RESEARCH

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

Mandl *et al. BMC Medical Research Methodology* (2024) 24:152

Page 2 of 11

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 (*paO*2) 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 *paO*2 continuously is not feasible, in contrast to other vital parameters. To obtain

a reliable assessment of hyperoxemia during the surgical procedure, the *paO*2 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 *paO*2 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 *paO*2-values, (iii) parameter choice approach, (iv) aggregation procedure, and (v) coding of the exposure variable *paO*2 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 *paO*2-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 *paO*2 values through surrogate modelling, for each surgery the *paO*2 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 *paO*2 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
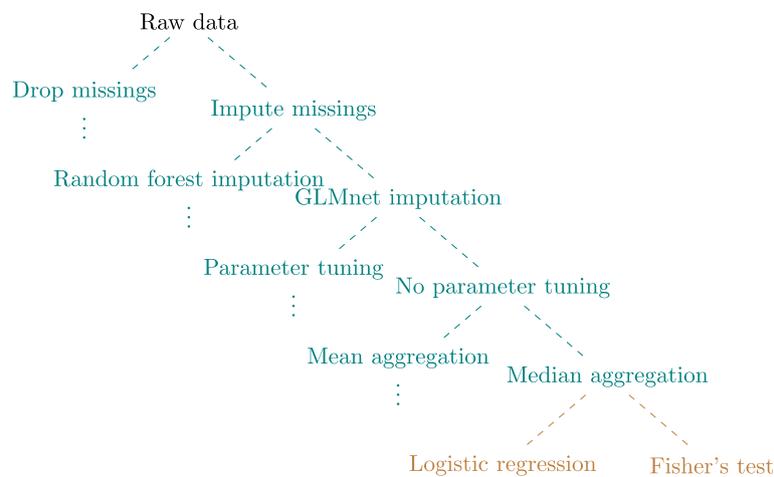
Fig. 1 Overview of the different researcher degrees of freedom. All in all 48 specifications were analyzed. Green depicts the data pre-processing decisions while brown depicts the method choices

of the test is related to the transformation of the variable *paO2*.

All in all, we consider a total of 48 specifications of the analysis strategy: 2 (missing values) $\times$ 2 (surrogate model) $\times$ 2 (parameter choice) $\times$ 2 (aggregation) $\times$ 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 "*paO2* 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—"*paO2* has an impact on post-operative complications" in our example. One may decide to formalize this research hypothesis as "H_0 : the mean *paO2* is equal in the groups with and without post-operative complications versus H_1 : the mean *paO2* 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(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(Reject at least one true $H_0^i \mid \cap_{i=1}^m H_0^i$) 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(Reject at least one true H_0^i)

Mandl *et al. BMC Medical Research Methodology*   (2024) 24:152

Page 7 of 11

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

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

with P_ℓ being the random variable for the unadjusted *p*-value for the ℓth 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 (*paO*2 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 *paO*2 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 (*paO*2 and the outcome "post-operative complications") by using a *paO*2 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 *paO*2. 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 *paO*2 and the outcome to be present, but do not formally know the truth. For each of the three

Mandl *et al. BMC Medical Research Methodology*   (2024) 24:152

Page 8 of 11
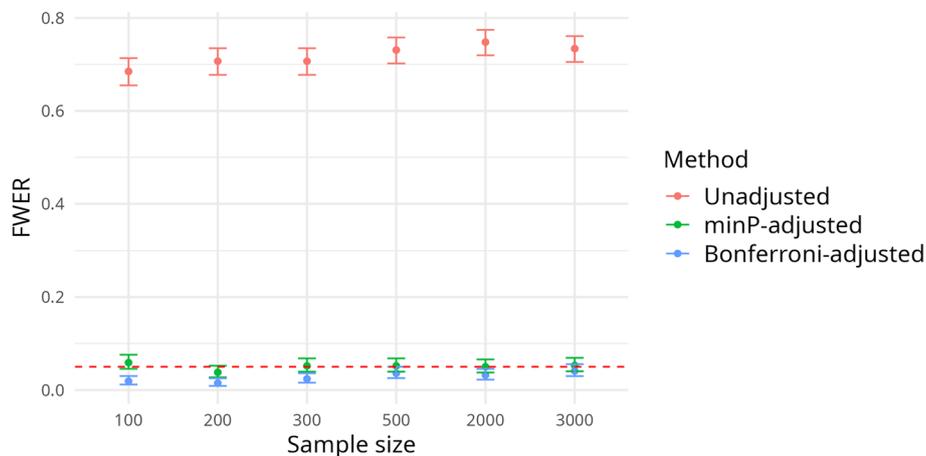
Fig. 2 FWER with Newcombe confidence intervals (computed over 1000 simulation runs) for different sample sizes without an association between post-operative complications and *paO*2. Dashed red line indicates 5% significance level
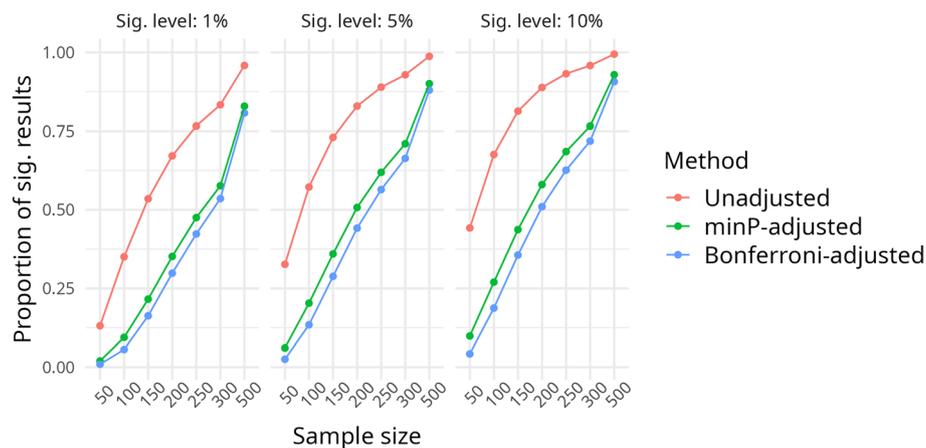
Fig. 3 Proportion of significant results for all 48 specifications for $\alpha \in (0.01, 0.05, 0.1)$ and sample size $n \in (50, 100, 150, 200, 250, 300, 500)$. Line colors indicate results based on unadjusted (red), minP-adjusted (green) and Bonferroni-adjusted (blue) *p*-values

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

Mandl *et al. BMC Medical Research Methodology*   (2024) 24:152

Page 9 of 11

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 *paO*2 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.

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.

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.

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

Mandl *et al. BMC Medical Research Methodology* (2024) 24:152

Page 11 of 11

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.1708274114.

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