org/10.1016/j.ajhg.2020.01.017

Marín-Franch, I. (2018). Publication bias and the chase for statistical significance. *Journal of Optometry*, *11*(2), 67–68. https://doi.org/10.1016/j.optom.2018.03.001

Mayo, D. G. (1980). The philosophical relevance of statistics. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, *1980*(1), 96–109. https://doi.org/10.1086/psaprocbienmeetp.1980.1.192556

Mayo, D. G., and Hand, D. (2022). Statistical significance and its critics: Practicing damaging science, or damaging scientific practice? *Synthese*, *200*(3), 220. https://doi.org/10.1007/s11229-022-03692-0

Merton, R. K. (1993). *On the shoulders of giants: The post-italianate edition.* University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/O/bo3641858

Mokry, L. E., Ross, S., Ahmad, O. S., Forgetta, V., Smith, G. D., Leong, A., Greenwood, C. M., Thanassoulis, G., and Richards, J. B. (2015). Vitamin d and risk of multiple sclerosis: A mendelian randomization study. *PLoS Medicine*, *12*(8), e1001866. https://doi.org/10.1371/journal.pmed.1001866

Molnar, C. (2025). *Interpretable machine learning: A guide for making black box models explainable* (3rd ed.). https://christophm.github.io/interpretable-ml-book

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. https://doi.org/10.1038/s41562-016-0021

Murali, N., Browne, R. W., Fellows Maxwell, K., Bodziak, M. L., Jakimovski, D., Hagemeier, J., Bergsland, N., Weinstock-Guttman, B., Zivadinov, R., and Ramanathan, M. (2020). Cholesterol and neurodegeneration: Longitudinal changes in serum cholesterol biomarkers are associated with new lesions and gray matter atrophy in multiple sclerosis over 5 years of follow-up. *European Journal of Neurology*, *27*(1), 188–e4. https://doi.org/10.1111/ene.14055

Naudet, F., Patel, C. J., DeVito, N. J., Le Goff, G., Cristea, I. A., Braillon, A., and Hoffmann, S. (2024). Improving the transparency and reliability of observational studies through registration. *BMJ*, *384*, e076123. https://doi.org/10.1136/bmj-2023-076123

Nießl, C., Herrmann, M., Wiedemann, C., Casalicchio, G., and Boulesteix, A.-L. (2022). Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results. *WIREs Data Mining and Knowledge Discovery*, *12*(2), e1441. https://doi.org/10.1002/widm.1441

Nießl, C., Hoffmann, S., Ullmann, T., and Boulesteix, A.-L. (2024). Explaining the optimistic performance evaluation of newly proposed methods: A cross-design validation experiment. *Biometrical Journal*, *66*(1), 2200238. https://doi.org/10.1002/bimj.202200238

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., and Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.170827411

Nosek, B. A., and Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nuzzo, R. (2015). Fooling ourselves. *Nature*, *526*, 182–185. https://doi.org/10.1038/526182a

Olsson-Collentine, A., van Aert, R., Bakker, M., and Wicherts, J. (2023). Meta-analyzing the multiverse: A peek under the hood of selective reporting. *Psychological Methods*. https://dx.doi.org/10.1037/met0000559

Patel, C. J., Burford, B., and 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*(9), 1046–1058. https://doi.org/10.1016/j.jclinepi.2015.05.029

Pawel, S., Kook, L., and Reeve, K. (2024). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, *66*(1), 2200091. https://doi.org/10.1002/bimj.202200091

Probst, P., Boulesteix, A.-L., and Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *The Journal of Machine Learning Research*, *20*(1), 1934–1965.

Rehms, R., Ellenbach, N., Rehfuess, E., Burns, J., Mansmann, U., and Hoffmann, S. (2024). A bayesian hierarchical approach to account for evidence and uncertainty in the modeling of infectious diseases: An application to covid-19. *Biometrical Journal*, *66*(1), 2200341. https://doi.org/10.1002/bimj.202200341

Ristl, R., Hothorn, L., Ritz, C., and Posch, M. (2020). Simultaneous inference for multiple marginal generalized estimating equation models. *Statistical Methods in Medical Research*, *29*(6), 1746–1762. https://doi.org/10.1177/0962280219873005

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. https://doi.org/10.1177/0956797617723726

Rubin, M. (2023). Questionable Metascience Practices. *Journal of Trial & Error*, *4*(1). https://doi.org/10.36850/mr4

Sanderson, E., Davey Smith, G., Windmeijer, F., and Bowden, J. (2019). An examination of multivariable mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, *48*(3), 713–727. https://doi.org/10.1093/ije/dyy262

Sauerbrei, W., Abrahamowicz, M., Altman, D. G., le Cessie, S., Carpenter, J., and on behalf of the STRATOS initiative. (2014). Strengthening analytical thinking for observational studies: The stratos initiative. *Statistics in Medicine*, *33*(30), 5413–5432. https://doi.org/https://doi.org/10.1002/sim.6265

Schneider, J. W. (2015). Null hypothesis significance tests. a mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics*, *102*(1), 411–432. https://doi.org/10.1007/s11192-014-1251-5

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature*, *515*, 9. https://doi.org/10.1038/515009a

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., et al. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

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. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., and Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Smaldino, P. E., and McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Steegen, S., Tuerlinckx, F., Gelman, A., and Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Strobl, C., and Leisch, F. (2024). Against the "one method fits all data sets" philosophy for comparison studies in methodological research. *Biometrical Journal*, *66*(1), 2200104. https://doi.org/10.1002/bimj.202200104

Turner, E. H., Knoepflmacher, D., and Shapley, L. (2012). Publication bias in antipsychotic trials: An analysis of efficacy comparing the published literature to the us food and drug administration database. *PLoS Medicine*, *9*(3), e1001189. https://doi.org/10.1371/journal.pmed.1001189

Turner, S. A., Jung, H. S., and Scarborough, J. E. (2019). Utilization of a specimen retrieval bag during laparoscopic appendectomy for both uncomplicated and complicated appendicitis is not associated with a decrease in postoperative surgical site infection rates. *Surgery*, *165*(6), 1199–1202. https://doi.org/10.1016/j.surg.2019.02.010

Ullmann, T., Beer, A., Hünemörder, M., Seidl, T., and Boulesteix, A.-L. (2023). Over-optimistic evaluation and reporting of novel cluster algorithms: An illustrative study. *Advances in Data Analysis and Classification*, *17*(1), 211–238. https://doi.org/10.1007/s11634-022-00496-5

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(91), 1–8. https://doi.org/10.1186/1471-2105-7-91

Wagenmakers, E.-J., Sarafoglou, A., and Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*, 423–425. https://doi.org/10.1038/d41586-022-01332-8

Wasserstein, R. L., and Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, *73*(2), 129–133. https://doi.org/10.1080/00031305.2019.1583913

Westfall, P. H., Young, S. S., and Wright, S. P. (1993). On adjusting p-values for multiplicity. *Biometrics*, *49*(3), 941–945. http://www.jstor.org/stable/2532216

Westfall, P. H., and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment* (Vol. 279). John Wiley & Sons.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., and 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. https://doi.org/10.3389/fpsyg.2016.01832

Woehrle, T., Pfeiffer, F., Mandl, M. M., Sobtzick, W., Heitzer, J., Krstova, A., Kamm, L., Feuerecker, M., Moser, D., Klein, M., Aulinger, B., Dolch, M., Boulesteix, A.-L., Lanz, D., and Choukér, A. (2024). Point-of-care breath sample analysis by semiconductor-based e-nose technology discriminates non-infected subjects from sars-cov-2 pneumonia patients: A multi-analyst experiment. *MedComm*, *5*(11), e726. https://doi.org/10.1002/mco2.726

Zuber, V., Colijn, J. M., Klaver, C., and Burgess, S. (2020). Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian Randomization. *Nature Communications*, *11*(29). https://doi.org/10.1038/s41467-019-13870-3

A Contributions

A.1 Contribution 1

»Addressing researcher degrees of freedom through minP adjustment«

Citation

Mandl, M.M., Becker-Pennrich, A.S., Hinske, L.C., Hoffmann, S., Boulesteix, A.-L. (2024) Addressing researcher degrees of freedom through minP adjustment. *BMC Med Res Methodol* **24**, 152. https://doi.org/10.1186/s12874-024-02279-2

Authors' contributions

MM, ALB, and SH designed the study. MM implemented the method, performed the simulations, analyzed the data and interpreted the results. ABP and LH provided the motivating example. MM, ALB, and SH prepared the initial manuscript draft. ALB directed the project. MM, ALB, SH, ABP, and LH reviewed and edited the manuscript. All authors have read and approved the manuscript.

Rights and permissions

BMC Medical Research Methodology

RESEARCH Open Access

Addressing researcher degrees of freedom through minP adjustment

Maximilian M. Mandl[1,5*], Andrea S. Becker-Pennrich[1,2], Ludwig C. Hinske[2,3], Sabine Hoffmann[4] and Anne-Laure Boulesteix[1,5]

Abstract

When different researchers study the same research question using the same dataset they may obtain different and potentially even conflicting results. This is because there is often substantial flexibility in researchers' analytical choices, an issue also referred to as "researcher degrees of freedom". Combined with selective reporting of the smallest *p*-value or largest effect, researcher degrees of freedom may lead to an increased rate of false positive and overoptimistic results. In this paper, we address this issue by formalizing the multiplicity of analysis strategies as a multiple testing problem. As the test statistics of different analysis strategies are usually highly dependent, a naive approach such as the Bonferroni correction is inappropriate because it leads to an unacceptable loss of power. Instead, we propose using the "minP" adjustment method, which takes potential test dependencies into account and approximates the underlying null distribution of the minimal *p*-value through a permutation-based procedure. This procedure is known to achieve more power than simpler approaches while ensuring a weak control of the family-wise error rate. We illustrate our approach for addressing researcher degrees of freedom by applying it to a study on the impact of perioperative paO_2 on post-operative complications after neurosurgery. A total of 48 analysis strategies are considered and adjusted using the minP procedure. This approach allows to selectively report the result of the analysis strategy yielding the most convincing evidence, while controlling the type 1 error—and thus the risk of publishing false positive results that may not be replicable.

Keywords Multiplicity, Open science, Replication crisis, Researcher degrees of freedom, Uncertainty

Introduction

In recent years, the scientific community has become increasingly aware that there is a high analytical variability when analysing empirical data, i.e. there are plenty of sensible ways to analyse the same dataset for addressing a given research question, and they may yield (substantially) different results [1, 2]. If combined with selective reporting, this variability may lead to an increased rate of overoptimistic results, e.g.—depending on the context—false positive test results and inflation of effect sizes [3–5], or, beyond the context of testing and effect estimation, to exaggerated measures of predictive performance [6] or clustering validity [7].

Hoffmann et al. [8] outline six sources of uncertainty that are omnipresent in empirical sciences and lead to

*Correspondence:
Maximilian M. Mandl
mmandl@ibe.med.uni-muenchen.de
[1] Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, LMU Munich, Marchioninistr. 15, Munich 81377, Germany
[2] Department of Anaesthesiology, LMU University Hospital, LMU Munich, Marchioninistr. 15, Munich 81377, Germany
[3] Institute for Digital Medicine, University Hospital of Augsburg, University of Augsburg, Stenglinstr. 2, Augsburg 86156, Germany
[4] Department of Statistics, LMU Munich, Ludwigstr. 33, Munich 80539, Germany
[5] Munich Center for Machine Learning (MCML), Munich, Germany

variability of results in empirical research regardless of the considered discipline, namely sampling, measurement, model, parameter, data pre-processing, and method uncertainty. Failure to take these various uncertainties into account may lead to unstable, supposedly precise, but overoptimistic and thus potentially unreplicable results. Most importantly, model, parameter, data preprocessing and method uncertainties lead to the analytical variability mentioned above. In this context, Simmons et al. [3] denote the flexibility researchers have regarding the different aspects of the analysis strategy as "researcher degrees of freedom".

While it is clear that selective reporting of the "most favorable results" out of a multitude of results is a questionable research practice that invalidates statistical inference, it is less clear how researchers should deal with their degrees of freedom in practice. In this study, we suggest to tackle this issue from the perspective of multiple testing. More precisely, for analyses based on hypothesis testing we formalize researcher degrees of freedom as a multiple testing problem. We further propose to use an adjustment procedure to correct for the over-optimism resulting from the selection of the lowest p-value out of a variety of analysis strategies.

As the results of different analysis strategies addressing the same research question with the same data are usually highly dependent, a naive approach such as the Bonferroni correction is inappropriate. It would indeed lead to an unacceptable loss of power. Instead, we propose resorting to the single-step "minP" adjustment method [9, 10] and discuss its use in this context. The power achieved by the minP procedure is typically larger than with simpler approaches while ensuring a weak control of the family-wise error rate. This is because the procedure is based on the distribution of the minimal p-value, which is obviously affected by the level of correlation between the tests.

The minP procedure has the major advantage that it has a relatively intuitive principle, as illustrated by the following example. In a comment on a study by Mathews et al. [11] claiming that breakfast cereal intake before pregnancy is positively associated with the probability to conceive a male fetus, Young et al. [12] reinterpret the small p-value of 0.0034 obtained in the original article. They notice that Mathews et al. [11] did not only analyse the association between fetal sex and the consumption of breakfast cereals, but also many other food items—a typical case of multiple testing. Based on the analysis of permuted data (i.e. data with randomly shuffled fetal sex status), Young et al. [12] argue that "one would expect to see a p-value as small as 0.0034 approximately 28 percent of the time when nothing is going on". Implicitly, they apply the minP procedure for adjusting the smallest raw p-value of 0.0034 to 0.28 in this context where multiple tests are performed to investigate multiple food items. Our suggestion consists of translating this approach into the context of the analytical researcher degrees of freedom towards addressing the statistical factors of the replication crisis.

The minP procedure as used in the example by Young et al. [12] and considered in this paper is based on an approximation of the null distribution of the minimal p-value through a permutation-based procedure. We note, however, that such a permutation-based procedure is not always possible, and that resorting to theoretical asymptotical results on the distribution of the minimal p-value (or maximal statistic) is more appropriate in some cases, as will be discussed later.

The goal of this paper can be seen as building bridges between two scientific communities. On one hand, the metascientific community has long recognized that the replication crisis in science is partly related to multiplicity issues, but has to date neither formalized the issue in terms of multiple testing nor applied known adjustment procedures for reducing the occurrence of false positive results. On the other hand, the multiple testing community is increasingly developing theoretically founded general approaches to multiple testing taking into account the dependence of the tests; see Ristl et al. [13] for a recent important milestone. These approaches are however not yet routinely used to adjust for researcher degrees of freedom in practice. The reasons are manifold. The lack of communication between the two communities and the methodological complexity of these methods certainly play an important role. Another reason is that these approaches, even if increasingly efficient and general, do not address all types of analyses but only regression models, and require assumptions regarding the data format that may not always be fulfilled in practice. In this context, the present paper aims to formalize and demonstrate the use of minP to adjust for researcher degrees of freedom in simple situations not only involving linear models, while hopefully creating a common basis fostering communication between the two communities towards the development (by statistical researchers) and routine use (by applied data analysts) of more complex approaches. This paper aims to establish an easy approach designed to prevent the detection of false-positive findings in the context of fishing expeditions.

The rest of this paper is structured as follows. Problems related to researcher degrees of freedom are outlined in more detail in Background: researcher degrees of freedom section, including potential approaches for handling it in practice that were proposed in the literature. As a motivating example, Motivating example section presents a study on the impact of perioperative partial arterial pressure (paO_2) on post-operative complications after neurosurgery that uses routinely collected real-world data. Our suggested approach is described in Method section, while Illustration section shows its results on the example dataset and Discussion section briefly discusses limitations of the approach and possible extensions. Furthermore, we have added a brief tutorial to our GitHub repository to make the method's dissemination and application simple and understandable[1].

Background: researcher degrees of freedom

Overview

When analysing biomedical data, researchers are often confronted with a number of decisions that may appear trivial at first view, but often have a considerable impact on study results. Which confounders should we adjust for? How should we handle missing values and outliers? Should we log-transform a continuous variable? What about categorical variables with categories that include no more than a handful of patients? Should these small categories be merged? Is a parametric or non-parametric test more appropriate? The term "researcher degrees of freedom" [3] denotes, in a broad sense, this flexibility arising from the many analytical choices researchers face when analysing data in practice.

In most cases, neither theory nor precise practical guidance from the literature can reliably point researchers to the "best way" to analyse their data. Model selection techniques based, e.g., on the Akaike Information Criterion (AIC) and diagnostic tools (e.g., to assess whether a variable is normally distributed) may be helpful in some cases. However, they most often do not provide definitive clear-cut answers to all the arising questions. Furthermore, the choice of these techniques is itself affected by uncertainty: there usually exist several suitable variants of them. For example, should we prefer the AIC or the Bayesian Information Criterion (BIC) for model selection? Should we use a QQ-plot or apply a test (if yes, which one and at which level?) to assess normality of a variable?

Combined with selective reporting, researcher degrees of freedom can lead to an increased rate of false positive results, inflation in effect sizes, and overoptimistic results [3–5, 8]. The terms "p-hacking" and "fishing for significance" have been used in the context of hypothesis testing to denote the selective reporting of the most significant results out of a multitude of results arising through the multiplicity of analysis strategies. The resulting optimism is however not limited to the context of hypothesis testing. "Fishing expeditions" (also termed "cherry-picking" or "data dredging") are common issues in all types of analyses beyond hypothesis testing [7].

The multiplicity of possible analysis strategies particularly affects studies involving electronic health records and administrative claims data, which currently raise hopes and promises of "real-world" evidence and personalized treatment regimes. With data that have not been primarily collected for research purposes, uncertainties related to the analysis strategies may indeed be even more pronounced compared to the analysis of classical observational research data. In the last few years, contradictory results have been published in this setting, which can be viewed as a consequence of the uncertainties in a broad sense. See for example the conflicting results on infectious complications associated with laparoscopic appendectomies [14–17] and on the association between cardiovascular disease and marijuana consumption [18, 19]. In both cases, different teams of researchers used the same data set to answer the same research question and found contradictory results which can be explained by seemingly trivial choices.

Partial solutions and related work

There are a number of approaches that have been proposed to deal with uncertainty regarding the analysis strategy and are preferable to the selective reporting of the preferred results.

A natural approach is to fix the analysis strategy in advance, i.e. prior to running the analyses, to avoid obtaining multiple results in the first place. For more transparency, this may be done within a publicly available pre-registration document [20–22], thus preventing result-dependent selective reporting [23]. This type of pre-registration is the standard for clinical trials [24]. However, even in the strictly regulated context of clinical trials, there is some controversy about the question whether statistical analysis plans of clinical trials are detailed enough [25] to prevent potential selective reporting. Fixing the analysis strategy in advance tends to be even more difficult for exploratory research questions and for complex data sets and research questions.

The opposite approach consists of transparently acknowledging uncertainty and reporting the variety of results obtained with the considered analysis strategies. This concept has been proposed in different variants in the last decade: it encompasses, e.g., the vibration of effect framework [26, 27], multiverse analyses [28] and the specification curve analysis [29, 30]. With these approaches, the multiple reported results might be conflicting, sometimes yielding a confusing picture and a paper without clear-cut take-home message. In other words, the pitfalls of selective reporting are obviously avoided, but this comes at a high price in terms of interpretability and clarity.

[1] https://github.com/mmax-code/researcher_dof

Finally, let us mention the approach of conducting various analyses, selecting the preferred results but—instead of reporting it in a cherry-picking fashion—publishing it only if it can be qualitatively confirmed by running the exact same analysis on independent "validation" data [31]. This is the approach Ioannidis [32] indirectly recommends when claiming *"Without highly specified a priori hypotheses, there are hundreds of ways to analyse the dullest dataset. Thus, no matter what my discovery eventually is, it should not be taken seriously, unless it can be shown that the same exact mode of analysis gets similar results in a different dataset."* This approach, however, requires to set apart (or subsequently obtain) a validation dataset of adequate size. This might not always be possible, and even in cases where it is possible, splitting the data may imply a substantial loss of power compared to the analyses that would have been performed using the totality of the data [31].

In the context of analyses strongly affected by uncertainties where none of these simple approaches seems applicable, we suggest an alternative approach based on multiple testing correction. More specifically, we view researcher degrees of freedom from a multiple testing perspective and propose to apply correction for multiple testing to the preferred result to reduce the risk of type 1 error, as outlined in Researcher degrees of freedom as a multiple testing problem and Controlling the Family-Wise Error Rate (FWER) sections.

Motivating example

Data

As a motivating example, we use a current research project on the effect of partial arterial pressure of oxygen ($paO2$) during craniotomy on post-operative complications among neurosurgical patients. This study is based on a routinely collected dataset from a Munich University Hospital preprocessed as described in Becker-Pennrich et al. [33].

While the irreversible damage to the brain caused by reduced levels of oxygen in the blood (hypoxemia) has been the topic of extensive research, the potential harm caused by an increased amount of oxygen (hyperoxemia) is comparatively not well understood. The dangers of over-supplementation of oxygen during surgical procedures are still debated among anesthesiologists and a topic of current research [34, 35].

The dataset under consideration was extracted from routine clinical care data of $n = 3,163$ surgical procedures performed on lung healthy neurosurgical patients. Vital data was measured at several timepoints during surgery for each surgical procedure. As outlined in Becker-Pennrich et al. [33], measuring $paO2$ continuously is not feasible, in contrast to other vital parameters. To obtain a reliable assessment of hyperoxemia during the surgical procedure, the $paO2$ values thus have to be imputed using a surrogate model based on proxy variables that can be measured continuously using non-invasive techniques. Becker-Pennrich et al. [33] suggest to use machine learning methods for this purpose and identify random forest, and regularized linear regression as well-performing candidates.

In this paper, we consider the assessment of the effect of $paO2$ on the binary outcome defined as the occurrence of post-operative complications after surgery. Even if we ignore model choice issues arising from the selection of a set of potential confounders, this analysis is characterized by a large number of uncertain choices. They are described in more detail in Researcher degrees of freedom section along with the options considered in our illustrative study in Illustration section.

Researcher degrees of freedom

In our study, we focus on the following choices, depicted in the form of a decision tree in Fig. 1: (i) missing value imputation, (ii) surrogate model for the unobserved $paO2$-values, (iii) parameter choice approach, (iv) aggregation procedure, and (v) coding of the exposure variable $paO2$ and testing method. Uncertainty (ii) is discussed in more details by Becker-Pennrich et al. [33]. In this study, we use the data preprocessed as described in Becker-Pennrich et al. [33] resulting from the different surrogate modelling strategies.

Uncertainties (i) to (iv) can be seen as *preprocessing uncertainty* in the terminology of Hoffmann et al. [8]. For the missing value imputation (i) the two considered options are to either drop or impute the missing values using multiple imputation in the 'mice' package [36]. For surrogate modelling of the unobserved $paO2$-values (ii) we either use random forest or a regularized general linear model, either using the default parameter values or the parameter values obtained through tuning via random search using predefined tuning spaces (iii) as implemented in the 'mlr3' package [37].

After obtaining a prediction of unobserved $paO2$ values through surrogate modelling, for each surgery the $paO2$ measurements are aggregated to a single value over multiple measurements for a single patient: either the mean or the median (iv). Finally (v), we either consider $paO2$ as a continuous variable and use a logistic regression model to assess its effect on the binary outcome, we dichotomize it using the clinically meaningful cutoff value of 200mmHg, or we categorize it into a three-category variable using the clinically meaningful cutoff values of 200mmHg and 250mmHg and use Fisher's exact test. The latter choice can be seen as referring both to preprocessing and method uncertainty, since the choice

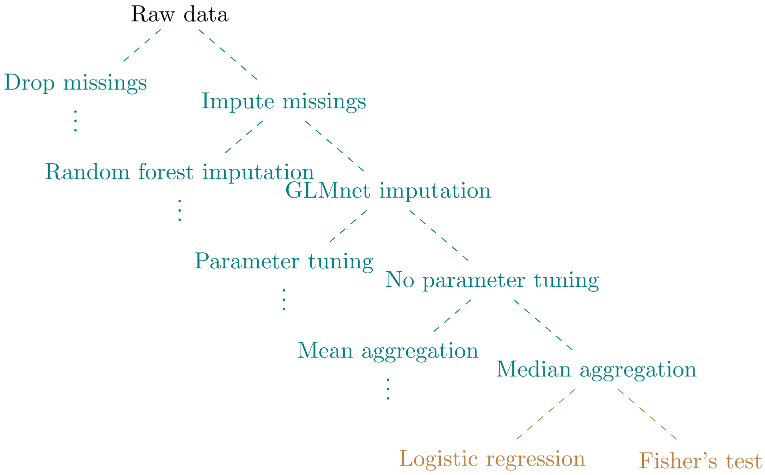

Fig. 1 Overview of the different researcher degrees of freedom. All in all 48 specifications were analyzed. Green depicts the data pre-processing decisions while brown depicts the method choices

of the test is related to the transformation of the variable *paO*2.

All in all, we consider a total of 48 specifications of the analysis strategy: 2 (missing values) × 2 (surrogate model) × 2 (parameter choice) × 2 (aggregation) × 3 (method) = 48.

Method

Researcher degrees of freedom as a multiple testing problem

In the remainder of this paper, we will focus on analyses that consist of statistical tests. We consider a researcher investigating a—possibly vaguely defined—research hypothesis such as "*paO*2 has an impact on post-operative complications", as opposed to the null- and alternative hypotheses of a formal statistical test, which are precisely formulated in mathematical terms. From now on, we assume that the research hypothesis the researcher wants to establish corresponds to the formal alternative hypothesis of the performed tests.

In this context, the term "analysis strategy" refers to all steps performed prior to applying the statistical test as well as to the features of the test itself. The following aspects can be seen as referring to *preprocessing uncertainty* in the terminology by Hoffmann et al. [8]: transformation of continuous variables, handling of outliers and missing values, or merging of categories. Aspects related to the test itself refer to *model and method uncertainty* in the terminology of Hoffmann et al. [8]. They include, for example, the statistical model underlying the test, the formal hypothesis under consideration, or the test (variant) used to test this null-hypothesis.

In the context of testing, an *analysis strategy* can be viewed as a combination of such choices. Obviously, different analysis strategies will likely yield different p-values and possibly different test decision (reject the null-hypothesis or not). Applying different analysis strategies successively to address the same research question amounts to performing multiple tests. From now on, we denote m as the number of analysis strategies considered by a researcher. The null-hypotheses tested through each of the m analyses are denoted as $H_0^i, i = 1, \ldots, m$.

These null-hypotheses and the associated alternative hypotheses can be seen as—possibly different—mathematical formalizations of the vaguely defined research hypothesis—"*paO*2 has an impact on post-operative complications" in our example. One may decide to formalize this research hypothesis as "H_0 : the mean *paO*2 is equal in the groups with and without post-operative complications versus H_1 : the mean *paO*2 is not equal in these two groups". But it would also be possible to formalize it as "H_0: the post-operative complication rates are equal for patients with $paO2 < 200$mmHg and those with $paO2 \geq 200$mmHg" versus "H_1 : the post-operative complication rates are not equal for patients with $paO2 < 200$mmHg and those with $paO2 \geq 200$ mmHg". Analysis strategies may thus differ in the exact definition of the considered null- and alternative hypotheses.

They may, however, also differ in other aspects, some of which were mentioned above (for example the handling of missing values or outliers). If two analysis strategies i_1 and i_2 (with $1 \leq i_1 < i_2 \leq m$) consider exactly the same null-hypothesis, we have $H_0^{i_1} = H_0^{i_2}$. Of course, it may also happen that the research hypothesis is not vaguely defined but already formulated mathematically as null- and alternative hypotheses, and that the m analysis strategies thus only differ in other aspects such as the handling of missing values or outliers. In this case the m null-hypotheses would all be identical.

Regardless whether the hypotheses H_0^i $(i = 1, \ldots, m)$ are (partly) distinct or all identical, a typical researcher who exploits the degree of freedom by "fishing for significance" performs the m testing analyses successively. They hope that at least one of them will yield a significant result, i.e. that the smallest p-value, denoted as $p_{(1)}$, is smaller than the significance level α. If it is, they typically report it as convincing evidence in favor of their vaguely defined research hypothesis. It must be noted that in this hypothetical setting the researcher is not interested in identifying the "best" model or analysis strategy but only in reporting the lowest p-value that supports the hypothesis at hand.

Considering this scenario from the perspective of multiple testing, it is clear that the probability to thereby make at least one type 1 error, denoted as Family Wise Error Rate (FWER), is possibly strongly inflated. In particular, even if all tested null-hypotheses are true, we have a probability greater than α that the smallest p-value $p_{(1)}$ is smaller than α; this is precisely the result researchers engaged in fishing for significance will report. This problem can be seen as one of the explanations as to why the proportion of false positive test results among published results is substantially larger than the considered nominal significance level of the performed tests [5].

A related concept that has often been discussed in the context of the replication crisis is "HARKing", standing for Hypothesing After Results are Known [38]. Researchers engaged in HARKing also perform multiple tests, but to test (potentially strongly) different hypotheses rather than several variants of a common vaguely defined hypothesis. While related to the concept of researcher degrees of freedom, HARKing is fundamentally different in that the rejection of these different null-hypotheses would have different (scientific, practical, organizational) consequences. In the sequel of this article, we consider sets of hypotheses that can be seen as variants of a single vaguely defined hypothesis, whose rejections would have the same consequences in a broad sense.

Controlling the Family-Wise Error Rate (FWER)

Following the formalization of researcher degrees of freedom as a multiple testing situation, we now consider the problem of adjusting for multiple testing in order to control the FWER. More precisely, we want to control the probability $P(\text{Reject at least one true } H_0^i)$ to make at least one type 1 error when testing $H_0^1, \ldots, H_0^m$, i.e. the FWER.

More precisely, we primarily want to control the FWER in case all null-hypotheses are true. Imagine a case where some of the null-hypotheses are false and there is at least one false positive result. On one hand, if $p_{(1)}$ is not among the falsely significant p-values, the false positive test result(s) typically do(es) not affect the results ultimately reported by the researchers (who focus on $p_{(1)}$). This situation is not problematic.

On the other hand, if $p_{(1)}$ is falsely significant, $H_0^{(1)}$ is *wrongly* rejected, and strictly speaking a false positive result ("$p_{(1)} < \alpha$") is reported. However, some of the $m - 1$ remaining null-hypotheses, which are closely related to $H_0^{(1)}$ (because they formalize the same vaguely defined research hypothesis), *are* false. Thus, rejecting $H_0^{(1)}$ is not fundamentally misleading in terms of the vaguely defined research hypothesis. As assumed at the end of Researcher degrees of freedom as a multiple testing problem section, the rejection of $H_0^{(1)}$ has the same consequence as the rejection of the hypotheses that are really false.

For example, in a two-group setting when studying a biomarker B, we may consider the null-hypotheses "H_0^1: the mean of B is the same in the two groups" and "H_0^2: the median of B is the same in the two groups". H_0^1 and H_0^2 are different, but both of them can be seen as variants of "there is no difference between the two groups with respect to biomarker B", and rejecting them would have similar consequences in practice (say, further considering biomarker B in future research, or—in a clinical context—being vigilant when observing a high value of B in a patient).

If biomarker B features strong outliers, the result of the two-sample t-test (addressing H_0^1) and the result of the Mann-Whitney test (addressing to H_0^2) may differ substantially. However, rejecting H_0^2 if it is in fact true and only H_0^1 is false would not be dramatic (and vice-versa). This is because, if H_0^1 is false, there is a difference between the two groups, even if not in terms of medians. The practical consequences of a rejection of H_0^1 and a rejection of H_0^2 are typically the same (as opposed to the HARKing scenario).

To sum up, in the context of researcher degrees of freedom, false positives have to be avoided primarily in the case when all null-hypotheses are true. In other words, we need to control the probability $P(\text{Reject at least one true } H_0^i | \cap_{i=1}^m H_0^i)$ to have at least one false positive result *given* that all null-hypotheses are true, i.e. we want to achieve a weak control of the FWER. Various adjustment procedures exist to achieve strong or weak control of the FWER; see Dudoit et al. [39] for concise definitions of the most usual ones (including those mentioned in this section).

The most well-known and simple procedure is certainly the Bonferroni procedure. It achieves strong control of the FWER, i.e. it controls $P(\text{Reject at least one true } H_0^i)$

under any combination of true and false null hypotheses. This procedure adjusts the significance level to $\tilde{\alpha} = \alpha/m$; or equivalently it adjusts the p-values p_i ($i = 1, \ldots, m$) to $\tilde{p}_i = \min(mp_i, 1)$. However, the Bonferroni procedure is known to yield low power in rejecting wrong null-hypotheses in the case of strong dependence between the tests. The so-called Holm stepwise procedure, which is directly derived from the Bonferroni procedure, has a better power. However, the Holm procedure adjusts the smallest p-value $p_{(1)}$ exactly to the same value as the Bonferroni procedure. It implies that, if none of the m tests lead to rejection with the Bonferroni procedure, it will also be the case with the Holm procedure. The latter can thus not be seen as an improvement over Bonferroni in terms of power in our context, where the focus is on the smallest p-value $p_{(1)}$.

The minP-procedure

The permutation-based minP adjustment procedure for multiple testing [9] indirectly takes the dependence between tests into account by considering the distribution of the *minimal* p-value out of $p_1, \ldots, p_m$. This increases its power in situations with high dependencies between the tests, and thus makes it a suitable adjustment procedure to be applied in the present context. In the general case it controls the FWER only weakly, but as outlined above we do not view this as a drawback in the present context.

The rest of this section briefly describes the single-step minP adjustment procedure based on the review article by Dudoit et al. [39]. The following description is not specific to researcher degrees of freedom considered in this paper. However, for simplicity we further use the notations (p_i, H_0^i, for $i = 1, \ldots, m$) already introduced in Researcher degrees of freedom as a multiple testing problem section in this context.

In the single-step minP procedure, the adjusted p-values $\tilde{p}_i, i = 1, \ldots, m$ are defined as

$$\tilde{p}_i = P\left(\min_{1 \le \ell \le m} P_\ell \le p_i \mid \cap_{i=1}^m H_0^i\right), \tag{1}$$

with P_ℓ being the random variable for the unadjusted p-value for the ℓ^{th} null-hypothesis H_0^ℓ [39]. The adjusted p-values are thus defined based on the distribution of the minimal p-value out of $p_1, \ldots, p_m$, hence the term “minP”. In the context of researcher degrees of freedom considered here, the focus is naturally on $\tilde{p}_{(1)} = P\left(\min_{1 \le \ell \le m} P_\ell \le p_1 \mid \cap_{i=1}^m H_0^i\right)$.

In many practical situations, including the one considered in this paper, the distribution of $\min_{1 \le \ell \le m} P_\ell$ is unknown. The probability in Eq. (1) thus has to be approximated using permuted versions of the data that mimic the global null-hypothesis $\cap_{i=1}^m H_0^i$. More precisely, the adjusted p-value $\tilde{p}_i$ is approximated as the proportion of permutations for which the minimal p-value is lower or equal to the p-value p_i observed in the original data set. Obviously, the number of permutations has to be large for this proportion to be estimated precisely. In the example described in Motivating example section involving only two variables ($paO2$ and post-operative complications), permuted data sets are simply obtained by randomly shuffling one of the variables. More complex cases will be discussed in Discussion section.

Illustration

Study design

The study aims at illustrating the use and behavior of the minP-based approach when used to adjust for the multiplicity arising through researcher degrees of freedom. We use the original as well as permuted versions of the $paO2$ data set. The 48 specifications of the analysis strategy outlined in Motivating example section are successively applied. P-values are either left unadjusted, or adjusted using the Bonferroni procedure, or adjusted using the recommended minP procedure with 1000 permutations. All analyses are performed for different sample sizes. Subsets of each considered size are randomly drawn from the original data set without replacement.

The study consists of two distinct parts. In the first part, we assess the family-wise error rate (FWER) for different sample sizes with the three approaches (no adjustment, Bonferroni adjustment, and minP adjustment). For this purpose, we generate data without association between the two variables of interest ($paO2$ and the outcome “post-operative complications”) by using a $paO2$ covariate vector drawn without replacement from the true dataset but randomly generating the binary outcome variable from a binomial distribution ($p = 0.5$) to break the association between the outcome and $paO2$. This procedure is repeated 1000 times for every $n \in \{100, 200, 300, 500, 2000, 3000\}$. For each run, we calculate unadjusted, minP-adjusted, and Bonferroni-adjusted p-values as outlined above and check whether there is at least one false positive, i.e. whether at least one of the respective p-values of the 48 specification is significant at the 5% level. The proportion of the 1000 runs for which this happens yields an estimate of the FWER of the three approaches.

In the second part, the original data set is analysed. Based on medical knowledge we expect a strong relationship between $paO2$ and the outcome to be present, but do not formally know the truth. For each of the three

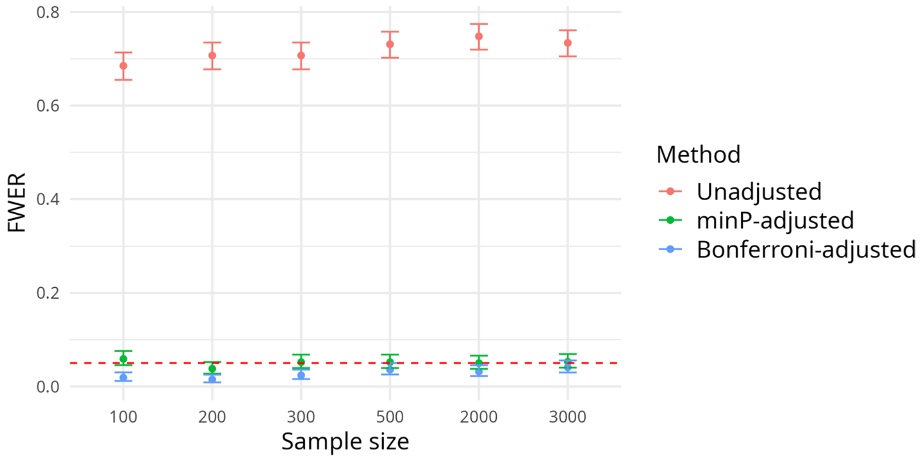

Fig. 2 FWER with Newcombe confidence intervals (computed over 1000 simulation runs) for different sample sizes without an association between post-operative complications and *paO2*. Dashed red line indicates 5% significance level

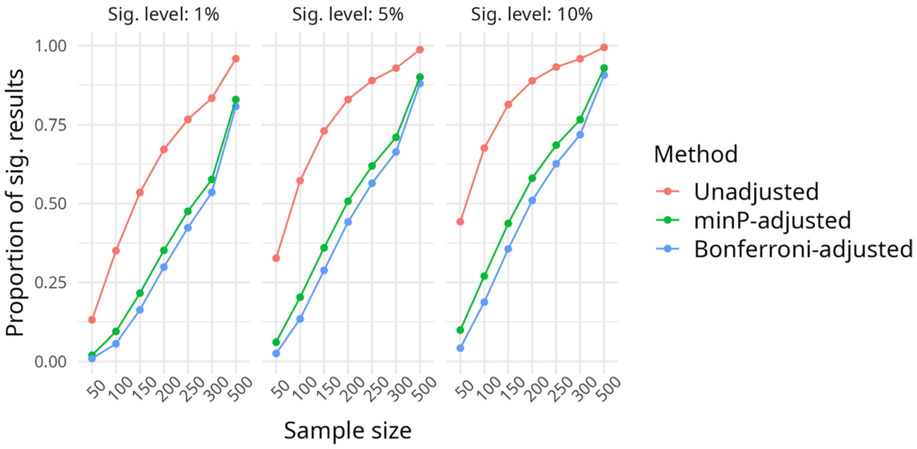

Fig. 3 Proportion of significant results for all 48 specifications for $\alpha \in (0.01, 0.05, 0.1)$ and sample size $n \in (50, 100, 150, 200, 250, 300, 500)$. Line colors indicate results based on unadjusted (red), minP-adjusted (green) and Bonferroni-adjusted (blue) *p*-values

approaches (no adjustment, Bonferroni adjustment, and minP adjustment), we calculate the proportion of significant *p*-values at the 1%, 5% and 10% level among the 48 specifications. This was repeated 1000 times for each sample size $n \in (50, 100, 150, 200, 250, 300)$. As in our example study, the association becomes highly significant for larger sample sizes and all *p*-values are then very close to zero, we only focus on these small sample sizes here. The code for reproducing the analyses can be found on GitHub[2].

Results

Figure 2 shows the estimated FWER for different sample sizes along with the Newcombe confidence intervals [40]. In the absence of adjustment, false-positive results appear to be present in at least one of the 48 specifications for about 70% of the data sets of size $n = 100$ and 76% of the data sets of size $n = 3000$, which aligns with the results of Simonsohn et al. [30]. If we adjust the *p*-values using the minP-approach (green), the 5% level is held for all considered sample sizes. As expected the Bonferroni adjustment (blue) is more conservative: the confidence intervals for FWER, which do not include 0.05, only overlap with those of the minP procedure for a sample size of $n = 3000$.

Figure 3 presents the proportion of significant *p*-values at the 1%, 5% and 10% level over the 48 specifications for the three approaches and different sample sizes. These proportions are averaged over 1000 runs. As we expect a highly significant association between

[2] https://github.com/mmax-code/researcher_dof

the two variables of interest, we focus on small sample sizes only. The observed trend is not surprising: For all $n \in (50, 100, 150, 200, 250, 300, 500)$ it holds that

$$\overline{\sum_{i=1}^{48} \mathbf{1}(p_{i_{\text{unadjusted}}} < \alpha)/48} > \overline{\sum_{i=1}^{48} \mathbf{1}(p_{i_{\text{minP}}} < \alpha)/48} > \overline{\sum_{i=1}^{48} \mathbf{1}(p_{i_{\text{bonferroni}}} < \alpha)/48},$$

(where the overline stands for the average over 1000 runs and $\alpha \in (0.01, 0.05, 0.1)$), i.e. more significant results appear for the unadjusted p-values compared to the adjusted p-values. Furthermore, the Bonferroni approach is more conservative than the minP-adjustment.

Discussion

In this work, we described a framework for performing valid statistical inference in the presence of researcher degrees of freedom through adjustment for multiple testing. Our results on simulated data and in an application concerning $paO2$ and post-operative complications suggest that the minP procedure is appropriate for this purpose. They are in line with known general principles related to (multiple) testing: (i) the minP procedure is less conservative than the Bonferroni procedure—especially when the hypotheses are strongly dependent—and thus better suited in the context of the adjustment for researchers degree of freedom, (ii) both are appropriate to avoid type 1 error inflation, and (iii) statistical power grows with increasing sample size, which is the reason why the attractive alternative to our approach—the two-stage split approach discussed below—is not a panacea.

The use of permutation-based procedures has already been recommended by Simonsohn et al. [30] to address researcher degrees of freedom. There are, however, fundamental differences between this approach and ours. Simonsohn et al. [30] address the problem of researcher degrees of freedom by specifying all plausible specifications (analysis strategies in our terminology) and ultimately evaluating the joint distribution of the estimated effects of interest across these model specifications. This evaluation is done graphically through the so-called specification curve, but also through a permutation test addressing null-hypotheses such as "the median effect across the specifications is zero".

This approach, while similar to ours at first view and interesting, is different in several aspects. Firstly, permutations are used by Simonsohn et al. [30] as part of a permutation-based test and not within a multiple testing adjustment procedure. Our suggestion is precisely to formalize the multiplicity of analysis strategies as a multiple testing problem—and to benefit from various methodological results obtained in the field, for example on the weak control of the FWER through the minP procedure. That said, minP adjustment can be viewed as a simple permutation test for the test statistic "minimal p-value", hence the apparent similarity with the permutation test for the median effect.

Secondly, and more importantly, the focus on the *median effect* makes the procedure by Simonsohn et al. [30] sensitive to misspecifications that do not model the data properly and thus fail to show an effect even if there is one. Imagine a fictive example where one runs 99 fully inappropriate analyses yielding non-significant results and one meaningful analysis that identifies a highly significant (truly existing) effect. The true median effect is zero, and the permutation test by Simonsohn et al. [30] will certainly not reject the null. In contrast, with our approach the truly existing effect is likely to be detected by the meaningful analysis. This is because the minP procedure focuses on the *minimal* p-value, which is very small in this fictive example. This focus on the minimal p-value better accounts for the fact that, in practice, one would often include some analysis strategies that are in fact inappropriate to detect the effect of interest. It also better reflects the common p-hacking practice that consists of selecting and reporting the smallest p-value. However, our approach raises a number of questions that may be addressed in future research.

Firstly, the specification of an appropriate permutation procedure taking the data and the specificity of the research question into account is not always easy/possible. Let us consider the following example: the null-hypothesis of interest is that the means of a variable are equal in two groups, while the variances may be different in the two groups. By permuting the group labels, one also inevitably enforces equality of the variances, which is a stronger assumption than the null-hypothesis of interest [39]. Defining a permutation scheme that reflects the global null-hypothesis $\cap_{i=1}^{m} H_0^i$ may also be intricate in the case of multivariable regression models involving confounders in addition to the exposure of interest whose effect on the dependent variable is to be investigated. On the one hand, permuting only the exposure of interest will destroy the association between this exposure and confounders. On the other hand, permuting the outcome will not only destroy the association between exposure variable and dependent variable, but also the association between the confounders and the outcome. In principle, none of these simple permutation procedures are suitable. Both enforce more than the considered null-hypothesis of no effect of the exposure on the outcome.

Complex alternative permutation procedures may be preferred [41, 42]. Alternatively, if all analysis strategies are based on marginal generalized estimating equation models, one may resort to asymptotical results on the distribution of the maximally selected statistic to derive adjusted p-values, thus avoiding time-consuming and methodologically complex permutation procedures; see for example Ristl et al. [13]. Even though this approach is extremely powerful for most cases and has the advantage that it can also adjust confidence intervals for multiplicity, it comes at the cost of some assumptions that are not applicable in our case (restrictions regarding the input data and focus on parametric tests).

Secondly, it would be interesting to investigate the behavior of our suggested approach compared to the validation approach mentioned in Partial solutions and related work section, that consists of splitting the data into two parts, applying all candidate analysis strategies to the first part, and validating the preferred result by applying the analysis strategy that was used to obtain it to the second part of the data. Both this splitting procedure and the adjustment for multiple testing suggested in this paper imply a loss of power compared to the unadjusted analysis one would perform with the selected analysis strategy on the whole dataset. Researchers may prefer to run analyses on the whole dataset without arbitrary splitting, which may be seen as an argument in favor of our adjustment approach. However, the concept of validation using independent data may also seem attractive. Importantly, this concept has the advantage that type 1 error inflation would be avoided even by researchers who are not yet aware of the dangers of researcher degrees of freedom or not willing (or able) to make a transparent list of the m tests that they conducted in the course of the project. Preference for one or the other approach is a matter of perspective. But the power resulting from these two approaches may yield a decisive argument in favor for one of them. Note that one might also combine the two approaches by applying the minP procedure in the first stage and proceeding with the second stage only if its results are promising.

Thirdly, one may also think about possible ways to make our approach more reliable in situations where researchers tend to "fool themselves" [43] and "forget" some of the hypothesis tests they performed, thus preventing full control of the type 1 error. Our approach may be particularly useful in combination with study registration including the elaboration of a detailed plan of the different analysis strategies to be applied *before seeing any result*—a concept that should in our view be more widely adopted in empirical scientific research for various reasons [23].

Finally, note that our paper should not be understood as a plea for the use of p-values in general. We merely claim that, if statistical testing is used and several analysis variants are performed, it certainly makes sense to adjust for multiplicity before interpreting these p-values. Our approach allows to selectively report the results of the analysis strategy yielding the most convincing evidence, while controlling the type 1 error—and thus the risk of publishing false positive results that may not be replicable. In future research, this approach could in principle be extended beyond the context of hypothesis testing. Provided a meaningful permutation scheme can be defined, minP-type approaches allow in principle to assess whether quantitative results of any type (such as, e.g., a cross-validated error [6] or a cluster similarity index [7]) selected out of many analysis variants may be the result of chance.

Acknowledgements
We thank Savanna Ratky for valuable language corrections and Julian Lange and Ludwig Hothorn for helpful comments.

Auhors' contributions

MM, ALB, and SH designed the study. MM implemented the method, performed the simulations, analyzed the data and interpreted the results. ABP and LH provided the motivating example. MM, ALB, and SH prepared the initial manuscript draft. ALB directed the project. MM, ALB, SH, ABP, and LH reviewed and edited the manuscript. All authors have read and approved the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. The authors gratefully acknowledge the funding by DFG grants BO3139/7 and BO3139/9 to Anne-Laure Boulesteix.

Availability of data and materials

The data that support the findings of this study are not publicly available due to privacy or ethical restrictions.

Code availability

The code for reproducing the analyses can be found on GitHub (https://github.com/mmax-code/researcher_dof).

Declarations

Ethics approval and consent to participate

Before accessing the data, our protocol (submission 19-539) received approval from the University of Munich's institutional review board and consent was waived because of the retrospective nature of the study.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 22 March 2024 Accepted: 9 July 2024
Published online: 17 July 2024

References

1. Gelman A, Loken E. The statistical crisis in science: data-dependent analysis-a "garden of forking paths"-explains why many statistically

significant comparisons don't hold up. Am Sci. 2014;102(6):460–6. https://doi.org/10.1511/2014.111.460.

2. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. Adv Methods Pract Psychol Sci. 2018;1(3):337–56. https://doi.org/10.1177/2515245917747646.
3. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 2011;22(11):1359–66. https://doi.org/10.1177/0956797611417632.
4. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. Am Stat. 2016;70(2):129–33. https://doi.org/10.1080/00031305.2016.1154108.
5. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124. https://doi.org/10.1371/journal.pmed.0020124.
6. Boulesteix AL, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. BMC Med Res Methodol. 2009;9:85. https://doi.org/10.1186/1471-2288-9-85.
7. Ullmann T, Peschel S, Finger P, Müller CL, Boulesteix AL. Over-optimism in unsupervised microbiome analysis: Insights from network learning and clustering. PLoS Comput Biol. 2023;19(1):e1010820. https://doi.org/10.1371/journal.pcbi.1010820.
8. Hoffmann S, Schönbrodt F, Elsas R, Wilson R, Strasser U, Boulesteix AL. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. R Soc Open Sci. 2021;8(4):201925. https://doi.org/10.1098/rsos.201925.
9. Westfall PH, Young SS, Wright SP. On Adjusting P-Values for Multiplicity. Biometrics. 1993;49(3):941–5. https://doi.org/10.2307/2532216.
10. Westfall PH, Young SS. Resampling-based multiple testing: Examples and methods for p-value adjustment, vol. 279. New York: Wiley; 1993.
11. Mathews F, Johnson PJ, Neil A. You are what your mother eats: evidence for maternal preconception diet influencing foetal sex in humans. Proc R Soc B Biol Sci. 2008;275(1643):1661–8. https://doi.org/10.1098/rspb.2008.0105.
12. Young SS, Bang H, Oktay K. Cereal-induced gender selection? Most likely a multiple testing false positive. Proc R Soc B Biol Sci. 2009;276(1660):1211–2. https://doi.org/10.1098/rspb.2008.1405.
13. Ristl R, Hothorn L, Ritz C, Posch M. Simultaneous inference for multiple marginal generalized estimating equation models. Stat Methods Med Res. 2020;29(6):1746–62. https://doi.org/10.1177/0962280219873005.
14. Fields AC, Lu P, Palenzuela DL, Bleday R, Goldberg JE, Irani J, et al. sDoes retrieval bag use during laparoscopic appendectomy reduce postoperative infection? Surgery. 2019;165(5):953–7. https://doi.org/10.1016/j.surg.2018.11.012.
15. Childers CP, Maggard-Gibbons M. Re: Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection? Surgery. 2019;166(1):127–8. https://doi.org/10.1016/j.surg.2019.01.019.
16. Childers CP, Maggard-Gibbons M. Same data, opposite results?: a call to improve surgical database research. JAMA Surg. 2021;156(3):219–20. https://doi.org/10.1001/jamasurg.2020.4991.
17. Turner SA, Jung HS, Scarborough JE. Utilization of a specimen retrieval bag during laparoscopic appendectomy for both uncomplicated and complicated appendicitis is not associated with a decrease in postoperative surgical site infection rates. Surgery. 2019;165(6):1199–202. https://doi.org/10.1016/j.surg.2019.02.010.
18. Jivanji D, Mangosing M, Mahoney SP, Castro G, Zevallos J, Lozano J. Association Between Marijuana Use and Cardiovascular Disease in US Adults. Cureus. 2020;12(12):e11868. https://doi.org/10.7759/cureus.11868.
19. Shah S, Patel S, Paulraj S, Chaudhuri D. Association of marijuana use and cardiovascular disease: A behavioral risk factor surveillance system data analysis of 133,706 US adults. Am J Med. 2021;134(5):614–20. https://doi.org/10.1016/j.amjmed.2020.10.019.
20. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proc Natl Acad Sci. 2018;115(11):2600–6. https://doi.org/10.1073/pnas.170827411.
21. Munafò MR, Nosek BA, Bishop DV, Button KS, Chambers CD, Percie du Sert N, et al. A manifesto for reproducible science. Nat Hum Behav. 2017;1:21. https://doi.org/10.1038/s41562-016-0021.
22. Hardwicke TE, Wagenmakers EJ. Reducing bias, increasing transparency and calibrating confidence with preregistration. Nat Hum Behav. 2023;7(1):15–26. https://doi.org/10.1038/s41562-022-01497-2.
23. Naudet F, Patel CJ, DeVito NJ, Goff GL, Cristea IA, Braillon A, et al. Improving the transparency and reliability of observational studies through registration. BMJ. 2024;384:e076123. https://doi.org/10.1136/bmj-2023-076123.
24. Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, et al. SPIRIT 2013 statement: defining standard protocol items for clinical trials. Ann Intern Med. 2013;158(3):200–7. https://doi.org/10.7326/0003-4819-158-3-201302050-00583.
25. Greenberg L, Jairath V, Pearse R, Kahan BC. Pre-specification of statistical analysis approaches in published clinical trial protocols was inadequate. J Clin Epidemiol. 2018;101:53–60. https://doi.org/10.1016/j.jclinepi.2018.05.023.
26. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J Clin Epidemiol. 2015;68(9):1046–58. https://doi.org/10.1016/j.jclinepi.2015.05.029.
27. Klau S, Patel CJ, Ioannidis JP, Boulesteix AL, Hoffmann S, et al. Comparing the vibration of effects due to model, data pre-processing and sampling uncertainty on a large data set in personality psychology. Meta Psychol. 2023;7(6). https://doi.org/10.15626/MP.2020.2556.
28. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspect Psychol Sci. 2016;11(5):702–12. https://doi.org/10.1177/1745691616658637.
29. Rohrer JM, Egloff B, Schmukle SC. Probing birth-order effects on narrow traits using specification-curve analysis. Psychol Sci. 2017;28(12):1821–32. https://doi.org/10.1177/0956797617723726.
30. Simonsohn U, Simmons JP, Nelson LD. Specification curve analysis. Nat Hum Behav. 2020;4(11):1208–14. https://doi.org/10.1038/s41562-020-0912-z.
31. Daumer M, Held U, Ickstadt K, Heinz M, Schach S, Ebers G. Reducing the probability of false positive research findings by pre-publication validation-experience with a large multiple sclerosis database. BMC Med Res Methodol. 2008;8(1):1–7. https://doi.org/10.1186/1471-2288-8-18.
32. Ioannidis JP. Microarrays and molecular research: noise discovery? Lancet. 2005;365(9458):454–5. https://doi.org/10.1016/S0140-6736(05)17878-7.
33. Becker-Pennrich AS, Mandl MM, Rieder C, Hoechter DJ, Dietz K, Geisler BP, et al. Comparing supervised machine learning algorithms for the prediction of partial arterial pressure of oxygen during craniotomy. medRxiv. 2022. https://doi.org/10.1101/2022.06.07.22275483.
34. McIlroy DR, Shotwell MS, Lopez MG, Vaughn MT, Olsen JS, Hennessy C, et al. Oxygen administration during surgery and postoperative organ injury: observational cohort study. BMJ. 2022;379:e070941. https://doi.org/10.1136/bmj-2022-070941.
35. Weenink RP, de Jonge SW, van Hulst RA, Wingelaar TT, van Ooij PJA, Immink RV, et al. Perioperative hyperoxyphobia: justified or not? Benefits and harms of hyperoxia during surgery. J Clin Med. 2020;9(3):642. https://doi.org/10.3390/jcm9030642.
36. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. J Stat Softw. 2011;45(3):1–67. https://doi.org/10.18637/jss.v045.i03.
37. Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, et al. mlr3: A modern object-oriented machine learning framework in R. J Open Source Softw. 2019;4(44), 1903. https://doi.org/10.21105/joss.01903.
38. Kerr NL. HARKing: Hypothesizing after the results are known. Personal Soc Psychol Rev. 1998;2(3):196–217. https://doi.org/10.1207/s15327957pspr0203_4.
39. Dudoit S, Shaffer JP, Boldrick JC. Multiple hypothesis testing in microarray experiments. Stat Sci. 2003;18(1):71–103. https://doi.org/10.1214/ss/1056397487.
40. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. Stat Med. 1998;17(8):857–72.
41. Berrett TB, Wang Y, Barber RF, Samworth RJ. The conditional permutation test for independence while controlling for confounders. J R Stat Soc Ser B Stat Methodol. 2020;82(1):175–97. https://doi.org/10.1111/rssb.12340.
42. Girardi P, Vesely A, Lakens D, Altoè G, Pastore M, Calcagnì A, et al. Post-selection inference in multiverse analysis (PIMA): An inferential framework based on the sign flipping score test. Psychometrika. 2024;89:542–68. https://doi.org/10.1007/s11336-024-09973-6.
43. Nuzzo R. Fooling ourselves. Nature. 2015;526(7572):182. https://doi.org/10.1038/526182a.

Publisher's Note

A.2 Contribution 2

»Outlier Detection in Mendelian Randomization«

Citation

Mandl, M.M., Boulesteix, A.-L., Burgess, S., Zuber, V. (2025) Outlier Detection in Mendelian Randomization. *Statistics in Medicine* 44, 15-17: e70143. https://doi.org/10.1002/sim.70143

Authors' contributions

MM and VZ designed the study. MM implemented the method, performed the simulations, analyzed the data and interpreted the results. MM and VZ prepared the initial manuscript draft. VZ directed the project. MM, ALB, VZ, and SB reviewed and edited the manuscript.

Rights and permissions

Statistics in Medicine

WILEY

Statistics in Medicine

RESEARCH ARTICLE

Outlier Detection in Mendelian Randomization

Maximilian M. Mandl[1,2] | Anne-Laure Boulesteix[1,2] | Stephen Burgess[3,4] | Verena Zuber[5,6,7]

[1]Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, Ludwig-Maximilians-Universität, München, Germany | [2]Munich Center for Machine Learning, Munich, Germany | [3]MRC Biostatistics Unit, School of Clinical Medicine, University of Cambridge, Cambridge, UK | [4]Department of Public Health and Primary Care, British Heart Foundation Cardiovascular Epidemiology Unit, University of Cambridge, Cambridge, UK | [5]Department of Epidemiology and Biostatistics, Imperial College London, London, UK | [6]MRC Centre for Environment and Health, School of Public Health, Imperial College London, London, UK | [7]Dementia Research Institute at Imperial College, Imperial College London, London, UK

Correspondence: Maximilian M. Mandl (mmandl@ibe.med.uni-muenchen.de)

Received: 28 April 2024 | **Revised:** 3 May 2025 | **Accepted:** 16 May 2025

Funding: This study was partly funded by DFG grant BO3139/7-1 to Anne-Laure Boulesteix. Verena Zuber is supported by the United Kingdom Research and Innovation Medical Research Council grant MR/W029790/1.

Keywords: instrumental variables | invalid instruments | Mendelian randomization | outlier detection | pleiotropy

ABSTRACT

Mendelian randomization (MR) uses genetic variants as instrumental variables to infer causal effects of exposures on an outcome. One key assumption of MR is that the genetic variants used as instrumental variables are independent of the outcome conditional on the risk factor and unobserved confounders. Violations of this assumption, that is, the effect of the instrumental variables on the outcome through a path other than the risk factor included in the model (which can be caused by pleiotropy), are common phenomena in human genetics. Genetic variants, which deviate from this assumption, appear as outliers to the MR model fit and can be detected by the general heterogeneity statistics proposed in the literature, which are known to suffer from overdispersion, that is, too many genetic variants are declared as false outliers. We propose a method that corrects for overdispersion of the heterogeneity statistics in uni- and multivariable MR analysis by making use of the estimated inflation factor to correctly remove outlying instruments and therefore account for pleiotropic effects. Our method is applicable to summary-level data.

1 | Introduction

Identification of causal effects in biomedical sciences is a challenging task. Most causal inference methods rely on specific assumptions which must be properly tested in practice. Mendelian randomization (MR) is an instrumental variable approach that uses genetic variants to infer causal effects of risk factors on an outcome [1]. Due to the randomization of the genetic variants during meiosis, these can be used as instrumental variables that can potentially meet the restrictive methodological requirements naturally. Thus, causal effects can be consistently inferred even if unobserved confounders are present. For example, relevant clinical questions that have been addressed using MR include the investigations of the effect of blood lipids on coronary heart disease (CHD), age-related macular degeneration (AMD) or Alzheimer's disease [2–4], and the effect of vitamin D levels on Multiple Sclerosis (MS) [5]. The instrumental variable assumptions underlying MR require that the genetic variants are independent of the outcome conditional on the risk factor and unobserved confounders, also known as the exclusion restriction assumption. Violations of this exclusion restriction assumption, that is, the effect of the instrumental variables on the outcome through a path other than the risk factor included in the model, can be caused by horizontal pleiotropy, which is a common phenomenon in human genetics [6].

Genetic variants which deviate from this assumption appear as outliers in the MR model fit and can be detected by general heterogeneity statistics proposed in the literature [7]. In MR analysis, these statistics are often inflated due to the heterogeneity of genetic variants exerting their downstream effects on the exposures of interest, mismatches of allele frequencies when data is integrated from distinct samples, or the variant-specific heterogeneity estimates not being normally distributed, as a ratio of two normal distributions does not follow a normal distribution. This excess heterogeneity may impede the detection of outlying instruments using the traditional methods and result in the removal of too many IVs which are not true outliers that impact the causal effect estimate and consequently the conclusions drawn from the MR analysis.

In this paper, we propose GC-Q, a simple method that corrects for overdispersion of the heterogeneity statistics in uni- and multivariable MR analysis by making use of the estimated inflation factor to correctly remove outlying instruments, therefore accounting for pleiotropic effects (Section 2). As we show in an extensive simulation study and analysis of real data examples, our proposed method is more conservative in detecting outliers than existing methods because it removes the minimum number of instruments necessary to retain unbiased effect estimates. Moreover, GC-Q leads to a reduction of the type I error in detecting outlying genetic variants used as instruments compared to the existing methods based on Cochran's Q.

Moreover, we provide a comprehensive review of different outlier detection methods in uni- and multivariable MR. The code for the simulation study and the real data example in this paper are provided on GitHub for the purpose of reproducibility.[1] Furthermore, our manuscript is accompanied by a dedicated R-function in the *MendelianRandomization* R-package for both our proposed method and the existing methods based on first and second order weights. In the recently introduced phases classification for methodological research [8], our contribution can be assigned to phase 2: it presents and demonstrates the use of a new method on real data and provides first simulation results suggesting that it is useful in some cases and worth being further considered in phase 3 studies.

2 | Methods

In this section, we first give a brief overview of univariable and multivariable MR, and how horizontal pleiotropy violates the exclusion restriction assumption of instrumental variable analysis. Next, we discuss how heterogeneity statistics can be used to detect violations of this assumption and how specific pleiotropic genetic variants can be detected as outliers. We further show limitations of existing implementations of heterogeneity statistics, and we introduce our novel method, GC-Q. Finally, we end with an overview and comparison of existing outlier detection methodologies for MR.

Regarding the notation, we examine the causal effect θ of a risk factor X on an outcome Y using genetic variants G_i for $i = 1, \dots, n$ as instrumental variables (IVs). Subsequently, in a multivariable MR model we consider multiple causal effects θ_j $(j = 1, \dots, d)$ for multiple risk factors X_j $(j = 1, \dots, d)$ on an outcome Y. Following the most common MR design [9], real data examples are based on two-sample summary-level data to take advantage of large sample sizes and thus improve the precision of the estimates [10]. Additionally, all of our derivations are based on summary-level data. We therefore assume that the associations of genetic variants with the risk factor(s) and the outcome, and the causal effect of the risk factor(s) on the outcome, are linear and homogeneous. These assumptions have already been discussed in the literature [11].

2.1 | Univariable Mendelian Randomization

In order to define a *valid* IV, the genetic variants in the univariable MR analysis require the following assumptions to hold [12]:

- IV1(U): Each genetic variant G_i for $i = 1, \dots, n$ is associated with the exposure.
- IV2(U): Each genetic variant G_i for $i = 1, \dots, n$ is not associated with any confounder of the risk factor-outcome association.
- IV3(U): Each genetic variant G_i for $i = 1, \dots, n$ is independent of the outcome Y conditional on the risk factor X and confounders U.

Figure 1 shows the causal DAG for the univariable MR setting. Each genetic variant G_i should only have an effect on the outcome via the risk factor. Pleiotropy is defined as the effect of any genetic variant G_i that contains an effect via an independent pathway, that is, not through the included risk factor in the MR model (red dashed lines in Figure 1). Therefore, IV3 would be violated.

If IV1(U)–IV3(U)[2] hold, the consistent estimate of the causal effect θ is the inverse-variance weighted (IVW) estimate [13]

$$\hat{\theta} = \frac{\sum_i^n \omega_i \hat{\theta}_i}{\sum_i^n \omega_i}, \tag{1}$$

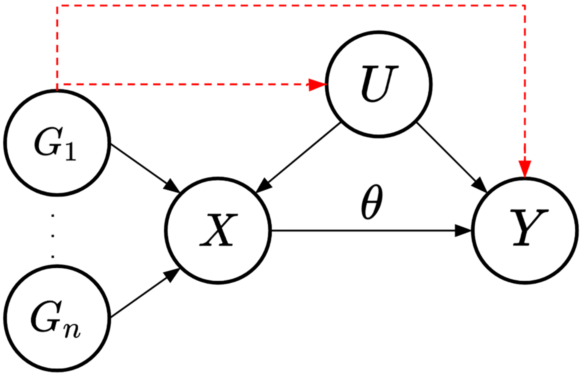

FIGURE 1 | Causal directed acyclic graph (DAG) for the univariable Mendelian randomization setting. Genetic variants are denoted as G_i for $i \in 1, \dots, n$, the set of confounders as U and the causal effect of the risk factor X on the outcome Y being θ. The red dashed lines represent the effect of the instrumental variable(s) on the outcome through paths other than the risk factor included in the model, for example, caused by pleiotropy.

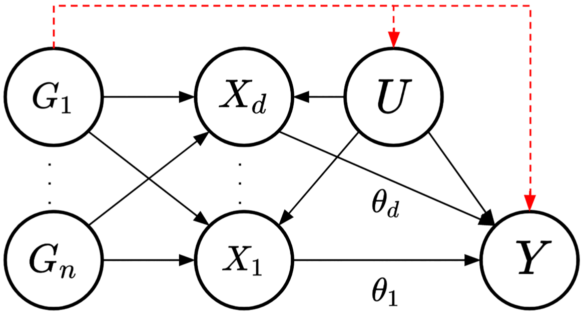

FIGURE 2 | Causal directed acyclic graph (DAG) for the multivariable Mendelian randomization setting. Genetic variants G_i ($i \in 1 \dots n$), set of confounders U and causal effects of the risk factors X_j ($j \in 1 \dots d$) on the outcome Y being θ_j. The red dashed lines represent the effect of the instrumental variable(s) on the outcome through paths other than the risk factors included in the model, for example, caused by pleiotropy.

where $\hat{\theta}_i$ is the ratio estimate of the i-th IV, defined as $\hat{\theta}_i = \frac{\hat{\beta}_{Y_i}}{\hat{\beta}_{X_i}}$, where in the summary-level data setting $\hat{\beta}_{X_i}$ and $\hat{\beta}_{Y_i}$ are the genetic effects of IV G_i on X and Y for variant i respectively. The IV-specific inverse-variance weight ω_i is the precision of the respective ratio estimate. The estimate based on individual-level data can be obtained via the two-stage least-squares (2-SLS) approach [14]. The 2-SLS estimate is equivalent to the IVW estimate [15]. However, in finite samples this is only true if all of the instruments are perfectly uncorrelated with each other.

2.2 | Multivariable Mendelian Randomization

As an extension to the standard MR approach, multivariable MR includes multiple potential risk factors in one joint model accounting for measured pleiotropy (see Figure 2). In order to define a *valid* IV, the genetic variants in the multivariable MR analysis require the following assumptions to hold for each genetic variant G_i where $i = 1, \dots, n$ [16]:

- IV1(M): Each genetic variant G_i for $i = 1, \dots, n$ is associated with at least one of the risk factors X_j.
- IV2(M): Each genetic variant G_i for $i = 1, \dots, n$ is not associated with any confounder of the risk factor-outcome associations.
- IV3(M): Each genetic variant G_i for $i = 1, \dots, n$ is independent of the outcome Y conditional on the risk factors X_j for $j = 1, \dots, d$ and confounders U.

Moreover, the following assumptions relate to which risk factors X_j for $j = 1, \dots, d$ can be included in a multivariable MR model:

- RF1(M) Each risk factor X_j for $j = 1, \dots, d$ needs to be strongly instrumented by at least one genetic variant G_i for $i = 1, \dots, n$, also denoted as relevance assumption.
- RF2(M) Each risk factor X_j for $j = 1, \dots, d$ considered in the analysis cannot be linearly explained by the genetic associations of any other risk factor X_j for $j = 1, \dots, d$ or by the combined genetic associations of several other risk factors included in the analysis, also denoted as no mulit-collinearity assumption.

If IV1(M)–IV3(M)[3] hold, the consistent estimates of the direct causal effects θ_j can be obtained from individual-level data via a 2-SLS approach or through the multivariable two-sample summary-level IVW method, with weights $se\left(\hat{\beta}_{Y_i}\right)^{-2}$ being the inverse of the estimated variance for genetic variant i [17] and $\hat{\beta}_{X_{ij}}$, and $\hat{\beta}_{Y_i}$ being the genetic effects of G_i on X_{ij} and Y_i for variant i and risk factor j, respectively

$$\hat{\beta}_{Y_i} = \sum_{j=1}^{d} \theta_j \hat{\beta}_{X_{ij}} + \epsilon_i. \tag{2}$$

2.3 | Heterogeneity Statistics

Tests for heterogeneity in the MR setting examine the null hypothesis that all genetic variants follow the same causal pathways through which the risk factors $X_1, \dots, X_d$ act on the outcome Y. The following heterogeneity statistics are based on Cochran's Q and compute weighted sums of squared residuals and differ in the variance factors they use for weighting [18]. Cochran's Q was established in meta-analysis for the detection of heterogeneity between studies. Two-sample MR can be viewed as a meta-analysis over genetic variants used as IVs. Analogously, the sample size or the number of studies included equals the number of genetic variants used as IVs. Previous research has shown that the power of Cochran's Q increases with the number of studies and the total information available (total weight or inverse variance) and decreases substantially if a large proportion of the total information is based on one study [19]. On the other hand, the test arguably shows "excessive" power when the number of large studies increases [20]. Translating this into the MR framework means that the Q-statistic is likely to miss true heterogeneity (not rejecting the null hypothesis) when there are few genetic variants as IVs available and detects false heterogeneity (rejecting the null hypothesis) when there are many genetic variants available. Usually, genome-wide association studies (GWAS) entail high statistical power given their huge case numbers (n > 1 million), that is, we are more concerned about detecting too many SNPs as outliers. Therefore, we introduce a modification of the Q-statistic which allows the calibration of the standard Q-statistic, in order to reduce the "excessive" power of the test and decrease the type I error.

Cochran's Q-statistic was first applied as a global test to identify the presence of any invalid instruments in two-sample summary data MR with a single exposure by del Greco et al. [18]. A generalized version of the Q-statistic for multivariable MR [21] is defined as

$$Q = \sum_{i=1}^{n} \left(\frac{1}{\omega_i}\right) \left(\hat{\beta}_{Y_i} - \sum_{j}^{d} \hat{\theta}_j \hat{\beta}_{X_{ij}}\right)^2 \sim \chi^2_{(n-d)}, \tag{3}$$

with i being the SNP index, j being the risk factor index, and ω_i being the SNP-specific weight which can be approximated either using first or second order weights [22]. Under the null hypothesis, Cochran's Q follows a χ^2_{n-d} distribution with $n - d$ degrees of freedom. First order weights are simply defined as

$\omega_i = \sigma_{Y_i}^2$ with σ_{Y_i} being the standard error of $\hat{\beta}_{Y_i}$. The first order weights are an approximation relying on the so-called no measurement error (NOME) assumption which assumes that the standard errors of the exposures associations are negligible [22]. These first-order weights are known to lead to an overdispersion in the heterogeneity statistic resulting in an inflation of the type 1 error rate, that is, detecting heterogeneity when it is not present [22].

In applied analysis, using the two-sample summary-level MR setting, there are additional sources of excess heterogeneity due to:

- Wide-spread but negligible pleiotropic effects
- Small disagreements in allele frequencies between the first sample used to derive the exposure association and the second independent sample used to derive the outcome associations.
- The variant-specific estimates being not normally distributed (as the ratio of two normal distributions is not normal)

Bowden et al. [22] and Sanderson et al. [21] propose an adjusted weighting scheme of the Cochran Q-statistic to test for invalid instruments in the two-sample univariable [22] and multivariable [21] summary-level data setting using the following modified second-order weights based on a Taylor expansion of the ratio estimate. In contrast to the first-order weights that are proportional to the uncertainty of the standard error of the genetic associations with the outcome, the second-order weights also account for the standard error of the genetic associations with the exposure and thus model uncertainty in both the numerator and denominator of the ratio estimate.

The second order weights for multivariable MR are defined as $\omega_i = \sigma_{Y_i}^2 + \sum_j^d \hat{\theta}_j^2 \sigma_{X_{ij}}^2 + \sum_{jk} 2\hat{\theta}_j \hat{\theta}_k \sigma_{X_{ijk}}$, with i being the SNP index, j being the risk factor index, σ_{Y_i} and $\sigma_{X_{ij}}$ being the standard error of $\hat{\beta}_{Y_i}$ and $\hat{\beta}_{X_{ij}}$ respectively, and $\sigma_{X_{ijk}}$ being the covariance for all pairs $\hat{\beta}_{X_{ij}}$ and $\hat{\beta}_{X_{ik}}$ for $j, k \in 1, \ldots, d$ and $j \neq k$. Importantly, the covariance term $\sigma_{X_{ijk}}$ for exposures j and k, which is necessary in multivariable MR cannot be estimated from the data at hand and can only be calculated from individual-level data, which is often not available in practice [21].

2.4 | Outlier Detection in MR

The main aim of outlier detection in MR is the detection of invalid IVs with strong pleiotropic effects that can when included into the MR model bias the causal effect estimate $\hat{\theta}$ and consequently distort conclusions drawn from the MR analysis. The objective is to identify these individual genetic instruments with pleiotropic effects which appear as outliers to the MR model fit. The local test statistic q_i of SNP i defined as:

$$q_i = \left(\frac{1}{\omega_i}\right)\left(\hat{\beta}_{Y_i} - \sum_j^d \hat{\theta}_j \hat{\beta}_{X_{ij}}\right)^2 \tag{4}$$

has been proposed for outlier detection [7]. Under the null hypothesis, q_i asymptotically follows a $\chi^2_{(1)}$ distribution with one degree of freedom. Note, that we must correct for multiple testing if we test individual genetic instruments, for example, using a Bonferroni correction, that is, dividing the significance level by the number of instruments n [22]. A conservative multiple testing procedure is recommended in order to only remove clear outliers and to retain as many genetic variants as possible as instrumental variables. Yet, outlier detection using the current implementation of the local q-statistic is impeded by overinflation when using first-order weights. In contrast, when using second-order weights in multivariable MR, one essential parameter (covariance term $\sigma_{X_{ijk}}$) is not readily available when working with summary-level data and consequently often set to zero in practice.

2.5 | Correction for Overdispersion With Genomic Control (GC-Q)

We suggest correcting for overdispersion of the first-order weighted heterogeneity statistics in MR-analysis by making use of the estimated inflation factor to remove outlying instruments which may be invalid due to horizontal pleiotropy. The idea of correcting for overdispersion is based on the Genomic Control approach which was originally used in the context of GWAS [23]. More precisely, the Genomic Control approach was developed for testing if a large set of genomic markers or SNPs are associated with a quantitative trait of interest. Typically, when performing genome-wide testing of genetic markers, like in GWAS, only a small proportion of genetic markers are associated with a trait of interest, and the large majority of genetic markers can be considered as following the null model. Yet, Devlin and Roeder observed that even these null genetic markers do not follow the theoretical null distribution of the statistical association test, but display overdispersion which is constant across the genome [23]. In GWAS, this observed overdispersion is due to population stratification, cryptic relatedness, or unobserved confounding [24].

In analogy with genetic association tests as described in Devlin and Roeder [23], the local heterogeneity statistic q_i of instrument i follows a χ^2 distribution with one degree of freedom and a non-centrality parameter 0 under the Null ($\chi^2_1(0)$), that is:

$$q_i/\lambda \overset{H_0}{\sim} \chi^2_1(0), \tag{5}$$

where λ is the overdispersion parameter and constant for all SNPs. Thus the empirical distribution of the local q_i-statistic is inflated from $\chi^2_1(0)$ to $\lambda\chi^2_1(0)$. This means that no heterogeneity is present or in other words, the instrument is valid. This follows from the general Cochran's statistic in Section 2.3 that is, generally inflated.

From a Bayesian perspective the distribution of the local q-statistic—with outlying SNPs being present—can be modeled using a mixture model of two χ^2 distributions, where the distribution under the Null ($\chi^2_1(0)$) is representing the valid IVs and $\chi^2_1(A_i^2)$ is the distribution with non-centrality parameter $A_i^2 > 0$ associated with the i-th outlier

$$q_i/\lambda \overset{H_1}{\sim} \rho\chi^2_1(A_i^2) + (1-\rho)\chi^2_1(0), \tag{6}$$

where ρ is the prior probability that a given SNP is an outlier as indicated by excess heterogeneity and consequently invalid.

As Devlin and Roeder [23] propose, a simple frequentist estimate of the inflation parameter $\hat{\lambda}$ can be derived from the data as

$$\hat{\lambda} = \frac{\tilde{q}}{0.675^2}, \tag{7}$$

with $\tilde{q}$ being the median of q_i for all $i = 1, \ldots, n$ SNPs and 0.675^2 being the median of the theoretical χ^2_1 distribution.

An important assumption when estimating $\hat{\lambda}$ according to Equation (7) is that at least half of the genetic variants used as IVs are valid instruments. This assumption is common to the median MR approach [25] and more general for any type of outlier detection approach [26]. For all $i = 1, \ldots, n$ SNPs, we reject the Null, that is, SNP i is considered as outlier, if $q_i/\lambda > \chi^2_{1,\alpha^*}$ with χ^2_{1,α^*} being the critical value at level α^* and $\alpha^* = \alpha/n$ being the Bonferroni adjusted significance level to provide a conservative multiple testing adjustment.

The local heterogeneity statistic q_i^{adj} can subsequently formulated as

$$q_i^{adj} = \left(\frac{1}{\lambda}\right)\left(\frac{1}{\omega_i}\right)\left(\hat{\beta}_{Y_i} - \sum_j^d \hat{\theta}_j \hat{\beta}_{X_{ij}}\right)^2. \tag{8}$$

2.6 | Other Methods to Detect Outliers in Summary-Level MR

2.6.1 | Heterogeneity Statistics With Second Order Weights

Sanderson et al. [21] propose an adjustment to Cochran's Q in the two-sample summary setting for testing the presence of horizontal pleiotropy as the standard version of Cochran's Q merely has a weighting of the variance of β_{Yi} denoted as σ^2_{Yi}, and is thus not asymptotically χ^2 distributed. Therefore, they make use of second-order weights,

$$\omega_i = \sigma^2_{Yi} + \sum_j^d \hat{\beta}_{X_{ij}} \sigma^2_{X_{ij}} + \sum_j^d \sum_{\substack{k \\ j \neq k}}^d \sigma_{ijk} \tag{9}$$

where $\hat{\beta}_{X_{ij}}$ are efficient estimators of the causal effects and σ_{ijk} are the covariances of the exposures which need to be estimated from individual-level data.

2.6.2 | MR-PRESSO

Verbanck et al. [27] developed the MR-PRESSO method to detect pleiotropy (global test), the correction for pleiotropy via outlier removal (outlier test), and test for significant distortions in the causal estimates before and after the outlier removal (distortion test). The MR-PRESSO global test is defined as the following residual sum of squares (RSS)

$$\text{RSS} = \sum_{i=1}^n \left(\frac{1}{\omega_i}\right)\left(\hat{\beta}_{Y_i} - \sum_j \hat{\theta}_j^{-i} \hat{\beta}_{X_{ij}}\right)^2 \sim \chi^2_{(n-d)} \tag{10}$$

where ω_i are the first-order weights and $\hat{\theta}_j^{-i}$ is the causal effect estimate from an IVW MR model without variant i. The respective p-values are calculated using a simulation procedure. The main difference between the RSS of MR−PRESSO and the Q-statistic is that the causal effect estimate of MR−PRESSO is calculated excluding the i-th IV and that p-values are derived in a non-parametric fashion using a simulation procedure which scales with the number of IVs and becomes prohibitively slow when including hundreds of IVs.

As with the local q-statistics, the outlier test aims at detecting individual SNPs as outliers. For a given genetic variant i, the observed RSS defined as $\left(\frac{1}{\omega_i}\right)\left(\hat{\beta}_{Y_i} - \sum_j \hat{\theta}_j^{-i} \hat{\beta}_{X_{ij}}\right)^2$ is compared to the distribution of the expected simulated residual sum of squares. The detection mechanism can be described as follows: For each variant the causal effect $\hat{\theta}_{-i}$ is computed without variant i. Afterwards, the observed residual sum of squares is compared to the expected residual sum of squares. Finally, an empirical p-value that is, Bonferroni adjusted is computed to decide whether variant i is an outlier or not.

2.6.3 | Radial MR

The Galbraith Radial plot, adapted for MR, plots the z-statistics for genetic variant i, which is the ratio estimate $\hat{\theta}_i$ divided by its standard error, against the precision of the ratio estimate which is equal the inverse standard error [28]. This is particularly relevant when different IVs have varying precision and consequently contribute with different weights to the final causal estimate. Moreover, the Radial MR approach allows for a flexible use of first or second order weights and can adapt an intercept in the MR model fit which is also known as the MR-Egger approach [29]. Outlier detection is based on the heterogeneity statistics as described above.

2.6.4 | Summary and Comparison of Outlier Detection Methodologies for MR

To conclude the Methods section, we present an overview of outlier detection methodologies for MR. The Q-statistic with first order weights is easy to estimate, but relies on the NOME assumption. Consequently, it is not well calibrated and shows an overinflation. In contrast, the Q-statistic with second order weights is well calibrated, but needs additional parameters which cannot be estimated from summary-level data alone. Moreover, since the second order weights include the causal effect estimate and iterative estimation procedure needs to be implemented [22]. MR-PRESSO relies on a permutation procedure which is computationally expensive when the number of genetic variants used as IVs increases.

Our newly proposed GC-Q approach is a recalibrated version of the first-order weights Q-statistic, which can be estimated from the data at hand and does not rely on computationally intensive permutation procedures. As we are going to show in an extensive simulation study in the next section, GC-Q selects the minimum number of potential outliers necessary to achieve unbiased MR effect estimates. All outlier detection methods perform a Bonferroni correction for multiple testing of the individual IV heterogeneity statistics with the aim to be as conservative as possible and to only remove the minimum number of outliers necessary to avoid any bias of the MR model.

3 | Simulation Study

The primary objective of this simulation study is to compare the performance of different methods to detect outlying genetic variants used as IVs, that is, SNPs that entail pleiotropic effects and violate the exclusion restriction assumption. For this purpose, we mainly consider a scenario reflecting directional pleiotropy, that is, some of the genetic variants G are consistently positively associated with the outcome Y through a different causal pathway than the risk factors for both uni- and multivariable MR. However, we show results for the balanced pleiotropy setting in Table 4.

The simulation study is set up as follows[4] for each individual $x \in 1, \ldots, 500{,}000$, we simulate 100 genetic variants from a binomial distribution where the minor allele frequency is a probability that is, drawn uniformly between [0.01, 0.5]. In the following, three β-coefficients are simulated from a normal distribution with $\beta_{X_j} \sim N(1, 2)$ for the first-stage regression, with $j \in 1, \ldots, d$ being the index of risk factors. We fix the variance explained for the first stage regression at 15% for all risk factors and the confounder. The calculated variances are used to simulate correlated error terms for the first-stage regression from a multivariate normal distribution, that is, $\epsilon_i \sim MVN(\mu, \Sigma)$ with $\mu = (0{,}0{,}0)$ and Σ being a positive-definite covariance matrix to simulate a medium correlation between the risk factors. The outlying SNPs are simulated as an additional unknown risk factor with a coefficient ρ which is equal to zero for $(1 - p)\%$ of the SNPs and drawn from a uniform or normal distribution for the remaining $p\%$. The variance explained for the second-stage regression is fixed at 50% with causal effects set to 0, 1, and −0.5, for the three risk factors and to 1 for the additional variable that represents the unmeasured pleiotropic pathway that creates outliers (i.e., the unobserved risk factor). From the individual-level data, summary-level data on genetic associations is generated and the different methods are compared for different measures, namely the sensitivity and specificity of the outlier classification, the mean bias of the causal effect estimates $(\frac{1}{z}\sum_{i=1}^{z}\hat{\theta}_i - \theta_i)$, the mean squared error $(\frac{1}{z}\sum_{i=1}^{z}(\hat{\theta}_i - \theta_i)^2)$, the average number of detected outliers in relation to the true outlier rate, and the average absolute number of detected outliers. In total $z = 1{,}000$ simulation runs were performed for each setting. Note that the parameter settings are inspired by Sanderson et al. [21]. The code for the simulation and real data analysis is available on GitHub.[5]

The competing methods are referred to as follows: *Full model* denotes the estimated model with all SNPs, *Standard* denotes the standard Cochran's q-statistic based on first-order weights, *Sanderson* denotes the adjusted q-statistic by Sanderson et al. based on second-order weights [21], *MR-Presso* refers to the outlier test by Verbanck et al. [27], *MR-Radial* describes the MR-Radial method using modified second order weights [28], and *GC-Q* refers to the newly proposed method based on the calibrated first-order weights.

As Tables 1 and 2 show, GC-Q outperforms the other methods in terms of specificity (true negative rate) with nearly 100% for the uni- and multivariable simulation settings. The specificity is similar for MR-Presso and the GC-Q version in the multivariate setting. With regard to the sensitivity (true positive rate), the GC-Q method performs as well as the other methods for most univariable settings and nearly as well as the Standard and Sanderson methods for the multivariable setting (except for the 20% outliers case). Note that the GC-Q method performs still better than MR-Presso (57%). With an increasing outlier rate, the sensitivity of the GC-Q method decreases. This is due to an increasing bias of the median of the q-statistics and is expected since by design GC-Q works up to 25% outlying SNPs for directional pleiotropy and analogously up to 50% for the balanced case.

This means that our method is more conservative in terms of detecting outlying SNPs than the Standard and Sanderson approaches, that is, we can observe less power but a smaller type I error. The average number of detected outliers is always closer to 1 for GC-Q. This illustrates the conservative behavior of GC-Q to only remove the minimum number of outliers necessary to obtain an unbiased causal effect estimate. Our method outperforms the other methods with regard to the bias for most settings, showing that the smaller number of outliers removed by GC-Q also provides the causal effect estimate that is, closest to the actual simulated effect. Even though this is true for a small number of outliers, the Sanderson method should be preferred if more than 25% outlying SNPs are expected. The violin plots in Figure 3 show that even though the GC-Q method exhibits a higher variance than the Standard and Sanderson methods, on average the bias is centered close to zero in contrast to the other methods that have a positive bias. The advantage of GC-Q becomes obvious with stronger outlier effects. As Table 3 shows, GC-Q still performs well in terms of bias, MSE, and average number of detected outliers while the results of the other methods seem to be more harmed by pleiotropic effects that are stronger. In addition, the behavior of GC-Q is as expected if balanced pleiotropy occurs. Table 4 once again depicts its conservatism in terms of sensitivity and its features with regard to bias and MSE.

4 | Real Data Application

In this section, we compare the results of different heterogeneity measures for uni- and multivariable MR based on real data with regard to Vitamin D as exposure for Multiple Sclerosis and blood lipids as candidate exposures for coronary heart disease, age-related macular degeneration, and Alzheimer's.

4.1 | Univariable MR: Vitamin D as Exposure for Multiple Sclerosis

The following application example considers circulating vitamin D levels as exposures for multiple sclerosis (MS) in the univariable MR setting. Summary-level data on genetic associations with vitamin D are derived from 361,194 individuals and taken from UK Biobank.[6] Summary-level data on genetic associations, with the outcome MS including 14,498 European ancestry cases and 24,091 European ancestry controls, was taken from the International Multiple Sclerosis Genetics Consortium (IMSGC)[7] [30].

As instrumental variables we selected $n = 22$ independent (clumping threshold of $r^2 < 0.001$) genetic variants associated with vitamin D at genome-wide significance (p-value $< 5 \times 10^{-8}$). As shown in Table 5 the GC-Q q-values did not detect any outlier,

TABLE 1 | Simulation results for outlier detection in the univariable MR scenario.

Measure	Full model	Standard	Sanderson	MR-presso	MR-radial	GC-Q
5% outliers						
Sensitivity	—	1.00	1.00	1.00	1.00	0.99
Specificity	—	0.97	0.99	0.99	0.90	1.0
Mean bias	0.041	0.002	−0.001	0.0001	−0.0001	−0.002
MSE	0.004	0.0002	0.0001	0.004	0.0005	0.0001
$\bar{p}$	—	1.40	1.11	1.18	2.81	1.01
$\bar{a}$	—	7.01	5.55	5.88	14.04	5.05
10% outliers						
Sensitivity	—	1.00	1.00	1.00	1.00	0.99
Specificity	—	0.94	0.97	0.97	0.84	1.0
Mean bias	0.09	0.01	0.003	0.005	0.03	−0.002
MSE	0.01	0.0004	0.0002	0.01	0.001	0.0001
$\bar{p}$	—	1.55	1.23	1.29	2.47	0.99
$\bar{a}$	—	15.45	12.27	12.89	24.72	9.93
15% outliers						
Sensitivity	—	1.00	1.00	1.00	1.00	0.97
Specificity	—	0.88	0.94	0.93	0.76	1.0
Mean bias	0.13	0.02	0.01	0.01	0.04	0.0007
MSE	0.02	0.001	0.0003	0.02	0.003	0.0002
$\bar{p}$	—	1.66	1.34	1.39	2.37	0.97
$\bar{a}$	—	24.92	20.06	20.88	35.54	14.52
20% outliers						
Sensitivity	—	1.00	1.00	1.00	1.00	0.91
Specificity	—	0.82	0.89	0.88	0.68	1.0
Mean bias	0.18	0.03	0.02	0.02	0.06	0.02
MSE	0.04	0.002	0.0006	0.04	0.005	0.002
$\bar{p}$	—	1.73	1.42	1.47	2.26	0.91
$\bar{a}$	—	34.60	28.38	29.40	45.20	18.23

Note: Sensitivity and specificity for detecting outliers. Mean bias and MSE for the causal effect estimate $\hat{\theta}$, average number of detected outliers in relation to the true outlier rate ($\bar{p}$), and average number of detected outliers ($\bar{a}$).

while the Standard and Sanderson's adjusted q-values detected the same single outlier, rs4944958. MR-Presso and MR-Radial detected two outliers, rs4944958 and rs7041. Genetic variants associated with vitamin D are known to belong to three tiers, the direct vitamin D pathway, U/V absorption and the cholesterol metabolism [31]. Of interest, the one outlier identified by MR-Presso, MR-Radial, the Standard and Sanderson q-values, rs4944958, is an intron of the *NADSYN1* gene which affects a precursor of cholesterol and is part of cholesterol metabolism which may indicate horizontal pleiotropy. In contrast, rs7041, which was only identified by MR-Presso and MR-Radial, is located in the *GC* gene, also known as vitamin D-binding protein, which is part of the direct vitamin D pathway and unlikely to reflect other biological pathways and may represent a false positive finding.

Interestingly, the effect size of the MR analysis depends on the outlier removal approach, with the MR-Presso and MR-Radial approaches that removed more genetic variants having the strongest protective effect estimate, as can be seen in Table 5.

4.2 | Multivariable MR: Blood Lipids as Candidate Exposures

As a second application example, we consider blood lipids as exposures in a multivariable MR setting following Burgess and Davey Smith [4] who analysed if genetically predicted levels of low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides are associated with age-related macular degeneration (AMD) and used coronary heart disease (CHD) as a positive control where clear evidence for lipids being a causal risk factor for CHD exists [32]. In contrast, evidence for lipids being causal for Alzheimer's disease is mixed and not supported by MR studies [33]. Independent

TABLE 2 | Simulation results for outlier detection in the multivariable MR scenario.

Measure	Full model	Standard	Sanderson	MR-presso	GC-Q
5% outliers					
Sensitivity	—	1.00	1.00	0.71	1.00
Specificity	—	0.98	0.99	1.00	1.00
Mean bias	0.03	0.002	0.0005	0.01	−0.0004
MSE	0.003	0.0002	0.0002	0.002	0.0002
$\overline{p}$	—	1.31	1.12	0.71	1.01
$\overline{a}$	—	6.57	5.61	3.57	5.04
10% outliers					
Sensitivity	—	1.00	1.00	0.66	1.00
Specificity	—	0.94	0.96	1.00	1.00
Mean bias	0.06	0.007	0.004	0.03	−0.0004
MSE	0.008	0.0004	0.0003	0.004	0.0002
$\overline{p}$	—	1.58	1.32	0.66	1.00
$\overline{a}$	—	15.82	13.17	6.65	10.01
15% outliers					
Sensitivity	—	1.00	1.00	0.62	0.98
Specificity	—	0.86	0.91	1.00	1.00
Mean bias	0.09	0.02	0.01	0.05	0.006
MSE	0.02	0.0007	0.0005	0.009	0.0008
$\overline{p}$	—	1.81	1.53	0.62	0.98
$\overline{a}$	—	27.14	22.92	9.28	14.65
20% outliers					
Sensitivity	—	1.00	1.00	0.57	0.77
Specificity	—	0.77	0.83	1.00	1.00
Mean bias	0.13	0.02	0.02	0.09	0.06
MSE	0.03	0.001	0.0008	0.02	0.01
$\overline{p}$	—	1.92	1.67	0.57	0.78
$\overline{a}$	—	38.46	33.38	11.46	15.62

Note: Sensitivity and specificity for detecting outliers. Mean bias and MSE for the causal effect estimates $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$, average number of detected outliers in relation to the true outlier rate ($\overline{p}$), and average number of detected outliers ($\overline{a}$).

genetic variants were selected as IVs if they were associated with any of the blood lipids at genome-wide significance [4] resulting in $n = 185$ IVs.

Table 6 shows the genetic variants used as IVs and detected as outliers by the different approaches. In general, the GC-Q approach detected the fewest outliers, followed by MR-Presso, which is in line with the simulation results for the methods based on Cochran's Q. The Standard q-value approach and Sanderson's q-value detected exactly the same outliers. The difference in detection of outliers is reflected in the respective MR estimates as shown in Table 7. For CHD, all methods yield similar effect sizes independent of the removal for outliers. In contrast, for the outcomes AMD and Alzheimer's disease, the MR effect estimates not only differ in their effect sizes but also in their significance. For example, the effect estimate of HDL-cholesterol on AMD is significant at the 5% level for the full model without outlier removal, whereas it doubles its effect size and is even significant at the 0.001 level after outlier removal with the other methods. The benefit of outlier removal is most striking for the effect of LDL-cholesterol on Alzheimer's disease. Here, including all genetic variants associated with any blood lipid, provided significant evidence for genetically predicted levels of LDL-cholesterol to be associated with Alzheimer's disease. The removal of outliers, in particular of one genetic variant, rs6859, in the *APOE* gene region, leads to an insignificant MR effect estimate.

5 | Related Work

Some researchers prefer to use robust methods, such as the MR-Median approach to avoid dealing with outliers. But firstly, we are more interested in detecting outlying SNPs and thus robust estimation methods such as the Median-MR approach would not be helpful in this endeavor. And secondly, these approaches have

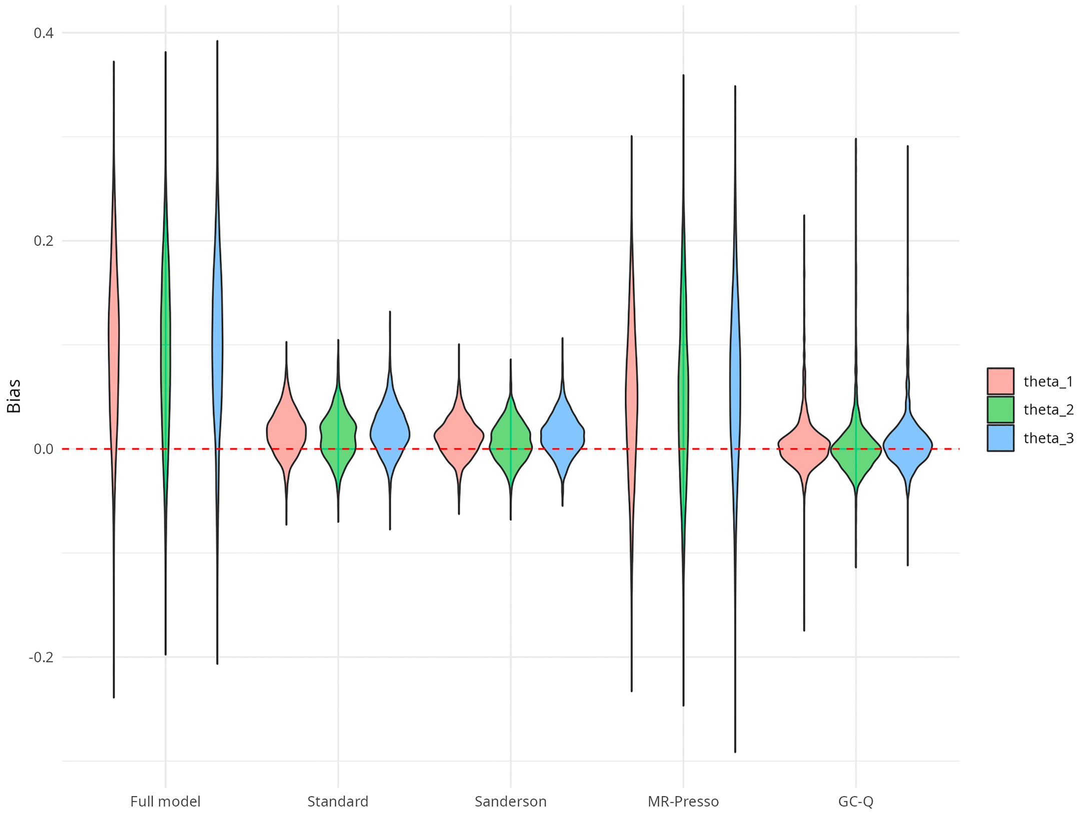

FIGURE 3 | Violin plots for the bias of the causal effect estimates of θ_1 (red), θ_2 (green), and θ_3 (blue) in the multivariate simulation setting (15% outliers) after outlier adjustment and for the full model.

TABLE 3 | Simulation results for outlier detection in the univariable MR scenario with a strong outlier effect ($p = 15\%$).

Measure	Full model	Standard	Sanderson	MR-presso	MR-Radial	GC-Q
Sensitivity	—	1.00	1.00	1.00	1.00	0.98
Specificity	—	0.60	0.67	0.66	0.47	1.0
Mean bias	0.54	0.11	0.07	0.07	0.19	0.007
MSE	0.40	0.02	0.01	0.40	0.05	0.003
$\overline{p}$	—	3.25	2.85	2.90	4.02	0.98
$\overline{a}$	—	48.70	42.69	43.50	60.37	14.68

Note: Sensitivity and specificity for detecting outliers. Mean bias and MSE for the causal effect estimate $\hat{\theta}$, average number of detected outliers in relation to the true outlier rate ($\overline{p}$), and average number of detected outliers ($\overline{a}$).

TABLE 4 | Simulation results for outlier detection in the univariable MR scenario with balanced pleiotropy ($p = 20\%$).

Measure	Full model	Standard	Sanderson	MR-presso	MR-Radial	GC-Q
Sensitivity	—	1.00	1.00	1.00	1.00	0.99
Specificity	—	0.96	0.98	0.97	0.86	1.0
Mean bias	−0.0005	−0.002	−0.003	−0.003	−0.003	−0.002
MSE	0.009	0.0004	0.0002	0.009	0.001	0.0003
$\overline{p}$	—	1.18	1.09	1.11	1.57	0.99
$\overline{a}$	—	23.59	21.78	22.24	31.37	19.73

Note: Sensitivity and specificity for detecting outliers. Mean bias and MSE for the causal effect estimate $\hat{\theta}$, average number of detected outliers in relation to the true outlier rate ($\overline{p}$), and average number of detected outliers ($\overline{a}$).

TABLE 5 | Real data analysis: Causal effect estimates, standard errors, p-values, and outlying SNPs for vitamin D on multiple sclerosis for the univariable MR scenario.

Method	Causal estimate $\hat{\theta}$	Std. Error	p	SNP
Full model	−0.44	0.10	0.0003	—
Standard	−0.30	0.09	0.0048	rs4944958
Sanderson	−0.30	0.09	0.0048	rs4944958
MR-presso	−0.53	0.12	0.0002	rs4944958, rs7041
MR-radial	−0.54	0.10	0.00001	rs4944958, rs7041
GC-Q	−0.44	0.10	0.0003	—

TABLE 6 | Real data analysis: Outlying SNPs in the multivariate MR scenario.

Method	CHD	AMD	ALZ
Standard	rs1250229, rs4530754, rs579459, rs12801636, rs653178, rs6489818, rs952044	rs1883025, rs653178, rs1532085, rs261342 rs9989419, rs6859, rs492602	rs1883025, rs17788930, rs6859
Sanderson	rs1250229, rs4530754, rs579459, rs12801636, rs653178, rs6489818, rs952044	rs1883025, rs653178, rs1532085, rs261342 rs9989419, rs6859, rs492602	rs1883025, rs17788930, rs6859
MR-presso	rs4530754, rs12801636, rs653178, rs952044	rs1883025, rs1532085, rs261342, rs9989419, rs6859	rs17788930, rs6859
GC-Q	rs653178	rs1532085, rs261342	rs6859

Note: Data from Burgess and Davey Smith [4], originally from the Global Lipids Genetics Consortium [34] and Fritsche et al. [35].

their own specific disadvantages, for example, MR-Median has less power compared to other methods. If we use the MR-median method in our real data application in Section 4 on the effect of vitamin D on multiple sclerosis, we get an estimate of −0.296 (0.164) with a p-value of 0.07. Even though the Standard and Sanderson method perform similarly in the effect size, we end up with higher standard errors and a change of significance of the causal estimate. Nonetheless, these methods offer an alternative to circumvent the issue of outliers in MR analyses with respect to less biased causal effect estimates. Table 8 we benchmarked GC-Q with the Median-based method [25] and the MR-RAPS method [36]. For the setting with 10% outliers, all methods perform similarly well. However, the bias and MSE get very large in the setting with 80% outliers for MR-RAPS and GC-Q, with an advantage for the Median-based method with respect to the bias and the MSE. GC-Q performs similarly to the full model without any outliers removed.

6 | Limitations

An important aspect of introducing new methods is to highlight their respective limitations. As we have already mentioned, GC-Q works for up to 25% outlying SNPs for directional pleiotropy and analogously up to 50% for the balanced case. Table 9 shows its performance in extreme cases with 50% and 80% of the SNPs being outliers in the univariate MR setting for directional pleiotropy, that is, the assumptions of the methods do not hold anymore.

As the estimation of $\hat{\lambda}$ is dependent on the median of the unadjusted local q-statistics, we observe an extremely inflated distribution over the simulation runs. $\hat{\lambda}$ is on average inflated by a factor of 100, which results in a deflation of the local q-statistics and subsequently no outlying SNPs can be found. GC-Q thus performs similar to the full model without any outliers removed.

As Table 9 the other outlier detection methods still work better in the 50% setting—the bias, however, is not negligible and seems to gradually approach the bias of the full model with higher outlier proportions. Since $\hat{\lambda}$ of the GC-Q method can no longer be estimated correctly and is based on the median of the χ^2_1 distribution, there might be a way to adjust it with a simple correction factor for the hyper-inflated median of the unadjusted q-statistic for settings where its assumptions do not hold. As a reliable rule of thumb for the correction factor is beyond the scope of this manuscript, we leave it open for future investigations.

7 | Discussion

Overdispersion in the heterogeneity statistic is a common problem in meta-analysis [19, 20] and is not limited to MR. For example, in meta-analysis of clinical trials, heterogeneity may arise because of a diverse range of factors including diversity in doses, lengths of follow-up, study quality, and inclusion criteria for participants [20]. In MR, heterogeneity can be caused by different molecular pathways affecting the exposure. For example, there genetic variation acts via many different biological pathways on obesity including the metabolism, cholesterol transport, fat storage, appetite regulation, food preference, reward mechanisms, and physical exercise. The heterogeneity test statistic is known to depend on the sample size [19] which is the number of studies included in a meta-analysis or in MR, the number of genetic variants used as IVs. The power is low when there are few SNPs; in contrast, the heterogeneity statistic shows substantial overdispersion when there are many SNPs. Powerful GWAS have

TABLE 7 | Real data analysis: Causal effect estimates for LDL- ($\widehat{\theta}_1$), HDL-cholesterol ($\widehat{\theta}_2$), and triglycerides ($\widehat{\theta}_3$) on coronary heart disease, macular degeneration, and Alzheimer's.

Method	**CHD**			**AMD**			**ALZ**		
	Estimate	**Std. Err.**	***p***	**Estimate**	**Std. Err.**	***p***	**Estimate**	**Std. Err.**	***p***
Full model									
$\widehat{\theta}_1$	0.39	0.04	2×10^{-16}	−0.04	0.07	0.55	0.24	0.08	0.002
$\widehat{\theta}_2$	−0.07	0.05	0.116	0.18	0.08	0.028	−0.11	0.09	0.206
$\widehat{\theta}_2$	0.14	0.06	0.012	−0.07	0.10	0.437	−0.14	0.11	0.186
Standard									
$\widehat{\theta}_1$	0.41	0.04	2×10^{-16}	−0.03	0.06	0.646	0.09	0.05	0.055
$\widehat{\theta}_2$	−0.06	0.04	0.165	0.35	0.08	1.7×10^{-5}	−0.06	0.05	0.238
$\widehat{\theta}_3$	0.14	0.05	0.005	0.10	0.08	0.250	−0.09	0.06	0.161
Sanderson									
$\widehat{\theta}_1$	0.41	0.04	2×10^{-16}	−0.03	0.06	0.646	0.09	0.05	0.055
$\widehat{\theta}_2$	−0.06	0.04	0.165	0.35	0.08	1.7×10^{-5}	−0.06	0.05	0.238
$\widehat{\theta}_3$	0.14	0.05	0.005	0.10	0.08	0.250	−0.09	0.06	0.161
MR-presso									
$\widehat{\theta}_1$	0.41	0.04	2×10^{-16}	−0.03	0.06	0.640	0.08	0.05	0.082
$\widehat{\theta}_2$	−0.05	0.04	0.201	0.34	0.08	1.7×10^{-5}	−0.10	0.05	0.074
$\widehat{\theta}_2$	0.13	0.05	0.010	0.14	0.08	0.106	−0.12	0.07	0.078
GC-Q									
$\widehat{\theta}_1$	0.40	0.04	2×10^{-16}	−0.06	0.06	0.324	0.09	0.05	0.081
$\widehat{\theta}_2$	−0.06	0.05	0.155	0.47	0.08	1.7×10^{-5}	−0.08	0.06	0.148
$\widehat{\theta}_3$	0.14	0.05	0.011	0.18	0.09	0.045	−0.11	0.07	0.091

Note: Data from Burgess et al. [4].

TABLE 8 | The simulation results for outlier detection in the univariable MR scenario.

Measure	**Full model**	**GC-Q**	**Median**	**MR-RAPS**
10% outliers				
Mean bias	0.087	−0.002	0.002	0.090
MSE	0.011	0.0001	0.0002	0.011
80% outliers				
Mean bias	0.72	0.65	0.39	0.72
MSE	0.54	0.50	0.26	0.54

Note: Mean bias and MSE for the causal effect estimates $\widehat{\theta}$ for GC-Q, MR-RAPS (with overdispersion and L2 loss), and the Median-based method. For details on the parameter settings see Section 3.

identified hundreds of regions in the genome associated with potential exposures and provide a large number of IVs, making the calibration of the heterogeneity statistic an important statistical problem.

Here, we propose GC-Q, an adjusted version of the local q-statistic to detect outliers. GC-Q has the potential to decrease the type I error at the price of a reduced power (see Tables 1 and 2). With this method, we correct for overdispersion of the heterogeneity statistics by making use of the estimated inflation factor using a mixture model approach.

GC-Q is using the first-order weights, which in contrast to the second-order weights do not include the precision of the genetic association with the exposure, which is also known as the no measurement error (NoME) assumption. Another important assumption of GC-Q is that less than half of the IVs are invalid (in the balanced pleiotropy setting), an assumption that is, necessary to guarantee the identifiability of the mixture model. In order to estimate the over-dispersion parameter, GC-Q requires a minimum number of IVs and is only recommended for polygenic exposures where there are many genetic variants available as IVs. The mixture model on which GC-Q is formulated has been shown in simulations to perform well on 50 observations and conservative when fewer observations are available [37]. In addition, GC-Q performs especially well if the outlier effect is strong (see Table 3). Another advantage of GC-Q is that it does not require additional parameters, as the observational covariance between exposures, and it does not require a two-step procedure for estimation (note the second-order weights require the causal effect estimate and can only be obtained in an iterative procedure). MR-PRESSO uses a computationally expensive simulation procedure to define the Null distribution, which becomes computationally more expensive as the number of instruments grows. In contrast, GC-Q is based on a fast and simple computational implementation which uses first-order weights and relies on a closed-form mixture model formulation. Thus, no iterative procedure to calculate the weights or simulations are needed to define the Null distribution. A disadvantage of GC-Q is that it relies on

TABLE 9 | A simulation results for outlier detection in the univariable MR scenario.

Measure	Full model	Standard	Sanderson	MR-presso	MR-radial	GC-Q
50% outliers						
Sensitivity	—	0.99	0.99	0.99	1.00	0.01
Specificity	—	0.54	0.65	0.64	0.42	1.0
Mean bias	0.45	0.14	0.13	0.12	0.18	0.45
MSE	0.22	0.04	0.05	0.22	0.05	0.22
$\overline{p}$	—	1.45	1.33	1.35	1.57	—
$\overline{a}$	—	72.67	66.74	67.30	78.70	—
80% outliers						
Sensitivity	—	0.96	0.94	0.94	0.97	0
Specificity	—	0.39	0.50	0.49	0.31	1.0
Mean bias	0.72	0.63	0.67	0.66	0.60	0.72
MSE	0.54	0.49	0.54	0.54	0.46	0.54
$\overline{p}$	—	1.12	1.06	1.07	1.15	—
$\overline{a}$	—	89.25	85.14	85.47	91.67	—

Note: Sensitivity and specificity for detecting outliers. Mean bias and MSE for the causal effect estimates $\hat{\theta}$, average number of detected outliers in relation to the true outlier rate ($\overline{p}$), and average number of detected outliers ($\overline{a}$).

the assumption that less than half of the IVs are valid, but this assumption is common to all other methods which rely on outlier detection, including MR-PRESSO, Radial MR, or the Median MR method.

However, we do not claim that our method outperforms the existing methods in all cases. We see our approach as complementary to the outlier detection methods in MR analysis.

When removing outliers in MR models, it is necessary to strike a balance between removing all invalid IVs that may bias the causal effect estimate and retaining the largest number of IVs to retain the largest sample size possible.

As we show in our simulation study and in the real data analysis, GC-Q removes the smallest number of outliers while obtaining unbiased causal effect estimates, which highlights that GC-Q is conservative and removes only the minimum number of invalid IVs necessary to obtain the unbiased causal effect.

Let us finish by a recent quote from Strobl and Leisch [38]. They claim that "the research question 'What is the best method in general' is ill-posed" and warn methodological researchers who present new methods against the "one method fits them all" philosophy. In this spirit, we emphasize that our method certainly cannot be recommended universally for all datasets and in all contexts (especially in an early phase paper such as ours [8]), but shows a promising behavior in practically relevant situations. With this in mind, we do not want to claim that our newly introduced method outperforms existing methods in every analytical setting and can be seen as complementary within the MR literature.

Author Contributions

M.M.M. and V.Z. designed the study. M.M.M. implemented the method, performed the simulations, analyzed the data, and interpreted the results. M.M.M. and V.Z. prepared the initial manuscript draft. V.Z. directed the project. M.M.M., A.L.B., V.Z., and S.B. reviewed and edited the manuscript.

Acknowledgments

This study was partly funded by DFG grant BO3139/7-1 to Anne-Laure Boulesteix. Verena Zuber is supported by the United Kingdom Research and Innovation Medical Research Council grant MR/W029790/1. All data used in our study is in the public domain. Our study is based on publicly available summary-level data on genetic associations from the International AMD Genetics Consortium http://amdgenetics.org/, the GWAS catalog https://www.ebi.ac.uk/gwas/, MAGNETIC NMR-GWAS http://www.computationalmedicine.fi/data, the International Multiple Sclerosis Genetics Consortium https://imsgc.net/, and UK Biobank http://www.nealelab.is/uk-biobank. Open Access funding enabled and organized by Projekt DEAL.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data and code can be found on **GitHub:** https://github.com/mmax-code/MR_outliers.

Endnotes

[1] **GitHub:** https://github.com/mmax-code/MR_outliers.

[2] Note that also linearity and homogeneity assumptions must hold.

[3] Note that also linearity and homogeneity assumptions must hold.

[4] The simulation setup is similar for the univariate case with only one risk factor.

[5] **GitHub:** https://github.com/mmax-code/MR_outliers.

[6] https://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=30890.

[7] https://www.ebi.ac.uk/gwas/studies/GCST005531.

References

1. G. Davey Smith and S. Ebrahim, “Mendelian Randomization: Can Genetic Epidemiology Contribute to Understanding Environmental Determinants of Disease?,” *International Journal of Epidemiology* 32, no. 1 (2003): 1–22.

2. T. G. Richardson, E. Sanderson, T. M. Palmer, et al., “Evaluating the Relationship Between Circulating Lipoprotein Lipids and Apolipoproteins With Risk of Coronary Heart Disease: A Multivariable Mendelian Randomisation Analysis,” *PLoS Medicine* 17, no. 3 (2020): e1003062.

3. V. Zuber, D. Gill, M. Ala-Korpela, et al., “High-Throughput Multivariable Mendelian Randomization Analysis Prioritizes Apolipoprotein B as Key Lipid Risk Factor for Coronary Artery Disease,” *International Journal of Epidemiology* 50, no. 3 (2021): 893–901.

4. S. Burgess and G. Davey Smith, “Mendelian Randomization Implicates High-Density Lipoprotein Cholesterol–Associated Mechanisms in Etiology of Age-Related Macular Degeneration,” *Ophthalmology* 124, no. 8 (2017): 1165–1174, https://doi.org/10.1016/j.ophtha.2017.03.042.

5. L. E. Mokry, S. Ross, O. S. Ahmad, et al., “Vitamin D and Risk of Multiple Sclerosis: A Mendelian Randomization Study,” *PLoS Medicine* 12, no. 8 (2015): e1001866.

6. N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller, “Pleiotropy in Complex Traits: Challenges and Strategies,” *Nature Reviews Genetics* 14, no. 7 (2013): 483–495.

7. J. Bowden, G. Hemani, and G. Davey Smith, “Invited Commentary: Detecting Individual and Global Horizontal Pleiotropy in Mendelian Randomization—A Job for the Humble Heterogeneity Statistic?,” *American Journal of Epidemiology* 187, no. 12 (2018): 2681–2685.

8. G. Heinze, A. L. Boulesteix, M. Kammer, T. P. Morris, I. R. White, and STRATOS initiative, “Phases of Methodological Research in Biostatistics—Building the Evidence Base for New Methods,” *Biometrical Journal* 66, no. 1 (2024): 2200222.

9. F. P. Hartwig, N. M. Davies, G. Hemani, and G. Davey Smith, “Two-Sample Mendelian Randomization: Avoiding the Downsides of a Powerful, Widely Applicable but Potentially Fallible Technique,” *International Journal of Epidemiology* 45, no. 6 (2016): 1717–1726, https://doi.org/10.1093/ije/dyx028.

10. S. Burgess, R. A. Scott, N. J. Timpson, G. Davey Smith, and S. G. Thompson, “Using Published Data in Mendelian Randomization: A Blueprint for Efficient Identification of Causal Risk Factors,” *European Journal of Epidemiology* 30, no. 7 (2015): 543–552, https://doi.org/10.1007/s10654-015-0011-z.

11. S. Burgess, F. Dudbridge, and S. G. Thompson, “Combining Information on Multiple Instrumental Variables in Mendelian Randomization: Comparison of Allele Score and Summarized Data Methods,” *Statistics in Medicine* 35, no. 11 (2016): 1880–1906.

12. S. Greenland, “An Introduction to Instrumental Variables for Epidemiologists,” *International Journal of Epidemiology* 29, no. 4 (2000): 722–729.

13. T. Johnson, “Efficient Calculation for Multi-SNP Genetic Risk Scores,” Technical Report, The Comprehensive R Archive Network, 2013.

14. J. D. Angrist and G. W. Imbens, “Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity,” *Journal of the American Statistical Association* 90, no. 430 (1995): 431–442.

15. S. Burgess and J. Bowden, “Integrating Summarized Data From Multiple Genetic Variants in Mendelian Randomization: Bias and Coverage Properties of Inverse-Variance Weighted Methods,” preprint, arXiv:1512.04486, 2015.

16. S. Burgess, A. Butterworth, and S. G. Thompson, “Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data,” *Genetic Epidemiology* 37, no. 7 (2013): 658–665.

17. S. Burgess and S. G. Thompson, “Multivariable Mendelian Randomization: The Use of Pleiotropic Genetic Variants to Estimate Causal Effects,” *American Journal of Epidemiology* 181, no. 4 (2015): 251–260.

18. F. Del Greco, C. Minelli, N. A. Sheehan, and J. R. Thompson, “Detecting Pleiotropy in Mendelian Randomisation Studies With Summary Data and a Continuous Outcome,” *Statistics in Medicine* 34, no. 21 (2015): 2926–2940.

19. R. J. Hardy and S. G. Thompson, “Detecting and Describing Heterogeneity in Meta-Analysis,” *Statistics in Medicine* 17, no. 8 (1998): 841–856.

20. J. P. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman, “Measuring Inconsistency in Meta-Analyses,” *British Medical Journal* 327, no. 7414 (2003): 557–560.

21. E. Sanderson, G. Davey Smith, F. Windmeijer, and J. Bowden, “An Examination of Multivariable Mendelian Randomization in the Single-Sample and Two-Sample Summary Data Settings,” *International Journal of Epidemiology* 48, no. 3 (2019): 713–727.

22. J. Bowden, F. M. Del Greco, C. Minelli, et al., “Improving the Accuracy of Two-Sample Summary Data Mendelian Randomization: Moving Beyond the NOME Assumption,” preprint, bioRxiv, 2018: 159442.

23. B. Devlin and K. Roeder, “Genomic Control for Association Studies,” *Biometrics* 55, no. 4 (1999): 997–1004.

24. B. Devlin, K. Roeder, and L. Wasserman, “Genomic Control, A New Approach to Genetic-Based Association Studies,” *Theoretical Population Biology* 60, no. 3 (2001): 155–166.

25. J. Bowden, G. Davey Smith, P. C. Haycock, and S. Burgess, “Consistent Estimation in Mendelian Randomization With Some Invalid Instruments Using a Weighted Median Estimator,” *Genetic Epidemiology* 40, no. 4 (2016): 304–314.

26. P. J. Rousseeuw and M. Hubert, “Robust Statistics for Outlier Detection,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1, no. 1 (2011): 73–79.

27. M. Verbanck, C. Y. Chen, B. Neale, and R. Do, “Detection of Widespread Horizontal Pleiotropy in Causal Relationships Inferred From Mendelian Randomization Between Complex Traits and Diseases,” *Nature Genetics* 50, no. 5 (2018): 693–698.

28. J. Bowden, W. Spiller, M. F. Del Greco, et al., “Improving the Visualization, Interpretation and Analysis of Two-Sample Summary Data Mendelian Randomization via the Radial Plot and Radial Regression,” *International Journal of Epidemiology* 47, no. 4 (2018): 1264–1278, https://doi.org/10.1093/ije/dyy101.

29. J. Bowden, G. Davey Smith, and S. Burgess, “Mendelian Randomization With Invalid Instruments: Effect Estimation and Bias Detection Through Egger Regression,” *International Journal of Epidemiology* 44, no. 2 (2015): 512–525, https://doi.org/10.1093/ije/dyv080.

30. A. H. Beecham, N. A. Patsopoulos, International Multiple Sclerosis Genetics Consortium (IMSGC), et al., “Analysis of Immune-Related Loci Identifies 48 New Susceptibility Variants for Multiple Sclerosis,” *Nature Genetics* 45, no. 11 (2013): 1353–1360, https://doi.org/10.1038/ng.2770.

31. D. Manousaki, R. Mitchell, T. Dudding, et al., “Genome-Wide Association Study for Vitamin D Levels Reveals 69 Independent Loci,” *American Journal of Human Genetics* 106, no. 3 (2020): 327–337.

32. M. V. Holmes, F. W. Asselbergs, T. M. Palmer, et al., “Mendelian Randomization of Blood Lipids for Coronary Heart Disease,” *European Heart Journal* 36, no. 9 (2015): 539–550.

33. S. C. Larsson, M. Traylor, R. Malik, M. Dichgans, S. Burgess, and H. S. Markus, “Modifiable Pathways in Alzheimer’s Disease: Mendelian Randomisation Analysis,” *BMJ* 359 (2017): j5375, https://doi.org/10.1136/bmj.j5375.

34. C. J. Willer, E. M. Schmidt, S. Sengupta, et al., “Discovery and Refinement of Loci Associated With Lipid Levels,” *Nature Genetics* 45, no. 11 (2013): 1274–1283, https://doi.org/10.1038/ng.2797.

35. L. G. Fritsche, W. Igl, J. N. C. Bailey, et al., “A Large Genome-Wide Association Study of Age-Related Macular Degeneration Highlights Contributions of Rare and Common Variants,” *Nature Genetics* 48, no. 2 (2016): 134–143, https://doi.org/10.1038/ng.3448.

36. Q. Zhao, J. Wang, G. Hemani, J. Bowden, and D. S. Small, “Statistical Inference in Two-Sample Summary-Data Mendelian Randomization Using Robust Adjusted Profile Score,” *Annals of Statistics* 48, no. 3 (2020): 1742–1769, https://doi.org/10.1214/19-AOS1866.

37. S. A. Bacanu, B. Devlin, and K. Roeder, “The Power of Genomic Control,” *American Journal of Human Genetics* 66, no. 6 (2000): 1933–1944, https://doi.org/10.1086/302929.

38. C. Strobl and F. Leisch, “Against the “One Method Fits All Data Sets” Philosophy for Comparison Studies in Methodological Research,” *Biometrical Journal* 66, no. 1 (2024): 2200104.

A.3 Contribution 3

»The impact of the *storytelling fallacy* on real data examples in methodological research«

Citation

Mandl, M.M., Weber, F., Wöhrle, T., Boulesteix, A.-L. (2025) The impact of the *storytelling fallacy* on real data examples in methodological research. *arxive:2503.03484*. https://doi.org/10.1371/journal.pcbi.1011936

Authors' contributions

Conceptualization: MM, ALB; Formal analysis: MM; Funding acquisition: ALB; Investigation: MM, TW, FW; Project Administration: ALB; Supervision: ALB; Writing – original draft: MM; Writing – review & editing: MM, ALB, TW, FW.

Rights and permissions

The impact of the *storytelling fallacy* on real data examples in methodological research

Maximilian M. Mandl*[1,2], Frank Weber[1], Tobias Woehrle[3] and Anne-Laure Boulesteix[1,2]

[1]Institute for Medical Information Processing, Biometry and Epidemiology, Faculty of Medicine,
Ludwig-Maximilians-Universität München, Germany
[2]Munich Center for Machine Learning, Germany.
[3]Department of Anesthesiology, LMU University Hospital, Ludwig-Maximilians-Universität München, Germany

Abstract

The term "researcher degrees of freedom" (RDF), which was introduced in metascientific literature in the context of the replication crisis in science, refers to the extent of flexibility a scientist has in making decisions related to data analysis. These choices occur at all stages of the data analysis process, e.g., data preprocessing and modelling. In combination with selective reporting, RDF may lead to over-optimistic statements and an increased rate of false positive findings. Even though the concept has been mainly discussed in fields of application of statistics such as epidemiology or psychology, similar problems affect methodological statistical research. Researchers who develop and evaluate statistical methods are left with a multitude of decisions when designing their comparison studies. This leaves room for an over-optimistic representation of the performance of their preferred method(s) and false positive findings. In this context, the present paper defines and explores a particular RDF that has not been previously identified and discussed. When interpreting the results of real data examples that are most often part of methodological evaluations, authors typically tell a domain-specific "story" that best supports their argumentation in favor of their preferred method. However, there are often plenty of other plausible stories that would support different conclusions. We define the "storytelling fallacy" as the selective use of anecdotal domain-specific knowledge to support the superiority of specific methods in real data examples. While such examples fed by domain knowledge play a vital role in methodological research, if deployed inappropriately they can also harm the validity of conclusions on the investigated methods. The goal of our work is to create awareness for this issue, fuel discussions on the role of real data in generating evidence in methodological research and warn readers of methodological literature against naive interpretations of real data examples. We illustrate this newly introduced RDF through two examples related to pleiotropy detection in Mendelian Randomisation and a prediction model to detect SARS-CoV-2 infections, respectively.

Keywords— Metascience, real data application, selective reporting, over-optimism, comparison studies

*__Correspondence:__ Maximilian M. Mandl, Institute for Medical Information Processing, Biometry, and Epidemiology, Faculty of Medicine, Ludwig-Maximilians-Universität München, Marchioninistr. 15, 81377 München, Germany. mmandl@ibe.med.uni-muenchen.de

1 Introduction

The term "researcher degrees of freedom" (RDF) [1], which was introduced in metascientific literature in the context of the replication crisis in science [2, 3, 4], refers to the extent of flexibility a scientist has in making decisions related to data analysis. These choices occur at all stages of the data analysis process, e.g., data preprocessing and modelling [5]. In combination with selective reporting, RDF may lead to over-optimistic statements and an increased rate of false positive findings [5, 6]. Even though the concept has been so far mainly discussed in fields of application of statistics such as epidemiology or psychology, similar problems affect methodological statistical research. Here, we define methodological statistical research as research dedicated to the development and evaluation of statistical methods, for example statistical tests or regression modelling—as opposed to research addressing epidemiological questions using these methods.

Researchers developing and evaluating statistical methods are left with a multitude of decisions when designing their comparison studies. Jelizarow et al. [7] demonstrate how a new class prediction method that is in fact worse than existing methods can artificially seem superior through selective reporting. More precisely, they intentionally focus on the datasets, data preprocessing settings, and method variants that maximize the performance of the new method while downplaying others. They show that this selective reporting approach leads to an overoptimistic and misleading representation of their preferred method. Such bias induced by the RDF typically affects the evaluations presented as part of papers introducing new methods, including those based on statistical simulation studies [8, 9].

It should be emphasized that we do not insinuate that scientists deliberately engage in scientific misconduct. In fact, selective reporting often happens subconsciously without malicious intention, as a result of self-deception [10]. The reasons are manifold. One key factor that has been widely acknowledged in health and social sciences is publication bias that encourages the reporting of positive findings. It is likely that such a bias also affects methodological literature [11]—in the sense that researchers are required to propose methods that outperform existing methods to have their research published.

In this context, the present paper defines and explores a particular RDF that has not been previously identified and discussed. When interpreting the results in real data examples that are most often part of methodological evaluations, authors typically tell a domain-specific "story" that supports their argumentation in favor of their preferred method. However, there are often plenty of other plausible stories that would support different conclusions. We define the "storytelling fallacy" as the selective use of anecdotal domain-specific knowledge to support the superiority of specific methods in real data examples. While applications fed by domain knowledge play a vital role in methodological research, if deployed inappropriately they can also harm the validity of conclusions on the investigated methods and lead to the publication of non-replicable methodological results.

The goal of our paper is to create awareness for this issue, fuel discussions on the role of real data in generating evidence in methodological research and warn readers of methodological literature against naive interpretations of real data examples. After introducing the concept of the "storytelling fallacy" in more detail in Section 2, we illustrate it through two examples inspired from our own research. In Section 3, we consider methods for pleiotropy detection in Mendelian Randomisation (MR) for causal inference. The different methods yield different sets of results. For each of these sets of results, a plausible biological interpretation ("story") can be elaborated to strengthen the case of the corresponding method. In Section 4, we consider a new diagnostic method for SARS-CoV-2 infections based on machine learning models. Different models identify different predictor variables as relevant. Again, for each set of results, a plausible biological "story" based on prior knowledge on these predictor variables can be created to support the use of the corresponding model. The paper finally discusses these findings and formulates recommendations for authors and readers of methodological comparison studies with respect to real data examples (Section 5).

2 The "storytelling fallacy"

2.1 Definition

Building on Jelizarow et al. [7], Nießl et al. [12] systematically investigate the impact of RDF on the results of a showcase study comparing the accuracy of various survival prediction methods. Typical RDF included in their study are the choices of performance measures, datasets, summarization of results over datasets, and handling of method failures. Depending on these choices, different survival prediction methods can be identified as best performing. Assuming that authors often engage (consciously or not) in selective reporting, this may largely explain why papers introducing new methods are generally over-optimistic with respect to their performances, even if they use essentially objective criteria [13, 14, 15].

In this context, a natural reaction would be to give more importance to real data applications and to the plausibility of the results obtained therein from a domain-knowledge perspective. If—according to expert judgement derived from domain knowledge—the results obtained with method A are much more plausible and meaningful than those obtained with method B, it is often seen as an argument in favor of the superiority of method A.

Such interpretations are common in methodological articles presenting new methods. For example, let us consider the case of a new model selection approach for multivariable regression modelling (called method A and compared to a standard approach called method B). In a real data application, the inventors of method A who want to present it as superior to method B may scrutinize the variables it selects and argue that they are more meaningful than those selected by method B. They may for example argue that method A selects a variable that was identified as important predictor of the target variable in previous literature, and that method B fails to select it. This would be presented as an argument in favor of the superiority of method A.

The goal of the present paper is to outline that RDF also affect this type of comparisons in a broad sense, and that the interpretation of real data applications is consequently not immune against bias. This is because, for a given real data application, there are usually numerous possible sensible domain-specific interpretations. Focusing on the one that makes method A appear better than method B in some sense (while ignoring those that make method B look better) can be seen as a form of selective reporting. This happens in a particularly subtle way, because in practice the authors do not actively select their interpretation out of a collection of ready-to-use interpretations: instead, they elaborate their interpretation based on the results of methods A and B—but may have elaborated another one if their preferences had been different.

With this in mind, we define the "storytelling fallacy" as the use of anecdotal domain-specific knowledge in order to support the superiority of the preferred method in real data examples.

2.2 Related literature

The “storytelling fallacy” is related to the question of the reliability and objectivity of domain-specific expert knowldege, which has been widely discussed in different fields of research. We give a brief overview of this literature in connection with the “storytelling fallacy” in the rest of this section.

Experts are generally assumed to be immune to blind spots and generally impartial [16] and, as stated by O’Hagan [17], “*[e]xpert opinion and judgment enter into the practice of statistical inference and decision-making in numerous ways. Indeed, there is essentially no aspect of scientific investigation in which judgment is not required.*” Regarding subjectivity, they further argue that “*[j]udgment is necessarily subjective, but should be made as carefully, as objectively, and as scientifically as possible.*” Kaptchuk [18] points out that the interpretation of data is inevitably subjective and is not preserved from bias—referred to as interpretative bias. Furthermore, the well-known confirmation bias [19] refers to the fact that researchers may evaluate evidence in a way that supports their own prior beliefs.

In statistics, the notion of expert judgement is often considered in the debate opposing frequentist to Bayesian statistics, which often ends in discussions on the subjectivity and objectivity of decisions [20, 21] such as the choice of the prior distributions in Bayesian statistics. The “storytelling fallacy” in methodological literature is related to the subjectivity of expert judgement, but in a different manner. The judgement of an expert who elaborates an interpretation to strengthen the argument in favor of a newly introduced method is subjective—in the sense that another expert may have another judgement. However, the core issue is the RDF combined with selective reporting, i.e. the multiplicity of potential post-hoc stories and the fact that the one that best fits the authors’ hopes is chosen, leading to a biased evaluation of the methods.

With a different perspective, Boutron and Ravaud [22] address the problem of misrepresentation of research in biomedical literature and discuss the notion of “spin”. The spin is defined as a type of reporting that does not accurately represent the nature and scope of findings and may influence readers’ perceptions of the results. There are different shades of “spins”, such as the misreporting of methods and the misreporting of results. Similarly, methodological researchers may evaluate their results in a biased way by focusing on specific performance measures (i.e. methods), datasets (i.e. results) or, as outlined in this paper, specific biological theories supporting the veracity of results of real data analyses and thus the superiority of the method that produced them.

3 Use Case I: Detection of pleiotropic SNPs in Mendelian Randomisation

Mendelian Randomisation (MR) employs genetic variants as instrumental variables to deduce the causal effects of exposures on an outcome. A crucial assumption in MR is that these genetic variants, used as instrumental variables, are independent of the outcome, given the risk factor and any unobserved confounders.[23] More precisely, the goal of MR is to examine the causal effect θ of a risk factor X on an outcome Y using genetic variants G_i for $i \in 1, ..., n$ as instrumental variables (IV), see Figure 1. Pleiotropy is defined as the effect of any genetic variant G_i (IV) on the outcome Y through any other path than the risk factor X included in the MR model—see the red dashed lines in Figure 1. Different approaches to detect pleiotropic SNPs exist in the literature [24, 25, 26]. They are mainly based on

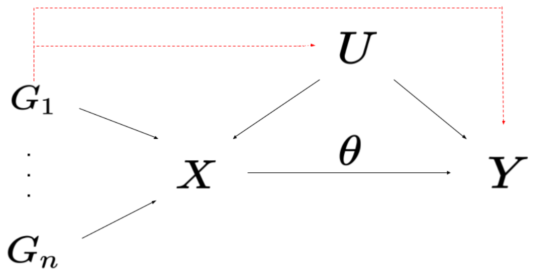

Figure 1: Causal directed acyclic graph (DAG) for univariable MR, based on Mandl et al. [27]. Genetic variants are denoted as G_i for $i \in 1, ..., n$, confounders as U, and the causal effect of the risk factor X on the outcome Y as θ. Red dashed lines represent the effect of the instrumental variable(s) G_i on the outcome through paths other than the risk factor—e.g., caused by pleiotropy.

adjusted versions of Cochran's Q.

In this section, we revisit parts of the results of our recent study comparing methods for the detection of pleiotropic SNPs [27] to illustrate the "storytelling fallacy". Beyond extensive statistical simulations, this comparison study includes a real data application assessing the effect of circulating vitamin D levels (exposure: X) on multiple sclerosis (MS, outcome: Y) through MR; see the original paper for details on data and methods [27]. Table 1 shows the results of two different pleiotropy detection methods: Method A (Standard) and method B (GC-Q). Method A detects one pleiotropic SNP, namely *rs4944958* while method B does not detect any pleiotropic SNPs.[1]

Method	pleiotropic SNP
Method A (Standard)	rs4944958
Method B (GC-Q)	—

Table 1: Real data analysis from [27]: pleiotropic SNPs for the analysis of vitamin D on MS in the univariable MR analysis.

Table 2 shows two different interpretations ("stories") that could be used to argue in favor of the superiority of the methods A and B, respectively, using prior biological domain knowledge. This clearly illustrates that the interpretation of real data results is uncertain or, in other words, affected by RDF, and should be considered with caution.

4 Use Case II: Breath sample analysis by semi-conductor based E-Nose technology

In order to distinguish SARS-CoV-2 infected from non-infected patients, an experimental analytical approach (E-Nose technology) can be applied, where volatile organic molecules, also termed gasotransmit-

[1] Details on code and data for the analysis can be found on GitHub: `https://github.com/mmax-code/MR_outliers`

Story 1 Method A (Standard)	**Story 2** Method B (GC-Q)
SNPs associated with vitamin D are known to belong to three tiers, the direct vitamin D pathway, U/V absorption, and the cholesterol metabolism [28]. The pleiotropic SNP identified by method A, *rs4944958*, is an intron of the *NADSYN1* gene which affects a precursor of cholesterol and is thus involved in cholesterol metabolism—which can be seen as confounder for MS [29]. Therefore, it may indicate horizontal pleiotropy. These arguments are in line with the result of method A.	The SNP identified by method A, *rs4944958*, is considered a perfect proxy for *rs12785878* [30]—a SNP that has been linked directly to vitamin D serum concentrations by several studies. Therefore, *rs4944958* has been explicitly included in similar studies, see, e.g, [30]. Even though *rs4944958* is involved in the cholesterol metabolism, the effect through this causal path on MS is not finally clarified [31]. *rs4944958* should be included in the MR analysis and is a false-positive finding of method A. These arguments are in line with the result of method B.

Table 2: Two different domain-specific "stories" in favor of two different methods for the MR use case.

ters, contained in human exhaled air are enriched and analysed by 10 different metal oxide semiconductor sensors [32]. Subsequently, these variations produce signal patterns that can be analysed using machine learning (ML) methods to detect SARS-CoV-2 infections. See the original paper for more details [32].[2]

Two prediction models with SARS-Cov-2 status as dependent variable and different sensor covariates are obtained using a Random Forest (RF: method A) and a Support Vector Machine (SVM: method B) [33], respectively. Note that for each sensor different covariates were extracted due to the complexity of the raw data. For the purpose of model interpretability, the covariates are then assessed using permutation importance measures. Table 3 shows the three top-ranked sensor-covariates according to the RF- and SVM-based variable importances, respectively. Even though sensor 9 is in the top three ranks for both methods, the other sensors differ.

Ranking	Method A (RF)	Method B (SVM)
1.	Sensor 9, covariate$_{9\text{-}3}$	Sensor 2, covariate$_{2\text{-}7}$
2.	Sensor 9, covariate$_{9\text{-}8}$	Sensor 10, covariate$_{10\text{-}1}$
3.	Sensor 9, covariate$_{9\text{-}2}$	Sensor 9, covariate$_{9\text{-}1}$

Table 3: Top three ranked sensor-covariates according to the permutation importance measure for methods A and B. Note that due to the complexity of the raw sensor data different covariates were extracted for each sensor. The first index of the covariate corresponds to the sensor, while the second stands for the specific covariate extracted from this sensor.

As outlined in Table 4, it is possible to interpret the results in such a way that method A appears to yield more plausible results. But it is also possible to make method B appear superior. As in Section 3 we can thus again use domain knowledge to favor one or the other method.

This example shows that despite identifying different sensors, both variable importance measures may have revealed valuable sensors for the detection of pathological processes during infections caused by SARS-CoV-2. From a medical point of view, both approaches yield highly interesting results, and especially in the medical setting with high inter-individual variations, several strategies should be considered and potentially combined, rather than focusing on a single analytical strategy. The results of the two

[2] Details on the code of the original analysis can be found on GitHub: `https://github.com/mmax-code/enose`

Story 1 Method A (RF)	**Story 2** Method B (SVM)
Method A only outputs sensor 9 in the first three ranks, which detects aromatic and sulphor organic compounds [32]. It is a highly sensitive broad range sensor, that cannot identify a single marker molecule alone. Sensor 9 is especially sensitive for sulphor compounds, with H_2S as its calibration gas. Like hydrogen and methane, H_2S represents a well-described gasotransmitter in lung disease [34], and endogenous H_2S production in humans may be increased to counteract viral infection and inflammation [35]. These arguments are in line with the ranking output by method A.	Method B identified sensor 2, which is a very sensitive broad range sensor, and sensor 10 which is selective for methane. Sensors identified by method B hint at methane and hydrogen as potential biomarkers of an underlying SARS-CoV-2 infection. Methane producing microbes can generate methane (CH_4) from carbon dioxide (CO_2) and hydrogen (H or H_2), often encountered in anoxic environments. Thus, elevated levels of hydrogen may facilitate increased production of methane, and both hydrogen and methane have been described as so-called gasotransmitters—small gas molecules that are endogenously generated, have well-defined functions, and play a role in respiratory diseases [34]. These arguments are in line with the ranking output by method B.

Table 4: Two different domain-specific “stories” in favor of two different methods for the E-Nose use case.

methods are different, but not necessarily incoherent. The fallacy lies solely in the *selective* reporting of arguments in favor of one of the methods.

5 Discussion

This manuscript discusses a new type of RDF that has not been described before. It was already well-known that methodological comparison studies can be biased in favor of the authors’ preferred method(s) through the selective reporting of, e.g., specific datasets or performance measures [9, 12]. We argue that the selective reporting of “stories” based on domain knowledge that make the results of the preferred method(s) seem more reliable than those of other methods also contributes to present a biased picture of the methods’ qualities.

This raises the question on the role of real data applications in methodological research in general and as piece of evidence on the behavior of methods in particular. We do not claim that real data applications are meaningless. In fact, they have a pivotal role in the various “phases of methodological research” [36]. In an early phase study presenting a new method idea, real data applications may be used to demonstrate that the method can be applied to real data and which type of results it yields. In a late phase study whose goal is to generate reliable evidence of the behavior of a method in various contexts, real data applications may be used to discuss special cases in which the method shows a particular behavior. In this context, it may also make sense to interpret the results based on multiple real data examples along with the results of simulation studies—where the ground truth is known.

If the biological plausibility of the results is considered a crucial criterion for the evaluation of methods, it is also conceivable, although not common, to define objective criteria for plausibility referring to literature or, say, biological databases, and to evaluate the plausibility of the methods’ results for several real datasets. Such an evaluation would not be without practical and conceptual difficulties, but would in principle address the flaw of the interpretations of real data examples discussed in this paper in two ways. Firstly, the evaluation would base on several datasets rather than on a single anecdotal dataset. Secondly, by defining objective criteria for plausibility one would reduce the RDF affecting interpretations.

Increasing awareness for the "storytelling fallacy" is especially important since the emergence of large language models (LLMs) in recent years. This stems from the fact that the flexibility researchers have when telling "stories" increases with the rise of user-friendly LLMs, which make it easier to generate literature-based consistent and plausible "stories" supporting the results of statistical methods—without even having to seek expert advice.

Based on the considerations outlined in this paper and previous literature on the design of methodological comparison studies, we formulate the following tentative recommendations. Real data applications are important and should remain an important part of methodological evaluations for illustrative purposes. However, anecdotal stories based on domain knowledge supporting the results of methods should not be considered as reliable evidence in favor of one or the other method—with only few exceptions involving several datasets and objective criteria. More generally, it is recommended to interpret real data applications in combination with those of simulation studies, and to abandon the "one beats them all philosophy" [37]. It should be acknowledged that no method is expected to yield uniformly "better results" in all situations. Relaxing the implicit expectation (of, e.g., editors and reviewers) that new methods should work clearly better than existing ones in all respects would certainly have a positive impact in terms of the incentive structure towards less biased interpretations.

Acknowledgments

This study was partly funded by DFG grants BO3139/7-2 and BO3139/9-1 to Anne-Laure Boulesteix.

Conflict of interest

The authors have declared no conflict of interest.

Data Availability Statement

The data that support the findings of this study are not publicly available due to privacy or ethical restrictions.

References

[1] Joseph P Simmons, Leif D Nelson, and Uri Simonsohn. "False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant". In: *Psychological Science* 22.11 (2011), pp. 1359–1366. DOI: 10.1177/0956797611417632.

[2] Monya Baker. "1,500 scientists lift the lid on reproducibility". In: *Nature* 533 (2016), pp. 452–454. URL: https://doi.org/10.1038/533452a.

[3] Andrew Gelman and Eric Loken. "The statistical crisis in science". In: *American Scientist* 102.6 (2014), pp. 460–465.

[4] Eric Loken and Andrew Gelman. "Measurement error and the replication crisis". In: *Science* 355.6325 (2017), pp. 584–585. URL: https://www.science.org/doi/abs/10.1126/science.aal3618.

[5] Sabine Hoffmann, Felix Schönbrodt, Ralf Elsas, Rory Wilson, Ulrich Strasser, and Anne-Laure Boulesteix. "The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines". In: *Royal Society Open Science* 8.4 (2021), p. 201925. URL: `https://doi.org/10.1098/rsos.201925`.

[6] John PA Ioannidis. "Why most published research findings are false". In: *PLoS Medicine* 2.8 (2005), e124. URL: `https://doi.org/10.1371/journal.pmed.0020124`.

[7] Monika Jelizarow, Vincent Guillemot, Arthur Tenenhaus, Korbinian Strimmer, and Anne-Laure Boulesteix. "Over-optimism in bioinformatics: an illustration". In: *Bioinformatics* 26.16 (2010), pp. 1990–1998. URL: `https://doi.org/10.1093/bioinformatics/btq323`.

[8] Theresa Ullmann, Anna Beer, Maximilian Hünemörder, Thomas Seidl, and Anne-Laure Boulesteix. "Over-optimistic evaluation and reporting of novel cluster algorithms: An illustrative study". In: *Advances in Data Analysis and Classification* 17.1 (2023), pp. 211–238. URL: `https://doi.org/10.1007/s11634-022-00496-5`.

[9] Samuel Pawel, Lucas Kook, and Kelly Reeve. "Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method". In: *Biometrical Journal* 66.1 (2024), p. 2200091. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202200091`.

[10] Regina Nuzzo. "Fooling ourselves". In: *Nature* 526 (2015), pp. 182–185. URL: `https://doi.org/10.1038/526182a`.

[11] Anne-Laure Boulesteix, Veronika Stierle, and Alexander Hapfelmeier. "Publication Bias in Methodological Computational Research". In: *Cancer Informatics* 14s5 (2015). URL: `https://doi.org/10.4137/CIN.S30747`.

[12] Christina Nießl, Moritz Herrmann, Chiara Wiedemann, Giuseppe Casalicchio, and Anne-Laure Boulesteix. "Over-optimism in benchmark studies and the multiplicity of design and analysis options when interpreting their results". In: *WIREs Data Mining and Knowledge Discovery* 12.2 (2022), e1441. URL: `https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1441`.

[13] Christina Nießl, Sabine Hoffmann, Theresa Ullmann, and Anne-Laure Boulesteix. "Explaining the optimistic performance evaluation of newly proposed methods: A cross-design validation experiment". In: *Biometrical Journal* 66.1 (2024), p. 2200238. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202200238`.

[14] Anne-Laure Boulesteix, Sabine Lauer, and Manuel JA Eugster. "A plea for neutral comparison studies in computational sciences". In: *PLoS ONE* 8.4 (2013), e61562. URL: `https://doi.org/10.1371/journal.pone.0061562`.

[15] Stefan Buchka, Alexander Hapfelmeier, Paul P Gardner, Rory Wilson, and Anne-Laure Boulesteix. "On the optimistic performance evaluation of newly introduced bioinformatic methods". In: *Genome Biology* 22.152 (2021). URL: `https://doi.org/10.1186/s13059-021-02365-4`.

[16] Itiel E. Dror. "Cognitive and Human Factors in Expert Decision Making: Six Fallacies and the Eight Sources of Bias". In: *Analytical Chemistry* 92.12 (2020), pp. 7998–8004. URL: `https://doi.org/10.1021/acs.analchem.0c00704`.

[17] Anthony O'Hagan. "Expert Knowledge Elicitation: Subjective but Scientific". In: *The American Statistician* 73.sup1 (2019), pp. 69–81. URL: `https://doi.org/10.1080/00031305.2018.1518265`.

[18] Ted J Kaptchuk. “Effect of interpretive bias on research evidence”. In: *BMJ* 326.7404 (2003), pp. 1453–1455. URL: `https://www.bmj.com/content/326/7404/1453`.

[19] Margit E. Oswald and Sabine Grosjean. “Confirmation Bias”. In: *Cognitive Illusions: A Handbook on Fallacies and Biases in Thinking, Judgement and Memory.* Ed. by Rüdiger F. Pohl. Hove and New York: Psychology Press, 2004, pp. 79–96.

[20] Andrew Gelman and Christian Hennig. “Beyond Subjective and Objective in Statistics”. In: *Journal of the Royal Statistical Society Series A: Statistics in Society* 180.4 (2017), pp. 967–1033. DOI: `10.1111/rssa.12276`. URL: `https://doi.org/10.1111/rssa.12276`.

[21] Naomi C Brownstein. *Perspective from the Literature on the Role of Expert Judgment in Scientific and Statistical Research and Practice.* 2018. arXiv: `1809.04721 [stat.OT]`. URL: `https://arxiv.org/abs/1809.04721`.

[22] Isabelle Boutron and Philippe Ravaud. “Misrepresentation and distortion of research in biomedical literature”. In: *Proceedings of the National Academy of Sciences* 115.11 (2018), pp. 2613–2619. URL: `https://www.pnas.org/doi/abs/10.1073/pnas.1710755115`.

[23] Stephen Burgess, Adam Butterworth, and Simon G. Thompson. “Mendelian Randomization Analysis With Multiple Genetic Variants Using Summarized Data”. In: *Genetic Epidemiology* 37.7 (2013), pp. 658–665. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.21758`.

[24] Marie Verbanck, Chia-Yen Chen, Benjamin Neale, and Ron Do. “Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases”. In: *Nature Genetics* 50.5 (2018), pp. 693–698. URL: `https://doi.org/10.1038/s41588-018-0099-7`.

[25] Eleanor Sanderson, George Davey Smith, Frank Windmeijer, and Jack Bowden. “An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings”. In: *International Journal of Epidemiology* 48.3 (2019), pp. 713–727. URL: `https://doi.org/10.1093/ije/dyy262`.

[26] Jack Bowden, George Davey Smith, and Stephen Burgess. “Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression”. In: *International Journal of Epidemiology* 44.2 (2015), pp. 512–525. URL: `https://doi.org/10.1093/ije/dyv080`.

[27] Maximilian M. Mandl, Anne-Laure Boulesteix, Stephen Burgess, and Verena Zuber. “Outlier Detection in Mendelian Randomisation”. In: *Statistics in Medicine (accepted)* (2025). URL: `https://arxiv.org/abs/2502.14716`.

[28] Despoina Manousaki et al. “Genome-wide association study for vitamin D levels reveals 69 independent loci”. In: *The American Journal of Human Genetics* 106.3 (2020), pp. 327–337. URL: `https://doi.org/10.1016/j.ajhg.2020.01.017`.

[29] N. Murali et al. “Cholesterol and neurodegeneration: longitudinal changes in serum cholesterol biomarkers are associated with new lesions and gray matter atrophy in multiple sclerosis over 5 years of follow-up”. In: *European Journal of Neurology* 27.1 (2020), 188–e4. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/ene.14055`.

[30] Lauren E Mokry et al. “Vitamin D and risk of multiple sclerosis: a Mendelian randomization study”. In: *PLoS Medicine* 12.8 (2015), e1001866. URL: `https://doi.org/10.1371/journal.pmed.1001866`.

[31] Balazs Lorincz, Elizabeth C. Jury, Michal Vrablik, Murali Ramanathan, and Tomas Uher. "The role of cholesterol metabolism in multiple sclerosis: From molecular pathophysiology to radiological and clinical disease activity". In: *Autoimmunity Reviews* 21.6 (2022), p. 103088. ISSN: 1568-9972. DOI: https://doi.org/10.1016/j.autrev.2022.103088. URL: https://www.sciencedirect.com/science/article/pii/S1568997222000581.

[32] Tobias Woehrle et al. "Point-of-care breath sample analysis by semiconductor-based E-Nose technology discriminates non-infected subjects from SARS-CoV-2 pneumonia patients: a multi-analyst experiment". In: *MedComm* 5.11 (2024), e726. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mco2.726.

[33] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[34] Simin Jiang et al. "Gasotransmitter Research Advances in Respiratory Diseases". In: *Antioxidants & Redox Signaling* 40.1-3 (2024), pp. 168–185. URL: https://doi.org/10.1089/ars.2023.0410.

[35] Valentina Citi et al. "Anti-inflammatory and antiviral roles of hydrogen sulfide: Rationale for considering H2S donors in COVID-19 therapy". In: *British Journal of Pharmacology* 177.21 (2020), pp. 4931–4941. URL: https://bpspubs.onlinelibrary.wiley.com/doi/abs/10.1111/bph.15230.

[36] Georg Heinze, Anne-Laure Boulesteix, Michael Kammer, Tim P. Morris, Ian R. White, and the Simulation Panel of the STRATOS initiative. "Phases of methodological research in biostatistics—Building the evidence base for new methods". In: *Biometrical Journal* 66.1 (2024), p. 2200222. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.202200222.

[37] Carolin Strobl and Friedrich Leisch. "Against the "one method fits all data sets" philosophy for comparison studies in methodological research". In: *Biometrical Journal* 66.1 (2024), p. 2200104. DOI: https://doi.org/10.1002/bimj.202200104.

A.4 Contribution 4

»Raising awareness of uncertain choices in empirical data analysis: A teaching concept toward replicable research practices«

Citation

Mandl, M.M., Hoffmann S., Bieringer S., Jacob A.E., Kraft M., Lemster, S., Boulesteix, A.-L. (2024) Raising awareness of uncertain choices in empirical data analysis: A teaching concept toward replicable research practices. *PLOS Computational Biology* **20**(3): e1011936. https://doi.org/10.1371/journal.pcbi.1011936

Authors' contributions

Conceptualization: MM, SH, ALB. Formal analysis: MM. Funding acquisition: ALB. Investigation: MM. Project administration: ALB. Software: MM. Supervision: ALB. Visualization: MM. Writing – original draft: MM, SH, ALB. Writing – review & editing: MM, SH, SB, AJ, MK, SL, ALB.

Rights and permissions

PLOS COMPUTATIONAL BIOLOGY

EDUCATION

Raising awareness of uncertain choices in empirical data analysis: A teaching concept toward replicable research practices

Maximilian M. Mandl[1,2,3¤]*, **Sabine Hoffmann**[1,3,4], **Sebastian Bieringer**[4], **Anna E. Jacob**[1], **Marie Kraft**[4], **Simon Lemster**[1], **Anne-Laure Boulesteix**[1,2,3]

1 Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, München, Germany, **2** Munich Center for Machine Learning (MCML), München, Germany, **3** LMU Open Science Center, München, Germany, **4** Department of Statistics, Ludwig-Maximilians-Universität München, München, Germany

¤ Current address: Institute for Medical Information Processing, Biometry and Epidemiology, Medical Faculty, Ludwig-Maximilians-Universität München, München, Germany
* mmandl@ibe.med.uni-muenchen.de

OPEN ACCESS

Citation: Mandl MM, Hoffmann S, Bieringer S, Jacob AE, Kraft M, Lemster S, et al. (2024) Raising awareness of uncertain choices in empirical data analysis: A teaching concept toward replicable research practices. PLoS Comput Biol 20(3): e1011936. https://doi.org/10.1371/journal.pcbi.1011936

Editor: B.F. Francis Ouellette, bioinformatics.ca, CANADA

Published: March 28, 2024

Funding: The authors gratefully acknowledge the funding by DFG grants BO3139/7-1 and BO3139/9-1 to A-LB. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Author summary

Throughout their education and when reading the scientific literature, students may get the impression that there is a unique and correct analysis strategy for every data analysis task and that this analysis strategy will always yield a significant and noteworthy result. This expectation conflicts with a growing realization that there is a multiplicity of possible analysis strategies in empirical research, which will lead to overoptimism and nonreplicable research findings if it is combined with result-dependent selective reporting. Here, we argue that students are often ill-equipped for real-world data analysis tasks and unprepared for the dangers of selectively reporting the most promising results. We present a seminar course intended for advanced undergraduates and beginning graduate students of data analysis fields such as statistics, data science, or bioinformatics that aims to increase the awareness of uncertain choices in the analysis of empirical data and present ways to deal with these choices through theoretical modules and practical hands-on sessions.

Introduction

Statistics and data analysis education frequently focuses on acquiring skills and techniques concerning specific topics that are covered successively and in isolation. Students may, for instance, first take a course on general techniques for regression modeling without considering the challenges associated with missing data, outliers, or nonrepresentative sampling mechanisms. They may then acquire skills to specifically address these additional challenges in a later course. In the classroom, students are often presented with clear examples and with clean data sets to practice these skills and techniques on. These exercises typically have unique, correct solutions to the analysis task and often yield significant results, possibly conditioning students to expect the same from real-world data. In this vein, problems arising during the analysis are considered in isolation, even though they occur simultaneously and may be interrelated. While the simplified and sequential treatment of specific topics certainly makes sense from a pedagogical standpoint, it may convey the unrealistic expectation that for any data analysis task, there is a unique and correct analysis approach that will always yield a significant or interesting finding. This expectation is further strengthened when reading published research articles in which the authors commonly describe a single analysis strategy and report a significant finding without a detailed discussion of alternative analysis options.

This impression conflicts with a growing realization that there is a multiplicity of possible analysis strategies when analyzing empirical data [1–3] and that data analysts require the ability to make subjective decisions and acknowledge the multiplicity of possible perspectives [4]. In particular, so-called multianalyst projects [5–7] show that different teams of researchers make very different choices when they are asked to answer the same research question on the same data set. These uncertain choices, which are also referred to as researcher degrees of freedom [8,9], can be combined with result-dependent selective reporting to obtain the "most noteworthy" or impressive results. This is a practice known as "p-hacking" or "fishing for significance" in the context of hypothesis testing and, more generally, "fishing expeditions" or "cherry-picking." These practices lead to overconfident and nonreplicable research findings in the literature and, ultimately, to situations where some may argue that "most published research findings are false," especially in combination with a low prior probability of the hypothesis being true [10,11]. Computational biology as a field is, unfortunately, not immune to these types of problems [3,12].

For example, Ullmann et al. [3] show how the combination of researchers' expectations and selective reporting may lead to overoptimistic results in the context of unsupervised microbiome analysis. Their paper highlights the relevance of open science practices in the field of computational biology.

Here, we argue that if students always encounter clean data sets with a correct unique analysis strategy yielding a significant and/or noteworthy finding during their training, they are ill-equipped for real-world data analysis tasks and unprepared for the dangers of selectively reporting the most promising results. In particular, data analysis courses commonly teach students to understand and apply statistical models, but in order to equip them against the cherry-picking, we need to strengthen awareness and understanding of uncertainties in the analysis of empirical research data. To address this point, we present a seminar course intended for advanced undergraduates and beginning graduate students of data analysis fields such as statistics, data science, or bioinformatics that aims to increase awareness of the multiplicity of analysis strategies and of ways to deal with this multiplicity through the introduction of theoretical concepts and practical hands-on sessions.

The remainder of the article is organized as follows: Section "Teaching concept" presents the general teaching concept of the proposed seminar course. Section "Implementation and student feedback" provides evidence on the instructional value of the proposed course. Section "Potential adaptations" discusses potential adaptations of the course, and in Section "Conclusion," we highlight key skills and takeaways that we hope students will gain.

Teaching concept

Overview

The course consists of theoretical modules and practical hands-on sessions. It starts with two short lectures, providing a brief introduction to the concepts of reproducibility and replicability. Subsequently, it focuses on reproducibility by introducing the students to version control software and R-Markdown to make their analyses reproducible, i.e., they learn to prepare their code in a way that all results can be reproduced "by mouse click." In this paper, we follow the definition by Nosek et al. [13], i.e., reproducibility involves verifying the reliability of a previous discovery by employing the identical data and analysis strategy.

The second part of the course is devoted to replicability in a broad sense, where a result is said to be replicable if one obtains a similar result when repeating the same study including the collection of independent data. More specifically, the students participate in a hands-on session, in which each student is asked to perform a regression analysis on the same data set. After this first hands-on session, they are presented with a second theoretical module that focuses on uncertain choices in the analysis of empirical data, the consequences of result-dependent selective reporting, and ways to address these issues. While the hands-on session can be seen as an evaluation of the extent of selective reporting in the classroom, this second theoretical module can be seen as an intervention. It aims to prevent the students from selectively reporting the most promosing results arising through the multiplicity of possible analysis strategies. The effect of this intervention can, to some extent, be measured by comparing the results of the first phase of the hands-on session with a second phase, which follows the theoretical module on researcher degrees of freedom, in which the students are again asked to analyze a data set that has been generated according to the same model and parameter values as the data set in phase 1. The students' experience with the two hands-on sessions, the results concerning this intervention effect, and their takeaways are discussed in the last two sessions of the course. A sample weekly schedule for a 10-week academic term is shown in Table 1. Note that the course might alternatively be conducted as an intensive course in one or few days as discussed in the section on potential adaptations.

Practical hands-on sessions

In the two hands-on sessions, which should ideally take around 3 hours and be onsite to guarantee that there is no exchange between the students, each student receives the same simulated data set and is asked to estimate the effect of a predictor of interest in a linear regression model and to provide a point estimate and a 95% confidence interval. See Section C "Instructions for the students" in S1 Appendix for more details on the exact instructions received by the students.

The analysis task is designed in such a way that several uncertain choices related to model selection, treatment of missing values, and handling of outliers are required. Although we realize that such questions should ideally be tackled at the design stage of a study, in practice many researchers unfortunately address these difficulties post hoc.

To help the students with these choices, they are provided with literature that gives an overview of methods and guidance on these choices (see, for instance, [14,15]) and they are able to ask the lecturer for advice during the entire session. Additionally, the students are given information on the "likely range" of the effect of interest, while the true effect is somewhat below this range. The goal is to mimic a realistic data analysis situation in which the life scientist may

Table 1. Sample weekly schedule for a 10-week academic term.

Topic	Details
Week 1–2: Reproducibility	Introduction to version control software (Git, GitLab) and R-Markdown.
Week 3: Phase 1 hands-on session	First assignment: 3-hour onsite task.
Week 4–7: Introduction of theoretical concepts	Lectures on uncertain choices in the analysis of empirical data, consequences of result-dependent reporting of analysis strategies, and ways to address these issues.
Week 8: Phase 2 hands-on session	Second assignment: 3-hour onsite task.
Week 9–10: Debriefing:	Review and discussion of results and the data generation process of the simulation setup.

https://doi.org/10.1371/journal.pcbi.1011936.t001

hope for a large effect and exert gentle pressure on the data analyst toward observing it in the data. For each of the hands-on sessions, students are asked to analyze the data in the best possible way (which is not necessarily the same for both phases) and to hand in their results and reproducible analysis code.

Theoretical module on uncertain choices in the analysis of empirical data and ways to address them

The theoretical module consists of lectures that address the ubiquity of uncertain choices in the analysis of empirical data, their consequences on the validity of statistical inference if they are combined with selective reporting, and solutions to address this issue. In particular, the lectures detail how result-dependent selective reporting (cherry-picking, HARKing [16], and selective publication of significant findings) can lead to overoptimism. Further, they outline that there is increasing evidence that this practice is both common and detrimental for the replicability and credibility of the scientific literature.

Finally, as an outlook, the theoretical module also presents general strategies to deal with the multiplicity of possible analysis strategies while preserving the validity of statistical inference. This can include preregistration, blind analysis, and multiverse-style analyses. A list of articles that can be used to design this theoretical module can be found in Section A "Details on the implementation" in S1 Appendix.

Debriefing

The last two sessions leave space for the discussion of the results of the two hands-on sessions, of the students' experience with the course, and of student takeaways. In the first session, the students are presented with the results of the first hands-on session in which they analyzed the same data set. Due to the uncertain choices in the analysis of this data set, it is likely that the students chose a variety of analysis strategies and obtained different results, providing them with first-hand experience that there is not a single correct analysis strategy for every data analysis task. These results are then compared with the true parameter value that was used to generate the data, providing insight to the extent of selective reporting that was performed during the analysis. Instructors may stress that true parameter values are not known in real data analysis and point out the principles of statistical simulations and their importance for data analysis methods by mimicking real-world scenarios with known truth.

In the second debriefing session, the results of the two hands-on sessions are compared to assess the intervention effect of the theoretical module on uncertain choices in the analysis of empirical data. As seminar courses tend to be small (with less than 30 students) and some students might lack motivation or skills to either perform multiple analyses (and selective reporting) in the first hands-on session or to change their analysis strategy in the second hands-on session, it is unlikely that a statistically significant intervention effect would be observed. Such a nonsignificant finding opens the discussion to reasons for this "failed experiment," including lack of power, imperfect adherence and, more generally, that this nonsignificant finding cannot be interpreted as evidence that the intervention is useless since "absence of evidence is not evidence of absence" [17] and that practical importance and significance are distinct concepts [18]. After discussing the realities of experimental design, the lecturer can present the students with alternative possible results on this intervention effect resulting, for instance, from more or less plausible inclusion and exclusion criteria or outcome switching that would lead to a statistically significant intervention effect. This could raise student awareness of their own preconceived expectations that it is only a matter of finding the right analysis strategy to produce an intended result. This is a common fallacy that can arise, especially in the analysis of underpowered studies.

Implementation and student feedback

We implemented a version of the course concept described in Section “Teaching concept” as a seminar course for advanced undergraduate students in statistics at Ludwig-Maximilians-Universität München (Germany) in 2021/2022.

The overall feedback from the students was very positive and indicated that the course had the intended effect of raising awareness of uncertain choices in the analysis of empirical data and of the dangers of result-dependent selective reporting.

The following 2 student statements, which we received after asking the students for more detailed feedback, further support this conclusion:

> “I think that the learning effect of the seminar was greater than in a classical seminar, which consists exclusively of frontal teaching and presentations. [. . .] This also made me aware of how difficult it is to make statistical decisions on the basis of the available information.”

> “The seminar was very practical compared to other seminars, which made itself and the experience unique. This seminar and the experiment have had a sustainable effect on the way I do statistics. For example, it is okay to get an inconclusive result when analysing data, not everything has to be significant.”

Fig 1 shows the difference between the estimated and true effects (represented as relative under- or overestimation) in phases 1 and 2 for the full sample (n = 26) and 3 further selected subsamples. In phase 1, the students reported a parameter estimate that was on average 17.55% larger than the true parameter value (one-sided t test: p = 0.03; Wilcoxon: p = 0.04), indicating that our instructions may indeed have incited the students to selectively report promising results.

In phase 2, the reported effect was on average 11.67% larger than the true effect (one-sided t test: p = 0.18; Wilcoxon: p = 0.05), providing less evidence for result-dependent selective reporting after the theoretical module on uncertain choices and their consequences for the validity of statistical inference. Even if there was a significant overestimation of the effect in phase 1 (17.55%) but not in phase 2 (11.67%), the 2 phases did not significantly differ with respect to this difference (paired one-sided t test: p = 0.35; Wilcoxon: p = 0.40), a result that may appear counterintuitive to students and is certainly worth pointing out.

An aspect worth being discussed with the students is shown in Fig 1. The intervention effect becomes significant (or very close to the 5% level) if we (slightly) change our analysis strategy, for instance, by performing the analysis only on students who overestimated the effect in phase 1 (Fig 1(c): n = 18, paired one-sided t test: p = 0.04; Wilcoxon: p = 0.06) or only on female students (Fig 1(d): n = 11, paired one-sided t test: p = 0.06; Wilcoxon: p = 0.09), leaving room for the selective reporting of promising intervention effects in this highly underpowered experiment. Conversely, the p-value of the intervention effect can also increase if we include only the students who performed well in terms of grades in the course (Fig 1(b): n = 15, paired one-sided t test: p = 0.57; Wilcoxon: p = 0.68).

For more details, see Sections A “Details on the implementation,” B “Data simulation,” and C “Instructions for the students” in S1 Appendix. The code and data can be found on GitHub (https://github.com/mmax-code/teaching_concept).

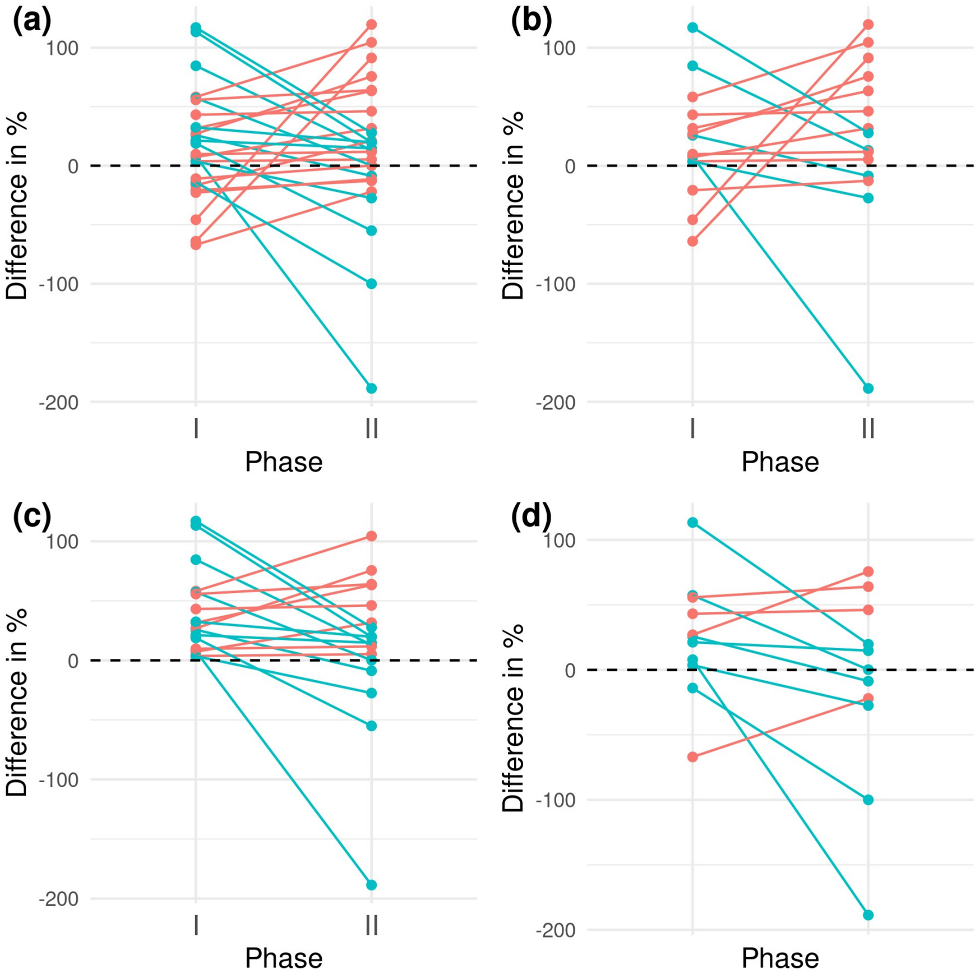

Fig 1. Difference in % between the estimated and true effects (represented as relative under- or overestimation) in phases 1 and 2. (a) Full sample (n = 26); (b) students with higher grades (n = 15); (c) students who overestimated the true effect in phase 1 (n = 18); (d) female students (n = 11). Connected points represent the values for phases 1 and 2 for each student. Red lines indicate an increased estimated effect size in phase 2 compared to phase 1, and blue lines indicate the reverse.

https://doi.org/10.1371/journal.pcbi.1011936.g001

Potential adaptations

Since the multiplicity of possible analysis strategies and result-dependent selective reporting are complex issues with many different aspects, there are several potential adaptations that can be made to tailor the course to varying preferences and needs.

In our implementation of the course, we chose to have the students work on simulated data sets, but it is of course possible to choose real data sets for the hands-on-sessions. To decide between these two options, it is important to decide whether one merely intends to raise awareness for the multiplicity of possible analysis strategies or to caution students against the dangers of result-dependent selective reporting. More generally, questionable research practices that may result from this multiplicity of possible analysis strategies include HARKing, fishing for significance, and data dredging. In the case where the aim is to caution against result-dependent selective reporting, it is indispensable to use simulated data sets in the hands-on session to be able to show how these practices lead to an overestimation of the true parameter value (which would be impossible on a real data set since the true parameter value is unknown). If, on the other hand, the course only focuses on raising awareness of uncertain choices and the multiplicity of possible analysis strategies, it seems more advisable to use real data sets with all their "ugly" features including, for instance, complex patterns of missing data and outliers since they offer a more realistic framework to achieve this teaching purpose, in the vein of the multiverse analysis in the classroom suggested by Heyman and Vanpaemel [19].

A second important decision in the teaching concept concerns the question of whether to focus on long-term strategies to address the multiplicity of possible analysis strategies or to present students with short-term solutions whose effects will be more observable when comparing the results from the first and the second phase of the hands-on session. The course concept that we presented here was designed to be instructive in the long term (such an effect being impossible to demonstrate in the course setting) rather than to show a large intervention effect. In this sense, the strategies that we presented to prevent result-dependent selective reporting included preregistration, blind analysis, and multiverse analyses. While these strategies are indubitably very helpful for students to address the multiplicity of possible analysis strategies in future projects, they may be of rather limited value in the second hands-on session of the course.

Related to this latter point, we chose the timing of the course to be rather early in the students' curriculum to inoculate them against result-dependent selective reporting among a multiplicity of possible analysis strategies. This is hopefully before they were even aware of the wealth of methods and modeling strategies that they could choose from. While we believe that this may very well increase the long-term effectiveness of the teaching intervention, it will inevitably reduce the size of the intervention effect that we can observe when comparing the first and the second phase of the hands-on session because this lack of awareness reduces the number of analysis strategies that the students can choose from. In contrast, one could choose a later timing of the course in the students' curriculum or provide the students with abundant literature on various methods and include additional lectures on methods (for instance, on model selection or missing values) in the course. In our implementation of the course of limited volume, we deliberately decided not to handle methodological issues beyond a brief introduction, in order to focus on reproducibility and replicability. The fact that students used (mostly the same) rather simple methods (for instance, AIC-based model selection) in the implementation suggests that they were probably not aware of the many possibilities they had —which may de facto prevent them from fishing for significance. Presenting the students with a multiplicity of methods before or during the hands-on sessions, on the other hand, might increase their fishing behavior, at least in the first hands-on session. Finally, we did not explicitly ask the students to change their analysis strategies, which may have led students with limited motivation to keep the same analysis strategy for both phases.

This focus on the long-term effectiveness of the course rather than on short-term strategies that may be perceivable in the comparison of the first and the second hands-on session might very well explain why we did not observe a significant reduction in result-dependent selective reporting between the two phases. However, as pointed out in Section "Debriefing," we would consider this nonsignificant result less of a bug and more of a feature since it opens the discussion to topics including lack of power, imperfect adherence and, more generally, reminds the students that a nonsignificant finding cannot be interpreted as evidence that an intervention did not work.

On a completely different level, the course could be adapted to other types of data analyses in a broad sense beyond the generic example of effect estimation with regression models considered here. Selective reporting is relevant and may be considered in various contexts such as supervised learning [20], cluster and network analysis [3], or gene set analysis [21] rather than statistical testing in regression models. Examples inspired from these studies may be appropriate for students majoring in fields related to computational biology. Note that even though a prerequisite for our course is the use of an interpreted programming language such as R or Python and at least basic knowledge of regression models, the general concept of the course can, in principle, also be applied to students with a weaker computational background. For example, one could implement the course with a simple hypothesis test setting using a statistical software framework including a user interface (for instance, SPSS).

Finally, depending on the complexity of the considered analyses and the amount of effort required from students to understand and execute the analyses, the course concept could also be adapted to a one or multiple day intensive course. With such a shorter format, the complexity of the hands-on task and the width of the covered theoretical topics (see section A in S1 Appendix) should probably be reduced compared to our original version of the course. For example, one could address primarily the multiplicity of analysis strategies and put less focus on specific software aspects (such as the use of R-Markdown).

Conclusion

There has been growing evidence in recent years that the current use (and misuse) of data analysis methods has contributed to what has been referred to as a "replication crisis" or "statistical crisis" in science. We argue that we need to address these problems in the way we teach statistics and data analysis [22]. In particular, we need to raise awareness regarding the potential dangers of selective reporting in the education of computational scientists. With the concept of the presented course, we address this issue through practical hands-on sessions and theoretical modules. Going beyond selective reporting, the course also provides the opportunity to teach students reproducible research practices [23] and to discuss important issues in the design and analysis of experimental studies, including lack of statistical power, nonadherence, and the common misinterpretation of absence of evidence as evidence of absence.

While the combination of a multiplicity of possible analysis strategies with selective reporting is an important issue today, it is likely to pose even more challenges in the future with the increasing availability of large complex data sets. In the analysis of these data sets, researchers are faced with even more uncertain choices than in data that are collected within simple focused experiments, as there is far less knowledge of the data generating mechanisms and control over measurement procedures. To avoid what Meng [24] calls "Big data paradoxes" in the analysis of these data sets ("the more the data, the surer we fool ourselves"), we urgently need to prepare our students for the realities of empirical data analysis by fostering their awareness and understanding of uncertain choices and ways to address these choices that preserve the validity of statistical inference.

Supporting information

S1 Appendix. Details on the implementation, data simulation, and instructions for the students.
(PDF)

Acknowledgments

We thank Roman Hornung, Christina Sauer, and Theresa Ullmann for their comments and helpful discussions and Savanna Ratky for valuable language corrections.

Author Contributions

Conceptualization: Maximilian M. Mandl, Sabine Hoffmann, Anne-Laure Boulesteix.

Formal analysis: Maximilian M. Mandl.

Funding acquisition: Anne-Laure Boulesteix.

Investigation: Maximilian M. Mandl.

Project administration: Anne-Laure Boulesteix.

Software: Maximilian M. Mandl.

Supervision: Anne-Laure Boulesteix.

Visualization: Maximilian M. Mandl.

Writing – original draft: Maximilian M. Mandl, Sabine Hoffmann, Anne-Laure Boulesteix.

Writing – review & editing: Maximilian M. Mandl, Sabine Hoffmann, Sebastian Bieringer, Anna E. Jacob, Marie Kraft, Simon Lemster, Anne-Laure Boulesteix.

References

1. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspect Psychol Sci. 2016; 11(5):702–712. https://doi.org/10.1177/1745691616658637 PMID: 27694465
2. Hoffmann S, Schönbrodt F, Elsas R, Wilson R, Strasser U, Boulesteix A L. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. R Soc Open Sci. 2021; 8 (4):201925. https://doi.org/10.1098/rsos.201925 PMID: 33996122
3. Ullmann T, Peschel S, Finger P, Müller C L, Boulesteix A L. Over-optimism in unsupervised microbiome analysis: Insights from network learning and clustering. PLoS Comput Biol. 2023; 19(1):e1010820. https://doi.org/10.1371/journal.pcbi.1010820 PMID: 36608142
4. Gelman A, Hennig C. Beyond subjective and objective in statistics. J R Stat Soc Ser A Stat Soc. 2017:967–1033.
5. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. Adv Methods Pract Psychol Sci. 2018; 1(3):337–356.
6. Aczel B, Szaszi B, Nilsonne G, Van Den Akker O R, Albers C J, Van Assen M A, et al. Consensus-based guidance for conducting and reporting multi-analyst studies. Elife. 2021; 10:e72185. https://doi.org/10.7554/eLife.72185 PMID: 34751133
7. Wagenmakers E J, Sarafoglou A, Aczel B. One statistical analysis must not rule them all. Nature. 2022; 605(7910):423–425. https://doi.org/10.1038/d41586-022-01332-8 PMID: 35581494
8. Simmons J P, Nelson L D, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci. 2011; 22(11):1359–1366. https://doi.org/10.1177/0956797611417632 PMID: 22006061
9. Wicherts J M, Veldkamp C L, Augusteijn HE, Bakker M, van Aert R C, Van Assen M A. Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. Front Psychol. 2016; 7:1832. https://doi.org/10.3389/fpsyg.2016.01832 PMID: 27933012
10. Ioannidis J P. Why most published research findings are false. PLoS Med. 2005; 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124 PMID: 16060722
11. Leek J T, Jager L R. Is most published research really false? Annu Rev Stat Appl. 2017; 4:109–122.
12. Boulesteix A L. Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLoS Comput Biol. 2015; 11(4):e1004191. https://doi.org/10.1371/journal.pcbi.1004191 PMID: 25905639
13. Nosek B A, Hardwicke T E, Moshontz H, Allard A, Corker K S, Dreber A, et al. Replicability, robustness, and reproducibility in psychological science. Annu Rev Psychol. 2022; 73:719–748. https://doi.org/10.1146/annurev-psych-020821-114157 PMID: 34665669

14. Sauerbrei W, Perperoglou A, Schmid M, Abrahamowicz M, Becher H, Binder H, et al. State of the art in selection of variables and functional forms in multivariable analysis–outstanding issues. Diagn Progn Res. 2020; 4:3. https://doi.org/10.1186/s41512-020-00074-3 PMID: 32266321

15. Little R J, Carpenter J R, Lee K J. A comparison of three popular methods for handling missing data: complete-case analysis, inverse probability weighting, and multiple imputation. Sociol Methods Res. 2022; 0(0):00491241221113873.

16. Kerr N L. HARKing: Hypothesizing after the results are known. Pers Soc Psychol Rev. 1998; 2(3):196–217. https://doi.org/10.1207/s15327957pspr0203_4 PMID: 15647155

17. Altman D G, Bland JM. Statistics notes: Absence of evidence is not evidence of absence. Br Med J. 1995; 311(7003):485.

18. Witmer J. Editorial. J Stat Educ. 2019; 27(3):136–137.

19. Heyman T, Vanpaemel W. Multiverse analyses in the classroom. Meta-Psychology. 2022; 6: MP.2020.2718.

20. Boulesteix A L, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. BMC Med Res Methodol. 2009; 9(1):1–14.

21. Wünsch M, Sauer C, Callahan P, Hinske L C, Boulesteix A L. From RNA sequencing measurements to the final results: a practical guide to navigating the choices and uncertainties of gene set analysis. WIREs Comp Stats. 2024; 16(1):e1643.

22. Gelman A. The problems with p-values are not just with p-values. In: Wasserstein Ronald L and Lazar Nicole A, The ASA Statement on p-values: Context, Process, and Purpose. Am Stat. 2016; 70(2):129–133.

23. Sandve G K, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS Comput Biol. 2013; 9(10):e1003285. https://doi.org/10.1371/journal.pcbi.1003285 PMID: 24204232

24. Meng X L. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. Ann Appl Stat. 2018; 12(2):685–726.

S1 Appendix

A Details on the implementation

The course and its hands-on sessions that we present in this paper were conducted in the context of a mandatory seminar course for bachelor of science in statistics students at Ludwig-Maximilians-Universität München (Germany). Its experimental setting, i.e., the measurement of a potential intervention effect, was approved by the ethical committee of the faculty for mathematics, informatics and statistics at the Ludwig-Maximilians-Universität München, Germany (EK-MIS-2021-065). Furthermore, the students gave their informed consent to participate in the experiment.

Table A shows an overview of the analytical choices of the students for phase 1 and 2. Most students (phase 1 and 2: 81% and 69%, respectively) used a stepwise AIC approach for model selection. For outlier detection, the preferred method was to visually detect them via boxplots (phase 1 and 2: 50%) and subsequently drop or adjust them according to the 97.5% and the 2.5% quantiles. Missing values were either naively dropped or mean imputed in most cases (phase 1 and 2: 84% and 79%, respectively). The rest of the students either imputed the median, imputed a value using parametric assumptions, replaced the missings with zeros, or dropped the entire variable containing missing values. Only one student in each phase implemented lasso regression for model selection in addition to the stepwise AIC approach. A few students based their model selection on univariate selection via Pearson's correlation.

The analytical choices in phase 2 were overall quite similar to phase 1. Interestingly, some students applied p-splines for the potential non-linear effect or a train/test split approach, which was not necessary in this setting. Only one student discussed the problem of the multiplicity of possible analysis strategies. This student reported two results and decided against the AIC criterion and in favor of the smaller effect size.

Phase	Set of Problems	Model selection	Outliers	Missings	Others
Phase 1	Unbalanced category, missings, interaction	no selection	Boxplots	Drop	selectively check for quadratic terms, interactions
Phase 1	Unbalanced category, missings, interaction	Manually	-	Drop	-
Phase 1	Unbalanced category, missings, interaction	Manually, AIC	Z-score	Mean imputation	-
Phase 1	Unbalanced category, missings, interaction	Manually, Stepwise AIC	-	Drop	selectively check for quadratic terms, interactions
Phase 1	Unbalanced category, missings, interaction	Stepwise AIC	Boxplots	Drop	-
Phase 1	Unbalanced category, missings, interaction	Stepwise AIC	Boxplots, Cooks-distance	Drop, Mean imputation	log-transformation
Phase 1	Unbalanced category, missings, interaction	Stepwise AIC, BIC	Boxplots	Drop	selectively check for quadratic terms, interactions, p-spline
Phase 1	Interaction, outliers	Manually, Stepwise AIC, BIC	Boxplots	-	selectively check for quadratic terms, interactions
Phase 1	Interaction, outliers	Pearson correlation	Boxplots	-	ANOVA, train-test split
Phase 1	Interaction, outliers	Pearson correlation	Boxplots, Grubbs-test	-	log-transformation, ANOVA
Phase 1	Interaction, outliers	Stepwise AIC	-	-	selectively check for quadratic terms, interactions
Phase 1	Interaction, outliers	Stepwise AIC	Boxplots	-	-
Phase 1	Interaction, outliers	Stepwise AIC	Boxplots	-	selectively check for quadratic terms, interactions
Phase 1	Interaction, outliers	Stepwise AIC	Boxplots	-	selectively check for quadratic terms, interactions, ANOVA
Phase 1	Outliers, missings	Manually, AIC	Manually	Imputation by distribution	-
Phase 1	Outliers, missings	Stepwise AIC	-	Drop	ANOVA
Phase 1	Outliers, missings	Stepwise AIC	Boxplots	Median Imputation	log-transformation
Phase 1	Outliers, missings	Stepwise AIC, BIC	Boxplots	Drop	selectively check for quadratic terms, interactions, p-spline
Phase 1	Outliers, missings	Stepwise AIC, Lasso	Boxplots	Drop	selectively check for quadratic terms, interactions
Phase 1	Interaction, quadratic term, missings	Manually, AIC	-	Mean imputation, MICE	-
Phase 1	Interaction, quadratic term, missings	no selection	-	Drop	-
Phase 1	Interaction, quadratic term, missings	Stepwise AIC	-	Mean imputation	-
Phase 1	Interaction, quadratic term, missings	Stepwise AIC	-	Replaced by 0	-
Phase 1	Interaction, quadratic term, missings	Stepwise AIC	Manually	Drop	selectively check for quadratic terms, interactions
Phase 1	Interaction, quadratic term, missings	Stepwise AIC	Z-score	Mean imputation	selectively check for quadratic terms, interactions
Phase 1	Interaction, quadratic term, missings	Stepwise AIC, Pearson Correlation	-	Drop full variable	selectively check for quadratic terms, interactions
Phase 2	Unbalanced category, missings, interaction	no selection	Boxplots	Drop	selectively check for quadratic terms, interactions
Phase 2	Unbalanced category, missings, interaction	Manually	-	Drop	-
Phase 2	Unbalanced category, missings, interaction	Manually, AIC	Z-score	Mean imputation	-
Phase 2	Unbalanced category, missings, interaction	Lasso, Stepwise AIC	-	Drop	p-spline
Phase 2	Unbalanced category, missings, interaction	Stepwise AIC	Boxplots	Drop	-
Phase 2	Unbalanced category, missings, interaction	Stepwise AIC	Boxplots	Drop	log-transformation
Phase 2	Unbalanced category, missings, interaction	Stepwise AIC, BIC	Boxplots	Drop	selectively check for quadratic terms, interactions, p-spline
Phase 2	Interaction, outliers	Manually, Stepwise AIC, BIC	Boxplots	-	selectively check for quadratic terms, interactions, p-spline
Phase 2	Interaction, outliers	Pearson correlation	Boxplots	-	ANOVA, train-test split, mixed model
Phase 2	Interaction, outliers	Bayesian model averaging	Boxplots, Grubbs-test	-	Cross validation
Phase 2	Interaction, outliers	Stepwise AIC	-	-	selectively check for quadratic terms, interactions
Phase 2	Interaction, outliers	Stepwise AIC	Boxplots	-	p-splines
Phase 2	Interaction, outliers	Stepwise AIC	Boxplots	-	selectively check for quadratic terms, interactions
Phase 2	Interaction, outliers	Stepwise AIC	Boxplots	-	selectively check for quadratic terms, interactions, ANOVA
Phase 2	Outliers, missings	Manually, AIC	Manually	Imputation by distribution	-
Phase 2	Outliers, missings	Stepwise AIC	-	Drop	ANOVA
Phase 2	Outliers, missings	Stepwise AIC	Boxplots	Median Imputation	log-transformation
Phase 2	Outliers, missings	Stepwise AIC, BIC	Boxplots	Drop	selectively check for quadratic terms, interactions, p-spline
Phase 2	Outliers, missings	Stepwise AIC	Boxplots	Drop	-
Phase 2	Interaction, quadratic term, missings	Manually, AIC	-	Mean imputation, MICE	-
Phase 2	Interaction, quadratic term, missings	no selection	-	Drop	-
Phase 2	Interaction, quadratic term, missings	Stepwise AIC	-	Median Imputation	-
Phase 2	Interaction, quadratic term, missings	Stepwise AIC	-	Replaced by 0	-
Phase 2	Interaction, quadratic term, missings	Stepwise AIC	Manually	Drop	selectively check for quadratic terms, interactions
Phase 2	Interaction, quadratic term, missings	Hypotheses testing and p-splines	Z-score	Mean imputation	-
Phase 2	Interaction, quadratic term, missings	Stepwise AIC, Pearson Correlation	-	Drop full variable	selectively check for quadratic terms, interactions

Table A. Analytical decisions in phase 1 and 2. Empty cells (-) represent no actions taken by the students. Unbalanced category means that students were left with only few observations in some categories of the categorical variable and thus needed to recode the variable.

Table B gives an overview of articles that can be used to design the theoretical module of the course. The selected topics include, among others, the multiplicity of analysis strategies, different sources of uncertainty in the analysis of empirical data, researcher degrees of freedom, p-hacking, and HARKing.

Topic	Details	Literature
Multiplicity	Introduce the multiplicity of analysis strategies and illustrate sources of uncertainties, namely measurement, data pre-processing, parameter, model, and method uncertainty	[1–5]
Researcher degrees of freedom	Introduce the topic and show how these free analytical choices may lead to an inflated type I error rate	[6]
P-Hacking	Introduce and define p-hacking and show how p-hacking leads to a change of the distribution within the area of significance	[7, 8]
Strategies against p-hacking and coping with sources of uncertainty	Adjusting for multiple comparisons and create a statistical analysis plan before the analysis or data collection takes place. Present the notion of confirmatory study and pre-specified/registered data analysis protocol. Reduce uncertainty (e.g., increase sample size), report uncertainty (e.g., vibration of effects framework), accept uncertainty (e.g. replication studies), and integrate uncertainty (e.g. Bayesian framework)	[2, 9–18]
Other literature notices and topics	Vibration of effects framework; Measuring sampling, model and measurement uncertainty; multiverse analysis; HARKing; p-values; publication bias; replication	[19–28]

Table B. Articles that can be used to design the theoretical module.

B Data simulation

All data sets were simulated from a similar data generating process (DGP) with slightly different parameter values. The data included different methodological difficulties that can be addressed in different ways, yielding so-called researcher degrees of freedom. These difficulties were: interaction effects, non-linear effects, missing values, outliers, and unbalanced classes in the categorical variables. See Table A for a detailed description. Both the effect and the methodological difficulties were the same for each student in phase 1 and 2.

For the assignment in phase 1 and 2, the students received instructions on the data and the (fictive) problem at hand (see Section C Instructions for the students).

The code for the simulation of the data sets can be found on Github (`https://github.com/mmax-code/teaching_concept`). The simulation was straightforward; the covariates were drawn from a (multivariate) normal, cauchy, uniform, t-, log-normal, beta, multinomial, and binomial distribution and the response variable was built as a linear combination of some of these covariates. The covariates that did not have an effect were also included in the data set to make the model selection more complex.

We randomly allocated each of the $n = 26$ students to one of four groups, whereby the characteristics of the data sets were the same within each group; see the code for more details. The group structure was implemented to avoid collaboration between the students. In an idealized version of the course, the students should not work on the exercises remotely, i.e., the allocation to groups can be avoided.

To analyze the potential for cherry-picking present in the data sets, we analyzed the different simulated data sets within the vibration of effects framework introduced by [19]. As an example, Fig A and B show a vibration of effects plot for a representative data set, i.e., a plot that represents the -log10 p-value against the effect estimate of interest (for X_3) obtained for different model choices, where the density of the points is coded using different colors. Yellow represents the highest and purple, the lowest density. The respective quantiles (2.5%, 50%, 97.5%) are represented by the violet dashed lines for both axes. The black lines additionally mark different levels of significance (0.001 and 0.05).

10,000 randomly sampled model combinations including the variable of interest were included. Fig A shows that the density concentrates around the true effect, $\beta_3 = 0.7$, which is depicted by the vertical red dashed line. However, many points lie between $x = 0.85$ and $x = 3.1$, indicating that it was possible to selectively report results towards the given interval $I = (0.85, 3.1)$ suggested in the instructions, which is depicted by the blue shaded area.

Fig B, on the other hand, displays the results for 10,000 randomly sampled model combinations, if missing values were naively dropped and outliers were not addressed. As can be seen from the many points with abscissa larger than 0.85, it was possible to obtain overoptimistic effect estimates in I even with a naive procedure ignoring missing values and outliers. The density of the estimates has two modes within the range of our intended interval I, both corresponding to significant results. This shows that reporting overoptimistic results could be achieved quite easily within our experimental setting.

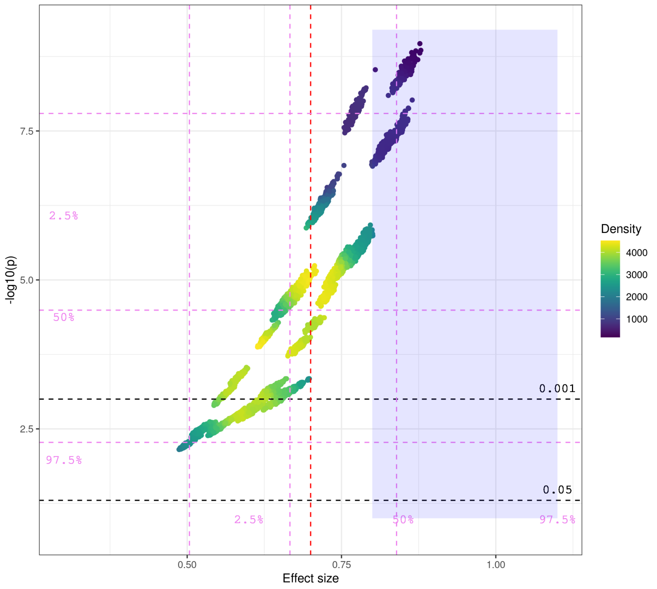

Fig A. Vibration of effects plot for 10,000 possible models (randomly sampled). Simulated data for the complete dataset without any difficulties such as interaction and non-linear effects, missing values, and outliers.

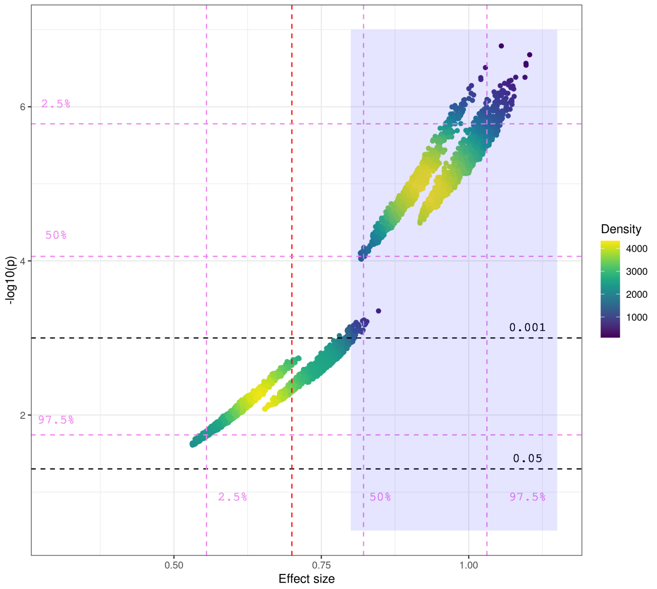

Fig B. Vibration of effects plot for 10,000 possible models (randomly sampled). Simulated data including outliers and MAR data. Missings were dropped and outliers ignored.

C Instructions for the students

Imagine yourself in the following situation[1]: You work at Ludwig-Maximilians-Universität München (Germany) and have been assigned as a statistician at Klinikum Großhadern (teaching hospital) to provide your statistical knowledge to a group of physicians. You receive a dataset with 12 variables (Y, X_1,...,X_{11}) and $n = 350$ observations. The physicians assume there is a linear effect of variable X_3 on Y, which was previously reported in scientific publications to be in the range $(0.85, 3.1)$. You also receive the following relevant information:

- X_{10} and X_{11} are categorical variables without intrinsic ordering to the categories.
- There may be further interaction effects, especially with the binary variable X_{11}. However, the literature is inconclusive on this interaction.
- Based on the variable X_6, one might suspect that the relationship is non-linear. Some studies have modeled it as non-linear, however others have modeled it linearly.
- Physicians are unsure of the effect or presence of an effect for the remaining variables.

(a) Estimate a linear regression model or related model for the situation described above. Make sure your results are reproducible, i.e., your model must always lead to the same results when you run your R-Markdown file.
(b) Explain the decisions you made during model selection and any data pre-processing procedures you followed. Typical data pre-processing procedures include, for example, handling missing values and outliers.
(c) Report the regression coefficient $\widehat{\beta}_3$ (including the confidence interval) for the variable X_3.

Note: Please avoid collaboration with classmates. Each participant has received a unique dataset, no conclusions can be drawn regarding other data sets. Your results and/or the reproducible code will be checked for similarities with your peers' work.

References

1. Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. Perspectives on Psychological Science. 2016;11(5):702–712.
2. Hoffmann S, Schönbrodt F, Elsas R, Wilson R, Strasser U, Boulesteix AL. The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines. Royal Society Open Science. 2021;8(4):201925.
3. Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. Advances in Methods and Practices in Psychological Science. 2018;1(3):337–356.
4. Childers CP, Maggard-Gibbons M. Re: Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection? Surgery. 2019;166(1):127–128.

[1] Note: The situation is fictitious and data is simulated.

5. Fields AC, Lu P, Palenzuela DL, Bleday R, Goldberg JE, Irani J, et al. Does retrieval bag use during laparoscopic appendectomy reduce postoperative infection? Surgery. 2019;165(5):953–957.

6. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychological Science. 2011;22(11):1359–1366.

7. Gelman A, Loken E. 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. 2013;.

8. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of p-hacking in science. PLoS Biology. 2015;13(3):e1002106.

9. Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, Kievit RA. An agenda for purely confirmatory research. Perspectives on psychological science. 2012;7(6):632–638.

10. Klein JR, Roodman A. Blind analysis in nuclear and particle physics. Annual Review of Nuclear and Particle Science. 2005;55(1):141–163.

11. MacCoun R, Perlmutter S. Blind analysis: Hide results to seek the truth. Nature. 2015;526(7572):187–189.

12. MacCoun RJ, Perlmutter S. 15. In: Blind Analysis as a Correction for Confirmatory Bias in Physics and in Psychology. John Wiley & Sons, Ltd; 2017. p. 295–322. Available from: https://doi.org/10.1002/9781119095910.ch15.

13. Chambers CD. Registered reports: A new publishing initiative at Cortex. Cortex. 2013;49(3):609–610.

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

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

16. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proceedings of the National Academy of Sciences. 2018;115(11):2600–2606.

17. Nosek BA, Beck ED, Campbell L, Flake JK, Hardwicke TE, Mellor DT, et al. Preregistration Is Hard, And Worthwhile. Trends in Cognitive Sciences. 2019;23(10):815–818.

18. Hardwicke TE, Wagenmakers EJ. Reducing bias, increasing transparency and calibrating confidence with preregistration. Nature Human Behaviour. 2023;7(1):15–26.

19. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. Journal of Clinical Epidemiology. 2015;68(9):1046–1058.

20. Klau S, Martin-Magniette ML, Boulesteix AL, Hoffmann S. Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. Biometrical Journal. 2020;62(3):670–687.

21. Klau S, Hoffmann S, Patel CJ, Ioannidis JP, Boulesteix AL. Examining the robustness of observational associations to model, measurement and sampling uncertainty with the vibration of effects framework. International Journal of Epidemiology. 2021;50(1):266–278.

22. Olsson-Collentine A, van Aert R, Bakker M, Wicherts J. Meta-analyzing the multiverse: A peek under the hood of selective reporting. PsyArXiv. 2020;.

23. Kerr NL. HARKing: Hypothesizing after the results are known. Personality and Social Psychology Review. 1998;2(3):196–217.

24. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "$p < 0.05$". The American Statistician. 2019;73(2):129–133.

25. Turner EH, Knoepflmacher D, Shapley L. Publication bias in antipsychotic trials: an analysis of efficacy comparing the published literature to the US Food and Drug Administration database. PLoS Medicine. 2012;9(3):e1001189.

26. Nosek BA, Errington TM. What is replication? PLoS Biology. 2020;18(3):e3000691.

27. Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Dreber A, et al. Replicability, robustness, and reproducibility in psychological science. Annual review of psychology. 2022;73:719–748.

28. Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Science translational medicine. 2016;8(341):341ps12–341ps12.

Nutzung von *Large Language Models*

Zur Anfertigung dieser Dissertation wurden Large Language Models ausschließlich dafür herangezogen, um vereinzelt Vorschläge für sprachliche und grammatikalische Korrekturen auf Basis bereits verfasster Inhalte zu generieren. Das folgende Modell kam zum Einsatz: GPT-4o (OpenAI).

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.

München, 22.05.2025

. . . .

Maximilian Michael Mandl