Mariusz Maziarz, Conflicting results and statistical malleability: embracing pluralism of empirical results
This is the Author Accepted Manuscript (AAM) version of the article published by *Perspectives on Science*:
https://doi.org/10.1162/posc_a_00627

Conflicting results and statistical malleability: embracing pluralism of empirical results

Mariusz Maziarz (corresponding author)

Jagiellonian University, Faculty of Philosophy, Institute of Philosophy & Interdisciplinary Centre for Ethics

mariusz.maziarz@uj.edu.pl

ORCiD: 0000-0003-1979-0746

Grodzka 52, Kraków, Poland

Abstract

Conflicting results undermine making inferences from the empirical literature. So far, the replication crisis is mainly seen as resulting from honest errors and questionable research practices such as p-hacking or the base-rate fallacy. We discuss the malleability (researcher degrees of freedom) of quantitative research and argue that conflicting results can emerge from two studies using different but plausible designs (e.g., eligibility criteria, operationalization of concepts, outcome measures) and statistical methods. We also explore how the choices regarding study design and statistical techniques bias results in a way that makes them more or less relevant for a given policy or clinical question.

Key-words: recalcitrant results, conflicting results, scientific pluralism

Conflict of Interest

This is the Author Accepted Manuscript (AAM) version of the article published by *Perspectives on Science*: https://doi.org/10.1162/posc_a_00627
No conflict of interest is reported by the authors.

1. Introduction

Conflicting results are frequent among clinical trials (RCTs) (Ioannidis 2005a; 2005b), psychological experiments (Wicherts et al. 2016), and statistical analyses of observational data in epidemiology (Tatsioni et al. 2007; Broadbent 2013), econometrics (Moosa 2019; Goldfarb 1997), ecological science (Robins et al. 2000), and global warming science (Power and Kociuba 2011). 'Conflicting results' refers to "studies reporting […] different sizes or even directions of effects" (Goldfarb 1995, p. 202) and can be considered as an aspect of the replication crisis (see Romero 2019). Recently, Bird (2021) suggested that the replication crisis emerges primarily from the base-rate fallacy and conflicting results can be reported even when individual studies do not include errors or fraudulent practices. Feest (2019) observed that some replication attempts are not identical to the original studies and the differences between studies account for the heterogeneity of reported results. Feest argued that "[s]uch research (of effects of different operationalizations), I suggest, can contribute to conceptual development by helping to explore and fine-tune the shape and scope of proposed or existing concepts. The fact that this is riddled with problems does not in and of itself constitute a crisis, let alone a replication crisis." (Feest 2019, p. 903).

However, even if each of the studies reporting conflicting results is reproducible, both scientists and philosophers view conflicting results as an obstacle to making plausible inferences from the empirical literature. For example, after discussing several examples of conflicting results of clinical trials, Ioannidis (2005a, p. 218) admitted that "such disagreements may upset clinical practice and acquire publicity in both scientific circles and in the lay press." Kulkarni (2005) highlighted that the movement of evidence-based medicine (EBM) waited for a solution to the problem of conflicting RCT results over a decade ago, but neither medicine nor other empirical disciplines have coped with

the problem of conflicting results. Existing approaches to systematic literature review and meta-analysis can shed some light on why studies report conflicting results but usually leave the heterogeneity unexplained and rely on averaging study outcomes (Feinstein 1995; Maziarz 2022) based on the assumption that differences arise due to random imbalances between the treatment and control groups. However, the problem of conflicting results is also frequent in disciplines dealing with observational data, with econometrics being the prime example for its use of publicly available datasets but still including studies reporting conflicting results. The prevalence of such cases made Goldfarb (1995, p. 203) conclude his paper on conflicting econometric results by asking: "how are plausible inferences to be drawn from empirical literature?"

We approach this question and argue that conflicting results emerge from the permissible malleability (researcher degrees of freedom) in statistical analysis and research design. What follows is that these different designs deliver responses to differing research questions, and hence, the results of studies that seem to be in conflict are not so. Our argument proceeds as follows. First, we analyze the reasons why conflicting results emerge. In section 2, we argue that designing a study and analyzing data requires undertaking several methodological decisions that are (to some extent) arbitrary. In section 3, we investigate how statistical results are malleable in the sense that choices leading to different results are plausible and discuss why researchers face incentives to commit to such choices that make obtaining results conflicting with previous evidence more likely. In section 4, we argue that these choices regarding study design and statistical techniques bias results in a way that makes them more or less relevant for a given policy or clinical question so that the results that seem to be in conflict are not so as they address different research questions.

2. The plausibility of alternative decisions in data research

In recent years, empirical researchers across several disciplines observed that data research (i.e., collecting data and using statistical methods) is 'malleable' in the sense that some choices (possibly leading to different results) are similarly warranted given statistical methodology and background

knowledge. While philosophers (Stegenga 2018; Williamson 2019) prefer the notion of malleability, the terminology used by researchers is inconsistent across the fields. For example, psychologists use the notion of researcher degrees of freedom (Wicherts et al. 2016). Medical researchers use the notion of biases (in the sense that one or some other choice biases result in one way or another. Economists discuss the problem of results' fragility, understood as their dependence on different choices (of, for example, specification or estimation technique) (Moosa 2019). The latter notion can also be encountered in the medical literature (Vargas and Servillo 2018). Taken together, these voices point out two intertwined problems. First, the notion of researcher degrees of freedom suggests that choosing one or another methodological commitment is (to some degree, at least) arbitrary. Second, choosing one or another alternative leads to obtaining different outcomes (what makes results fragile). Below, we shed some light on why background knowledge and statistical methodology do not fully determine researchers' decisions. In the next sections, we approach the question of what factors influence researchers' choices in data research.

In advance, we need to highlight that our claim that statistical analysis and research design involves such permissible malleable to some extent does not preclude some choices from being wrong (i.e., disagreeing with statistical methodology or standards accepted in a field) or unsuitable for the context or research question at hand. In other words, we agree that while there is a significant degree of arbitrariness in some methodological decisions, certain choices are methodologically incorrect, and committing to them undermines study integrity. But we mainly exclude the latter from subsequent analysis. Not much can be said about the distinction between such permissible malleability in research design and statistical analysis and untenable commitments that undermine study integrity at the general level, and the distinction between permissible malleability and errors or fraudulent practices should best be studied on a case-by-case basis.

For instance, statistical hypotheses can be tested in several ways. Usually, more than one test can be used, but the assumptions of some of the tests are not fulfilled, and hence their application is not

warranted by statistical methodology. As Hodges et al. (2022, p. 1) put it, "[t]he use of inappropriate statistical tests leads to unreliable measurement and threatens statistical conclusion validity, a special form of internal validity that concerns sources of random error and the appropriate use of statistics and statistical tests." Another example of such a methodologically unsound decision is including several explanatory variables that are collinear or measure concepts in an unreliable way (Osborne 2011). In such cases, the effect size estimates will vary depending on the set of explanatory variables, and the confidence intervals will be too narrow, leading to reporting false-positive results. However, the matter of fact about some research design and statistical analysis choices being untenable does not undermine claims regarding permissible malleability, when researchers face such methodological decisions that are not constrained sufficiently to assert procedural objectivity (see Jukola 2015).

To have a glimpse at what methodological decisions are involved in designing a study and analyzing data, let us discuss methodological decisions involved in planning, executing, and analyzing an interventional study such as a clinical trial, psychological experiment, or field experiment in economics). While our argumentation concerned with statistical analysis applies equally well to observational studies such as an econometric analysis of macroeconomics data or an epidemiological study of electronic health records, interventional studies involve more degrees of freedom as they require not only gathering and analyzing data but also making decisions about study design that influence what data are produced. Later, we will consider the specific characteristics of interventional and noninterventional studies across medicine, psychology, and economics. However, such a general discussion can only illustrate the problem of the malleability of statistical research as "it is impossible to map out all degrees of freedom that exist in all methods of data analysis in the quantitative sciences" (Stefan & Schönbrodt 2023, p. 4).

Conducting an interventional study begins with designing it. This requires specifying an intervention, defining the control group, measuring outcomes, choosing sample size, and formulating inclusion

and exclusion criteria. All these choices not only can but, sometimes, do influence results. Let us consider the toy example of two conflicting studies of the benefits of supplementing vitamin E discussed by Ioannidis (2005a). The disagreement between treatment effects estimated by Yusuf et al. (2000) and Stephens et al. (1996) can be ascribed to a different approach to measuring outcomes (non-lethal vs. lethal myocardial infarctions[1]) and a higher dose (800 vs. 400 IU). The results of the RECORD study (Home et al. 2007) assessing the cardiovascular safety of rosiglitazone contrast with the meta-analytic finding of Nissen and Wolski (2007) due to differences in sample size, outcome (all cardiovascular attacks vs. coronary heart failure), and inclusion criteria.

Some designs can be considered more informative *for a given purpose*. For example, one may argue that knowing the difference in all-cause mortality between two treatments of diabetes is more beneficial for clinical practice than comparing surrogate outcomes (such as the incidence of coronary heart failure). However, this alone is insufficient to conclude that studies focusing on coronary heart failure are useless or misguided. They simply respond to a different research question. Therefore, in our view, one cannot convincingly argue that some of these choices are *objectively* superior to others. The decisions involved in designing and executing clinical trials was recently documented by van Drimmeln's et al. (2023) ethnographic study of medical researchers working in the end-of-life medicine. They showed that there are many researcher degrees of freedom despite the requirement of study preregistration because research plans are not sufficiently specific to constrain researchers' choices.

But such a malleability in designing research is not specific to medical RCTs. In her popular press article about flexibility in psychological research, Rohrer (2018, p. 1) focused on defining the independent variable, whose effects are studied and showed that there is no direct way from a theoretical concept to its operationalization:

[1] Given that the number of lethal incidents is by definition lower, such a methodological decision lowers the power of study to discover significant effect when the actual effect exists but is considerably weak.

> "Imagine you are trying to figure out whether the personality traits of firstborns systematically differ from those of younger siblings. You set about planning your analyses, a seemingly straightforward task that quickly raises a multitude of questions. Is there any need to control for third variables? How do you handle the fact that the number of siblings varies? What exactly does "firstborn" mean when some people have half- or stepsiblings? And what about the age gaps between siblings — does it make a difference if the firstborn is barely a year older than the younger sibling compared with siblings who are separated by a gap of 10 years? Different answers to such questions will lead to different analyses. What began as a simple question leads to a large number of potential ways to analyze the data, a large number of so-called 'researcher degrees of freedom'. The right data analytic strategy might hinge on details of the hypothesis or on additional assumptions. If the hypothesis is vague, or if we lack crucial pieces of theoretical knowledge to decide which set of assumptions is more plausible, various approaches to running an analysis might be justifiable."

After conducting an interventional study or collecting observational data, one must preprocess the dataset before analyzing them. This includes developing a strategy for dealing with missing data and outliers, differentiating (or not) possibly cointegrated time series, normalizing (or not) distributions, choosing numerical measures for vague concepts, etc. All these commitments, while plausible, shape future results. For example, using each popular method of outlier identification leads to different results (Kianifard and Swallow 1990) despite significant "ambiguity surrounding the definition of outlier values" (Stefan & Schönbrodt (2023, p. 8). Simmons et al. (2011, 1360) discussed the example of defining outliers in psychological studies of reaction time:

> "In a perusal of roughly 30 *Psychological Science* articles, we discovered considerable inconsistency in, and hence considerable ambiguity about, this decision. Most (but not all) researchers excluded some responses for being too fast, but what constituted "too fast" varied enormously: the fastest 2.5%, or faster than 2 standard deviations from the mean, or

> faster than 100 or 150 or 200 or 300 ms. Similarly, what constituted "too slow" varied enormously: the slowest 2.5% or 10%, or 2 or 2.5 or 3 standard deviations slower than the mean, or 1.5 standard deviations slower from that condition's mean, or slower than 1,000 or 1,200 or 1,500 or 2,000 or 3,000 or 5,000 ms. None of these decisions is necessarily incorrect, but that fact makes any of them justifiable and hence potential fodder for self-serving justifications."

Alternative data imputation techniques also bias results in different directions (Gold and Bentler 2000). Dividing a continuous variable into cardinal categories can be done in an infinite number of ways, and neither is supported by arguments other than computational simplicity or conventions practiced in a field. However, using alternative methods in the statistical analysis may deliver inconsistent causal hypotheses. A prime example is a recent change in the definition of hypertension by the American Heart Association in 2017 (Whelton et al. 2018). According to the new standard, blood pressure of 130 mm Hg or higher is considered 'high,' while the earlier definition sets the threshold at the level of 140 mm Hg. Given the linear dependence between blood pressure and cardiovascular outcomes (Choi 2018), applying one or another definition of hypertension to epidemiological research leads to estimating different values of relative risk (RR) for patients suffering from hypertension compared to those unaffected. However, the linearity of the dependence indicates that any threshold level is arbitrary in the sense that a slightly higher or a bit lower value is equally plausible. The discussion of dealing with data cointegration is also far from being settled. As Ashley and Verbrugge (2009, p. 242) put it, "[i]t is often unclear whether time series (…) should be modeled as a vector autoregression in levels (…) or in differences" even though the decision is known to shape the results of Granger-causality tests and other methods of time series analysis.

Decisions concerned with data preprocessing techniques and statistical analysis, which includes choosing model specification, its functional form, and estimation technique, are malleable. The

problem of multicollinearity leads to the situation when adding an explanatory variable (X_{n+1}) to a regression ($Y=a_0+ a_1* X_1+ a_2* X_2+…+ a_n* X_n$) changes the sign of the coefficient of interest (e.g., a_1). This allows for estimating two statistical models that suggest inconsistent associations between the two variables under study (Y and X_1). The persistence of the problem in some areas of econometrics (e.g., development economics) has led econometricians to develop methods of model averaging (see Sala-I-Martin 1997; Hoover and Perez 1999), which resembles multiverse analysis, a method of growing importance in psychology. However, the problem also occurs in the health sciences. For instance, regressing on the causes of autism spectrum disorder, one can find that maternal age is positively related to its prevalence, but adding paternal age to the set of explanatory variables makes maternal age unrelated (Reichenberg et al. 2006). Even though the models with more explanatory variables are usually preferred on the grounds that they control for more confounders, the simpler ones can also be useful even if they only deliver correlational evidence.

Furthermore, data scientists need to choose the functional form of the regression. The solutions to the curve-fitting problem (see Turney 1990) rely on a conventional choice between simplicity and accuracy, even though using alternative functional forms leads to obtaining distinct parameter estimates. The use of different estimation techniques may also shape results. This can also be said about the two basic methods used to estimate linear regressions. The least absolute deviations (LAD) estimation has been offered as a robust estimator that minimizes the effect of outliers that bias the estimates obtained using ordinary least squares (OLS). Still, testing for the presence of outliers in a sample is itself malleable because each of several standard methods returns different results (Kianifard and Swallow 1990). Given this, except for some obvious cases, both estimation procedures are plausible but deliver different results.

An interesting study exemplifying the problem with the malleability of statistical analysis has been conducted by Silberzahn et al. (2018). To simulate how alternative methodological decisions influence results, they gathered 61 analysts divided into 29 teams and asked them to study the

effect of a soccer player's skin color on a referee's propensity to give a red card. The teams of analysts used the same data and assessed the same hypothesis, but the results differed greatly due to differences in analytic approaches. They summarized the decisions involved in this research task as follows:

> "Would you treat each red-card decision as an independent observation? How would you address the possibility that some referees give more red cards than others? Would you try to control for the seniority of the referee? Would you take into account whether a referee's familiarity with a player affects the referee's likelihood of assigning a red card? Would you look at whether players in some leagues are more likely to receive red cards compared with players in other leagues, and whether the proportion of players with dark skin varies across leagues and player positions? As these questions suggest, many analytic decisions are required. Moreover, for a given question, different decisions might be defensible and simultaneously have implications for the findings observed and the conclusions drawn. You and another researcher might make different judgment calls (regarding statistical method, covariates included, or exclusion rules) that, prima facie, are equally valid." (Silberzahn et al. 2018, p. 339).

Stefan & Schönbrodt (2023) delivered a list of methodological decisions required for running a quantitative study that is not field-specific. While their analysis focuses on using these researcher degrees of freedom to p-hack and obtain favorable results, it is also useful for studying the repertoire of alternative choices. Accordingly, the list of researcher degrees of freedom includes the choice of the dependent variable (outcome measure), independent variable (a measure of an intervention), and confounders added to a regression, decision about an optional stopping rule, an approach to dealing with outliers, defining measurement scale, whether to transform a variable (e.g., by calculating its logarithm or discretizing a continuous variable), choice of a statistical significance test (e.g., between Pearson chi-squared test, likelihood ratio test, independent samples

t-test, Welch test, Wilcoxon test, Yuen test), dealing with missing data, i.e., the choice among many data imputation techniques or excluding instances including missing observations.

A skeptic could oppose the view that each of these decisions is equally plausible (and hence the choice arbitrary) by pointing out some arguments (if possibly only weak) to make the choices. It is indeed true that the choice among some alternatives is not arbitrary but a matter of better or worse decisions. For example, if the exogeneity assumption that the sum of error terms is zero in expectancy is not fulfilled, then the ordinary least squares (OLS) estimator is known to be biased and some alternatives such as the instrumental variables method should be used. But, in other cases, statistical methodology is not sufficiently univocal to limit researcher degrees of freedom and determine decisions. Still another possibility is that while more than one decision is correct according to statistical methodology, the context of research (e.g., the research or policy question at hand) suffices for claiming that only one of the alternatives is the right one.

For example, Yarkoni (2022, p. 3) argued that the replication crisis in psychology should rather be labeled the crisis of generalizability because conflicting results are usually reported by studies assessing effects of either modified interventions or the same interventions in different contexts. He observed that "the measured variables must be suitable operationalizations of the verbal constructs of interest, and the relationships between the measured variables must parallel those implied by the logical structure of the verbal statements.". In fact, even in medicine, some measures used as primary outcomes are questioned despite being in widespread use. For example, the HAM-D scale of depression severity has been criticized for overestimating effectiveness and depending on patient characteristics other than the severity of their symptoms (Stegenga 2015). If concepts are operationalized in such a way that what is measured does not reflect theoretical concepts under study and reported results do not reflect the research hypotheses under study, what deflates the epistemic significance of such studies? If one of a pair of seemingly conflicting studies relies on a misguided concept operationalization, uses an inappropriate statistical approach, or a questionable

research practice, its result obviously does not undermine the previous finding as it lacks internal validity. However, the matter of fact that some studies include inferior methods does not undermine the claim about researcher degrees of freedom in quantitative research.

Another approach to opposing the arbitrariness of some methodological commitments would be to distinguish between uncertainty regarding which of two methods is superior from equal plausibility of the two. That is the argumentative strategy DelGuidice and Gangestad (2021) employed to criticize the multiverse analysis. This approach has recently emerged in psychology (Steegen et al. 2016) in response to the arbitrariness of data preprocessing and statistical analysis. Roughly, multiverse analysis estimates all possible data sets (alternatively preprocessed raw data) and all plausible statistical analyses so that the overall estimate of effect size is representative of all plausible choices. As Gelman (2018, p. 21) put it, "[t]he point of multiverse analysis is [(…) to] recognize that all these possible analyses are legitimately of interest" for the reason that those 'possible' analyses are methodologically sound, i.e., permissible on the ground of statistical methodology and disciplinary standards. In a similar vein, DelGuidice and Gangestad (2021) observed that "[t]he central notion of these methods is that the alternatives included in the multiverse are 'arbitrary' or equally 'reasonable.'" However, there is little guidance or consensus on how to evaluate arbitrariness and virtually no consideration of the potential pitfalls of multiverse-style methods. (DelGuidice & Gangestad 2021). They distinguished among principled equivalence (arbitrary choice among different statistical approaches), principled nonequivalence (some methods being superior), and uncertainty (no reason to apply the former notions or insufficient information to suspect which of the alternative statistical techniques is superior). The notion of principled equivalence resembles what is meant by researcher degrees of freedom or 'arbitrary decisions'. Simonsohn et al. (2020) defined the latter as choices for which theory and statistical methodology offer little justification. Those decisions where one of the alternatives can be preferred based on some background knowledge are nonarbitrary (Simonsohn et al. 2020).

It seems helpful to think about the arbitrariness of decisions concerned with statistical analysis and research design in terms of a plausibility threshold. The choice between alternatives is arbitrary if each option is plausible, i.e., they are warranted from the perspective of statistical methodology and the good practices of research design. Thinking in terms of a threshold (plausible/implausible) rebuts DelGuidice and Gangestad's (2021) criticism relying on the distinction between principled equivalence and uncertainty and arguments often voiced by researchers taking part in empirical controversies that one of two alternatives is more justified since weighting arguments for one and another commitment raises many problems. It also aligns with the metascientific studies focusing on the phenomenon of researcher degrees of freedom (Wicherts et al. 2016). In the following section 3, we discuss why researchers commit to different methodological decisions using examples from econometrics, psychology, and medicine and show how these commitments bias results.

3. The impact of committing to different alternatives in research design and statistical analysis

Above, we have argued that data research involves undertaking many decisions and, in some situations, different commitments are plausible. Each of these choices, separately, is unlikely to overturn the result. However, studies that differ in several aspects may report substantially different (and sometimes contrasting) effect sizes or significance levels despite relying on the same or similar datasets. As Šoškić et al. (2022, p. 5) observed, „it is still unknown to what extent choosing one pre-processing and analysis pipeline over another can affect the conclusions that are drawn from the same dataset". Below, we argue that studies that differ in only a few methodological commitments may report results contradicting previous findings.

This can be established by studying empirical controversies where statistical models are estimated on the same or very similar datasets. In such cases, the difference in results can be exclusively ascribed to different commitments regarding data preprocessing techniques, model specification, and estimation strategies. Below, we discuss some examples from econometrics, supplement them with case studies of empirical controversies in psychology, and discuss the results of Silberzahn et al.

(2018), who asked several statisticians to study the same dataset covering soccer players and red cards. This contrasts with interventional and observational studies in medicine, and psychological experiments: using own and differing datasets and sampling is another source of result variability and claims that results differ due to alternative designs or statistical analyses can only be established by conducting exact replications of the experiments reporting conflicting results. But the conclusions drawn from the controversies arising despite the same datasets are used can be extrapolated to those disciplines that rely on experimental data as research design creates more rather than less flexibility.

Econometrics is the discipline where a significant number of controversies (particularly in macroeconometrics) emerge despite the use of the same dataset. For example, our systematic literature review of three top economic journals (*American Economic Review, Journal of Political Economy,* and *Quarterly Journal of Economics*) shows that about 5% of studies contradict other studies published in the same journals. Goldfarb's (1997) estimate based on the most prestigious economic journal (*American Economic Review*) is twice as high due to the focus of AER on empirical research. Doucouliagos & Stanley (2013) delivered a long list of 65 topics in empirical economics that include conflicting results.

The standard view is that conflicting results emerge due to mistakes or biases (such as publication bias, p-hacking, etc.) resulting from the malleability of statistical techniques or outright fraud, and only one of two conflicting results can be accurate. However, suppose two statistical techniques (e.g., weighted and unweighted averaging scheme, see Maziarz 2017) are plausible. In that case, nothing in the data would indicate which conflicting results are accurate. In those cases, the commonsensical view that the results of two studies disagree because one of them is biased due to the malleability of methods (e.g., Stegenga 2018, p. 82) fraudulent, or simply erroneous is not justified. Moosa (2019) argued that, at least in some cases, none of the conflicting econometric studies uses a superior statistical technique but econometric modeling is malleable in the sense that

several methodological decisions are plausible. This view can also be found in the writings of econometricians. For example, Reinhart and Rogoff (2010) admitted that there is some degree of subjectivity in econometric modeling: "[t]hose who have done data work know that mapping vague concepts like "high debt" or "overvalued exchange rates" into workable definitions requires arbitrary judgments about where to draw lines; there is no other way to interpret the facts and inform the discussion". But their opinion applies also to other quantitative disciplines. Especially psychology has experienced a heated methodological debate about the malleability of research design and statistical analysis with the concept 'researcher degrees of freedom' presuming that methodological decisions are not constrained and "usually there is more than one right way through the proverbial garden of forking paths" (Stefan and Schönbrodt 2023, p. 2).

Estimating an econometric model can be compared to running an experiment: the statistics describing the quality of the obtained model (such as the statistical significance of parameters) are only interpretable if a regression was estimated once. Similarly to the file-drawer problem in the randomized controlled trial (RCT) literature (Dalton et al. 2012; Bird 2021), p-hacking and torturing the data for other purposes influence the probability distribution of estimated coefficients and makes the values of parameters meaningless and fallacious. On this ground, De Long and Lang (1992) argued that most econometric results are spurious. According to their analysis, the journal editor's preference to publish results confirming theory makes econometricians torture the data to obtain the results in demand that are spurious. Some questionable practices (such as increasing the sample size to get smaller p-values) allow for obtaining statistically significant results despite their spuriousness (Ziliak and McCloskey 2008) and can lead to the emergence of conflicting results.

Still, looking for confirmatory results does not explain the phenomenon of recalcitrant results unless the economic theory is also split. However, the persistence of different schools of thought creates a situation where scientific disagreement is also present at the theoretical level. Doucouliagos and Stanley (2013) delivered evidence that in the fields where theory is more split, econometric results

also differ. They take their results as evidence that econometricians are biased in their research and aim at confirming the theory that they accept a priori. However, due to cross-sectional data, the opposite direction of influence is also possible: differing econometric results can inspire theoretical economists to construct axiomatic models resembling a chosen empirical result.

Nevertheless, split economic theory is not the only factor accountable for creating disagreements in econometric results. Another plausible cause is the institutional setting that promotes publishing novel (opposing previous literature) econometric models. Goldfarb (1995) reformulated De Long and Lang's (1992) hypothesis regarding publication bias. Accordingly, the bias changes in time as follows. When econometricians start to study a new topic (for example, the relation between two variables (*X* and *Y*), whose values were recently estimated), publishing results confirming the existence of a (statistically significant) association is more manageable. When the literature on a topic matures, then journal editors, striving for novelty, are more likely to accept for publication evidence that shows that the partial correlation is spurious or the sign of the dependency differs from what was previously believed. In other words, Goldfarb (1995) hypothesized that it is not p-hacking or torturing the data aimed at obtaining results confirming a theory but striving for novel results that creates reversals in econometric literature.

Goldfarb's (1995) analysis identifies several topics in which econometric models deliver evidence for conflicting or contradictory hypotheses regarding the relation among variables under consideration. However, despite its detrimental effect on drawing policy-oriented conclusions and informing theoretical discourse, the existence of conflicting results in empirical literature did not get the broader attention of the philosophers of science/economics or methodologically oriented economists. In his subsequent article, Goldfarb (1997) added more strains of econometric literature that included conflicting results and hypothesized additional reasons why econometricians disagree (pp. 231-232). Accordingly, conflicting results emerge due to (1) more/different data, (2) different or 'fancier' techniques, (3) changing the exogenous/endogenous variables, (4) publication bias at the

journal level, (5) torturing data aimed at obtaining novel results, (6) statistical/calculation errors, (7) a change in setting/context, (8) a change in theoretical presumption, and (9) theory development affecting empirical work. Goldfarb (1997, pp. 233-234) analyzed several cases of conflicting econometric results and concluded that, in each case, more than one explanation is plausible given gathered (article-based) evidence. Goldfarb's systematization of the reasons underlying empirical disagreements includes methodological and sociological/conflict-of-interest reasons. While novel or different data, different/fancier statistical techniques, changing the exogenous/endogenous variables, data-dredging, and errors/fraudulent practices are reasons concerned with research methods, publication bias, searching for novel results to achieve academic success, and inspiration by a different theory are not directly linked to statistical estimation.

The standard (textbook) approach to econometrics requires estimating a theory-inspired, single model. However, such an account of econometric modeling is unlikely to be adequate for the actual research practice of academic econometricians. Granger and Jeon (2004, p. 324) admitted that "[t]he pristine sequence of events in selecting, estimating, and using a single specification sounds well in textbooks but is not realistic". There is a difference between the 'standard' (handbook) description of econometric modeling and how the process actually proceeds. "The standard method of model building is to consider just a single specification and then discuss its estimation, interpretation, and output. In practice, many specifications are available, so the usual technique is to choose the best" (Granger and Jeon 2004, p. 523).

In contrast to this idealized view, statistical analysis of either experimental or observational data requires many decisions concerning research design, data preprocessing techniques, and statistical analysis. The actual practice leads to the situation where econometricians obtain several different models, choose one or up to a few, and describe it in an article. The problem is usually solved by employing "a combination of diagnostic checking and significance testing of the coefficients, in the hope that one of these variants will emerge as the "best" model" (Spanos 2012, p. 381). Such an

error-fixing strategy leads to estimating numerous econometric models. Only a minor fraction of these models is described in the published articles. Considering the strive for novelty (Goldfarb 1997), researchers are incentivized to choose from the menu of plausible econometric models those estimations that contrast with previous findings. Hence, the audience for empirical literature often observes that subsequent studies disagree with prior results.

The problem of the institutional surrounding is also persistent in other disciplines. Bryan et al. (2019, p. 25543) argued that the same non-epistemic values of striving for novelty at both journal editors' and researchers' sides lead to a situation when many replication attempts report false-negative results despite the effect reported by the original study being accurate (see also Dweck and Yeager 2019). In such cases, p-hacking changes into 'null hacking':

> "There is a clear incentive for replicators to obtain results that conflict with the original study […] that should be presumed to exert as much influence on how replicating investigators exercise degrees of freedom as the incentive to find significant results influences original investigators. Even before this incentive can shape how investigators carry out replication tests, it might influence which effects they choose to replicate. That is, the incentive to find null results in replication tests could bias investigators to select effects for replication testing that they already believe are false."

To support their claim Bryan et al. (2019) described the controversy concerning whether slight changes in polls (e.g., using the phrase 'being a voter' instead of 'voting') may change voter's behavior and make them participate in elections. According to their data re-analysis, the replication of Garber et al. (2016) differs from the original study of Bryan et al. (2011) in such a way to report results conflicting with the original study. The replicators controlled for a substantial number of (co-linear) covariates, the inclusion of interaction items, and questioning voters up to four days before the election day but also on the day of the elections. While Bryan et al. (2019) argued that not only the methodological decisions of the replication study are aimed at obtaining results contradicting

previous findings, but they are also inferior compared to the choices made by the researchers conducting the original study, our focus here is on the fact that a slight change in study design (a larger time frame for questioning voters) and a more complicated statistical analysis leads to obtaining a statistically insignificant result interpreted by the replicators as an indication of no effect. Furthermore, Bryan et al. (2019) conducted a specification-curve analysis showing that about 50% of all methodologically sound analyses deliver a statistically significant result.

However, it is not only the incentive created by the present-day academic environment to obtain novel results that is the driving force of conflicting results. Researchers disagree about the most appropriate data preprocessing techniques and statistical methods, even in cases when there are neither incentives for obtaining a statistically significant result or contradicting previous findings nor ambiguity about the research question. This is the conclusion emerging from Silberzahn et al. (2018) experiment about researcher degrees of freedom. They gathered two dozen statisticians and asked them to deliver a response to a research question concerned with soccer players' skin color and referees' propensity to give red cards. All researchers participating in the experiment knew that their results would be published regardless of the results they obtained and their statistical significance. Despite no incentives to bias results, the results reported by individual teams of statisticians varied significantly: the effect size estimates covered both slightly negative and moderately positive odds ratios (0.89 – 2.93). The same can be said about statistical techniques as they "ranged from simple linear regression to complex multilevel regression and Bayesian approaches. The teams also varied greatly in their decisions regarding which covariates to include" (Silberzahn et al. 2018, p. 334). Unfortunately, neither researchers' aprioristic views about bias in soccer referee decisions nor the methodological soundness of statistical analysis (assessed by other peers) correlated to results.

The high variability of results reported by the teams of statisticians participating in the experiment made Silberzahn et al. (2018, p. 351) conclude that "highly defensible analytic decisions made without direct incentives to achieve statistical significance can still produce wide variability in effect-

size estimates". Furthermore, the comparison of individual results and methodological commitments allowed for identifying those choices (including league and club as additional explanatory variables in a regression) that led to statistically insignificant results. Considering that a "debate emerged regarding whether the inclusion of these covariates was quantitatively defensible given that the data on league and club were available for the time of data collection only and these variables likely changed over the course of many players' careers" (Silberzahn et al. 2018, p. 343)", none of the choices was univocally supported by the statisticians participating in the experiments. Such a disagreement about this and other methodological decisions emerged despite no institutional incentives promoting novelty but purely on the grounds of different presuppositions about what research method is best in a given context. But Silberzahn et al. (2018, p. 352) admitted that such situations of no incentives promoting disagreements are rare in the actual science as "[o]riginal authors have strong incentives to find positive results so that their work will be published, and commenters have strong incentives to find different (usually negative) results for the same reason." In a similar vein, Simmons et al. (2011, p. 1359) observed that, in psychology, "it is uncommon for prestigious journals to publish null findings or exact replications, researchers have little incentive even to attempt them".

Given the incentives researchers face to publish novel findings contradicting previous results and the malleability of quantitative research that allows for obtaining different results by changing research design and statistical analysis, conflicting results will likely emerge across all data-intensive disciplines. In such situations, the publication of a result overturning previous evidence drives rapid policy and clinical practice changes, which may undermine trust in the effectiveness of the policy measures and treatment guidance. The Reinhart-Rogoff controversy is an excellent example of the damaging influence of conflicting empirical results on policy. The publication of a study (Herndon et al. 2014) contradicting existing evidence on the relationship between debt and growth contributed to a change in fiscal policy in a few developed countries (with the United States being the primary example). Another situation where policymakers followed the most recent results was the case of

deworming policy in Kenya (Aiken et al. 2015). Recently, rapid changes in clinical guidance regarding COVID treatments were driven by following the most recent results of clinical trials. For example, Food and Drug Administration has endorsed and later revoked their emergency use authorization for hydroxychloroquine and remdesivir (Maziarz & Stencel 2022).

4. Is the replication crisis driven by too broad generalizations?

Replication projects indicate that calculation errors or fraudulent practices are present in as many as 1 in 7 studies (Bakker and Wicherts 2011). This shows that the discrepancies between published research and actual results corrected for errors can account only for a fraction of disagreements present in the empirical literature. Another reason for the differences in published findings is the emergence of random differences between treatment and control groups (Bird 2021). Statistical hypothesis testing is inextricably related to the risk of false-positive results. While the institutional context of academic publishing, with its focus on positive results, creates the file drawer problem by itself, researchers' motivation to fish for statistical significance by computing several test statistics and reporting only those results that support the hypothesis under test exaggerates the false positive report probability (Wacholder et al. 1995). Furthermore, Gelman and Loken (2013, p. 2) argued that the inflation of false positive report probability is driven not only by researchers purposefully testing many alternative model specifications but also by the possibility of conducting a different analysis if data produced by an experiment or an observational dataset were different: "there is the misleading implication that researchers were consciously trying out many different analyses on a single data set; and, second, because it can lead researchers who know they did not try out many different analyses to mistakenly think they are not so strongly subject to problems of researcher degrees of freedom."

Gelman and Loken (2013) argument relies on the notion of 'potential comparisons' that arise from the multitude of potential analyses dependent on decisions concerned with data preprocessing and statistical analysis. Since each decision (researcher degree of freedom) divides an analytical path, all

those possible paths create a 'garden of forking paths' where "whatever route you take seems predetermined, but that's because the choices are done implicitly. The researchers are not trying multiple tests to see which has the best p-value; rather, they are using their scientific common sense to formulate their hypotheses in reasonable way, given the data they have. The mistake is in thinking that, if the particular path that was chosen yields statistical significance, that this is strong evidence in favor of the hypothesis." Gelman and Loken (2013, p. 10). The implication of their viewpoint is that nominal p-values are meaningless, and the actual probability of a result being a false positive is much higher because correcting the level of statistical significance for multiple comparisons should take into account not only those analyses that were calculated but also those alternatives that could have been chosen. As Gelman (2016, p. 1) put it, "[t]he whole point of "the garden of forking paths" [(…)] is that to compute a valid p-value you need to know what analyses would have been done had the data been different."

The simulations run by Stefan and Schönbrodt (2023) show that p-hacking using the researcher degrees of freedom related to each separate decision easily elevates the false-positive rate to about 30%, but using several researcher degrees of freedom in the process of p-hacking has much more severe consequences for the interpretation of p-values. Additionally, the problem of false-positive results is further elevated by the developments in science: since most interventions producing large effect sizes have been studied and are known, researchers turn their interests towards smaller and smaller effects (Gelman 2018). Considering that journal editors prefer to publish articles reporting statistically significant results and null results are hidden from the audience for empirical literature, some portion of positive results published in psychology journals (Simmons et al. 2011) and other disciplines such as medicine (Maziarz and Stencel 2022) might result from the base-rate fallacy.

So, some instantiations of the replication crisis are produced by honest errors, a misuse of statistics, or fraudulent practices or are false-positive or false-negative results. In each of these cases, further research can show that either a study reporting an effect or another one showing no effect or a

negative sign of the effect reported the accurate result. This seems to be the mainstream view on the replication crisis among researchers, as they are unwilling to endorse mutually conflicting results or results that seem to be in conflict when extrapolated to the same policy or clinical context. For instance, facing conflicting epidemiological literature on the relation between dietary fat and cancer, Prentice and Sheppard (1990) motivated their research project by aiming "to understand the apparent discrepancy between these observations [and…] provide an insight into the magnitude of cancer risk reduction that may follow from a practical reduction in dietary fat" (p. 81). Even if one appraises scientific disagreements on the grounds that they foster research in a field, the view that further studies will establish more accurate and consistent effect size estimates persists. Based on a case study from anthropology, De Cruz and De Smedt (2013) argued that conflicting hypotheses are valuable in science because they promote further research leading to 'acquiring true beliefs about' the problematic issue.

However, other cases of conflicting results emerge from different but plausible commitments concerning research design and statistical analysis. In such cases, both studies may report an accurate effect size, but those actually differing effect sizes can only be observed in different contexts or in different subpopulations. For example, both claims, 'supplementing vitamin E reduced mortality' and 'supplementing vitamin E was unrelated to mortality' are accurate (since no flaws in the two studies considered by Ioannidis (2005a) have been detected so far). However, the two results (extrapolated to the same target population, i.e., the patients at risk of cardiovascular disease) are evidence suggesting undertaking opposing actions. If two or more mutually conflicting results apply to the same decision problem or generalize to the same population of patients, only one of them can accurately describe the true effect size of the intervention under consideration. However, it may be the case that two or more studies reporting conflicting results estimate accurate effect sizes of the intervention in different contexts.

In clinical research, different inclusion and exclusion criteria have pretty similar effects compared to the choices of statistical techniques that emphasize different types of observations. For instance, a methodological analysis of the two studies on the effect of vitamin E supplementation considered by Ioannidis (2005a) suggests that the results differ due to a different approach to measuring outcome (non-lethal vs. lethal myocardial infractions) and a higher dose (800 vs. 400 IU). These commitments may lead to estimating different treatment effects of vitamin E supplementation, but none of the results can be considered more accurate. They address two different research questions concerned with other doses and outcomes. In a similar vein, the RECORD study (Home et al. 2007), which has been criticized for being conducted on a sample of the Caucasian population, delivers evidence relevant to that population (but not for other ethnicities). If a causal effect under study is heterogeneous, using samples of different populations or employing statistical techniques that put different weights on some types of instances allows for estimating different treatment effects. This, in effect, deems conflicting results more or less like the true effect sizes in different subgroups.

However, not only different inclusion and exclusion criteria and using statistical procedures that differentiate the weight of observations on results or using specific samples leads to inconsistent estimates of treatment effect. Otherwise, the problem of conflicting results could be reformulated into the reference class problem. One of the main decisions in statistical modeling concerns the inclusion of other explanatory variables. Obviously, including all possible confounders in a model specification is impossible due to concerns regarding the degrees of freedom of estimated regression. For this and other reasons, one can be interested in the relation between Y and X_1 conditional on some other factors X_2, and X_3. In contrast, other purposes may require knowledge of the regression of Y on X_1, X_2, X_3, and X_4. Two such statistical models M_1 and M_2 can, despite being estimated on the same dataset (or samples of the same population) deliver evidence for inconsistent hypotheses regarding the relation between X_1 and Y so that M_1 suggests a positive dependence and M_2 either a negative or insignificant dependence. The two models supporting inconsistent conclusions can be considered as alternative representations of a broader causal structure, where

M_1 represents the dependencies among X_1, X_2, X_3, and Y whereas M_2 represents the influence of X_1, X_2, X_3, and X_4. Both models can be seen as partial representations of a broader causal structure, with Y also being determined by other factors X_n-X_m (see Fig. 1).

[Figure 1 should be put somewhere here]

For instance, consider the example of epidemiological studies assessing the influence of maternal age on the risk of autism spectrum disorder (ASD) in offspring. This strain of literature includes several conflicting results. Some studies find a positive influence, and others claim no statistically significant effect. Still, the meta-analysis of 16 studies conducted by Sandin et al. (2012) clearly shows that maternal age is associated with the risk of a child's autism: children of mothers aged over 35 are 52% more likely to develop ASD than mothers aged 25-29. However, Sandin et al. (2012) employed the random-effects model assuming two sources of inter-study result variability: between-study variation u_i and within-study variation e_i. That is, their study does not control for any other confounding factor. This result contrasts with the observation of Reichenberg et al. (2006) that maternal age is associated with ASD risk in offspring only if paternal age is excluded from the analysis. Otherwise, the relation between maternal age and ASD risk proves spurious, with maternal age being only a proxy for paternal age, which is the actual risk factor.

Recently, Yarkoni (2022) argued that conflicting results of psychological experiments can be interpreted in terms of 'the generalizability crisis', since researchers fail to define the limits of the generalizability of their results and replicators often test hypotheses under different experimental conditions (in different contexts). Furthermore, he pointed out that result stability across context is more important than reproducibility from the practical perspective: "the current focus on reproducibility and replicability risks is distracting us from more important, and logically antecedent, concerns about generalizability. The root problem is that when the manifestation of a phenomenon is highly variable across potential measurement contexts, it simply does not matter very much whether any single realization is replicable or not" (Yarkoni 2022, p. 17). There have been some

practical propositions that rely on the assumption that different statistical analyses tell us different things about the datasets. For example, Dragicevic et al. (2019) supported the use of interactive reports for presenting the results of multiverse analysis. The readers of such reports could interact with the statistical analysis and make their own analytical choices, potentially obtaining results relevant to the context they are interested in.

All in all, both decisions regarding research design and statistical analysis have an impact on the reported results, and studies that differ regarding a few such commitments may report results that differ significantly or are conflicting. But this can only be proven if each individual study with differing commitments replicates for otherwise two other factors (random differences between treatment and control groups or errors) might be responsible for the difference in outcomes.

5. Concluding Remarks

While thinking that questions such as what the effect of an intervention is on an outcome have only one answer is appealing, nothing in the world or the data would guarantee that all repertoire of research designs and statistical methods will lead to the same result. Also, there is no ground to believe that an effect size is homogenous across all policy or clinical contexts. For many cases of empirical controversies, there is no methodological or theoretical ground to choose the statistical techniques and research designs that are superior or which variables to include in a statistical model. In other cases, some studies use inappropriate methods or fraudulent practices. We have argued that in cases when two statistical models implying inconsistent hypotheses result from the application of plausible research design and statistical methods, the audience for empirical results should be pluralist and endorse the two estimates of effect size instead of fruitlessly disputing the superiority of one or another study.

However, even if there is no ground to argue that one of a menu of methodological commitments is better than others, some research designs and choices concerning statistical analysis are better

suited to address a particular research question. Future research is needed to develop more straightforward approaches to choosing studies that report effect sizes estimating the actual effect size of an intervention in each context. Statistical modeling and research design are malleable because they require numerous choices and commitments, and it is rare for only one decision to be right. This permissible malleability should not be considered a drawback of empirical research. Instead, the multitude of different studies reporting conflicting results should be used to raise our (incomplete) knowledge of the world and improve our ability to intervene in it.

Acknowledgment

The author thanks for the insightful comments received during the European Philosophy of Science Association Fellowship at the University of Amsterdam. The author acknowledges the supportive comments received from the reviewers for the journal.

Funding

The work of Mariusz Maziarz on the first draft of the manuscript has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 805498). The additional research required by the Reviewers for *Perspectives of Science* was funded by The National Science Centre, Poland, under grant number 2022/45/B/HS1/00183. For the purpose of Open Access, the author has applied a CC-BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

Data Availability Statement

No new data have been produced by this research.

Ethical Approval Statement

Ethical approval is not required for the study.

6. Literature

Aguinis, H., R. K. Gottfredson, and H. Joo. 2013. "Best- Practice Recommendations for Defining, Identifying, and Handling Outliers." *Organizational Research Methods* 16:270–301.

Aiken, Alexander M., Calum Davey, James R. Hargreaves, and Richard J. Hayes. 2015. "Re-analysis of Health and Educational Impacts of a School-based Deworming Programme in Western Kenya: A Pure Replication." *International Journal of Epidemiology* 44(5):1572–1580.

Armijo-Olivo, Susan, Jorge Fuentes, Maria Ospina, Humam Saltaji, and Lisa Hartling. 2013. "Inconsistency in the Items Included in Tools Used in General Health Research and Physical Therapy to Evaluate the Methodological Quality of Randomized Controlled Trials: A Descriptive Analysis." *BMC Medical Research Methodology* 13(1):1–19.

Ashley, Richard A., and Randal J. Verbrugge. 2009. "To Difference or Not to Difference: A Monte Carlo Investigation of Inference in Vector Autoregression Models." *International Journal of Data Analysis Techniques and Strategies* 1(3):242–274.

Bird, Alexander. 2021. "Understanding the Replication Crisis as a Base Rate Fallacy." *The British Journal for the Philosophy of Science*. Retrieved February 5, 2024 (https://doi.org/10.1093/bjps/axaa038).

Broadbent, Alex. 2013. "Why Philosophy of Epidemiology?" Pp. 1–9 in *Philosophy of Epidemiology*, edited by A. Broadbent. London: Palgrave Macmillan.

Bryan, Christopher J., David S. Yeager, and Joseph M. O'Brien. 2019. "Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate." *Proceedings of the National Academy of Sciences* 116(51):25535–25545.

Bryan, Christopher J., David S. Yeager, and Joseph M. O'Brien. 2019. "Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate." Proceedings of the National Academy of Sciences 116(51):25535–45.

Cartwright, Nancy, and Jeremy Hardie. 2012. *Evidence-based Policy: A Practical Guide to Doing It Better*. Oxford: Oxford University Press.

CDC COVID Response Team. 2021. "Allergic Reactions Including Anaphylaxis After Receipt of the First Dose of Pfizer-BioNTech COVID-19 Vaccine—United States, December 14–23, 2020." *Morbidity and Mortality Weekly Report* 70(2):46.

Choi, You-Jung, Sun-Hwa Kim, Si-Hyuck Kang, Chang-Hwan Yoon, Hae-Young Lee, Tae-Jin Youn, In-Ho Chae, and Cheol-Ho Kim. 2019. "Reconsidering the Cut-off Diastolic Blood Pressure for Predicting Cardiovascular Events: A Nationwide Population-based Study from Korea." *European Heart Journal* 40(9):724–731.

De Long, J. Bradford, and Kevin Lang. 1992. "Are All Economic Hypotheses False?" *Journal of Political Economy* 100(6):1257–1272.

Del Giudice, Marco, and Steven W. Gangestad. 2021. "A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions." *Advances in Methods and Practices in Psychological Science* 4(1):2515245920954925.

Doucouliagos, Chris, and Tom D. Stanley. 2013. "Are All Economic Facts Greatly Exaggerated? Theory Competition and Selectivity." *Journal of Economic Surveys* 27(2):316–339.

Dragicevic, P., Y. Jansen, A. Sarma, M. Kay, and F. Chevalier. 2019. "Increasing the Transparency of Research Papers with Explorable Multiverse Analyses." Pp. 1–15 *in Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Dweck, Carol S., and David S. Yeager. 2019. "A Simple Re-analysis Overturns a 'Failure to Replicate' and Highlights an Opportunity to Improve Scientific Practice: Commentary on Li and Bates (in press)." Retrieved February 5, 2024 (https://doi.org/10.31234/osf.io/8qfzj).

Dweck, Carol S., and David S. Yeager. 2019. "A Simple Re-analysis Overturns a 'Failure to Replicate' and Highlights an Opportunity to Improve Scientific Practice: Commentary on Li and Bates (in press)." Retrieved February 5, 2024 (https://doi.org/10.31234/osf.io/8qfzj).

Feest, Uljana. 2019. "Why Replication Is Overrated." *Philosophy of Science* 86(5):895–905.

Feinstein, Alvan R. 1995. "Meta-analysis: Statistical Alchemy for the 21st Century." *Journal of Clinical Epidemiology* 48(1):71–79.

Fraser, David. 2008. "Understanding Animal Welfare." *Acta Veterinaria Scandinavica* 50(1):1–7.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'P-hacking' and the Research Hypothesis Was Posited Ahead of Time." Department of Statistics, Columbia University 348:1–17.

Gelman, Andrew, and Eric Loken. 2013. "The Garden of Forking Paths: Why Multiple Comparisons Can Be a Problem, Even When There Is No 'Fishing Expedition' or 'P-hacking' and the Research Hypothesis Was Posited Ahead of Time." Department of Statistics, Columbia University 348:1–17.

Gelman, Andrew. 2018. "The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It." *Personality and Social Psychology Bulletin* 44(1):16–23.

Goldfarb, Robert S. 1995. "The Economist-as-audience Needs a Methodology of Plausible Inference." *Journal of Economic Methodology* 2(2):201–22.

Goldfarb, Robert S. 1997. "Now You See It, Now You Don't: Emerging Contrary Results in Economics." *Journal of Economic Methodology* 4(2):221–44.

Herndon, Thomas, Michael Ash, and Robert Pollin. 2014. "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff." *Cambridge Journal of Economics* 38(2):257–79.

Hodges, Cooper B., Bryant M. Stone, Paula K. Johnson, James H. Carter III, Chelsea K. Sawyers, Patricia R. Roby, and Hannah M. Lindsey. 2022. "Researcher Degrees of Freedom in Statistical Software Contribute to Unreliable Results: A Comparison of Nonparametric Analyses Conducted in SPSS, SAS, Stata, and R." *Behavior Research Methods.* Retrieved February 5, 2024 (https://doi.org/10.3758/s13428-021-01654-0).

Home, Philip D., Stuart J. Pocock, Henning Beck-Nielsen, Ramón Gomis, Markolf Hanefeld, Nigel P. Jones, Michel Komajda, and John JV McMurray. 2007. "Rosiglitazone Evaluated for Cardiovascular Outcomes—An Interim Analysis." *New England Journal of Medicine* 357(1):28–38.

Hoover, Kevin D., and Stephen J. Perez. 1999. "Data Mining Reconsidered: Encompassing and the General-to-specific Approach to Specification Search." *The Econometrics Journal* 2(2):167–91.

Howick, Jeremy, Paul Glasziou, and Jeffrey K. Aronson. 2013. "Problems with Using Mechanisms to Solve the Problem of Extrapolation." *Theoretical Medicine and Bioethics* 34(4):275–291.

Howick, Jeremy, Paul Glasziou, and Jeffrey K. Aronson. 2013. "Problems with Using Mechanisms to Solve the Problem of Extrapolation." *Theoretical Medicine and Bioethics* 34(4):275–291.

Ioannidis, John PA. 2005a. "Contradicted and Initially Stronger Effects in Highly Cited Clinical Research." *JAMA* 294(2):218–28.

Ioannidis, John PA. 2005b. "Why Most Published Research Findings Are False." *PLoS Medicine* 2(8):e124.

Jukola, Saana. 2015. "Meta-analysis, Ideals of Objectivity, and the Reliability of Medical Knowledge." *Science & Technology Studies* 28(3):101–21.

Kianifard, Farid, and William H. Swallow. 1990. "A Monte Carlo Comparison of Five Procedures for Identifying Outliers in Linear Regression." *Communications in Statistics-Theory and Methods* 19(5):1913–1938.

Kulkarni, Abhaya V. 2005. "The Challenges of Evidence-based Medicine: A Philosophical Perspective." *Medicine, Health Care and Philosophy* 8(2):255–260.

Maziarz, Mariusz, and Adrian Stencel. 2022. "The Failure of Drug Repurposing for COVID-19 as an Effect of Excessive Hypothesis Testing and Weak Mechanistic Evidence." *History and Philosophy of the Life Sciences* 44(4):1–26.

Maziarz, Mariusz. 2017. "The Reinhart-Rogoff Controversy as an Instance of the 'Emerging Contrary Result' Phenomenon." *Journal of Economic Methodology* 24(3):213–225.

Maziarz, Mariusz. 2022. "Is Meta-analysis of RCTs Assessing the Efficacy of Interventions a Reliable Source of Evidence for Therapeutic Decisions?" *Studies in History and Philosophy of Science* 91:159–167.

Moosa, Imad A. 2019. "The Fragility of Results and Bias in Empirical Research: An Exploratory Exposition." *Journal of Economic Methodology* 26(4):347–360.

Nissen, Steven E., and Kathy Wolski. 2007. "Effect of Rosiglitazone on the Risk of Myocardial Infarction and Death from Cardiovascular Causes." *New England Journal of Medicine* 356(24):2457–2471.

Osborne, Jason W. 2002. "Effect Sizes and the Disattenuation of Correlation and Regression Coefficients: Lessons from Educational Psychology." *Practical Assessment, Research, and Evaluation* 8(1):11.

Polack, Fernando P., Stephen J. Thomas, Nicholas Kitchin, Judith Absalon, Alejandra Gurtman, Stephen Lockhart, John L. Perez et al. 2020. "Safety and Efficacy of the BNT162b2 mRNA Covid-19

Vaccine." *New England Journal of Medicine*. Retrieved February 5, 2024 (https://doi.org/10.1056/NEJMoa2034577).

Post, Piet N., Hans de Beer, and Gordon H. Guyatt. 2013. "How to Generalize Efficacy Results of Randomized Trials: Recommendations Based on a Systematic Review of Possible Approaches." *Journal of Evaluation in Clinical Practice* 19(4):638–643.

Power, Scott B., and Greg Kociuba. 2011. "The Impact of Global Warming on the Southern Oscillation Index." *Climate Dynamics* 37(9):1745–1754.

Prentice, Ross L., and Lianne Sheppard. 1990. "Dietary Fat and Cancer: Consistency of the Epidemiologic Data, and Disease Prevention That May Follow from a Practical Reduction in Fat Consumption." *Cancer Causes & Control* 1(1):81–97.

Reichenberg, Abraham, Raz Gross, Mark Weiser, Michealine Bresnahan, Jeremy Silverman, Susan Harlap, Jonathan Rabinowitz et al. 2006. "Advancing Paternal Age and Autism." *Archives of General Psychiatry* 63(9):1026–1032.

Reinhart, Carmen M., and Kenneth S. Rogoff. 2010. "Growth in a Time of Debt." *American Economic Review* 100(2):573–578.

Robins, James M., Miguel Angel Hernan, and Babette Brumback. 2000. "Marginal Structural Models and Causal Inference in Epidemiology." *Epidemiology* 11(5):550–560.

Rohrer, Julia M. 2018. "Run All the Models! Dealing with Data Analytic Flexibility." *APS Observer* 31. Retrieved February 5, 2024 (https://www.psychologicalscience.org/observer/run-all-the-models-dealing-with-data-analytic-flexibility).

Romero, Felipe. 2019. "Philosophy of Science and the Replicability Crisis." *Philosophy Compass* 14(11):e12633.

Sala-i-Martin, Xavier. 1997. "I Just Ran Four Million Regressions." *NBER Working Paper* no. 6252.

Sandin, Sven, Christina M. Hultman, Alexander Kolevzon, Raz Gross, James H. MacCabe, and Abraham Reichenberg. 2012. "Advancing Maternal Age Is Associated with Increasing Risk for Autism: A Review and Meta-analysis." *Journal of the American Academy of Child & Adolescent Psychiatry* 51(5):477–486.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11):1359–1366.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. "False-positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." Psychological Science 22(11):1359–66.

Simonsohn, Uri, Joseph P. Simmons, and Leif D. Nelson. 2020. "Specification Curve Analysis." *Nature Human Behaviour* 4(11):1208–1214.

Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency through a Multiverse Analysis." *Perspectives on Psychological Science* 11(5):702–712.

Stefan, Angelika M., and Felix D. Schönbrodt. 2023. "Big Little Lies: A Compendium and Simulation of P-hacking Strategies." *Royal Society Open Science* 10(2):220346.

Stegenga, Jacob. 2015. "Measuring Effectiveness." *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 54:62–71.

Stegenga, Jacob. 2011. "Is Meta-analysis the Platinum Standard of Evidence?" *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 42(4):497–507.

Stegenga, Jacob. 2018. *Medical Nihilism*. Oxford: Oxford University Press.

Stephens, Nigel G., Ann Parsons, Morris J. Brown, P. M. Schofield, F. Kelly, K. Cheeseman, and M. J. Mitchinson. 1996. "Randomised Controlled Trial of Vitamin E in Patients with Coronary Disease: Cambridge Heart Antioxidant Study (CHAOS)." *The Lancet* 347(9004):781–786.

Turney, Peter. 1990. "The Curve Fitting Problem: A Solution." *British Journal for the Philosophy of Science* 41(4):509–30.

van Drimmelen, Tom, Nienke Slagboom, Ria Reis, Lex Bouter, and Jenny van der Steen. 2023. "Decisions, Decisions, Decisions: An Ethnographic Study of Researcher Discretion in Practice. Ethnographic Study of Researcher Degrees of Freedom in End-of-life Research." Retrieved February 5, 2024 (https://doi.org/10.31222/osf.io/7dh3t).

Vargas, Maria, and Giuseppe Servillo. 2018. "The End of Corticosteroid in Sepsis: Fragile Results From Fragile Trials." *Critical Care Medicine* 46(12):e1228.

Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Laure El Ghormli, and Nathaniel Rothman. 2004. "Assessing the Probability That a Positive Report Is False: An Approach for Molecular Epidemiology Studies." *Journal of the National Cancer Institute* 96(6):434–442.

Wacholder, Sholom, Stephen Chanock, Montserrat Garcia-Closas, Laure El Ghormli, and Nathaniel Rothman. 2004. "Assessing the Probability That a Positive Report Is False: An Approach for Molecular Epidemiology Studies." *Journal of the National Cancer Institute* 96(6):434–42.

Wicherts, Jelte M., Coosje LS Veldkamp, Hilde EM Augusteijn, Marjan Bakker, Robbie Van Aert, and Marcel ALM Van Assen. 2016. "Degrees of Freedom in Planning, Running, Analyzing, and Reporting Psychological Studies: A Checklist to Avoid P-hacking." *Frontiers in Psychology* 7:1832.

Williamson, Jon. 2019. "Establishing Causal Claims in Medicine." *International Studies in the Philosophy of Science* 32(1):33–61.

Yusuf, Salim, G. Dagenais, J. Pogue, J. Bosch, and P. Sleight. 2000. "Vitamin E Supplementation and Cardiovascular Events in High-risk Patients." *The New England Journal of Medicine* 342(3):154–160.

Ziliak, Steve, and Deirdre Nansen McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Ann Arbor: University of Michigan Press.

Fig. 1: Relevance of different results for different decision settings