Quality & Quantity (2024) 58:1447–1480
https://doi.org/10.1007/s11135-023-01698-5

Theory and methods of the multiverse: an application for panel-based models

Giulio Giacomo Cantone[1] · **Venera Tomaselli**[2]

Accepted: 2 June 2023 / Published online: 26 June 2023

Abstract
Multiverse analysis involves systematically sampling a vast set of model specifications, known as a multiverse, to estimate the uncertainty surrounding the validity of a scientific claim. By fitting these specifications to a sample of observations, statistics are obtained as analytical results. Examining the variability of these statistics across different groups of model specifications helps to assess the robustness of the claim and gives insights into its underlying assumptions. However, the theoretical premises of multiverse analysis are often implicit and not universally agreed upon. To address this, a new formal categorisation of the analytical choices involved in modelling the set of specifications is proposed. This method of indexing the specification highlights that the sampling structure of the multiversal sample does not conform to a model of independent and identically distributed draws of specifications and that it can be modelled as an information network instead. Hamming's distance is proposed as a measure of network distance, and, with an application to a panel dataset, it is shown how this approach enhances transparency in procedures and inferred claims and that it facilitates the check of implicit parametric assumptions. In the conclusions, the proposed theory of multiversal sampling is linked to the ongoing debate on how to weigh a multiverse, including the debate on the epistemic value of crowdsourced multiverses.

Keywords Multiversal modelling · Sensitivity analysis · Janus effect · Panel regression · COVID-19

1 Introduction

The community of scholars of Statistics, Data Analysis, and Quantitative Methods has always been worried about misuses of, misconceptions about, and errors of misspecification in statistical modelling (Rao 1971; Ross 1985; Lagakos 1988; Czado and

✉ Venera Tomaselli
venera.tomaselli@unict.it

Giulio Giacomo Cantone
g.g.cantone@sussex.ac.uk

1 Science Policy Research Unit - Business School, University of Sussex, Brighton, UK

2 Department of Economics and Business, University of Catania, 55, Corso Italia, 95129 Catania, Italy

Santner 1992; Verbeke and Lesaffre 1997; Olsson et al. 2000; Gardenier and Resnik 2002; Agresti et al. 2004; Fan and Sivo 2007). Aware of problems of overconfidence in one specification of the model, a precautionary sceptic demeanour resonates with the famous words of George Box (1976): “All models are wrong. Some are useful”. The theory of Multiverse of specification emerged as paradigmatic for those intended to provide reassurance to these concerns. A multiversal method uses the diversity of opinions and ambivalence of analytical choices as a means to measure the uncertainty behind a scientific hypothesis. In particular, the act of ‘modelling a multiverse of specifications’ consists of a peculiar method to draw and select random samples of “useful models” at massive scale (Gelman and Loken 2014; Steegen et al. 2016).

The present manuscript is aimed at making explicit assumptions behind the employment of a multiversal model for estimating the uncertainty around coefficients of regression between a dependent variable and a regressor. To achieve this result, a literature review of Multiverse Analysis is presented mainly in Sect. 2. This review highlights the process of convergence of different theoretical traditions and methods into a coherent methodological paradigm. The essential features of a Multiverse of specifications, as a derivation of the general theory of sampling, are presented in Sect. 3. Differently from the traditional representation of multiversal methods, a multiverse can be encoded as a string of information. This method of representation shows how a multiversal sample departs from canonical processes of independently identically distributed (i.i.d.) draws. It also allows a more thorough assessment of the sensitivity of the results, with a deeper linkage to the procedure of modelling the choices of the multiverse itself.

The theory is applied in Sect. 4, an example of a ’mapping’ of the model-induced uncertainty about the effectiveness of the vaccination plans against the COVID-19 pandemic in the year 2021. The sensitivity of the estimation procedure to relatively arbitrary modelling choices is assessed and a procedure to check a parametric assumption of the multiversal model is proposed.

A relevant result is that even if the observed effect of vaccination plans on the number of infected dead is generally significantly negative, adopting relatively reasonable choices it is possible to fabricate statistical results which would suggest, most likely erroneously, that vaccination plans induced the death of the infected people, instead. The technical causes for this concerning ambiguity are explained.

Final considerations are in Sect. 5: the debate surrounding weighting schemes for multiversal estimates is reignited with a focus on future research directions. Materials to reproduce the application, inclusive of the code in language *R*, are downloadable as indicated in Supplementary Information.

2 History of multiversal methods

In the history of criticism against bad practices in quantitative studies, three papers are the milestones of a broader scientific movement that crossed many labels: Open Science, Meta-research or Metascience (Christensen et al. 2019; Peterson and Panofsky 2020; Breznau 2021) These three papers are:

- Schor and Karten (1966), which is considered one of the first meta-analyses in Medicine. They found that 73% of 295 papers from 67 journals claimed results not correctly supported by the methods of the papers.

- Ioannidis (2005) is an explanatory summary of the methodological features predicting unreliability in scientific results: small sample sizes, small effect sizes, multiple hypotheses tested with unadjusted test statistics, etc. The paper also stressed the push from the system of the academic journals to submit novel findings to *peer review*, leading into *publication bias* (Rosenthal 1979; Simonsohn et al. 2014). If statistically significant results are over-published, there are incentives to employ weak methods to fish for false (but significant) discoveries in data analysis (Ioannidis et al. 2015).
- A large team led by Brian Nosek (Open Science Collaboration 2015) replicated 100 experimental and correlational studies published in three Psychology journals. Even if the 97 papers had significant results in their own data analysis, only 36 of replications had significant results.

Another study (Camerer et al. 2018) was conducted only on papers from the renowned journals *Nature* and *Science*. It reached significance only in 14 replications on 21. These papers launched the alarm for the "replication crisis", which soon became an important topic for the community of statisticians, who started to question the scientific validity of Null Hypothesis Statistical Testing (NHST), which reflected the application of the criterion significance to test statistics: the probability (the p-value) to observe equal or more extreme values of a test statistic under the assumption that the *null hypothesis* is true must be inferior to a threshold (α) in order for a result to be statistically significant (Gelman 2015; Earp and Trafimow 2015; Wasserstein and Lazar 2016; McShane et al. 2019; Wasserstein et al. 2019).

Often results failing to achieve statistical significance for $\alpha = .05$ are perceived as less intellectually relevant. It has been empirically established, indeed, that not statistically significant results are much less likely to be published in *peer reviewed* venues (van Zwet and Cator 2021). However, the likelihood to observe at least one falsely significant test result (a false positive result) increases for each attempt at testing, so a veracious result requires to report also the number of attempts before reaching significance, eventually adjusting p-values or α *per* the number of simultaneous test attempts (Hothorn et al. 2008). This problem is reflected in different specifications of the tests.

This is a problem in Science: there are incentives to not report not significant results and no material benefits from reporting them. The fact that α is fixed (e.g., $\alpha = .05$) sets the stage for a Goodhart–Campbell[1] phenomenon: the public knowledge of a *filter* in α pushes an incentive to p-hack the significance of its own results. p-hacking means that authors are disposed to sacrifice the substantial veracity of their scientific claims in order to decrease the p-values of their results and make them more credible (Nosek and Bar-Anan 2012; Nissen et al. 2016; West and Bergstrom 2021). Simmons et al. detailed this process (Simmons et al. 2011), and other studies demonstrated evidence of p-hacking in scientific production (Simonsohn et al. 2014; Head et al. 2015). p-hacking can be performed in more than one way, but one should deserve particular attention: when authors have the chance to opt for two or more j-specifications of the model conceptually equivalent, they may test

[1] Goodhart–Campbell Law: an excess of institutional relevance for a metric alters the well-functioning of the system that it was supposed to govern. This Law is summarised by the adage: "When a measure becomes a target, it ceases to be a good measure" (Rodamar 2018). Goodheart-Campbell is actually an empirical law on the second-order effects of quantitative policies.

both, choose the result of that one with lower computed *p*-values, or with an estimate more aligned to their theoretical predictions, and not report the alternative.[2]

Gelman and Loken (2014) frame these arbitrary choices as if they are "degrees of freedom" of a researcher. The more modelling options are left to researchers, the easier it is for them to be able to meet quantitative requirements by pure chance, thanks to a fortuitous combination of success factors. Steegen et al. (2016) elaborated this theory to account for the sensitivity of the models to what is actually featured in the specifications, and called their statistical methodology Multiverse Analysis. They adopted the *p*-curve, a tool originally developed for the detection of *p*-hacking and publication bias in meta-analyses (Simonsohn et al. 2014), and, given a unique conceptual model of effect size, they provided a visual summary of all the 'reasonable' *j*-specifications of that conceptual model. Indeed, through the *p*-curve of the specifications of the model, it is possible to visually infer the likelihood that the scientific claim represented in the model is *p*-hacked (Fig. 1): for each 'reasonable' specification, the *p*-value of the main effect is computed. The higher the proportion of $p < \alpha$ for a multiverse (or, a subset of it), the more robust the claim is.

Another visual layout presented by Steegen et al. (2016) is the *p*-Grid (Fig. 2): a multidimensional array where both columns and rows are modelling choices. These are crossed into cells, with one *p*-value for the cell. The *p*-Grid is useful because it allows to visualise the global sensitivity of the *p*-value to the modalities of the modelling choices, however, it is not very practical for large multiverses.

p-Curve and *p*-Grid are tools focused on displaying the statistical behaviour of *p*-values in the multiverse. To visualise the statistical behaviour of both the estimate of the coefficient of the regressor, is proposed the Volcano Plot (Fig. 3), which is commonly associated with the so-called "Vibration of Effect" (VoE) framework (Patel et al. 2015; Palpacuer et al. 2019; Tierney et al. 2021). Volcano Plot is useful because it allows to visualise the risk of what the authors call 'Janus effect', which is the condition for observing both a significant positive and a significant negative estimate of the same effect size.

The most advanced tool to visualise a Multiverse Analysis is the Specification Curve (Simonsohn et al. 2020). The Specification Curve is a plot in two sections (Fig. 4). The upper section is a plot with the estimates of the coefficient of the main regressor on the y-axis, and the rank of the coefficients on the x-axis. A flat curve is indicative of low variance in the population. If an estimate is associated with a statistically significant *p*-value and higher than the null value, it is coloured; if significantly lower than the null value, in the opposite colour. Estimates not associated with a statistically significant *p*-value are grey-coloured. If necessary, instead of plotting point estimates, intervals can be reported instead. In the lower section of the specification curve, the modalities of the most relevant analytical decisions are vertically piled as horizontal sequences of blocks. These blocks follow the curve of the upper section: when a point in the curve is associated with a modality of a feature, the corresponding block in the horizontal sequence is of the same colour.

[2] *p*-hacking is conceptually equivalent to Hypothesising only After Results are Known, or HARKing (Rubin 2017). HARKing is not necessarily a malicious activity: the fraud is there only if the number of tested specifications before significance is omitted. One solution to prevent *p*-hacking and HARKing is pre-registration of the specification of the model before data is collected. But pre-registration forces a limited set of features of the model, limiting serendipitous discoveries. Furthermore, pre-registrations do not technically prevent PARKing, or Pre-Registrations After Results are Known (Yamada 2018): authors can still lie about the moment of their data collection. Arguably, pre-registration is not immune from Goodhart–Campbell Law: since pre-registration is costly, it sets incentives to fabricate data that are coherent with the pre-registered specification (Gelman and Loken 2014; Pham and Oh 2021).

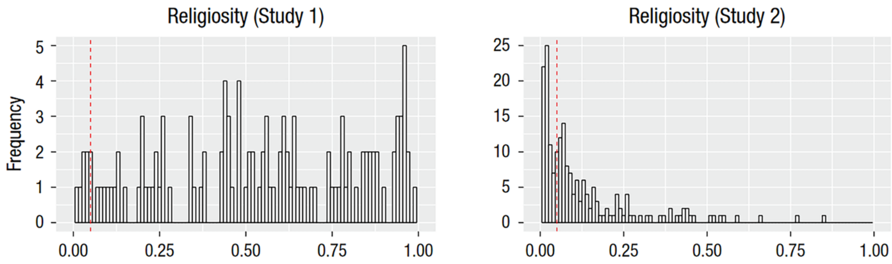

Fig. 1 Examples of p-curves as employed in Multiverse Analysis. A p-Curve is a histogram of a sample of p-values, where on the y-axis is counted the density or the frequency of the p-value in the sample. In Meta-analysis, a high concentration of p-values close to .05 is a red flag that the sample of studies is biased and not representative (Simonsohn et al. 2014). In Multiverse Analysis, the p-Curve is a visual summary of the likelihood in the multiverse of the statistical significance of the effect of the regressor on the dependent variable. In the exemplary figures are presented two p-curves of two studies with the same scientific claim: "not romantically involved women feel less religious when they are close to ovulation". The two samples are from Durante et al. (2013). The p-curves are generated by Steegen et al. (2016). In the first sample, only 8% ($n = 120$) of the specifications of the multiverse are statistically significant at $\alpha = .05$. In the second sample, this rate raises up to 44% ($n = 270$). Durante, Rae, and Griskevicius claim that their scientific theory on religiosity and ovulation is successfully replicated through these two studies, but their claim is based only on one significant specification of the model *per* sample. Applying Raftery's heuristic (> .5 of specifications must be statistically significant), both the studies cannot reject the *null* hypothesis that ovulation and religiosity are uncorrelated (Raftery 1995)

Religiosity (Study 2)

			R1 F1	R1 F2	R1 F3	R1 F4	R1 F5	R2 F1	R2 F2	R2 F3	R2 F4	R2 F5	R3 F1	R3 F2	R3 F3	R3 F4	R3 F5
NMO1	ECL1	EC1	0	0	0.03	0.02	0.08	0.02	0.02	0.03	0.05	0.06	0	0	0.02	0.01	0.04
NMO1	ECL1	EC2	0.01	0	0 .08	0.03	0.16	0.07	0.05	0.18	0.24	0.41	0.01	0	0.09	0.06	0.23
NMO1	ECL2	EC1	0	0	0.06	0.04	0.37	0.02	0.03	0.07	0.08	0.21	0	0	0.05	0.03	0.23
NMO1	ECL2	EC2	0.01	0	0 .13	0.08	0.44	0.06	0.03	0.22	0.24	0.52	0.01	0	0.14	0.09	0.43
NMO2	ECL1	EC1	0	0	0.03	0.01	0.08	0.15	0.07	0.17	0.07	0.14	0.02	0.01	0.06	0.02	0.07
NMO2	ECL1	EC2	0	0	0.02	0.01	0.06	0.2	0.05	0.42	0.23	0.44	0.03	0	0.14	0.05	0.19
NMO2	ECL3	EC 1	0.01	0.01	0.05	0.01	0.1	0.39	0.2	0.45	0.11	0.26	0.08	0.04	0.17	0.03	0.13
NMO2	ECL3	EC 2	0.01	0	0 .05	0.02	0.11	0.33	0.09	0.59	0.26	0.55	0.09	0.02	0.26	0.08	0.27
NMO3	ECL1	EC1	0.01	0.01	0.02	0.1	0.28	0.11	0.09	0.43	0.26	0.85	0.02	0.02	0.12	0.12	0.51
NMO3	ECL1	EC2	0.01	0.01	0	0 .07	0.06	0.07	0.1	0.11	0.14	0.23	0.01	0.02	0.02	0.06	0.08
NMO3	ECL2	EC1	0.02	0.01	0.06	0.11	0.36	0.06	0.04	0.3	0.13	0.66	0.02	0.01	0.13	0.07	0.46
NMO3	ECL2	EC2	0.02	0.01	0.02	0.15	0.13	0.04	0.05	0.07	0.07	0.16	0.01	0.02	0.03	0.05	0.09
NMO3	ECL3	EC1	0.07	0.04	0.12	0.09	0.16	0.16	0.11	0.54	0.32	0.77	0.07	0.04	0.25	0.13	0.39
NMO3	ECL3	EC2	0.02	0.02	0.01	0.06	0.02	0.06	0.1	0.07	0.16	0.17	0.02	0.03	0.02	0.07	0.05

Fig. 2 Example of p-Grid. It represents the multiverse of Study 2 of Fig. 1. Acronyms represent modelling choice. This scheme is useful for statements about the sensitivity of the p-values in the multiverse. For example, in this case, modality $F2$ is always significant conditionally to modalities $R1$ and $R3$, but almost never under $R2$. Given the whole picture, one is led to think that results are sensitive to the analytical choice R, which is the operative definition of the concept of being "not romantically involved"

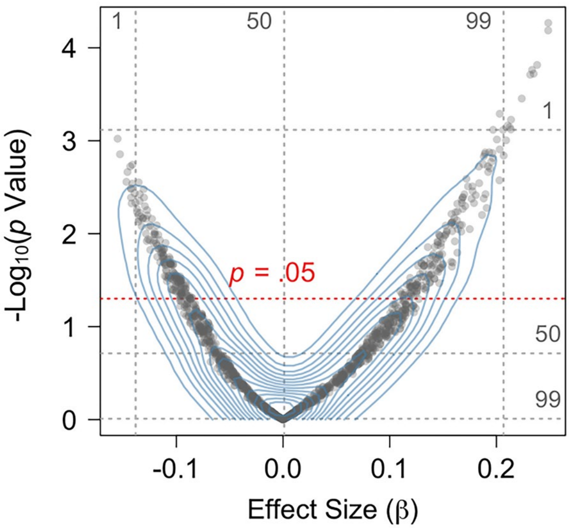

Fig. 3 An example of Volcano Plot about an artificial dataset made ad hoc to display a Janus effect. The null value of the coefficient is 1 because the estimate is of the hazard ratio between a biomarker and outcome. The y-axis of the scatterplot (p-value) has been scaled to $-log_{10}$ for proper visualisation. This multiverse is a simulation by Del Giudice and Gangestad (2021). Their aim was to demonstrate that is theoretically possible that the output statistics of the same general model can be so sensitive to Q to the point that to observe significantly positive or significantly negative estimates is almost equally likely to be in the same multiverse

The union of all the modalities of a feature always reconstructs the sequence of colours of the whole ranked curve. If a multiverse has too many specifications, it is convenient to just plot a representative sample (stratified *per* feature) of it. The lower section allows visual analysis of the distribution of ranked outputs (p-values and estimates). Clusters of blocks in the modalities are visually informative regarding the sensitivity of outputs to that feature.

For canonical applications of a Specification Curve to a big observational dataset, see Burton et al. (2021). Slightly alternative visual templates for Specification Curve are in Rohrer et al. (2017), Orben and Przybylski (2019), and Cosme and Lopez (2020).

A Specification Curve has been employed as the main visualisation tool in Breznau et al. (2022) for a multi-teams meta-analysis: Brezau and his teams provided 72 teams of analysts with the same dataset and he asked them a research question: on the basis of this dataset, what can be said about the effect of migratory flows on public opinion on welfare policies? Each team was tasked with testing their hypotheses and providing a set of model specifications that they deemed to be equally valid. Upon aggregating all the estimates, the experiment revealed concerning variability in the results, since emerged an evident case of Janus effect between positive and negative estimates.

To conclude this historical Section, it is appropriate to mention that there is pre-existing scientific literature from which multiversal methods explicitly draw upon: Leamer (1983, 1985), Raftery (1995), Sala-I-Martin (1997), Durlauf et al. (2012) and Athey and Imbens (2015). Saltelli et al. (2019) ascribe to Leamer the guilt of confusing uncertainty analysis, i.e. analysis of the robustness of a single result, with proper sensitivity analysis. For Saltelli et al. (2019), sensitivity analysis regards methods of assessment of the relevance of elements of inputs as determinants of the output, which is actually doable with the p-Grid (Fig. 2) or the lower section of the Specification Curve (Fig. 4) but not with p-Curve (Fig. 1) o Volcano Plot (Fig. 3). This criticism is not far-fetched and it must be taken into consideration for theoretical developments. Historically, authors of the multiversal methods did not focus primarily on the assessment of the causal impact of analytical choices on the numerical estimation of a parameter (e.g., what happens if we 'switch' this assumption?) but on the generalisation of a scientific claim after relaxation of its assumptions.

Fig. 4 An example of specification curve. Data is from the same simulation of Fig. 3, by Del Giudice and Gangestad (2021). In the upper section, significant positive estimates are blue, significant negative estimates are red, and non-significant estimates are grey. In the lower section, ranked estimates are decomposed into three features: `Predictors` (main regressors plus control structure of alternative main regressors), `Covariates` (additional control structure), and `Outliers` (preprocessing). Some results are visually immediate: biomarker `BM4` and `Fatigue` are robustly associated with not-positive effects, while the absence of control (no covariate) will lead to non-negative effects - even for `BM4`, because `BM4` is statistically significant only when controlled. Other results are less intuitive: `Genotype` leans towards positive effects, but once paired with `Fatigue`, it does not mediate the negative effects of Fatigue. Outputs are not sensible to data pre-processing (outliers)

The development of Multiverse Analysis is closely aligned with both the epistemology of robust analysis in Leamer and the program to achieve a method of estimation for what Nosek and Bar-Anan (2012) called a "conceptual replication" (p. 619) of a result, a terminology which was originally proposed by sociologist Collins (1992). Leamer's concern was that analysts might resort to an "easy way" of presenting their work as significant and insightful, while Collins was interested in the veracity of scientific discoveries.

3 Theory of the multiverse

Section 2 presented Multiverse Analysis and its variants as a paradigm of quantitative research methods that emerged in the social sciences and epidemiology. The historical sight fails to provide a systematic treatment of the core methodology, primarily due to a lack of coordination among its authors in developing unified definitions and principles. As a result, some problems of Multiverse modelling still remain unresolved, such as the concept of "reasonableness" of alternative choices, which is almost expressed in subjective terms. In this Section, we aim to address these issues by clarifying the theoretical assumptions that underlie the adoption of a multiversal method.

3.1 Multiversal methods and multiversal modelling

A multiversal method is a procedure involving (1) a phase of collection of a systematically differentiated multiplicity of alternative specifications of the same unitary regression model, and (2) a comparative evaluation of fit statistics regarding the estimation of the proprieties of the regression coefficient of the main regressor variable, commonly referred as 'effect size' in studies oriented towards causal inference[3]. Multiverse Analysis is not the only possible instance of a multiversal methodology, since one can collect specifications of the scientific modelling of a unitary scientific hypothesis from many teams (Schweinsberg et al. 2021; Breznau et al. 2022). However, Multiverse Analysis as exposed in Steegen et al. (2016) remains the canonical application of multiversal methodology, whereas Vibration of Effect (Patel et al. 2015) or many teams (Breznau et al. 2022) are alternative methods with different epistemological premises.

During phase (1) of Multiverse Analysis, a set of specifications is collected through the systematic differentiation of a single conceptual model into a multiplicity of specifications. This operation is itself a form of (multiversal) modelling: while a regression model is a representation of a scientific hypothesis, a multiversal model represents the knowledge of a group of analysts about testing a specific hypothesis through a regression method.

Gelman and Loken's metaphor of a "Garden of Forked Paths" provides an example of how decisions made by analysts, flowing from an abstract hypothesis into a specified model (often a line of code in software), follow an organic process of systemic differentiation; whose magnitude and range depends on their methodological knowledge and insight. For instance, when estimating the coefficient of a binomial regression between a singular regressor and a dependent binary outcome, an analyst has to choose between using a *logit* or *probit* link function. While these two link functions may be considered conceptually equivalent, there is always a numerical difference between the two coefficients, however

[3] These statistics are commonly the estimate of the effect size itself or p-values. Potentially, there may be others: the output of a linear model in base software `R` provides values for estimates of regression coefficients, their standard errors, confidence intervals, test statistics, and p-values for each, global R^2 (adjusted and unadjusted) for the whole specification of the model, the p-value of specification and the sum of the residuals. It includes also a more refined measure of fit as *Information Criteria* of the specification (Akaike's and Schwartz's), and the logarithm of the likelihood of the specification.

trivial. Another example of a binary decision concerns the absence or presence of a third variable as a "control". Such decisions are often ambiguous but can lead to a numerical differentiation in the resulting estimates. Gelman and Loken refer to these arbitrary decisions as "researcher degrees of freedom." When two binary decisions are combined, four coefficients are assumed to be equally valid a priori. With a sufficiently high number of "researcher degrees of freedom," the number of possible specifications can easily exceed a thousand, resulting in a proper sample of different numerical estimates. The assumptions on multiversal modelling draw on this analogy between the process of organic specification of a model and the process of sampling observations.

3.1.1 Assumptions of multiverse analysis

Let θ be an estimand on whose value depends the validity of a scientific claim (Muñoz and Young 2018a). For example, θ may represent an effect size of a unitary increase of a regressor x on a dependent y. $\hat{\theta}$ is the generic estimate of the θ from a regression model fit on a sample. The claim 'x causes y' substantiate its veracity if it can be demonstrated that $\hat{\theta}$ is sufficiently distant from 0.

Conventional sampling theory assumes that θ is a parameter of the involved data-generating process and also a parameter of the sampling distribution of representative samples of that data-generating process. If samples are unbiased (e.g. i.i.d. draws), then the error between the estimate $\hat{\theta}_k$ (fit a sample k) and θ follows a Normal distribution. Under these assumptions, the expected variability between the estimate and the parameter can be measured with the canonical estimator of sampling standard error $\sigma_k(\bar{\theta})$:

$$\sigma_k(\bar{\theta}) = \sqrt{\frac{\sum_{k=1}^{J}(\bar{\theta}_k - \hat{\theta}_k)^2}{K}} \tag{1}$$

whereas K is the number of samples. Assuming $\bar{\theta}_k \approx \theta$, then,

$$\epsilon_k \approx \bar{\theta}_k - \hat{\theta}_k \tag{2}$$

is equivalent to the sampling error, which can be characterised as the random component of the error of measurement of the estimate.

The theory of sampling error assumes that exists one correct specification j_θ of the model of the data-generating process, and that $\hat{\theta}_k$ are estimated through j_θ. A $\hat{\theta}_k$ is accepted as an approximation of θ by an analyst when he or she is confident that it is sufficiently close to the parameter, hence ϵ_k is small.

Young and Holsteen (2017) accept the formal reasoning behind sampling theory, but they argue that the assumption $j_{\hat{\theta}_k} = j_\theta$ in practice systematically neglects the component in the estimation error that is due to the misspecification of the model. Keeping in mind the expression: "All models are wrong, some are useful", they are correct in identifying that sampling theory alone does not measure the usefulness of a model specification, and that this usefulness should reflect how "less wrong" the specification is in the representation of the data-generating process involved in a scientific claim (Aronow and Miller 2019). Even if Young and Holsteen (2017) never explicitly mentions Multiverse Analysis, the theoretical foundations to refute that ϵ_k is a sufficient measure of the uncertainty regarding a scientific claim lies deeply on the same theoretical foundations of the works mentioned in Sect. 2. Their work, focused on effect sizes, is an extended attempt to re-frame the Multiversal methodology as a theory of variation of estimates across model specifications. Mirroring Eq. 1, they derive an estimator of what they refer to as model (standard) error $\sigma_j(\bar{\theta})$:

$$\sigma_j(\bar{\theta}) = \sqrt{\frac{\sum_{j=1}^{J}(\bar{\theta}_j - \hat{\theta}_j)^2}{J}} \tag{3}$$

which in the context of a theory of Multiverse Analysis can also be understood as the multiversal standard error.

From Eq. 3 it is possible to derive the assumptions of the theory of the Multiverse, and appreciate how it mirrors the theory of sampling, albeit with some not trivial differences. The primary assumption of Multiverse theory is that, if θ is a value looked for validating a scientific claim, and exists a population of $\theta + \epsilon_k = \hat{\theta}_k$ values that are equivalent to θ, then, it holds the following generalisation: it exists a population of equivalent parameters $\Theta : \{\theta_1, \theta_2, \dots\}$ such that the same scientific hypothesis originally modelled after θ would still be validated by the estimation of the approximation of θ as $\theta_j \in \Theta$. From this assumption derives the formal definition of model error (multiversal error) as:

$$\epsilon_j = \mathbb{E}(\Theta) - \hat{\theta}_j \tag{4}$$

The following assumption of Multiverse theory is that the Θ can be sampled through the identification of a finite set $\mathbf{j} : \{j_1, j_2, \dots, j_J\}$ of specifications that are 'reasonably' conceptually equivalent. Indeed, from this set of specifications is computed a finite sample of estimates $\hat{\Theta} : \{\hat{\theta}_{j_1}, \hat{\theta}_{j_2}, \dots, \hat{\theta}_J\}$. Of course, the corollary of this assumption is that it is not only possible to sample representative estimates through the identification of a finite scheme of specifications (a theoretical multiversal model), but also that this identification is possible through a reproducible procedure (a sample of specification).

Compared to Eq. 2, the concept of model error as expressed in Eq. 4 is not formalised after an approximation. It is just an abstract propriety of the set Θ that epistemologically derives from the unknown θ. But there are no assumptions on the distribution of Θ, so in this stage, there is no formal connection between θ, Θ and $\bar{\theta}$. In other words, these two assumptions are sufficient to define only the non-parametric proprieties of the Multiverse, and to derive non-parametric tests as those proposed in Simonsohn et al. (2020).

There is a third parametric assumption that is definitely worth mentioning, because it seems a necessary premise to accept the methodological toolbox of Young and Holsteen (2017), Muñoz and Young (2018b), and arguably of Steegen et al. (2016), too. This third assumption can be expressed as follows: given that both θ and j_θ are unknown, for each $j \in \mathbf{j}$, for a sufficiently large J, then the modeller of a sufficiently large Multiverse expects the average $\hat{\theta}_j$ to be closer to the latent parameter θ than the majority of the individual estimates $\hat{\theta}_j$.

Following this third assumption, conditional to no other prior information on θ and j_θ, even without other assumptions on the distribution of Θ, the approximation

$$\bar{\theta}_j - \hat{\theta}_j = \hat{\epsilon}_j \tag{5}$$

is the unconditional best estimator of model error ϵ_j. This assumption reflects the belief that the average of estimates from many sources of knowledge (many specifications) is a priori more reliable than a single opinion, as well-informed as it may be. Why? Because a priori it is not possible to know how informative $\hat{\theta}_j$ is. So the belief that it is not possible to weight a priori the relevance of specification is the theoretical origin of the parametric assumption in Multiverse Analysis.

The suggestion to expand to higher values the number of specifications J surely resonates with the approaches in Young and Holsteen (2017) or Breznau (2021), and arguably with all the other authors mentioned in Sect. 2. However, authors as Slez (2019) and Lundberg et al. (2021) proposed arguments for refusing to treat a large set of estimates as if all specifications were epistemologically equivalent.

3.2 Modelling the multiverse

It is convenient to represent the set of specifications as a vector $\mathbf{j} : \{j_1, j_2, \ldots, j_J\}$, of J length. After the specifications are fitted on the sample, the vector is matched to their regression statistics (estimates, p-values). The resulting database is the multiversal sample. The modelling of the Multiverse consists primarily of the procedure to identify $\mathbf{j}$.

3.3 Taxonomical issues: the analogy of the switch

Gelman and Loken (2014) presents analytical choices as "degrees of freedom"; from a modelling standpoint, a better analogy is that of the 'switches' connected to an engine. These switches dictate how the regression model (the "engine") should proceed to estimate parameters, conditioned by input data. In this analogy, a specification of a model is a scheme of the positions of all the switches. Another word for the position of a switch is modality.

Let q represent the generic analytical decision regarding a model. The set of analytical choices can be represented by a vector $\mathbf{q} = \{q_1, q_2, \ldots, q_Q\}$. The initial step in multiversal modelling involves the quantification of Q by identification of the ambivalent analytical choices encountered during the specification of the model. This is not a trivial task, as the literature presents multiple taxonomies and possibly contradicting methodological rules (Steegen et al. 2016; Simonsohn et al. 2020; Del Giudice and Gangestad 2021). Uncontroversial examples of analytical decisions are the adoption of alternative estimators and the inclusion of variables in the structures of controls.

In social sciences, different operative definitions (or 'proxies') of the same latent construct are generally admitted as 'switches' of a model, even when these operative definitions regard the dependent variable. In this case, the analysis of multiversal statistics truly depends on choices that the analysts regard as 'good sense'. This 'good sense' is mostly an expression of their own methodological knowledge and *finesse*. Indeed, a multiverse, if not a representative sample of Θ, is very informative on the methodological knowledge and insight of its own authors.

Some choices depend strictly on what exactly θ is, and how it is informative about the scientific claim: when estimating the variability in multiversal estimates of effect size, it is only reasonable to avoid mixing estimates from a linear model with those from a binomial, unless they are converted into a comparable scale. However, when observing the variability in p-values, a differentiation through types of regression models could be a valid analytical choice.

Once those analytical decisions are set, for each decision it is necessary to identify the set of admitted modalities, too. A detailed attempt to formalise multiversal modelling is in Hall et al. (2022), but it does not catch some emergent proprieties of model specifications within a multiverse. A different formalism is proposed, with the aim of a more radical identification of the type of analytical decisions in multiversal modelling. Three categories are identified:

- a logical decision is one that can be thought of as a switch set on ON or OFF. For example, deciding whether to include a covariate as a control involves the presence or absence of the covariates in the model. The proposed formalism is to record value 0 representing the absence of the feature, and 1 representing its presence.
- a multimodal decision regards two or more alternative options, of which if one is included in the model, no other is. An example is which estimator to adopt for a parameter. The proposed formalism is to use Latin letters to identify a modality of this kind of decision.
- a multimodal decision with absence regards the presence or absence of a feature, that can be present with alternative modalities, too. An example is whether to interpolate missing data at all and, if so, how to do it. The proposed formalism is to adopt the value 0 when the feature is absent, and Latin letters to identify a modality of the decision.

The correct identification of modalities is a cumbersome task because it should follow the guidance of the current state-of-the-art of the involved methodology. Three general situations can be identified:

- Some alternative choices are truly ambivalent in literature, e.g. the multimodal choice between *logit* and *probit*.
- In certain cases, there may be clear indications about the inclusion of a modality, which depends on just checking if an assumption holds. For instance, it has been argued in Sect. 4 that a strong assumption of a Poisson model regression is the absence of overdispersion, which refers to an inflation of variance in the observed values of the dependent variable y. If overdispersion is observed, it is highly advisable to adopt a corrected estimator. Not correcting is not a reasonable alternative, as uncorrected estimates will be less representative of the data-generating process (Lundberg et al. 2021).
- Yet, there are decisions where the literature seems to suggest a direction, but there are other factors that lead an analyst to still include some modalities. An example would be modelling a regression over samples presenting a panel structure (see, Sect. 4). Specific estimators have been developed to improve the estimation after observing the clustering of the observations across time and groups, but yet the literature suggests that is also worth checking results from conventional linear models (Gelman and Hill 2007).

Identified the m_q number of modalities for each q, it holds:

$$J = \prod_{q=1}^{Q} m_q \tag{6}$$

because $\mathbf{j}$ is the finite set of all the combinations of $j : \{m_1, m_2, \dots, m_Q\}$.

3.3.1 Model specifications as strings of information

It is now possible to introduce a fundamental difference between Eq. 1 and Eq. 3, which goes beyond any abstract assumption on the distribution of Θ. The theoretical k-draws of a sample are equally likely and mutually independent, so they conform to the i.i.d condition. Instead, the j of Eq. 3 are combinations of features that, by how they are differentiated, can not be assumed to be drawn by an i.i.d. sampling distribution (Western 2018).

This fact becomes even more clear with a representation of j as a string of Q symbols, each symbol representing a m_q. For example, following the rules of Sect. 3.2, a small multiverse of $J = 4$ made of a logical choice and a binary multimodal choice would consist of a list of four strings:

- j_1 : "0A"
- j_2 : "1A"
- j_3 : "0B"
- j_4 : "1B"

Two proprieties emerge after this example. The first is that strings are sampled in clusters, not as individual units. For example, if a multimodal $q_3 : m_{q_3} = 3$ is added to the model, then 8 new strings will be sampled together (see, Eq. 6). If this choice is then repelled, these will fall out together. A corollary is that the same string will be never sampled twice.

The second propriety is that the canonical structure of the multiverse of specifications is not exactly a "Garden of Forked Paths", but more akin to a regular network structure, where some strings are more similar to others. This similarity can be represented in two equivalent ways. The first is through a Hamming distance (d) between two j, which is the number of symbols that should be altered for one of the two to be an identical string to the other (Bookstein et al. 2002). Since the same string cannot be sampled twice, the minimal Hamming distance between two strings is $d = 1$, and the maximal is $d = Q$. The second representation of this quantity is the length of the shortest path between two nodes in a graph (Fig. 5).

3.4 Inference on multiversal statistics

An argument against the canonical employment of parametric inference through multiversal statistics is proposed by Del Giudice and Gangestad (2021): the choice q of including a control variable biases the estimates of half of the multiverse if that variable is a collider (Elwert and Winship 2014).

Through a simulation, Del Giudice and Gangestad (2021) shows how a collider (`Fatigue`, see Fig. 4 and also Fig. 3) can also induce a Janus effect in a multiverse. If instead this variable had been correctly identified as a collider, then it would have been excluded from controls of the multiverse. Given that the identification rules of q primarily rely on the prior knowledge of the analyst, they argue that nothing in Multiverse theory precludes the presence of collider variables, or analogous analytical decisions, within the multiverse.

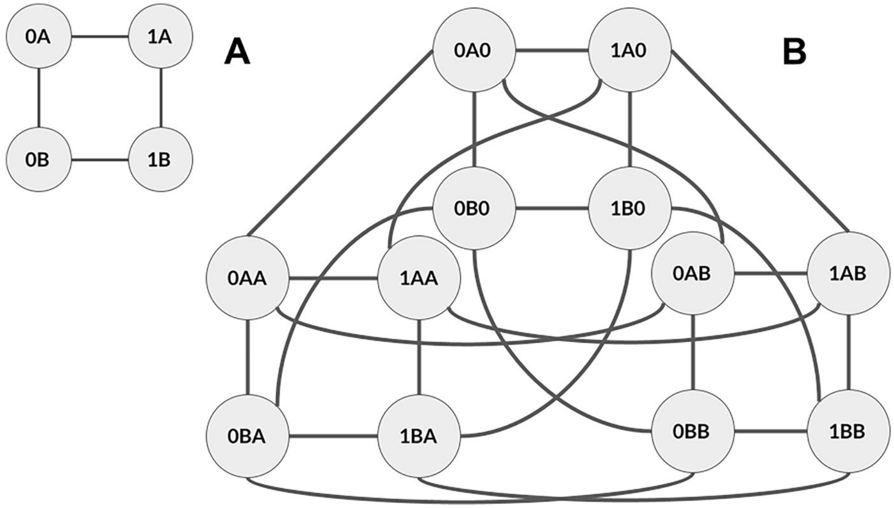

Fig. 5 The multiverse of strings represented as a graph. The object **A** represents the simpler case of multiverse, with $J = 4$. Strings are nodes. Each node is reachable from any other, but some pairs of nodes are closer: the number of links to cross (path length) is lower. The number of links to cross is the length of the shortest path and it is equivalent to Hamming's distance d between two strings. The object **B** is much more complex just by adding a third q with 3 modalities, yet the structure preserves a regular form

This issue is concerning because the modern theory of collider structures incorporated the concept of biased data collection: any selection bias is a collider bias (Munafò et al. 2018). While it is truly important to be aware of collider bias, the claim that large multiverses are inevitably biased should not rest undisputed. To better understand this issue, we can draw an analogy with the concept of M-bias. An M-bias arises from an M-shaped causal structure where a collider cannot be distinguished from a confounder (Shrier 2008; Rubin 2009). It is credible that, in the case of M-bias, a team of analysts could include the collider variable as a logical q of the multiverse, biasing half of the estimates as a consequence. However, further studies have shown that M-shaped causal structures are highly specific and embedded within more complex causal structures, reducing the magnitude of bias to trivial values (Liu et al. 2012; Ding and Miratrix 2015). In other words, as long as there is no clear intention to bias the estimates towards a specific outcome, independent sources of bias in the multiverse behave as random errors, annihilating each other out in the estimator of $\bar{\theta}_j$. Thus, contrary to what Lundberg et al. (2021) seems to suggest agnostic yet large multiverses may be more reliable than small well-tailored ones. This analogy could be an argument to not refute the parametric assumptions of multiversal methods.

A prudent approach to the application of multiversal statistics involves conducting sensitivity analyses to evaluate the effects of the modelling choices, such as understanding the conditions for which Janus effect manifests in a multiverse. The formal categorisation of the analytical choices in the modelling of the specifications as strings has the potential to facilitate such analyses by enabling differentiation between absent and included features. This differentiation is also foundational for a preliminary method to check the parametric assumption underlying the Multiverse approach.

3.4.1 Sensitivity analysis through multiversal models

In the sensitivity analysis within Multiverse Analysis, modalities act as grouping variables for the statistics of the multiverse. The aim is to assess which modality is the most relevant contributor to two proprieties of the multiverse: the variance of the estimates and the relative location within the sampling space of the multiverse. For the scopes of the present analysis, the variance of the estimates can be measured as the exponentiation of the estimator of the multiversal standard error (Eq. 3).

About the second property, locating the estimates when a modality is present, can be actually insightful in understanding the sources of biases, with the caveat that one does not really see a 'bias' but only the stability of the estimates in a certain portion of the sampling space of estimates. Despite this caveat, this operation is not at all useless. In fact, the problem represented by the Janus effect is quite evident. Generally speaking, and especially in social sciences, rarely models check quadratic relationships. Whether this is an epistemological limit or not, except when it is expressly provided for in a scientific claim, the coexistence of positive and negative effect size estimates is a strong red flag of an error of misspecification. A posterior selection of the modalities, under which the Janus effect does not occur, can be a method to reduce this effect.

Some purists may argue against a posterior selection of the modalities since it could be arbitrary. In Sect. 3.2, the analytical choices are classified into three categories: legitimate differentiations of the model; false options, that should not be considered when modelling the multiverse, and cases in-between, where one modality seems more valid than others, or where choosing one modality makes it seems contradictory to consider others. In this third case, it is recommended to still compute multiversal statistics for all combinations and then cluster them to reflect these ambiguities in a priori modelling. So, this operation of a priori clustering can be the preliminary step for validating a posterior selection as not arbitrary.

The significance of differences between modalities can be inferred through a statistical test of the difference between groups. Practically, in large multiverses, it is usually sufficient for the visual outlook of the lower section of the Specification Curve to detect sufficient differences among modalities (see, Fig. 4).

3.4.2 Check the parametric assumption

The following procedure is proposed to check the assumption that, the larger a multiverse, the less unbiased the estimate of $\bar{\theta}_j$ is.

In a string, the symbol "0" always represents the absence of a feature. A model specification represented as a string with a "0" is always equivalent to a specification of that model in a multiverse where that feature has not been included. From this equivalence, it follows that splitting a multiverse across the number of "0" in their strings, those with a high number of 0 are equivalent to more parsimonious models and, by converse, those with a low number of 0 are more complex models.

So, this procedure, akin to sensitivity analysis, consists in grouping specifications across the number of their 0 and checking their differences in variance and location. The assumptions of this procedure are derived from an analogy to model high-parameterised inferential models. Conventionally, raising the number of parameters in a model has the effect to reduce the bias of estimates, while inducing higher variance. This is also referred to as the trade-off between bias and variance in estimation (James et al. 2013; Belkin et al. 2019).

So, in this procedure, it is assumed that a shift in location is representative of a reduction of pre-existing biases. Also, as the number of 0 raises, the sample variance of the estimates is expected to rise.

4 Application: COVID-19 vaccination

A multiverse is modelled around the following scientific hypotheses:

- $\mathbf{H_0}$: The vaccination plans in 2021 had no significant impact on the risk of death in infected people with COVID-19.
- $\mathbf{H_1}$: The vaccination plans in 2021 reduced the risk of death in infected people.

These hypotheses are identified as such:

1. the *explanans* of death reduction is a collective social fact, the vaccination plan - and not the vaccine shot as an individual biological fact. The demographic effect is mediated by the biological effects of the vaccine shot but it accounts also for spillover effects of potential reduction of contagiousness of the virus due to a reduction in symptoms. This definition does not differentiate between different brands of vaccines for COVID-19. It also allows to account for behavioural responses to enacted policies (lockdowns, etc.).
2. the death reduction is imputed only on the infected, not to general mortality in the population. $\mathbf{H_1}$ does not regard a cost-benefit analysis of vaccination plans on general human mortality. For example, it can be hypothesised that before vaccination plans, lockdown policies actually reduced general mortality at cost of mobility. If after vaccination plan lockdown policies are ceased, general mortality could raise not because of the effects of vaccine shots, but only because mobility is regained (Islam et al. 2021).

From point 1. follows that this theory involves the *effectiveness* of vaccination, not its *efficacy* (Olliaro et al. 2021; Lipsitch et al. 2022). The literature on COVID-19 vaccine effectiveness against risk of death (Dagan et al. 2021; Haas et al. 2021; Jabłońska et al. 2021; Patel et al. 2021; Tregoning et al. 2021; Fiolet et al. 2022; Lipsitch et al. 2022; Pormohammad et al. 2022) reveals a certain variability. These sources report that the reduction spanned around 75–80% of the relative risk of death in the un-vaccinated, but with large confidence intervals.

4.1 Panel dataset

The dataset is a data linkage of sources:

- the dataset on the pandemic from Johns Hopkins University (Dong et al. 2020);
- “Our World in Data COVID-19” (oWiD) datasets (Mathieu et al. 2021)
- the Oxford COVID-19 Government Response Tracker (OxCGRT); (Hale et al. 2021), accessed through the COVID-19 DataHub (Guidotti and Ardia 2020).

These are fused into a panel dataset, which is a multivariate sample indexed by a time covariate with regular time intervals, and by one or more grouping covariates. In the sample, the time covariate ($\mathbf{t}$) consists of 409 days and contains all year 2021, and the grouping

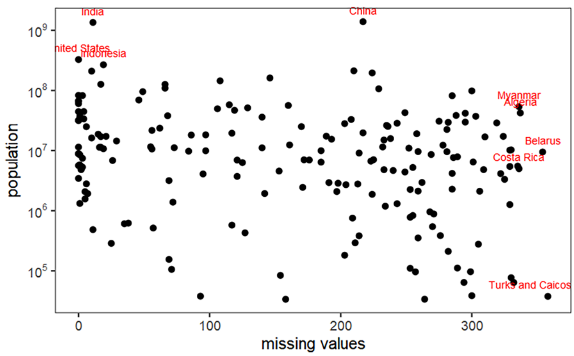

Fig. 6 Missing values of the proportion of people with at least a vaccine shot. In this plot, the days before the first day of reporting do not count as `na`, even if in some cases, they should. China is the most problematic case since it scarcely reported data on vaccination plans only

variable (**g**) contains 186 countries.[4] There is a total of $186 \cdot 509 = 76.074$ rows in the dataset.

The response **y** is the count of deaths in infected. **y** has not a problem of missing values, however, as a count variable, it is more overdispersed than a Poisson distribution[5].

In the dataset, there are two candidate variables to represent the main regressor mentioned in $\mathbf{H_1}$, the state of a "vaccination plan". These two are:

A Minimal: The percentage of people with at least one shot of vaccine. This quantity can never conceptually decrease over time.

B Full: The percentage of people that are counted by the country administration of their State as "fully vaccinated", e.g. because they took up-to-date booster doses. This number can also decrease over time.

This is the first degree of freedom the researcher, or q_1, and it is multimodal. Differently from the case for identifying a vector **y**, these two potential candidates for x have relevant issues of missing values (`na`), clustered around some countries, China being the most problematic (Fig. 6).

The issue of missing data can be disjoint in two considerations:

- The vaccination plan does not start before the first vaccination, hence the natural percentage of the vaccinated population before the first non-`na` values are ~ 0 (China is an exception). These '0' are natural zeroes.
- All of `na` between two non-`na` are logically non-zeroes, instead.

It follows that q_2 is the logical decision to impute natural zeroes before the first not-missing value in the location group, and q_3 is the logical decision to interpolate the others missing values[6].

[4] 46 other locations have been removed for being substantially uninformative e.g. ships like the "Diamond Princess", micro-nations and semi-independent administrative territories with less than 30.000 inhabitants, countries that reported data on COVID-19 deaths only once in 2021 (e.g. Eritrea).

[5] The assumption of a Poisson distribution is $\frac{\bar{y}}{s^2(y)} \sim 1$. In the dataset, $\frac{\bar{y}}{s^2(y)} = 962$ and $\frac{Median(y)}{s^2(y)} = 11$.

[6] There is a regular progression of the vaccination plan, hence linear interpolation is appropriate. Linear interpolation is also the standard model of interpolation with the command `zoo::na.approx` in R. One may argue that instead q_3 should be a multimodal decision with an 'absence' modality (no interpola-

4.2 Modelling of the regression on panel data

Since **y** is a count, the most appropriate type of regression is the Poisson, however, since **y** it is also overdispersed a Negative Binomial (NB) or a Quasi-Poisson (QP) is used instead. These alternative methods of parametric estimation differ in how the parameter of overdispersion (ϕ) maps the relationship between the parameter of the mean value in Poisson (λ) and the expected value of the variance. q_4 is the multimodal choice of the correction of overdispersion for the Poisson regression. Details on this correction are provided in Appendix A.

A Poisson coefficient is usually not immediately interpreted as an effect size. However, conditional to some further assumptions, since the vector **x** is scaled within the unit interval (i.e., it is a percentage), then the coefficient b of a Poisson regression can be interpreted as the natural logarithm of the estimator of the hazard ratio of the binary choice to treat (get vaccinated) *vs* not. For clarity, $1 - \exp(\hat{b})$ can also be interpreted as an estimate of relative risk reduction, i.e. how likely is to expect a reduction in the risk of death after infection, if treated with the vaccine. For the assumptions holding this interpretation, see Appendix A.

The causal effect stated in $\mathbf{H_1}$ requires time before manifesting itself. To properly model this delay, the vector **y** must be lagged compared to both **x** and the controls Z. The most mentioned lags in the literature (Sect. 4) are 7 days, 14 days, and 21 days, and these will be the modalities of the fifth ‘switch’ (q_5), the lagging scheme, which is a multimodal decision.

$\mathbf{H_1}$ does not make a distinction among subgroups of human populations (gender, ethnic genetics, etc.). This makes sense for the causal effect of the vaccine shot: vaccines are not developed to be biased across human groups. Nevertheless, data on deaths (**y**) are collected by countries, and variability in **y** may depend on *how* different countries collect data, plus on specific unobserved features of these countries. There are two ways to account for the variability that is *within* the grouping covariate **g**:

A to pool all the rows of the dataset, ignoring **g**, while controlling for all the fixed determinants Z_g of **y**. Z_g are all the variables that can be assumed as fixed to only one numeric value within the grouping variable while having an impact on the value of y, i.e. all those variables whose variability is dependent mostly by differences *between* groups. For example, is practical to assume that in a time span of one year, the population of the country or the percentage of elderly people in that country is stable around only one numeric value.

B to model the assumption of the *fixed effect* directly through the mediation of the grouping variable **g** in itself. This can be done with the so-called Fixed Effect (FE) estimator[7]. To intuitively understand FE estimation: it is an implicit control for all of these unobserved

Footnote 6 (continued)

tion), and different models of interpolation would be the modalities of presence but the practical differences would be abysmal.

[7] For linear models, there are many estimators of FE e.g., the de-meaned (a.k.a *within*) estimator or the first difference estimator. These estimators do not generalise for GLMs. Hausman et al. (1984) proposed a specific FE estimator for Poisson based on Conditional Likelihood Estimation (CLE), but Allison and Waterman (2002), deemed its link function for Negative Binomial biased. However, a new unbiased FE estimator has been proposed for count data, this time based on MLE (Broström and Holmberg 2011).

features (covariates) that are fixed within **g** - even those that are not determinant for y. In this sense, FE is a shortcut to avoid an identification strategy for what is Z_g.

Should this decision be q_6 about how to control the variability between groups? Yes, but with a *caveat*: this is a complex decision because it is not necessarily limited to a binary choice.

The first approach requires an identification strategy for the Z_g: which ones to include, or not. The population of the country (actually its natural logarithm, since the regression type is Poisson) should always be included in Z_g. Without this operation, it would be not possible to claim a proper panel estimation, since y is a count of daily events happening exactly in the finite population. Two other relevant time-fixed variables in the dataset are the share of elderly people (> 65 years old) in the population and the population density (`pop_d`). So there are 4 different modalities for the pooled estimator: including only `age`, including only `pop_d`, including both, or neither. The number became 5 with the addition of the FE alternative.

q_6 can be framed as a multimodal decision with absence but put into practice the theoretical rules, there is still ambiguity. If one applies the rules of Sect. 3.3.1 *within* a simple multiverse where only a pooled model is considered, then there are two decisions: to include or exclude `age` and `pop_d`. The shape of this small multiverse would be a square, like the one in Fig. 5, and the specifications where both ("11") or neither ("00") are included would measure a Hamming's distance equal to 2. But if instead the same problem is re-framed as a problem of comparison of specifications of a pooled model to a FE model, then "00" should be re-coded as "0", and "11", "01", "10" and "FE" are conceptually re-coded as "A" "B", "C", and "D". In addition, the FE estimation is peculiar because it aims to mimic the coefficient that would be seen if all the fixed characteristics of the group features would be controlled. In FE estimation, differently from *pooled* models, Z_g are not added to not induce technical multicollinearity in the model (Allison 2009), so the distance between "FE" and "00" should be even higher than between "00" and "11". On the other side, the reason to include these differences as different modalities of the same q is exactly their role as fixed Z_g. They do not control for the variance in the effect of $X \rightarrow Y$, they control jointly the variance *within* **x**. These issues would be accounted for in the sensitivity analysis of the result.

This does not exhaust the analytical choices involved in a panel model. Should y be assumed as time-independent? This question is tricky: serial correlations could rise through behavioural dynamics of information cascades (e.g., "once I see a reduction in deaths, then I vaccinate myself"), but also because the vaccination plans could decrease the contagiousness of the virus.

One way to control the time dependencies is to extend FE to the **t** vector. **t** and **g** are orthogonal, hence the FE of **t** can be combined with pooled models to control for variance within g, too. **t** can be pre-processed with two different principles:

A conventionally: **t** is the distance from the first day in the dataset.
B after onset: **t** is the distance from the first day of the vaccination plan within **g**, the 'onset' of the vaccination plan for g. **t** would be a random variable pegged to **g**, and all days with 0% vaccination rate would be set as $t := 0$.

Table 1 Features and modalities of the multiverse

	Decision	Type	M_q
q_1	Measure for vaccination plan	Multimodal	2
q_2	Natural zeroes imputation	Logical	2
q_3	Linear interpolation of `na`	Logical	2
q_4	Overdispersion's correction	Multimodal	2
q_5	Lagging schemes	Multimodal	3
q_6	Panel Estimator	Multimodal with 0	5
q_7	Time control	Multimodal with 0	3
q_8	Infected rate control	Logical	2
q_9	Lockdown policies	Logical	2

The q_7, about how to control for time-dependencies, is not dual but threefold, with a modality for absence: $\}\}0''$ would stand for no control, A = *Conventional* for regular time count, and B = *Onset* for the alternative time counts from the first day of vaccination.

In a regression on a panel-structured sample, there are two reasons to add a control. Fixed numerical controls (Z_g) are added to identify the contributing causes for the variance between groups. Other controls ($Z_{t,g}$) are added to identify contributing causes of variance over time. q_8 is the inclusion/exclusion of the reported rate of infected (positive rate or `pos_rt`). This would be a natural inclusion in the multiverse, yet 40% of the rows of the dataset reports a `na`, so this q could be a considerable source of variance in the multiversal estimates.

A concept to explore within the formulation of $\mathbf{H_1}$ is the mediating effect of mobility restrictions. Oxford COVID-19 Government Response Tracker (OxCGRT) (Hale et al. 2021) provides both a composite normalised index in the unit interval (`OXSI`) and many indicators of anti-pandemic policies. The most pertinent indicator is an ordered multinomial measure of the severity of lockdown policies (`LKDW`). `OXSI` and `LKDW` are collinear proxies of the same concept (mobility restrictions) and should not be included together, so $q_1 0$ has two modalities: control for `OXSI` and control for `LKDW`,[8]. This is an example of a case where a theory leads to reject a 'false' modality.

4.3 Results

In Table 1 are reported the 9 q.

The maximum number of 0 in the strings is 6. The number of specifications in the multiversal posterior distribution is then $2^6 * 3^2 * 5 = 2280$.

The estimates of the regression coefficient b_x and their p-values are represented in the upper section of the specification curve of the multiverse in Fig. 7.

The specification curve shows two jumps in the slope at its extremes and a linear slope in the middle. The absence of jumps in the middle, plus the scarcity of not significant estimates leads to think that the hypothesis H_1 is plausible, even if not without uncertainty. The Janus effect is neither trivial nor alerting: only 14% (413) of the specifications have a significant positive b_x, associated with an increase in the risk of death

[8] All of `OXSI` `LKDW`, and `pos_rt` can be pre-processed as rates in the unit interval, hence they are not logged. Other pertinent covariates in the dataset have a high rate of missing values.

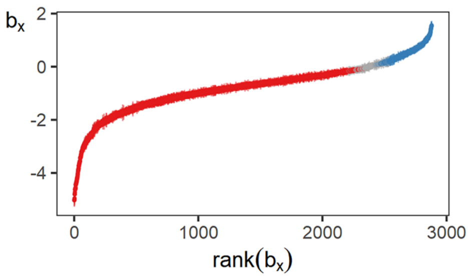

Fig. 7 Upper section of the specification curve of the multiverse. This curve represents the concept of variability across two dimensions: on the vertical axis is represented the range of the confidence interval of the estimate; on the horizontal axis, the slope represents the variability across specifications. p-values are not very informative, since almost any specification is statistically significant at $\alpha = .05$

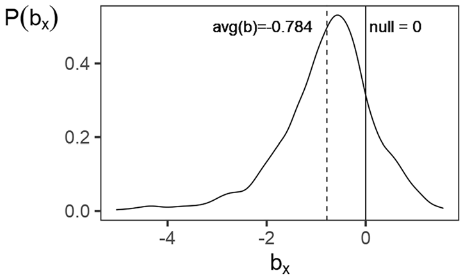

Fig. 8 Distribution of b in the multiverse of this application. Following Appendix A, coefficients can be interpreted as hazard ratios. A b_x inferior to 0 implies a reduction in the risk of death. Instead, interpreting the coefficients in the scale $1 - \exp(b_x)$ it is possible to estimate an effect size of the average coefficients of the vaccination treatment as a reduction of .54 of the risk of death. The median estimate is associated with an effect size of a reduction of .49

after vaccination. 79% (2285) have a significant negative estimate, associated with a decrease in the risk of death after vaccination. Given that the hypothesis of a reduction of the risk is much more credible, the multiverse data can be employed *a posteriori* to ascertain potential sources of bias, through a sensitivity analysis. The variation of the estimates across their own range of variation is represented in Fig. 8. Compared to the upper section of the specification curve, this representation is more attuned to a parametric interpretation of the estimates.

The estimate of the effect size $1 - \exp(\bar{b}_x) = .54$ is significantly lower than the majority values of vaccine effectiveness reported in the literature, although it does not fall outside the confidence intervals of most of the literature (which can also reach lower bounds equal to .2). Methodological differences in estimation methods between *in vivo* monitoring experiments and panel regression do not emerge in the literature (Jabłońska et al. 2021; Tregoning et al. 2021).

An alternative hypothesis is that authors over-focused on some countries associated with the administration of peculiar brands of vaccines and overestimated the overall effectiveness against death of the state-of-the-art of vaccine technology against COVID-19, especially against late virus strains.

Table 2 Sensitivity analysis: insensitive analytical decisions

q	Decision	Modality	n	$\bar{b}_x$	$s^2(b_x)$
1	Measure for vaccination plan	Full	1440	−0.80	0.69
1	Measure for vaccination plan	Minimal	1440	−0.77	1.12
2	Natural zeroes imputation	0	1440	−0.88	0.81
2	Natural zeroes imputation	1	1440	−0.68	0.98
3	Linear interpolation of `na`	0	1440	−0.84	0.85
3	Linear interpolation of `na`	1	1440	−0.72	0.95
5	Lagging schemes	7	960	−0.80	0.92
5	Lagging schemes	14	960	−0.78	0.90
5	Lagging schemes	21	960	−0.78	0.89
9	Lockdown policies	LKDW	1440	−0.84	0.98
9	Lockdown policies	OXSI	1440	−0.73	0.82

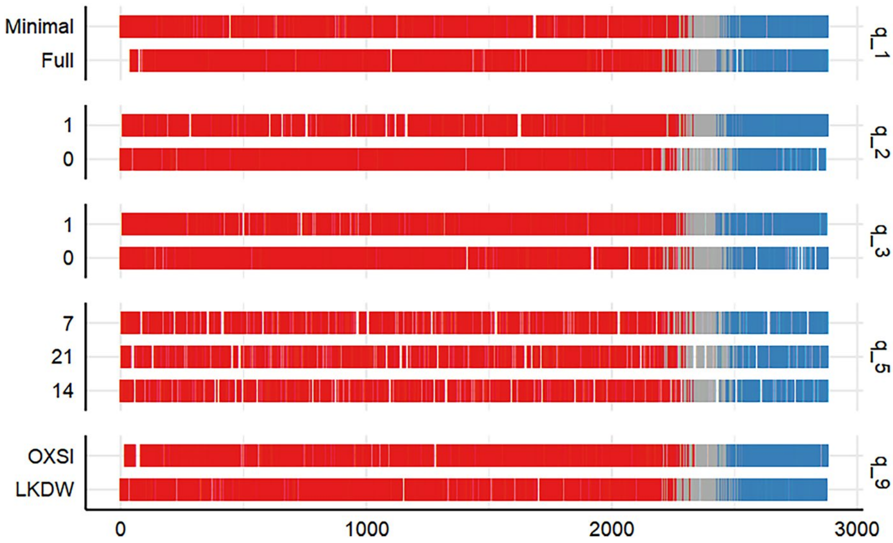

Fig. 9 Lower section of specification curve of the multiverse: insensitive analytical decisions

4.3.1 Sensitivity to modalities

The multiversal model has 9 analytical choices or q. These are divided into a first group of insensitive decisions, and a second group that shows more sensitivity to their own modalities.

The first group is represented in Fig. 9. Summary statistics are proposed in Table 2. q_1 and q_9 are operative definitions of abstract concepts, q_2 and q_3 are other operations of pre-processing. q_5 regards true conceptual differences in the regression model.

The second group is made of the other four analytical decisions relevant choices are provided in Table 3 and Fig. 10.

q_6 is the choice with more alternative modalities, and possibly the most complex, as highlighted in Sect. 4.2. "FE" is the only modality in the whole sample multiverse that is never associated with a significant positive estimate (see, Fig. 10), so a convenient

Table 3 Sensitivity analysis: sensitive analytical decisions

q	Decision	Modality	n	$\bar{b}_x$	$s^2(b_x)$
4	Overdispersion's correction	NB	1440	−0.62	0.50
4	Overdispersion's correction	QP	1440	−0.95	1.26
6	Panel Estimator	0	576	−0.14	0.51
6	Panel Estimator	`pop_d`	576	−0.30	0.47
6	Panel Estimator	`age`	576	−0.80	0.36
6	Panel Estimator	`pop_d` & `age`	576	−0.94	0.48
6	Panel Estimator	FE	576	−1.74	1.12
7	Time control	0	960	−0.52	0.24
7	Time control	Conventional	960	−0.57	1.37
7	Time control	After onset	960	−1.26	0.76
8	Infected rate control	0	1440	−0.95	1.08
8	Infected rate control	1	1440	−0.62	0.67

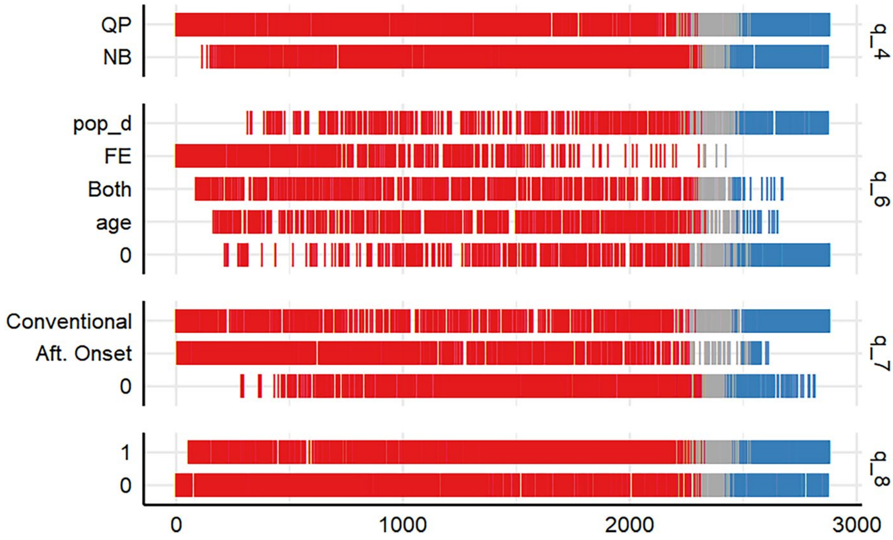

Fig. 10 Lower section of specification curve of the multiverse—sensitive analytical decisions

solution to avoid Janus effect would be to restrict the multiverse at the 576 specifications processed with a FE estimator. Indeed, the effect size in relative risk reduction is $1 - \exp(-1.74) = 82\%$, which is a value close to the prior of $75 \sim 80\%$.

Accepting the assumption that FE is a methodological approximation of including 'as much fixed-within-groups (Z_g) variables' as possible, then the progression of $\bar{b}_x$ in Table 3 assesses that in the sample multiverse the magnitude of the estimate increases with the inclusion of Z_g variables in the model.

Negative Binomial has significantly higher estimates (less risk reduction), but this may be an effect of the amplification of the magnitude, since overdispersion brings significantly higher variance, too. Similar considerations hold for controlling for the infection rate (q_8).

Finally, conventional FE control for the time variable fails to achieve a high magnitude of effect compared to no control, however the properly modelled control "After onset"

Table 4 Sensitivity to absent features

$n(m \neq 0)$	$n(j)$	$\bar{b}_x$	$s^2(b_x)$
0	24	−0.77	0.03
1	216	−0.66	0.35
2	696	−0.74	0.83
3	1032	−0.79	1.00
4	720	−0.82	1.02
5	192	−0.91	0.95
< 2	936	−0.72	0.70
2	1032	−0.79	1.00
> 2	912	−0.84	1.00

achieves to reach lower estimates, similarly to q_4. This is an analytical decision when diverging from the convention seems a much more sensible decision, because it makes no sense to account effects of a vaccination plan before it even started. This is also an example of an analytical decision that can easily be overlooked if the data-generating process is not very well understood, and including all three modalities in the Multiverse helps to assess the impact of this modelling choice.

Selecting only those 192 estimates when the variance within groups is corrected through the Fixed Effect estimator and when time-effects are controlled accounting for the onset of the vaccination campaign, too, the new effect size for the relative risk reduction is 88%. It is farther than 82% from the prior of $75 \sim 80\%$, but it is still closer to it than the average estimate of effect size in the whole multiverse (54%) or its median (49%).

4.3.2 Sensitivity to the absence of features

In Table 4 is checked the assumption that location and variance are correlated across classes of estimates grouped by the frequency of "0" in their string. In Table 4, groups proceed from the simplest to the most complex.

This negative correlation is indisputable, although it is more ambiguous when estimates are re-grouped into only three classes of similar size. Assuming as a reference that the multiverse has a positive bias, the lower values of the estimates in the latter classes confirm that more complex specifications of the multiverse, in this case, are slightly less biased. As a reference: $1 - \exp(-0.91) = 60\%$.

5 Conclusions, limitations and future developments

In this study, the methodological paradigm of Multiverse Analysis has been linked to a theoretical framework of sampling. It has been demonstrated that under a rigorous classification of the analytical decisions involved in the procedure of modelling the specifications of a multiverse, some non-trivial proprieties of the specifications emerge. These proprieties are: (1) specifications within a multiverse can be coarsely classified with a degree of complexity; (2) two specifications have a measurable distance between their associated result statistics but also the distance a priori is measurable.

In the application, it has been demonstrated that the classification of the analytical decision is not self-evident and that a typical procedure of sensitivity analysis through the observation of clustered multiversal statistics can lead to a reasonable selection of a more credible subset of specifications within the multiverse. A simple procedure to check the parametric assumption for adopting the mean estimate of a multiverse is discussed through the results of the application.

The paper aims to illustrate two divergent approaches to the application of multiverse methods. One approach, championed by Del Giudice and Gangestad (2021), involves the development of tailored multiverses that carefully reflect the causal assumptions underlying the data-generating process. In contrast, the other approach, advocated by Young and Holsteen (2017) or Patel et al. (2015), focuses on constructing large multiverse models. The current study demonstrates that both approaches have provided valid results, and both are objectively more useful than attempting to construct a single, optimal specification a priori. The balance between a priori feature inclusion and post-selection has successfully resolved the Janus effect and brilliantly shifted the average estimate closer to the prior value. Moreover, the paper provides a tentative confirmation that the arithmetic mean of estimates from a large multiverse cancels out different sources of bias, even if much more evidence is needed to assess this definitely.

To synthesise the contribution of the theory of multiversal methods and multiversal modelling, the following analogy is proposed: multiversal modelling of analytical decisions acts as a map of what Young and Holsteen (2017) refers to as the 'model space', while the operations of clustering and selection act as suggested 'paths' to follow in order to interpret the estimates. The selection of a subset of the multiverse should be interpreted as a suggestion from the authors that helps the reader to understand the technical and methodological conditions of validity of a scientific thesis. The 'multiverse-map' makes more transparent the theoretical link between a scientific claim and the proposed inferential procedures, helping to refute or to correct the procedures (or the claims!) if necessary.

5.1 Weighting schemes

Through the paper lurks a recurring theme that is often neglected in the literature on multiversal methods, robustness checks, and sensitivity analysis. This is the topic of the epistemological trade-off between prior model identification and posterior model selection in multimodel inference, where analysts identify a set of valid models to support or refute a scientific claim but prefer to account for comparative information from multiple models instead of focusing on only one. While this topic may not be at the core of multiverse literature, the exchange of papers between Slez (2019) and Young (2019) provides valuable insights into the limitations of multiversal models. As such, it has significant implications for the future development of the theory of the Multiverse.

Slez's argument starts as a criticism of a family of measures of the Robustness Ratio $\frac{\mathbb{E}(\theta)}{\sigma_{TOT}(\theta)}$, originally proposed by Young and Holsteen (2017). σ_{TOT} represents the 'total' standard error, that combines sampling and model errors.[9] Young and Holsteen propose different measures with slightly different assumptions. More or less any employment of it is based

[9] One way to think about the total error is as a function of the squares of the sampling error and model error. Specifically, it can be seen as the hypotenuse of a right triangle where the two components of the error are orthogonal segments: $\sigma_{TOT} = \sqrt{\sigma_k^2 + \sigma_j^2}$. The geometric interpretation of this measure is fertile soil to better understand advanced measures of heterogeneity variance between classes of specifications in a multiverse.

on epistemological premises similar to Multiverse Analysis (see, Fig. 1). In addition, the authors extend the heuristic to assert that the magnitude of a $\hat{\theta}_j$ or an average $\bar{\theta}_{CLASS}$ must be parameterised after σ_{TOT} in order to not reject *a posteriori* the associated specifications.

To understand the main lines of criticism of this approach, a strong argument is that a simple idea of 'robustness' implied in Young and Holsteen (2017) seems to place no value in any procedure to re-calibrate the multiversal estimates (Western 2018), which in practice translates into the refusal to weight the estimates through any scheme that is not uniform weights (unweighted estimates). For example, Young (2019) is critical of the proposal of Slez (2019) to calibrate the multiverse through a quantity that derives from the Information Criteria:

$$w_\theta = \frac{\exp\left(-0.5 \cdot \Delta(I_j)\right)}{\sum_{j=1}^{J} \exp\left(-0.5 \cdot \Delta(I_j)\right)} \tag{7}$$

whereas I is a statistic of information of the specification (Bayesian Information Criterion in the original) and Δ is a function of distance from the global minimum of I_j within the multiverse. Young (2019) argues that this method, being still based on fitting the theoretical model over the sample at disposal, does not overcome canonical problems like omitted variable bias. Furthermore, the exponential form of the function induces an extreme model selection, with few specifications counting for the majority of the weight.

In this debate, it is important to distinguish between two different topics. The first concerns the use of non-uniform weighting schemes *vs* uniform weights (unweighted estimates). This aspect of the debate largely centers around epistemological arguments and may reflect two different scientific cultures. As aforementioned, one culture is focused on transparently representing subjective multiversal models, while the other is concerned with interpreting evidence correctly. Those who prioritise accurate representations of their original ideas will favour an unweighted multiverse, to demonstrate that they are not 'hacking' a specific result. In this view, paradoxically a 'bad model' may even enrich the multiverse by demonstrating the conceptual robustness of scientific ideas, and therefore the multiverse should be large and unweighted.

The second dichotomy is about how to weigh the multiverse. This debate is still relatively new but potentially crucial for the future developments of the Multiverse paradigm. Slez's proposal, specifically, converges towards a supposedly excessive focus on a few specifications. If that is the case, it may be concerning if these 'strong' estimates do not reflect a theoretical coherence, i.e. they are strings with a high mutual Hamming's distance. Generally, the employment of prior distance to evaluate the coherence of a weighting scheme seems a valuable contribution.

Muñoz and Young (2018a) suggest, instead, to amplify the relevance of those specifications that have a great impact on the estimate. Following this line of thinking (the authors prefer to not connect the proposal to a specific functional form), the theoretical developments of this manuscript could be important for reconsidering the concept of 'impact' or 'importance' on local portions of the multiverse, instead of global. Referencing Eq. 7, rather than considering the global function Δ, it can be considered a local function δ_d in the subset of specifications at a certain d Hamming distance to j (see, Fig. 5). For those who refute weighting on fit or information statistics, other local statistics can be considered instead of I.

Of particular relevance is the subset for $d = 1$, because it would reflect a specific explicating model for the misspecification error: p-hacking. The most parsimonious and less detectable method to drive results towards the desired agenda is to alter minimally the specification of the model. In the exemplary application, it is demonstrated that the choice of treating fixed Z_g as random values, not opting for FE estimation, would have a huge impact on raising the estimate, to the point to allow significant positive estimates. In addition, Burnham and Anderson (2002) deter to compare information criteria with their 'quasi' counterparts. This is an occurrence in the application, whereas Negative Binomial is compared to Quasi Poisson. Although this suggestion is arguable, adjusted local statistics may overcome the strict imposition of a unique $\arg\min(I)$ in Eq. 7.

Considering that the number of zeroes in a specification string counts as a measure of the complexity of the model, these proposals should be in line with advanced frameworks for sensitivity analysis. Another future direction of multiverse analysis is towards understanding heterogeneity variance across the features involved in multiversal modelling (Saltelli and Annoni 2010; Veroniki et al. 2016; Saltelli et al. 2021; Langan et al. 2019). A natural development would focus on conditional modalities $(m_{q_1} \mid m_{q_2})$.

5.2 Multi-teams multiverses

The value of representing specifications as strings is most evident in multi-teams multiverses (Breznau et al. 2022). These multiverses are identified by pooling together other sets of specifications (and fit statistics) which are provided by different but coordinated, teams of analysts. The structure of the pooled sample of specifications across multi-teams follows the guidelines mandated by the coordinator, who can ask teams to follow strictly the rules reported in Sect. 3.3, or not. In the latter case, a team could decide to include strings "0AA" and "0AB" but not "0A0".

The temptation is to interpret the frequency of a modality within the pool of multiverses as an indication of its scientific relevance as if a prior weight emerges from the equivalence with this frequency. It is important to exercise prudence with this interpretation of the pooled multiverse. The frequency of certain modalities may be influenced by factors other than the nature of the scientific problem at hand. A potential issue is that teams may have the freedom to select more than one unique $\mathbf{q}$ set of analytical decisions. To address this, the absence of a certain q in a team's multiverse can be interpreted as if the value of the corresponding modality is set to 0. In this case, there may be an over-representation of modalities set to 0. This would reflect a tendency of teams to provide simpler models.

It is easier for a team to overlook a possible q, rather than to mistakenly add a q that they did not actually intend to include. For example, let's consider two scientific hypotheses. The first hypothesis asserts that (1) "vaccination plans reduce deaths in infected individuals", while the second hypothesis asserts that (2) "a vaccine shot reduces the probability of dying after infection". Suppose that the most appropriate specification for hypothesis (1) is a regression model that includes the number of vaccine shots (x) adjusted by a third variable (z) as control, with the number of infected deaths (y) as the dependent variable. On the other hand, the correct specification for hypothesis (2) would exclude controlling for z. Now, suppose that one team correctly includes a logical q regarding the presence and absence of z in the model for (i), while another team interprets "vaccination plan" as something that can be proxied by x alone. Thus they only always exclude z in the model. As a result, the pooled multiverse would show an inflated frequency of the analytical $q_z = 0$.

The latter can be accounted as a stochastically erratic occurrence: sometimes teams misinterpret the research questions. However, given the definition of hypothesis (2), much simpler and clean-cut than (i), it would be bizarre if one of the teams could opt to include q_z. So, it is much more unlikely that the frequency of $q_z = 1$ is inflated, compared to $q_z = 0$. A more comprehensive list of arguments regarding the correct interpretations of multiversal modelling in multi-teams is in Auspurg and Brüderl (2021).

6 Supplementary information

The contents of the manuscript can be reproduced by executing in R Studio the `RMarkdown` scripts downloadable at `https://figshare.com/s/cb9767ce5fd19c0c0c10`. The first author re-coded some commands from the package `specr`, Version 0.2.2 (Masur and Scharkow 2020).

To fully reproduce the results, firstly execute all the code-chunks of `Preprocessing.Rmd`, then execute all the code-chunks of `Multiverse_preprocessing.Rmd`.

The output of these pre-processing files should be the same contained in the file `multiverse.Rdata`; so, a shortcut is just to load `multiverse.Rdata` in R Studio. Finally, proceed to execute `Results.Rmd` to reproduce figures and tables of results.

Appendix: Poisson coefficients as hazard ratios

Relative risk R of binary outcome (y) after administration of a binary treatment (x) is:

$$R = \frac{Pr.(y = 1) \mid x = 1}{Pr.(y = 1) \mid x = 0} \tag{8}$$

Probabilities can be estimated with empirical frequencies, so R can be estimated with a hazard ratio h:

$$R \sim \hat{h} = \frac{\hat{y} \mid x = 1}{\hat{y} \mid x = 0} \tag{9}$$

This assumption allows y to be a count variable instead of a binary outcome. The assumption that allows to estimate $\hat{h}$ through the $\hat{b}$ of a Poisson regression is that x is not a binary input but a relative frequency that can assume any value in the unit interval. The technical assumption is: for a large population the response of a partial treatment on a full population is technically the same as the response of a full treatment on a part of the population. If this assumption holds, then the effect of the full treatment can be estimated through the difference of expectation in the count response in two hypotheses:

1. all population is treated (y_1)
2. nobody in the population is treated (y_0)

that is already expressed in Eq. 9. Since y is a count, it holds the Poisson assumption that the link function between response and treatment is $\exp(b \cdot x)$, which allows further technical operations of simplification:

$$\hat{h} = \frac{\hat{y_1}}{\hat{y_0}} \sim \frac{\exp(b \cdot x_1)}{\exp(b \cdot x_0)} \tag{10}$$

and because

$$
\begin{aligned}
x_1 &= 1 \\
x_0 &= 0
\end{aligned}
\tag{11}
$$

then from Eqs. 9, 10, and 11, it follows

$$
\hat{h} \sim \frac{\exp(b)}{\exp(0)} = \exp(b) \tag{12}
$$

This is the full set of assumptions to derive that the exponentiation of the Poisson coefficient is a hazard ratio h. If this is true, to facilitate interpretation of the regression, it is adopted the relative risk reduction of treatment:

$$
1 - h = 1 - \exp(b) \tag{13}
$$

Interpretation of risk reduction is straightforward because it can be reconnected to the individual case: the *null* is 0 (not 1 as for h) and the value is the likelihood of a reduction in unit outcome after a condition is switched from untreated to treated. A negative risk reduction would imply that one additional treatment would induce a surge of outcomes.

In presence of overdispersion in the counts of y, it is suggested to adopt an advanced estimator of the variance of y in the regression engine. The two standards in the literature are Negative Binomial (NB) and Quasi-Poisson (QP).

The negative Binomial's assumption of Variance in the count of y is:

$$
s^2(y) \sim \bar{y} + (\phi \cdot \bar{y})^2 \tag{14}
$$

Quasi-Poisson's assumption of Variance in the count of y is:

$$
s^2(y) \sim \phi \cdot \bar{y} \tag{15}
$$

The method of estimation of *phi* is different between the two: Negative Binomial allows Maximum Likelihood Estimation (MLE), but Quasi-Poisson recurs to a quasi-Maximum Likelihood Estimation (qMLE). As a result, Information Criteria cannot be correctly computed, and usually, they are estimated through their quasi-Information Criteria counterparts (Gay and Welsch 1988; Burnham and Anderson 2002). Yet, in the scientific literature about modelling, these corrected estimators are treated as equivalent, even if generally QP is reported to be slightly more robust (Land et al. 1996; Ver Hoef and Boveng 2007; Ibarra-Espinosa et al. 2022).

Funding Open access funding provided by Università degli Studi di Catania within the CRUI-CARE Agreement. The authors have not disclosed any funding.

Declarations

Conflict of interest The authors have not disclosed any conflict of interest.

References

Agresti, A., Caffo, B., Ohman-Strickland, P.: Examples in which misspecification of a random effects distribution reduces efficiency, and possible remedies. Comput. Stat. Data Anal. **47**(3), 639–653 (2004). https://doi.org/10.1016/j.csda.2003.12.009

Allison, P.: Fixed Effects Regression Models. SAGE Publications Inc., Thousand Oaks (2009). https://doi.org/10.4135/9781412993869

Allison, P.D., Waterman, R.P.: Fixed-effects negative binomial regression models. Sociol. Methodol. **32**(1), 247–265 (2002). https://doi.org/10.1111/1467-9531.00117

Aronow, P.M., Miller, B.T.: Foundations of Agnostic Statistics. Cambridge University Press, Cambridge (2019)

Athey, S., Imbens, G.: A measure of robustness to misspecification. Am. Econ. Rev. **105**(5), 476–480 (2015). https://doi.org/10.1257/aer.p20151020

Auspurg, K., Brüderl, J.: Has the credibility of the social sciences been credibly destroyed? Reanalyzing the "many analysts, one data set'' project. Socius **7**(23780231211024), 421 (2021). https://doi.org/10.1177/23780231211024421

Belkin, M., Hsu, D., Ma, S., et al.: Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proc. Natl. Acad. Sci. **116**(32), 15849–15854 (2019). https://doi.org/10.1073/pnas.1903070116

Bookstein, A., Kulyukin, V.A., Raita, T.: Generalized hamming distance. Inf. Retr. **5**(4), 353–375 (2002). https://doi.org/10.1023/A:1020499411651

Box, G.E.P.: Science and statistics. J. Am. Stat. Assoc. **71**(356), 791–799 (1976). https://doi.org/10.1080/01621459.1976.10480949

Breznau, N.: I saw You in the crowd: credibility, reproducibility, and meta-utility. PS Polit. Sci. Polit. **54**(2), 309–313 (2021). https://doi.org/10.1017/S1049096520000980

Breznau, N., Rinke, E.M., Wuttke, A., et al.: Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. Proc. Natl. Acad. Sci. **119**(44), e2203150,119 (2022). https://doi.org/10.1073/pnas.2203150119

Broström, G., Holmberg, H.: Generalized linear models with clustered data: fixed and random effects models. Comput. Stat. Data Anal. **55**(12), 3123–3134 (2011). https://doi.org/10.1016/j.csda.2011.06.011

Burnham, K.P., Anderson, D.R.: Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2nd edn. Springer, New York (2002)

Burton, J.W., Cruz, N., Hahn, U.: Reconsidering evidence of moral contagion in online social networks. Nat. Hum. Behav. **5**(12), 1629–1635 (2021). https://doi.org/10.1038/s41562-021-01133-5

Camerer, C.F., Dreber, A., Holzmeister, F., et al.: Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat. Hum. Behav. **2**(9), 637–644 (2018). https://doi.org/10.1038/s41562-018-0399-z

Christensen, G., Freese, J., Miguel, E.: Transparent and Reproducible Social Science Research: How to Do Open Science, 1st edn. University of California Press, Berkeley (2019)

Collins, H.: Changing Order: Replication and Induction in Scientific Practice, reprint, edition University of Chicago Press, Chicago (1992)

Cosme, D., Lopez, R.B.: Neural indicators of food cue reactivity, regulation, and valuation and their associations with body composition and daily eating behavior. Soc. Cogn. Affect. Neurosci. (2020). https://doi.org/10.1093/scan/nsaa155

Czado, C., Santner, T.J.: The effect of link misspecification on binary regression inference. J. Stat. Plan. Inference **33**(2), 213–231 (1992). https://doi.org/10.1016/0378-3758(92)90069-5

Dagan, N., Barda, N., Kepten, E., et al.: BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. N. Engl. J. Med. **384**(15), 1412–1423 (2021). https://doi.org/10.1056/NEJMoa2101765

Del Giudice, M., Gangestad, S.W.: A traveler's guide to the multiverse: promises, pitfalls, and a framework for the evaluation of analytic decisions. Adv. Methods Pract. Psychol. Sci. **4**(1), 2515245920954,925 (2021). https://doi.org/10.1177/2515245920954925

Ding, P., Miratrix, L.W.: To adjust or not to adjust? Sensitivity analysis of M-bias and butterfly-bias. J. Causal Inference **3**(1), 41–57 (2015). https://doi.org/10.1515/jci-2013-0021

Dong, E., Du, H., Gardner, L.: An interactive web-based dashboard to track COVID-19 in real time. Lancet. Infect. Dis. **20**(5), 533–534 (2020). https://doi.org/10.1016/S1473-3099(20)30120-1

Durante, K.M., Rae, A., Griskevicius, V.: The fluctuating female vote: politics, religion, and the ovulatory cycle. Psychol. Sci. **24**(6), 1007–1016 (2013). https://doi.org/10.1177/0956797612466416

Durlauf, S., Fu, C., Navarro, S.: Capital punishment and deterrence: understanding disparate results. J. Quant. Criminol. (2012). https://doi.org/10.1007/s10940-012-9171-0

Earp, B.D., Trafimow, D.: Replication, falsification, and the crisis of confidence in social psychology. Front. Psychol. **6**, 621 (2015). https://doi.org/10.3389/fpsyg.2015.00621

Elwert, F., Winship, C.: Endogenous selection bias: the problem of conditioning on a collider variable. Ann. Rev. Sociol. **40**(1), 31–53 (2014). https://doi.org/10.1146/annurev-soc-071913-043455

Fan, X., Sivo, S.A.: Sensitivity of fit indices to model misspecification and model types. Multivar. Behav. Res. **42**(3), 509–529 (2007). https://doi.org/10.1080/00273170701382864

Fiolet, T., Kherabi, Y., MacDonald, C.J., et al.: Comparing COVID-19 vaccines for their characteristics, efficacy and effectiveness against SARS-CoV-2 and variants of concern: a narrative review. Clin. Microbiol. Infect. **28**(2), 202–221 (2022). https://doi.org/10.1016/j.cmi.2021.10.005

Gardenier, J., Resnik, D.: The misuse of statistics: concepts, tools, and a research agenda. Account. Res. **9**(2), 65–74 (2002). https://doi.org/10.1080/08989620212968

Gay, D., Welsch, R.: Maximum likelihood and quasi-likelihood for nonlinear exponential family regression models. J. Am. Stat. Assoc. **83**(404), 990–998 (1988). https://doi.org/10.1080/01621459.1988.10478690

Gelman, A.: The connection between varying treatment effects and the crisis of unreplicable research: a Bayesian perspective. J. Manag. **41**(2), 632–643 (2015). https://doi.org/10.1177/0149206314525208

Gelman, A., Hill, J.: Data Analysis Using Regression and Multilevel/Hierarchical Models, 1st edn. Cambridge University Press, Cambridge (2007)

Gelman, A., Loken, E.: The statistical crisis in science. Am. Sci. **102**(6), 460–466 (2014)

Guidotti, E., Ardia, D.: COVID-19 data hub. J. Open Source Softw. **5**(51), 2376 (2020). https://doi.org/10.21105/joss.02376

Haas, E.J., Angulo, F.J., McLaughlin, J.M., et al.: Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. The Lancet **397**(10287), 1819–1829 (2021). https://doi.org/10.1016/S0140-6736(21)00947-8

Hale, T., Angrist, N., Goldszmidt, R., et al.: A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). Nat. Hum. Behav. **5**(4), 529–538 (2021). https://doi.org/10.1038/s41562-021-01079-8

Hall, B.D., Liu, Y., Jansen, Y., et al.: A survey of tasks and visualizations in multiverse analysis reports. Comput. Graph. Forum **41**(1), 402–426 (2022). https://doi.org/10.1111/cgf.14443

Hausman, J., Hall, B.H., Griliches, Z.: Econometric models for count data with an application to the patents-R & D relationship. Econometrica **52**(4), 909–938 (1984). https://doi.org/10.2307/1911191

Head, M.L., Holman, L., Lanfear, R., et al.: The extent and consequences of p-hacking in science. PLoS Biol. **13**(3), e1002,106 (2015). https://doi.org/10.1371/journal.pbio.1002106

Hothorn, T., Bretz, F., Westfall, P.: Simultaneous inference in general parametric models. Biom. J. **50**(3), 346–363 (2008). https://doi.org/10.1002/bimj.200810425

Ibarra-Espinosa, S., Dias de Freitas, E., Ropkins, K., et al.: Negative-binomial and quasi-Poisson regressions between COVID-19, mobility and environment in São Paulo, Brazil. Environ. Res. **204**(112), 369 (2022). https://doi.org/10.1016/j.envres.2021.112369

Ioannidis, J.P.A.: Why most published research findings are false. PLoS Med. **2**(8), e124 (2005). https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J.P.A., Fanelli, D., Dunne, D.D., et al.: Meta-research: evaluation and improvement of research methods and practices. PLoS Biol. **13**(10), e1002,264 (2015). https://doi.org/10.1371/journal.pbio.1002264

Islam, N., Shkolnikov, V.M., Acosta, R.J., et al.: Excess deaths associated with covid-19 pandemic in 2020: age and sex disaggregated time series analysis in 29 high income countries. BMJ **373**, n1137 (2021). https://doi.org/10.1136/bmj.n1137

Jabłońska, K., Aballéa, S., Toumi, M.: The real-life impact of vaccination on COVID-19 mortality in Europe and Israel. Public Health **198**, 230–237 (2021). https://doi.org/10.1016/j.puhe.2021.07.037

James, G., Witten, D., Hastie, T., et al.: An Introduction to Statistical Learning: with Applications in R, 1st edn. Springer, New York (2013)

Lagakos, S.: Effects of mismodelling and mismeasuring explanatory variables on tests of their association with a response variable. Stat. Med. **7**(1–2), 257–274 (1988). https://doi.org/10.1002/sim.4780070126

Land, K., McCall, P.L., Nagin, D.S.: A comparison of Poisson, negative binomial, and semiparametric mixed Poisson regression models: with empirical applications to criminal careers data. Sociol. Methods Res. **24**(4), 387–442 (1996). https://doi.org/10.1177/0049124196024004001

Langan, D., Higgins, J.P., Jackson, D., et al.: A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Res. Synth. Methods **10**(1), 83–98 (2019). https://doi.org/10.1002/jrsm.1316

Leamer, E.E.: Let's take the con out of econometrics. Am. Econ. Rev. **73**(1), 31–43 (1983)

Leamer, E.E.: Sensitivity analyses would help. Am. Econ. Rev. **75**(3), 308–313 (1985)

Lipsitch, M., Krammer, F., Regev-Yochay, G., et al.: SARS-CoV-2 breakthrough infections in vaccinated individuals: measurement, causes and impact. Nat. Rev. Immunol. **22**(1), 57–65 (2022). https://doi.org/10.1038/s41577-021-00662-4

Liu, W., Brookhart, M.A., Schneeweiss, S., Mi, X., Setoguchi, S.: Implications of M bias in epidemiologic studies: a simulation study. Am. J. Epidemiol. **176**(10), 938–948 (2012). https://doi.org/10.1093/aje/kws165

Lundberg, I., Johnson, R., Stewart, B.M.: What is your estimand? Defining the target quantity connects statistical evidence to theory. Am. Sociol. Rev. **86**(3), 532–565 (2021). https://doi.org/10.1177/00031224211004187

Masur PK, Scharkow M (2020) specr: conducting and visualizing specification curve analyses

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., et al.: A global database of COVID-19 vaccinations. Nat. Hum. Behav. **5**(7), 947–953 (2021). https://doi.org/10.1038/s41562-021-01122-8

McShane, B.B., Gal, D., Gelman, A., et al.: Abandon statistical significance. Am. Stat. **73**(sup1), 235–245 (2019). https://doi.org/10.1080/00031305.2018.1527253

Munafò, M.R., Tilling, K., Taylor, A.E., et al.: Collider scope: when selection bias can substantially influence observed associations. Int. J. Epidemiol. **47**(1), 226–235 (2018). https://doi.org/10.1093/ije/dyx206

Muñoz, J., Young, C.: Rejoinder: can we weight models by their probability of being true? Sociol. Methodol. **48**(1), 43–51 (2018). https://doi.org/10.1177/0081175018796841

Muñoz, J., Young, C.: We ran 9 billion regressions: eliminating false positives through computational model robustness. Sociol. Methodol. **48**(1), 1–33 (2018). https://doi.org/10.1177/0081175018777988

Nissen, S.B., Magidson, T., Gross, K., et al.: Publication bias and the canonization of false facts. eLife **5**, e21,451 (2016). https://doi.org/10.7554/eLife.21451

Nosek, B.A., Bar-Anan, Y.: Scientific Utopia: I. Opening scientific communication. Psychol. Inq. **23**(3), 217–243 (2012). https://doi.org/10.1080/1047840X.2012.692215

Olliaro, P., Torreele, E., Vaillant, M.: COVID-19 vaccine efficacy and effectiveness: the elephant (not) in the room. The Lancet Microbe **2**(7), e279–e280 (2021). https://doi.org/10.1016/S2666-5247(21)00069-0

Olsson, U., Foss, T., Troye, S., et al.: The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. Struct. Equ. Model. **7**(4), 557–595 (2000). https://doi.org/10.1207/S15328007SEM0704_3

OPEN Science COLLABORATION: Estimating the reproducibility of psychological science. Science **349**(6251), aac4716 (2015). https://doi.org/10.1126/science.aac4716

Orben, A., Przybylski, A.K.: The association between adolescent well-being and digital technology use. Nat. Hum. Behav. **3**(2), 173–182 (2019). https://doi.org/10.1038/s41562-018-0506-1

Palpacuer, C., Hammas, K., Duprez, R., et al.: Vibration of effects from diverse inclusion/exclusion criteria and analytical choices: 9216 different ways to perform an indirect comparison meta-analysis. BMC Med. **17**(1), 174 (2019). https://doi.org/10.1186/s12916-019-1409-3

Patel, C.J., Burford, B., Ioannidis, J.P.A.: Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. J. Clin. Epidemiol. **68**(9), 1046–1058 (2015). https://doi.org/10.1016/j.jclinepi.2015.05.029

Patel, M.K., Bergeri, I., Bresee, J.S., et al.: Evaluation of post-introduction COVID-19 vaccine effectiveness: summary of interim guidance of the World Health Organization. Vaccine **39**(30), 4013–4024 (2021). https://doi.org/10.1016/j.vaccine.2021.05.099

Peterson, D., Panofsky, A.: Metascience as a scientific social movement (2020). https://doi.org/10.31235/osf.io/4dsqa

Pham, M.T., Oh, T.T.: Preregistration is neither sufficient nor necessary for good science. J. Consum. Psychol. **31**(1), 163–176 (2021). https://doi.org/10.1002/jcpy.1209

Pormohammad, A., Zarei, M., Ghorbani, S., et al.: Effectiveness of COVID-19 vaccines against delta (B.1.617.2) variant: a systematic review and meta-analysis of clinical studies. Vaccines **10**(1), 23 (2022). https://doi.org/10.3390/vaccines10010023

Raftery, A.E.: Bayesian model selection in social research. Sociol. Methodol. **25**, 111–163 (1995). https://doi.org/10.2307/271063

Rao, P.: Some notes on misspecification in multiple regressions. Am. Stat. **25**(5), 37–39 (1971). https://doi.org/10.1080/00031305.1971.10477302
Rodamar, J.: There ought to be a law! Campbell versus Goodhart. Significance **15**(6), 9–9 (2018). https://doi.org/10.1111/j.1740-9713.2018.01205.x
Rohrer, J.M., Egloff, B., Schmukle, S.C.: Probing birth-order effects on narrow traits using specification-curve analysis. Psychol. Sci. **28**(12), 1821–1832 (2017). https://doi.org/10.1177/0956797617723726
Rosenthal, R.: The file drawer problem and tolerance for null results. Psychol. Bull. **86**(3), 638–641 (1979). https://doi.org/10.1037/0033-2909.86.3.638
Ross, J.: Misuse of statistics in social sciences. Nature **318**(6046), 514–514 (1985). https://doi.org/10.1038/318514a0
Rubin, D.B.: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? Stat. Med. **28**(9), 1420–1423 (2009). https://doi.org/10.1002/sim.3565
Rubin, M.: When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. Rev. Gen. Psychol. **21**(4), 308–320 (2017). https://doi.org/10.1037/gpr0000128
Sala-I-Martin, X.X.: I just ran two million regressions. Am. Econ. Rev. **87**(2), 178–183 (1997)
Saltelli, A., Annoni, P.: How to avoid a perfunctory sensitivity analysis. Environ. Model. Softw. **25**(12), 1508–1517 (2010). https://doi.org/10.1016/j.envsoft.2010.04.012
Saltelli, A., Aleksankina, K., Becker, W., et al.: Why so many published sensitivity analyses are false: a systematic review of sensitivity analysis practices. Environ. Model. Softw. **114**, 29–39 (2019). https://doi.org/10.1016/j.envsoft.2019.01.012
Saltelli, A., Tarantola, S., Campolongo, F., et al.: Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models, 1st edn. New Publisher, Hoboken (2021)
Schor, S., Karten, I.: Statistical evaluation of medical journal manuscripts. JAMA **195**(13), 1123–1128 (1966). https://doi.org/10.1001/jama.1966.03100130097026
Schweinsberg, M., Feldman, M., Staub, N., et al.: Same data, different conclusions: radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. Organ. Behav. Hum. Decis. Process. **165**, 228–249 (2021). https://doi.org/10.1016/j.obhdp.2021.02.003
Shrier, I.: Letter to the editor. Stat. Med. **27**(14), 2740–2741 (2008). https://doi.org/10.1002/sim.3172
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. **22**(11), 1359–1366 (2011). https://doi.org/10.1177/0956797611417632
Simonsohn, U., Nelson, L.D., Simmons, J.P.: P-curve: a key to the file-drawer. J. Exp. Psychol. Gen. **143**(2), 534–547 (2014). https://doi.org/10.1037/a0033242
Simonsohn, U., Simmons, J.P., Nelson, L.D.: Specification curve analysis. Nat. Hum. Behav. **4**(11), 1208–1214 (2020). https://doi.org/10.1038/s41562-020-0912-z
Slez, A.: The difference between instability and uncertainty: comment on Young and Holsteen (2017). Sociol. Methods Res. **48**(2), 400–430 (2019). https://doi.org/10.1177/0049124117729704
Steegen, S., Tuerlinckx, F., Gelman, A., et al.: Increasing transparency through a multiverse analysis. Perspect. Psychol. Sci. **11**(5), 702–712 (2016). https://doi.org/10.1177/1745691616658637
Tierney, B.T., Anderson, E., Tan, Y., et al.: Leveraging vibration of effects analysis for robust discovery in observational biomedical data science. PLoS Biol. **19**(9), e3001,398 (2021). https://doi.org/10.1371/journal.pbio.3001398
Tregoning, J.S., Flight, K.E., Higham, S.L., et al.: Progress of the COVID-19 vaccine effort: viruses, vaccines and variants versus efficacy, effectiveness and escape. Nat. Rev. Immunol. **21**(10), 626–636 (2021). https://doi.org/10.1038/s41577-021-00592-1
Ver Hoef, J.M., Boveng, P.L.: Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology **88**(11), 2766–2772 (2007). https://doi.org/10.1890/07-0043.1
Verbeke, G., Lesaffre, E.: The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. Comput. Stat. Data Anal. **23**(4), 541–556 (1997). https://doi.org/10.1016/S0167-9473(96)00047-3
Veroniki, A.A., Jackson, D., Viechtbauer, W., et al.: Methods to estimate the between-study variance and its uncertainty in meta-analysis. Res. Synth. Methods **7**(1), 55–79 (2016). https://doi.org/10.1002/jrsm.1164
Wasserstein, R.L., Lazar, N.A.: The ASA statement on p-values: context, process, and purpose. Am. Stat. **70**(2), 129–133 (2016). https://doi.org/10.1080/00031305.2016.1154108
Wasserstein, R.L., Schirm, A.L., Lazar, N.A.: Moving to a World Beyond “$p < 0.05$”. Am. Stat. **73**(sup1), 1–19 (2019). https://doi.org/10.1080/00031305.2019.1583913
West, J.D., Bergstrom, C.T.: Misinformation in and about science. Proc. Natl. Acad. Sci. (2021). https://doi.org/10.1073/pnas.1912444117

Western, B.: Comment: Bayes, model uncertainty, and learning from data. Sociol. Methodol. **48**, 39–43 (2018)

Yamada, Y.: How to crack pre-registration: toward transparent and open science. Front. Psychol. (2018). https://doi.org/10.3389/fpsyg.2018.01831

Young, C.: The difference between causal analysis and predictive models: response to "Comment on Young and Holsteen (2017)''. Sociol. Methods Res. **48**(2), 431–447 (2019). https://doi.org/10.1177/0049124118782542

Young, C., Holsteen, K.: Model uncertainty and robustness: a computational framework for multimodel analysis. Sociol. Methods Res. **46**(1), 3–40 (2017). https://doi.org/10.1177/0049124115610347

van Zwet, E.W., Cator, E.A.: The significance filter, the winner's curse and the need to shrink. Stat. Neerl. **75**(4), 437–452 (2021). https://doi.org/10.1111/stan.12241