The multiplicity of analysis strategies jeopardizes replicability: lessons learned across disciplines

Sabine Hoffmann,[1,2] Felix Schönbrodt,[1,3] Ralf Elsas,[1,4]
Rory Wilson,[5] Ulrich Strasser,[6] Anne-Laure Boulesteix[1,2,7*]

[1]Open Science Center of the Ludwig Maximilian University (LMU) Munich, Germany

[2]Institute for Medical Information Processing, Biometry, and Epidemiology, Medical School, LMU Munich, Germany

[3]Department of Psychology, Psychological Methods and Assessment, LMU Munich, Germany

[4]Institute for Finance & Banking, Munich School of Management, LMU Munich, Germany

[5]Research Unit of Molecular Epidemiology, Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany

[6]Department of Geography, University of Innsbruck, Austria

[7]Department of Statistics, Faculty of Mathematics, Computer Science and Statistics, LMU Munich, Germany

*To whom correspondence should be addressed; E-mail: boulesteix@ibe.med.uni-muenchen.de.

For a given research question, there are usually a large variety of possible analysis strategies acceptable according to the scientific standards of the field, and there are concerns that this multiplicity of analysis strategies plays an important role in the non-replicability of research findings. Here, we define a general framework on common sources of uncertainty arising in computational analyses that lead to this multiplicity, and apply this framework within an overview of approaches proposed across disciplines to address the issue. Armed with this framework, and a set of recommendations derived therefrom, researchers will be able to recognize strategies applicable to their field and use them to generate findings more likely to be replicated in future studies, ultimately improving the credibility of the scientific process.

In recent years, the scientific community has been rocked by the recognition that research findings often do not replicate on independent data, leading to what has been referred to as a replication crisis [1], reproducibility crisis [2] or statistical crisis in science [3]. In particular, a series of attempts to reproduce the results of published research findings in different disciplines found that these replication efforts produced much weaker evidence than the original study [4, 5, 6, 7]. It has been estimated that in preclinical research alone, approximately $28 billion dollars are spent every year on research findings that are not replicable[1] [10].

The crisis has consequences far beyond an insular world of scientists. Experts with strongly disagreeing viewpoints and publicized results that are subsequently contradicted are highly detrimental to the trust the general public and decision makers have in scientific results. This distrust endangers one of the key functions of science - providing robust research findings that can be built upon to help tackle important challenges to society [11]. The recent intense public debate surrounding key "global issues" identified and targeted by the United Nations [12] – such as, for example, climate change and migration – and the ambivalent public perception of scientific contributions to these issues illustrates the importance of the scientific community retaining credibility.

While there have been a number of widely publicized examples of fraud and scientific misconduct [13, 14], many researchers agree that this is not the major problem in this crisis [3, 15]. Instead, the problems seem to be more subtle and partly due to the multiplicity of possible analysis strategies [16, 17].

For a given research question of interest, there is usually great flexibility in the

---

[1]While the two terms replicability and reproducibility are sometimes used interchangeably, we will here adopt the definition of Peng [8] and Stodden et al. [9] according to which replicability is the ability to obtain a similar result when an independent data set is analyzed, while reproducibility is the ability to obtain identical results when the same data set is analyzed by an independent researcher using the same methods.

choice of analysis strategy, as many possible strategies are acceptable according to the scientific standards of the field [18, 19]. The resulting multiplicity of possible analysis strategies is nicely illustrated by a recent experiment performed by Silberzahn et al. [20]: they recruited 29 teams of researchers with strong statistical background and asked them to answer the same research question ("Are football referees more likely to give red cards to players with dark skin than to players with light skin?") with the same data set. The teams obtained highly varied results, as they approached the data with a wide array of analytical techniques. There is evidence that the combination of this multiplicity with selective reporting can systematically lead to an increase in false positive results, inflated effect sizes and overoptimistic measures of predictive performance [21, 22, 23, 24, 17]. Ignoring the multiplicity of analysis strategies can therefore lead to an overconfidence in the precision of results and to research findings that do not replicate on independent data.

While the social and biomedical sciences have been at the heart of the recent replication crisis in science, the multiplicity of analysis strategies has also contributed to credibility crises in other disciplines, e.g. - very prominently - in climatology. In 2009, emails and documents of leading climate scientists at the University of East Anglia became publicly available. Taken out of context, parts of these emails suggested that researchers felt it was "a travesty" they could not "account for the lack of warming", and included an allusion to "Mike's Nature trick" to "hide the decline" [25]. This incident, which received broad media attention, became known as "climategate" and led to an erosion of belief in climate change [26, 27] by creating the impression that climatologists are exploiting the multiplicity of possible analysis strategies to obtain overly alarmist results.

In response to the current crisis in science, a myriad of solutions to improve

the replicability of empirical findings have been developed in different disciplines. There are for instance a number of recently proposed approaches which assess the robustness of research findings to alternative analytical pathways by reporting the results of a large number of analysis strategies: the "vibration of effects" approach in epidemiology [28], "specification curve analysis" [29] and "multiverse analysis" in psychology [30], a "measure of robustness to misspecification" in economics [31] or "multimodel analysis" [32] and "computational robustness analysis" [33] in sociology. In other disciplines, including climatology, ecology and risk analysis, there is a long-standing tradition of addressing the robustness to alternative analysis strategies through sensitivity analyses, multimodel ensembles [34] and Bayesian model averaging [35, 36]. While the development of approaches addressing the issue of the multiplicity of possible analysis strategies remains important, we currently risk "reinventing the wheel" in each discipline. In order to avoid the proliferation of approaches that address the same problems with similar ideas, we consider it advisable to benefit from lessons learned in other disciplines by means of a multidisciplinary perspective. In this work, we define a framework on common sources of uncertainty arising in computational analyses across a broad range of disciplines, covering both the statistical analysis of empirical data and the prediction of complex systems of interest through mechanistic physically-based models. The aim of this framework is to provide a common language to efficiently translate ideas and approaches across disciplines. We illustrate how it can help researchers benefit from experiences gained in other fields by giving an overview of solutions and ideas that have been proposed to improve the replicability and credibility of research findings.

The multiplicity of analysis strategies: examples from epidemiology and hydroclimatology

In a large number of disciplines, an important part of a given research project is the generation of numerical results describing the association between p input variables (denoted by $X_1, X_2, \ldots, X_p$ in the following) and an outcome variable Y through a mathematical function $f()$. $f()$ is typically referred to as the model, while the input variables $X_1, X_2, \ldots, X_p$ are called independent, exogenous, or explanatory variables, or predictors, features, attributes or covariates, depending on context and discipline. The outcome variable Y is also known as the dependent, endogenous, or response variable, or the output, label, criterion, or predictand. In the following, we refer to this type of research, which relies to some extent on data, as empirical research, in contrast to research that is of a purely theoretical nature.

To illustrate the multitude of analyst decisions necessary in empirical research, we consider two examples from different disciplines. The first example is the analysis of data from an epidemiological study on the link between meat consumption and the risk of colorectal cancer to answer a research question in public health which has attracted considerable attention in recent years. The second example, from hydroclimatology, concerns the prediction of water mass stored in seasonal snowpack and its release as meltwater into the river runoff [37, 38], which is essential in the prediction of future flood occurrence and water availability for irrigation and hydropower generation. Figure 1 illustrates how the analysis decisions in empirical research, applied to these two examples, lead to a multiplicity of analysis strategies.

Both in observational epidemiology and hydroclimatology, the first step is to col-
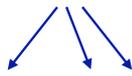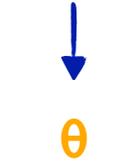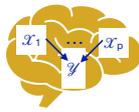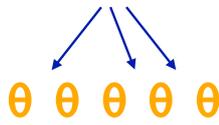
Figure 1: The multiplicity of analysis strategies to obtain an estimate of the parameter of interest θ and values of the outcome variable $\mathcal{Y}$ in epidemiology and hydroclimatology, respectively

lect data on the phenomenon of interest. In our example from epidemiology, these data come from n individuals who are assumed to form a representative sample from a specified population of interest. In our example from hydroclimatology, the system to be investigated is a valley of a certain size in which the winter snowpack is transformed into spring snowmelt-induced streamflow. The data, which are indexed by space and time, consist of two parts. The first part are historical data for which observations at one or more meteorological stations exist for the input

variables which include, e.g., time series data of measured air temperature and precipitation covering the previous 20 years. The outcome is the gauge streamflow at the outlet of the valley. The second part consists of values of the model input variables which reflect future changes in temperature and precipitation resulting from different greenhouse gas emission scenarios: with these model input variables in hand, and a developed model, one can predict the future evolution of the seasonal snowpack and hence, the resulting streamflow.

For both examples, before beginning analysis, the data must be preprocessed, a procedure involving numerous subjective choices. The flexibility in data preprocessing partly arises because the research hypotheses are generally not precise enough that they fully specify the input and the outcome variables [17]. Indeed, while measuring meat consumption and determining incidence of colon cancer may naïvely appear to be straightforward, the analyst has considerable flexibility in the definition of these two variables. For the same research question, one could consider meat consumption of all kinds, focus on red meat or processed meat, or distinguish between beef, pork, lamb and chicken [39]. Similarly, concerning the outcome, it is possible to concentrate on colon cancer, on rectal cancer or to include all types of colorectal cancer and even precancerous lesions like colorectal adenoma. In our example from hydroclimatology, we are faced with similar choices. For instance, in the absence of measurements of the input variables for all locations in the region of interest, the recordings from a single meteorological station have to be extrapolated. Further, the possible values of temperature and precipitation reflecting changes in greenhouse gas emission scenarios are themselves outputs of mechanistic models. There are a large number of strategies to obtain both spatial extrapolations and these possible future values.

Following data preprocessing, we next decide on a model to describe our phe-

nomenon of interest. In epidemiology, the aim is to control for all variables which might confound the association between meat consumption and colorectal cancer. These variables may for instance include body mass index, smoking, physical activity, socioeconomic status and the consumption of alcohol, fruits and vegetables [40, 41]. However, this is not an exhaustive list and there is no clear guidance on which variables should be considered mandatory in the model and which will lead to an unnecessarily complex description of the phenomenon of interest. Similarly, a model describing how snow accumulates, is stored and melts can be based on a variety of alternative model assumptions, potentially leading to "1701 snow models" [37]: examples include modelling the snow microstructure and evolution over time in physical detail, or a more simplified description of these processes in a snowpack representation with a single layer. Additionally, there are a number of model parameters to be specified in order to predict the seasonal snowpack evolution, parameters such as, e.g., the surface albedo and roughness length, and the initial density of the snow [38].

Once the model structure and its constituent variables are specified, there remain further decisions concerning the method to use to obtain the main result of interest. In our example from epidemiology, the result of interest is an estimate of a parameter θ describing the increase in colorectal cancer risk associated with meat consumption. Common estimation techniques for this model parameter include maximum likelihood estimation, Bayesian inference, least squares minimization and the method of moments. In our hydroclimatology example, the result of interest is the outcome $\mathcal{Y}$, i.e. the water equivalent in the seasonal snowpack and its melt. A first method choice concerns the discretization of the system to be investigated in space and time, i.e. the temporal and spatial resolution of the model setup. Further method choices arise due to the terabytes of simulation outputs

typically produced. These simulation outputs consist of a myriad of state variables describing for instance snow depth, snow surface temperature and snow density. To be able to interpret these outputs, they have to be aggregated, analyzed and illustrated; however, different transformations and spatial and temporal aggregation techniques may either mask or accentuate oscillations and trends which may be present in these results.

Through the illustration of two examples from different fields, we see that for a research question of interest, researchers across disciplines are faced with a multitude of choices when presented with data in a situation where there is no clear guidance on analysis strategy from a theoretical or a substantive point of view. Although some of the choices made could be considered 'wrong', many would also be justifiable. As all justifiable paths are likely to lead to different results, we see there is a source of variability attributable to the choice of analysis strategy. In the presentation of scientific results, however, this variability is not commonly accounted for or discussed.

Sources of uncertainty arising in empirical research

As the aim of research is to expand existing knowledge by operating on the edge of what is known, it is hardly surprising that there are numerous sources of uncertainty arising in scientific discovery. In this section, we will introduce a general, albeit inevitably incomplete, framework on common sources of uncertainty arising in computational analyses and show how the combination of these sources of uncertainty with selective reporting can lead to unreplicable research findings.

The idea behind the use of a mathematical model $f()$ is, in general, either ex-

planation or prediction [42, 43]. The modelling of the association between meat consumption and colorectal cancer in epidemiology can be seen as an instance of explanatory modelling. The main aims of explanatory modelling are to test a causal hypothesis [24], i.e. to assess to what extent a theoretical variable $\mathcal{Y}$ is influenced by the theoretical variables $\mathcal{X}_1, \ldots, \mathcal{X}_p$. As these theoretical variables are not directly observable, they have to be operationalized by defining measurable outcome Y and input variables $X_1, X_2, \ldots, X_p$ [43]. Once these observable variables are measured on a sample of observations, statistical methods can be used to estimate the value of an unknown parameter of interest θ which quantifies the association between Y and $X_1, \ldots, X_p$. The reporting of this parameter estimate is typically combined with a p-value and a confidence interval, which are used to test the research hypothesis concerning the association between $\mathcal{Y}$ and $\mathcal{X}_1, \ldots, \mathcal{X}_p$. Our example from hydroclimatology, the modelling of the future evolution of the seasonal snowpack, on the other hand, can be seen as an instance of what can be referred to as mechanistic predictive modelling. The idea of mechanistic predictive modelling is to predict the values of an outcome $\mathcal{Y}$ at new or untried values of the input variables. In contrast to explanatory modelling, the goal of mechanistic predictive modelling is hence to apply a model to predict the behavior of a system which is so complex that it would be difficult to predict and analyze otherwise [44]. Mechanistic predictive models (also referred to as physically- or process-based models [45, 46]) typically heavily rely on subject matter knowledge and the (physical) principles underlying the behavior of the studied phenomenon. Hence, model predictions are derived by relying on a number of physical laws or mechanistic assumptions, and the values of a certain number of model parameters $\theta = (\theta_1, \ldots, \theta_p)$, which are assumed to be known.

With the increasing availability of large data sets and improvements in com-
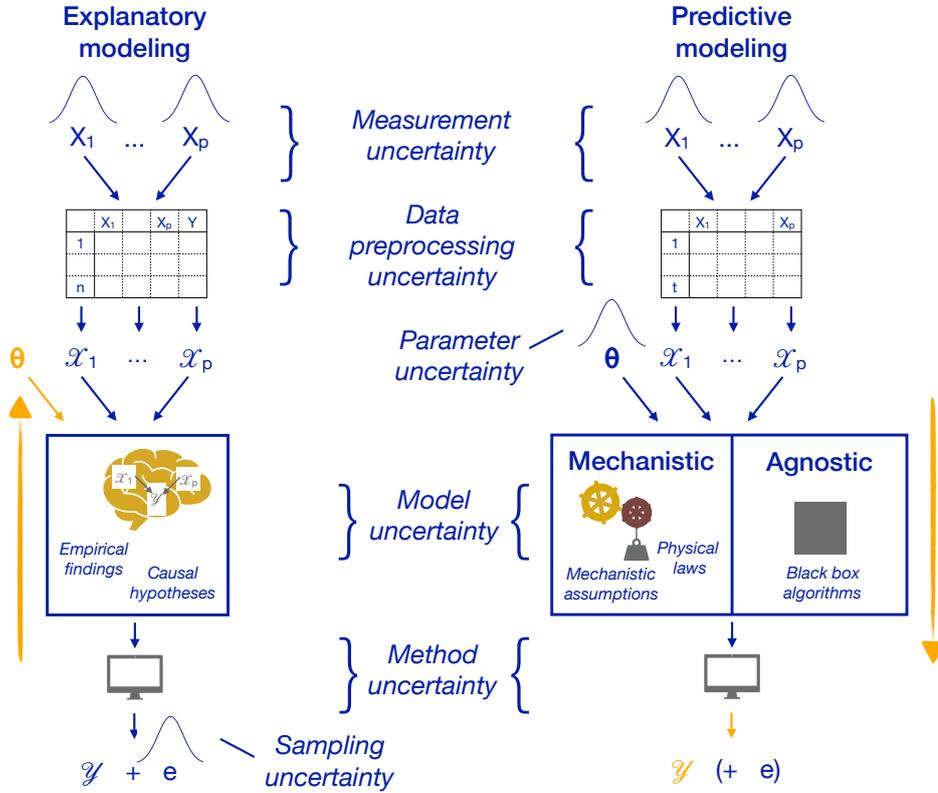
Figure 2: Sources of uncertainty in explanatory, mechanistic predictive and agnostic predictive modelling. The model structure describes the association between the p input variables $X_1, X_2, \ldots, X_p$ and the outcome of interest Y. θ is a parameter and e represents a probabilistic error term.

putational efficiency, a second type of data-driven predictive modelling coming from an algorithmic modelling culture - often from artificial intelligence and more specifically from machine learning - is growing in popularity in many disciplines [47, 48]. Here the function $f()$ is estimated by an algorithm rather than by fitting a pre-specified model class to the data [42]. These algorithms try to dispense with (potentially restrictive) assumptions on the association between $X_1, X_2, \ldots, X_p$ and Y [49] and do not typically rely on theoretical reasoning: they can thus be referred to as "agnostic" predictive models.

Despite the main focus of explanatory modelling being the estimation of an un-

known parameter and the main focus of predictive modelling being prediction, many analyses are concerned with both aims. For example, some parameter values in mechanistic predictive modelling can be determined by fitting the model to historical data in which both the input variables and the outcome are measured, a process referred to as calibration [44]. Conversely, in explanatory modelling, where the main focus is on explanation, the estimated parameter values can be used to predict new observations and to evaluate the adequacy of the chosen probability model [24].

While certain types of models are more popular in some disciplines than in others, there is no unique assignment of disciplines to modelling strategies. Mechanistic predictive models, which are popular in disciplines ranging from the geosciences and risk analysis to decision analytic modelling in health economics, can also be used in the prediction of infectious disease dynamics in epidemiology [50]. Explanatory modelling, on the other hand, which is popular in disciplines such as biology, psychology and economics [43, 51], can for example also be applied in climatology when assessing the extent to which an extreme event can be attributed to anthropogenic climate change [52].

As illustrated in our examples from epidemiology and hydroclimatology, the multiplicity of possible analysis strategies can arise from data preprocessing, parameter, model and method choices. *Data preprocessing uncertainty* is caused by all decisions needing to be made in the selection of the data to analyze and in the definition, the cleaning and the transformation of the input and the outcome variables. Additionally, there is usually *model uncertainty*, as the best or the "true" model structure to describe the phenomenon of interest is unknown. *Parameter uncertainty*, which is mainly present in predictive modelling, arises through model parameter values having to be specified when the analyst is armed with neither

13

precise theoretical knowledge nor direct measurements [44]. In mechanistic predictive modelling, estimates of these parameter values can be achieved by observation of the system, through calibration or incomplete expert knowledge [44, 46], but substantial uncertainty regarding the true values typically remains. Similarly, the performance of many algorithms for agnostic modelling is sensitive to hyperparameters [19], which have to be specified before running these algorithms. Examples include the minimal size of splitting nodes in random forests, the kernel and the cost parameter in support vector machines, and the number of neighbors in k-nearest neighbors [53]. As there is typically no clear guidance on which values to choose for these hyperparameters, their specification is sometimes considered to be more of an art than a science. Finally, specifying a model and parameter values is not sufficient to run the actual computations – a specific implementation and computational method must be chosen, or even developed, before a statistical model can be estimated or predictions from a mechanistic or agnostic model derived. Again, there is a multitude of options without clear guidance or a definitive choice on the method that will provide the most suitable answer to their research question: researchers encounter here *method uncertainty* [54].

The uncertainties detailed above – data preprocessing uncertainty, parameter uncertainty, model uncertainty and method uncertainty – are epistemic: they arise due to a lack of knowledge. As illustrated in Figure 2, these epistemic sources of uncertainty can be contrasted with two additional sources of uncertainty, namely *measurement uncertainty* and *sampling uncertainty.*

Measurement uncertainty is ubiquitous in empirical research as it is generally impossible to determine the input variables $X_1, \ldots, X_p$ and the outcome Y with absolute precision and accuracy. Depending on the discipline, information on these variables may be acquired through questionnaires, measurement devices or exper-

imental protocols, which are all, to some extent, prone to imprecision. The final type of uncertainty described here, especially prominent in explanatory modelling, results from the variability introduced when analyzing a data set assumed to be a random sample from a larger population of interest. This variability can be described as *sampling uncertainty*, and is often expressed through an error term e.

The interplay between these random sources of uncertainty and the multiplicity
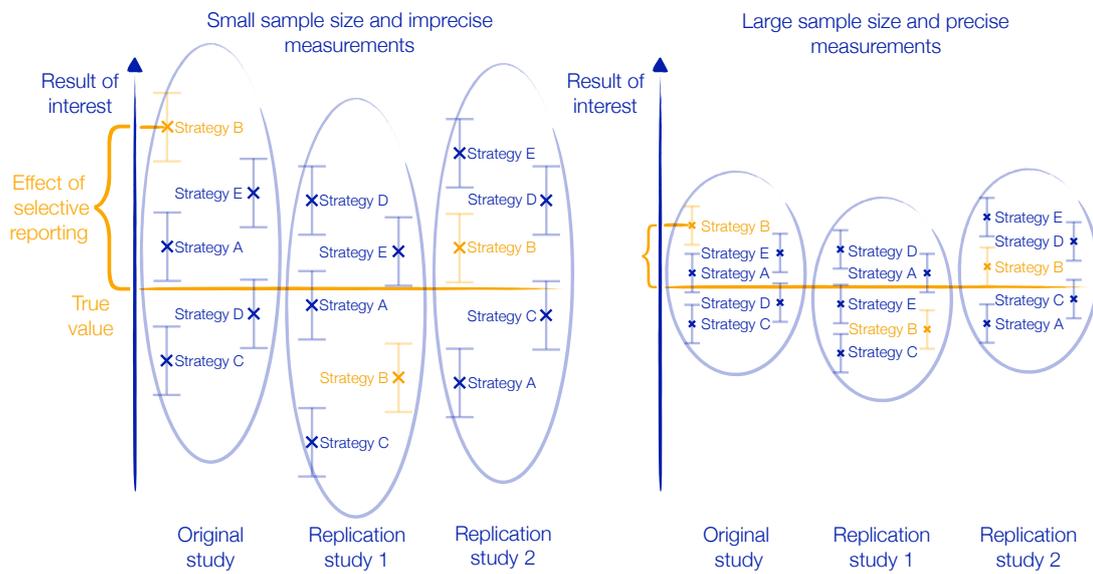
Figure 3: The impact of random sources of uncertainty and of the multiplicity of possible analysis strategies on the replicability of research findings. The result of interest is the parameter θ in explanatory modelling, the outcome $\mathcal{Y}$ in mechanistic predictive modelling and the predictive performance in agnostic predictive modelling.

of analysis strategies arising from the four epistemic sources of uncertainty can lead to unreplicable research findings when combined with selective reporting, as illustrated in Figure 3. If there is no restriction on the chosen analysis strategy, a researcher may try to compare the results of many strategies – each a path resulting from the given preprocessing, model, parameter and method choices – and then select the final analysis strategy based on the "nicest" result: a smaller

15

p-value, an effect in the "desired" direction, or predictions in accordance with the expectations of the researcher, for example. This "selective reporting" can lead to a substantial overestimation of the result of interest in empirical research, an effect heightened by small samples and imprecise measurements [1, 55].

In explanatory modelling, confidence intervals and hypothesis tests typically only account for sampling uncertainty, while ignoring measurement and epistemic uncertainty. This limited focus thereby leads to apparently precise results which are not robust to variations in the choice of analysis strategy; they therefore have a high probability of being contradicted in a replication study.

Lessons learned across disciplines

The solution to the replication crisis which has probably received the most attention in the scientific community and beyond is the abandonment of statistical significance within the scientific literature and its replacement with Bayes factors, confidence intervals or other inferential methods [56, 57, 58, 59, 60]. While there may be inferential paradigms that are easier to interpret and less prone to overconfidence than null hypothesis statistical testing [61, 62], the simple solution of jettisoning statistical significance can seem somewhat shortsighted in addressing the issues. Taking confidence intervals as an example, Figure 3 illustrates how they can be just as prone as p-values to selective reporting. More generally, no inferential paradigm in itself is immune to overconfidence and the result-dependent selection of an analysis strategy from among a multiplicity of possible strategies [63, 64].

Another prominent solution which has been proposed in response to the replication crisis in science is to increase transparency by promoting reproducible research

practices [65, 66]. While the publication of research data, code and materials can help build trust in science [67] and make the entire research process more efficient, it only indirectly addresses the multiplicity of possible analysis strategies. Transparency alone is again not enough to prevent selective reporting or eliminate overconfidence in results [68].

We therefore argue that we have to go beyond open science practices and the focus on statistical significance as the main culprit in the non-replicability of research findings by explicitly addressing the sources of uncertainty introduced in the previous section. A wealth of ideas and approaches to reduce, report, integrate or accept one or several of the sources of uncertainty have been discussed in the literature, leading to a myriad of solutions in different disciplines. In Figures 4 and 5, we give an overview of these solutions and derive simple steps researchers can take to make their own research more replicable and credible.
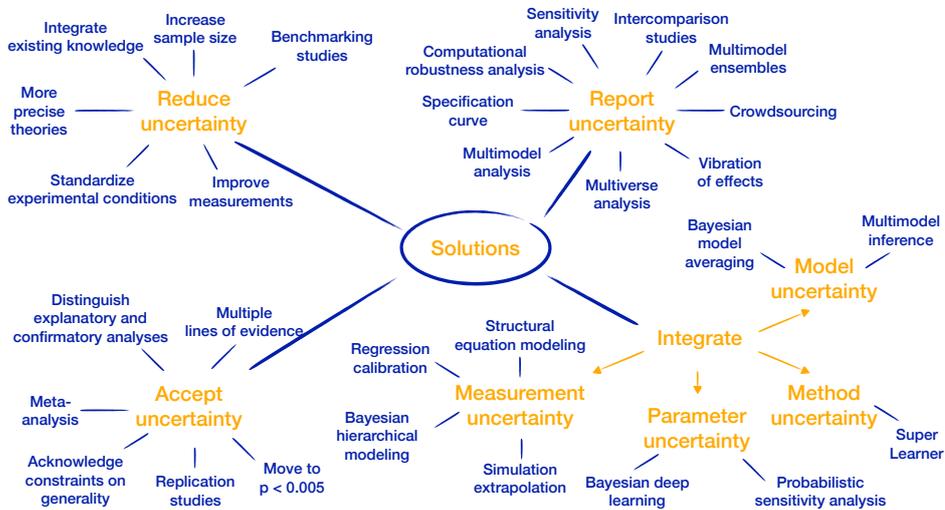
Figure 4: Overview of solutions to the replication crisis which address the multiplicity of analysis strategies by reducing, reporting, integrating or accepting uncertainty

Reduce uncertainty

There are a variety of strategies to reduce one or several sources of uncertainty. In explanatory modelling, a reduction in sampling and measurement uncertainty can for instance be achieved by increasing the sample size of studies [69, 70, 71], by improving the quality of measurements [1] or by standardizing experimental conditions [72, 73, 74]. To reduce model and data preprocessing uncertainty, Steegen et al. [30] and Schaller [75] call for more conceptual rigor and precise theories to reduce the number of possible analysis strategies. Method uncertainty, on the other hand, can be reduced through adoption of the results of "benchmarking" studies, which aim to identify a best method for a given research question of interest in a given setting [54, 76]. The integration of existing knowledge into explanatory modelling can also reduce uncertainty and help to obtain more precise parameter estimates. An example is the specification of informative prior distributions [77] in Bayesian inference, where the prior evidence can range from functional information in genome-wide association studies [78] to historical data in clinical trials [79].

Report uncertainty

In many disciplines, there is a long-standing tradition of reporting the results of a large number of possible analysis strategies, or the variability of these results, to assess their robustness to alternative assumptions and model specifications. Common examples of this strategy include extreme bounds analysis in econometrics [80], multimodel ensembles [34] in hydrology and climatology and sensitivity analyses, which are used across many disciplines. More recent approaches to report uncertainty include the "vibration of effects" approach [28], "specification curve analysis" [29], "multiverse analysis" [30] "multimodel analysis" [32] and "compu-

tational robustness analysis" [33], as discussed previously. Silberzahn et al. [20] go a step further and propose the reporting of the results of different teams of researchers analyzing the same research question on the same data set. In their "crowdsourcing approach", it is thereby possible to simultaneously report the variability in results due to data preprocessing, model and method uncertainty, as different teams of researchers are likely to follow different paths in formulating their analysis strategies. Considered from a multidisciplinary perspective, this approach is similar to intercomparison studies, which have a long tradition in mechanistic predictive modelling in hydroclimatology [81]. In cases where the data are made publicly available, methods to report uncertainty can be applied by readers and reviewers to assess to what extent the originally reported results are robust to alternative analysis choices.

Integrate uncertainty

There are a number of approaches which can generate broader and more realistic uncertainty intervals by integrating measurement, model, parameter or method uncertainty when deriving parameter estimates in explanatory modelling and predictions in mechanistic and agnostic predictive modelling. In explanatory modelling, it is possible to account for measurement uncertainty through, for example, structural equation models [82], Bayesian hierarchical approaches [83], simulation extrapolation or regression calibration [84]. With regard to model uncertainty, Bayesian model averaging [35, 36] and multimodel inference [85] go beyond the simple reporting of the results of all possible models by weighting the parameter estimates or predictions of all candidate models to produce a single summary measure and a measure of its uncertainty. In mechanistic predictive modelling, Bayesian melding [86] and probabilistic sensitivity analysis [87] can be used to

integrate parameter uncertainty. Similarly, in agnostic modelling, it has been suggested to account for parameter uncertainty through Bayesian deep learning, where the uncertainty in hyperparameters is described by a prior distribution [88], and to integrate method uncertainty by combining the weighted predictions of a great number of candidate methods through a so-called "Super Learner" [89].

Accept uncertainty

Many authors have argued that classical statistical methods used in explanatory modelling suggest a disproportionate level of certainty [90, 57] and that the replication crisis in science is in fact a "crisis of overconfidence in statistical results" [91]. In this sense, a solution to the current crisis is to acknowledge the inherent uncertainty in scientific findings. This can be achieved by recognizing that statistical inference within exploratory analyses should be interpreted with great caution and that scientific generalizations need to be based on cumulative knowledge rather than on a single study [91]. Strictly confirmatory analyses can be realized either through the pre-registration of analysis plans [92, 93] and registered reports [94], where the analysis strategy is specified in detail before observing the data; or through blind analyses, where researchers select an analysis strategy while being blinded to the outcome of interest [95]. Alternatively, it is common in agnostic modeling to perform exploratory and confirmatory analysis on the same data set through split analysis plans: one part of the data is used to determine the best analysis strategy, the other to fit the final algorithm and determine its predictive performance [51].

A focus on cumulative evidence can be found in calls for replications as post-publication quality control [96, 97, 17] and in the proposal of Benjamin et al. [98] to redefine statistical significance by considering a p-value of < 0.05 merely sug-

gestive (i.e. having to be confirmed in subsequent studies) and only p-values < 0.005 significant. In psychology, Simons et al. [99] emphasize the need for cumulative evidence and encourage authors to specify a "constraints on generality" statement, which clearly identifies and justifies the target population of reported research findings. In biology, cumulative knowledge can be achieved by providing multiple lines of convergent evidence from several independent experiments, experiments performed for instance with isolated molecules, in cultured cell lines, or using animal models. Lastly, in psychology and medicine we see the usefulness of meta-analyses, the summarization and aggregation of the results of similar studies.

Steps to take to make one's own research more replicable

Based on the lessons learned across disciplines discussed in the previous section, what are the steps an individual researcher can take to improve the replicability and credibility or his or her own research? A first step, which should not be underestimated, is simply to be aware of the multiplicity of possible analysis strategies and the potential for selective reporting. As pointed out by Nuzzo [100], even the most honest researcher is a master of self-deception and it is easy to jump to conclusions when finding patterns in randomness.

Once aware of the potential for increased uncertainty, one should evaluate and implement possibilities to reduce both the randomness in the data and the flexibility in analysis plan. In our example from epidemiology, we could for instance have reduced some of the sources of uncertainty by determining an adequate sample size through a power calculation, by integrating results from previous studies on meat consumption and colorectal cancer to specify an informative prior distribution and by clearly defining the research hypothesis and the input and outcome

21

Step 1:
Be aware of the multiplicity of possible analysis strategies

Step 2:
If possible, reduce sources of uncertainty before the analysis

Step 3a:
If possible, integrate remaining sources of uncertainty into the analysis

Step 3b:
Report the results of alternative analysis strategies to assess
the robustness of results

Step 4:
Acknowledge the inherent uncertainty in your findings

**Step 5:**
Publish all research code, data and material

Figure 5: Recommendations on steps researchers can take to make their research findings more replicable and credible

variables before collecting the data. In our hydroclimatology study, measurement and model uncertainty could have been reduced by, for example, including a very large number of confirmed measurements of the input variables, e.g. through the integration of remote sensing data, and by using only models that had been extensively validated elsewhere.

When deriving the result of interest, one should attempt to integrate all sources of uncertainty which could not be reduced in the previous step. Since to this point there is no all-encompassing method accounting for model, measurement, method,

data preprocessing and parameter uncertainty simultaneously, an alternative is to systematically report the robustness to alternative analysis strategies through one of the approaches presented in the last section. The next step – again, not to be underestimated – is to acknowledge the inherent uncertainty in the presented research findings and thereby avoid misleading readers into overinterpretation of the relevance of the results.

Finally, to make one's research findings more credible and improve the efficiency of the research process as a whole, one should publish all research code, data and material, both to allow others to try alternative analysis strategies and for reuse of the data in future studies.

Outlook

Despite growing evidence for its pervasive impact on the validity of research findings, current research practices largely fail to address the multiplicity of analysis strategies. Currently, it is a highly profitable strategy to analyze small data sets and to exploit the multiplicity of possible analysis strategies arising from data preprocessing, model, parameter, and method uncertainty to obtain significant and surprising results. These results have a high probability of getting published, but a low probability of being replicated in subsequent studies. In the short-term this lack of replication may simply be embarrassing, but in the long run this strategy has devastating consequences for the scientific community. While imprecise but convergent results are often readily accepted by the public, multiple apparently precise but contradictory results have a negative impact on the credibility of research findings [101, 102]: these contradictory results can easily be discredited as conflicting evidence to create the impression that scientific knowledge is unreliable

and that there is no scientific consensus on important research topics [103]. According to van der Linden et al. [104], this line of argumentation has been used for years to delay or prevent regulatory actions concerning climate change, maybe contributing to the low belief in anthropogenic climate change among the American public [105].

If, on the other hand, we address the multiplicity of possible analysis strategies arising through data preprocessing, model, parameter and method uncertainty through reporting, integrating and acknowledgement, we will obtain broad but more realistic measures of our uncertainty, and research findings that are robust to the choice of the analysis strategy.

It is important to raise awareness of the fact that the multiplicity of possible analysis strategies is an issue affecting many different disciplines in similar ways; this awareness will enable us to join forces in our efforts to increase the transparency, replicability and credibility of research findings. Integrating multidisciplinary experience and insights is not only essential in the further development of appropriate solutions and in the elaboration of guidelines to help researchers make their research more replicable, but also in generating enough momentum to bring about change. As long as the reward structure in academia favors significant, overly clear-cut, and hypothesis-consistent results, researchers might be tempted to exploit the multiplicity of possible analysis strategies instead of addressing this issue in a transparent way to make research findings more replicable. This creates a social dilemma structure where societal and scientific interests are at odds with the individual career interest of researchers.

The multiplicity of possible analysis strategies is likely to become an even bigger challenge with the advent of increasing amounts of data that are not originally recorded for research purposes in many disciplines, for instance in the form of

routine care data in medicine, of administrative data in the social sciences and of remote sensing data in ecology [106, 107, 108, 24]. These data are not the result of well-designed experiments where we have accurate knowledge on the data generation process, and a small set of research hypotheses of interest. Instead, the data may be imperfect, heterogeneous, noisy, and high-dimensional [47, 109]. When analyzing these data, sampling uncertainty, which has attracted a disproportionate amount of attention in the scientific community to this point, will be comparably small, but measurement, data preprocessing, model, and method uncertainty will be much larger than when dealing with more traditional data [110].

Given the importance and the urgency of the challenges we are facing today, we need scientific results that are veracious - both in their precision and in their (preliminary) imprecision. Novel and exciting but unreplicable results impede scientific progress and its societal translation. By addressing the multiplicity of possible analysis strategies through the framework and approaches suggested here, we can make the research process more efficient and improve the replicability, and ultimately the credibility, of research findings.

Bibliography

[1] Loken E, Gelman A. Measurement error and the replication crisis. Science. 2017;355(6325):584–585.

[2] Baker M. Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the'crisis rocking science and what they think will help. Nature. 2016;533(7604):452–455.

[3] Gelman A, Loken E. The Statistical Crisis in Science. American Scientist. 2014;102(6):460–465.

[4] Ioannidis JPA, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, et al. Repeatability of published microarray gene expression analyses. Nat Genet. 2009 Feb;41(2):149–55.

[5] Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? Nature Reviews Drug Discovery. 2011;10(9):712.

[6] Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012 Mar;483(7391):531–3.

[7] Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015 Aug;349(6251):aac4716.

[8] Peng R. The reproducibility crisis in science: A statistical counterattack. Significance. 2015;12(3):30–32.

[9] Stodden V, Leisch F, Peng RD. Implementing reproducible research. CRC Press, New York; 2014.

[10] Freedman LP, Cockburn IM, Simcoe TS. The economics of reproducibility in preclinical research. PLoS Biology. 2015;13(6):e1002165.

[11] Barnosky AD, Ehrlich PR, Hadly EA. Avoiding collapse: Grand challenges for science and society to solve by 2050. Elem Sci Anth. 2016;4:000094.

[12] Gaffney O. Sustainable development goals: improving human and planetary wellbeing. Global Change. 2014;82:20–23.

[13] Ince D. The Duke University scandalwhat can be done? Significance. 2011;8(3):113–115.

[14] van der Zee T, Anaya J, Brown NJ. Statistical heartburn: An attempt to digest four pizza publications from the Cornell Food and Brand Lab. BMC Nutrition. 2017;3(1):54.

[15] Ioannidis JP, Munafo MR, Fusar-Poli P, Nosek BA, David SP. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends in Cognitive Sciences. 2014;18(5):235–241.

[16] Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? Science Translational Medicine. 2016;8:341ps12.

[17] Harvey CR. Presidential address: The scientific outlook in financial economics. The Journal of Finance. 2017;72(4):1399–1440.

[18] Gelman A, Hennig C. Beyond subjective and objective in statistics. Journal of the Royal Statistical Society: Series A (Statistics in Society). 2017;180(4):967–1033.

[19] Hutson M. Artificial intelligence faces reproducibility crisis. Science. 2018;359:725–726.

[20] Silberzahn R, Uhlmann EL, Martin DP, Anselmi P, Aust F, Awtrey E, et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. Advances in Methods and Practices in Psychological Science. 2018;1(3):337–356.

[21] Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;2(8):e124.

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

[23] Wasserstein RL, Lazar NA. The ASA's statement on p-values: Context, process and purpose. The American Statistician. 2016;70(2):129–133.

[24] Hofman JM, Sharma A, Watts DJ. Prediction and explanation in social systems. Science. 2017;355(6324):486–488.

[25] Pearce F. How the 'climategate' scandal is bogus and based on climate sceptics' lies. The Guardian, Tue 9 Feb 2010. 2010;.

[26] Tollefson J. Climate science: An erosion of trust? Nature News. 2010;466(7302):24–26.

[27] Leiserowitz AA, Maibach EW, Roser-Renouf C, Smith N, Dawson E. Climategate, public opinion, and the loss of trust. American Behavioral Scientist. 2013;57(6):818–837.

[28] Ioannidis JP. Why most discovered true associations are inflated. Epidemiology. 2008;19(5):640–648.

[29] Simonsohn U, Simmons JP, Nelson LD. Specification curve: Descriptive and inferential statistics on all reasonable specifications. SSRN Electronic Journal. 2015;DOI: 10.2139/ssrn.2694998.

[30] Steegen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing Transparency Through a Multiverse Analysis. Perspect Psychol Sci. 2016 09;11(5):702–712.

[31] Athey S, Imbens G. A measure of robustness to misspecification. American Economic Review. 2015;105(5):476–80.

[32] Young C, Holsteen K. Model uncertainty and robustness: a computational framework for multimodel analysis. Sociological Methods & Research. 2017;46(1):3–40.

[33] Muñoz J, Young C. We ran 9 billion regressions: Eliminating false positives through computational model robustness. Sociological Methodology. 2018;48(1):1–33.

[34] Palmer TN, Alessandri A, Andersen U, Cantelaube P, Davey M, Delécluse P, et al. Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bulletin of the American Meteorological Society. 2004;85(6):853–872.

[35] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. Statistical Science. 1999;14:382–401.

[36] Raftery AE, Gneiting T, Balabdaoui F, Polakowski M. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review. 2005;133(5):1155–1174.

[37] Essery R, Morin S, Lejeune Y, Ménard CB. A comparison of 1701 snow models using observations from an alpine site. Advances in Water Resources. 2013;55:131–148.

[38] Günther D, Marke T, Essery R, Strasser U. Uncertainties in Snowpack Simulations-Assessing the Impact of Model Structure, Parameter Choice and Forcing Data Error on Point-Scale Energy-Balance Snow Model Performance. Water Resources Research. 2019;55:2779–2800.

[39] Carr PR, Walter V, Brenner H, Hoffmeister M. Meat subtypes and their association with colorectal cancer: systematic review and meta-analysis. International Journal of Cancer. 2016;138(2):293–302.

[40] Chan DS, Lau R, Aune D, Vieira R, Greenwood DC, Kampman E, et al. Red and processed meat and colorectal cancer incidence: meta-analysis of prospective studies. PLOS ONE. 2011;6(6):e20456.

[41] Aune D, Chan DS, Vieira AR, Rosenblatt DAN, Vieira R, Greenwood DC, et al. Red and processed meat intake and risk of colorectal adenomas: a systematic review and meta-analysis of epidemiological studies. Cancer Causes & Control. 2013;24(4):611–627.

[42] Breiman L. Statistical modeling: The two cultures. Statistical Science. 2001;16(3):199–231.

[43] Shmueli G. To explain or to predict? Statistical Science. 2010;25(3):289–310.

[44] Kennedy MC, O'Hagan A. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001;63(3):425–464.

[45] Sacks J, Welch WJ, Mitchell TJ, Wynn HP. Design and analysis of computer experiments. Statistical Science. 1989;4:409–423.

[46] Oppenheimer M, Little CM, Cooke RM. Expert judgement and uncertainty quantification for climate change. Nature Climate Change. 2016;6(5):445.

[47] Khoury MJ, Ioannidis JP. Big data meets public health. Science. 2014;346(6213):1054–1055.

[48] Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015;349(6245):255–260.

[49] Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nature Methods. 2018;15(4):233–234.

[50] Lessler J, Cummings DA. Mechanistic models of infectious disease and their impact on public health. American Journal of Epidemiology. 2016;183(5):415–422.

[51] Yarkoni T, Westfall J. Choosing prediction over explanation in psychology: Lessons from machine learning. Perspectives on Psychological Science. 2017;12(6):1100–1122.

[52] Stott PA, Christidis N, Otto FE, Sun Y, Vanderlinden JP, Van Oldenborgh GJ, et al. Attribution of extreme weather and climate-related events. Wiley Interdisciplinary Reviews: Climate Change. 2016;7(1):23–41.

[53] Probst P, Boulesteix AL, Bischl B. Tunability: Importance of Hyperparameters of Machine Learning Algorithms. Journal of Machine Learning Research. 2019;20(53):1–32.

[54] Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J, initiative S. Strengthening analytical thinking for observational studies: the STRATOS initiative. Statistics in Medicine. 2014;33(30):5413–5432.

[55] Macleod M. Why animal research needs to improve. Nature News. 2011;477(7366):511.

[56] Wagenmakers EJ. A practical solution to the pervasive problems of p values. Psychonomic Bulletin & Review. 2007;5(14):779–804.

[57] Nuzzo R. Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. Nature. 2014;506:150–152.

[58] Cumming G. The new statistics: Why and how. Psychological Science. 2014;25(1):7–29.

[59] Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019;567(7748):305.

[60] Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p<0.05". The American Statistician. 2019;73(sup1):1–19.

[61] Goodman S. A dirty dozen: twelve p-value misconceptions. Seminars in Hematology. 2008;45(3):135–40.

[62] Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. European Journal of Epidemiology. 2016;31(4):337–50.

[63] Gelman A. The problems with p-values are not just with p-values. The American Statistician, supplemental material to the ASA statement on p-values and statistical significance. 2016;10.

[64] McShane BB, Gal D, Gelman A, Robert C, Tackett JL. Abandon statistical significance. The American Statistician. 2019;73(S1):235–245.

[65] Nosek BA, Alter G, Banks GC, Borsboom D, Bowman S, Breckler S, et al. Promoting an open research culture. Science. 2015;348(6242):1422–1425.

[66] Chen X, Dallmeier-Tiessen S, Dasler R, Feger S, Fokianos P, Gonzalez JB, et al. Open is not enough. Nature Physics. 2018;15:113–119.

[67] Chandler RE, Thorne P, Lawrimore J, Willett K. Building trust in climate science: data products for the 21st century. Environmetrics. 2012;23(5):373–381.

[68] Gelman A. Ethics and Statistics: Honesty and Transparency Are Not Enough. Chance. 2017;30(1):37–39.

[69] Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience. 2013;14(5):365–376.

[70] Maxwell SE. The persistence of underpowered studies in psychological research: causes, consequences, and remedies. Psychological Methods. 2004;9(2):147–163.

[71] Schönbrodt FD, Perugini M. At what sample size do correlations stabilize? Journal of Research in Personality. 2013;47(5):609–612.

[72] Elson M, Mohseni MR, Breuer J, Scharkow M, Quandt T. Press CRTT to measure aggressive behavior: The unstandardized use of the competitive reaction time task in aggression research. Psychological Assessment. 2014;26(2):419–432.

[73] Bradbury A, Plückthun A. Reproducibility: Standardize antibodies used in research. Nature News. 2015;518(7537):27–29.

[74] Jarvis MF, Williams M. Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. Trends in Pharmacological Sciences. 2016;37(4):290–302.

[75] Schaller M. The empirical benefits of conceptual rigor: Systematic articulation of conceptual hypotheses can reduce the risk of non-replicable results (and facilitate novel discoveries too). Journal of Experimental Social Psychology. 2016;66:107–115.

[76] Boulesteix AL, Wilson R, Hapfelmeier A. Towards evidence-based computational statistics: lessons from clinical research on the role and design of real-data benchmark studies. BMC Medical Research Methodology. 2017;17(1):138.

[77] Goldstein M. Subjective Bayesian analysis: Principles and practice. Bayesian Analysis. 2006;1(3):403–420.

[78] Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genetics. 2014;10(10):e1004722.

[79] Viele K, Berry S, Neuenschwander B, Amzal B, Chen F, Enas N, et al. Use of historical control data for assessing treatment effects in clinical trials. Pharmaceutical Statistics. 2014;13(1):41–54.

[80] Leamer EE. Lets take the con out of econometrics. Modelling Economic Series. 1983;73:31–43.

[81] Etchevers P, Martin E, Brown R, Fierz C, Lejeune Y, Bazile E, et al. Validation of the energy budget of an alpine snowpack simulated by several snow models (Snow MIP project). Annals of Glaciology. 2004;38:150–158.

[82] Bagozzi RP, Yi Y. Specification, evaluation, and interpretation of structural equation models. Journal of the Academy of Marketing Science. 2012;40(1):8–34.

[83] Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. American Journal of Epidemiology. 1993;138(6):430–442.

[84] Bennett DA, Landry D, Little J, Minelli C. Systematic review of statistical approaches to quantify, or correct for, measurement error in a continuous exposure in nutritional epidemiology. BMC Medical Research Methodology. 2017;17(1):146.

[85] Burnham KP, Anderson DR. Multimodel inference: understanding AIC and BIC in model selection. Sociological Methods & Research. 2004;33(2):261–304.

[86] Poole D, Raftery AE. Inference for deterministic simulation models: the Bayesian melding approach. Journal of the American Statistical Association. 2000;95(452):1244–1255.

[87] Oakley JE, O'Hagan A. Probabilistic sensitivity analysis of complex models: a Bayesian approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2004;66(3):751–769.

[88] Kendall A, Gal Y. What uncertainties do we need in Bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. Curran Associates Inc.; 2017. p. 5580–5590.

[89] Van der Laan MJ, Polley EC, Hubbard AE. Super learner. Statistical Applications in Genetics and Molecular Biology. 2007;6(1):25.

[90] Goodman S. A comment on replication, p-values and evidence. Statistics in Medicine. 1992;11:875–879.

[91] Amrhein V, Trafimow D, Greenland S. Inferential statistics as descriptive statistics: There is no replication crisis if we dont expect replication. The American Statistician. 2019;73(sup1):262–270.

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

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

[94] Chambers CD. Registered reports: a new publishing initiative at Cortex. Cortex. 2013 Mar;49(3):609–10.

[95] MacCoun R, Perlmutter S. Blind analysis: hide results to seek the truth. Nature News. 2015;526(7572):187–189.

[96] Hartshorne J, Schachner A. Tracking replicability as a method of post-publication open evaluation. Frontiers in Computational Neuroscience. 2012;6:8.

[97] Ioannidis J. How to Make More Published Research True. PLoS Med. 2014;11(10):e1001747.

[98] Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nature Human Behaviour. 2018;2(1):6–10.

[99] Simons DJ, Shoda Y, Lindsay DS. Constraints on generality (COG): A proposed addition to all empirical papers. Perspectives on Psychological Science. 2017;12(6):1123–1128.

[100] Nuzzo R. How scientists fool themselves - and how they can stop. Nature. 2015;526(7572):182–185.

[101] Corner A, Whitmarsh L, Xenias D. Uncertainty, scepticism and attitudes towards climate change: biased assimilation and attitude polarisation. Climatic change. 2012;114(3-4):463–478.

[102] Lewandowsky S, Oreskes N, Risbey JS, Newell BR, Smithson M. Seepage: Climate change denial and its effect on the scientific community. Global Environmental Change. 2015;33:1–13.

[103] Freudenburg WR, Muselli V. Global warming estimates, media expectations, and the asymmetry of scientific challenge. Global Environmental Change. 2010;20(3):483–491.

[104] Van der Linden S, Leiserowitz A, Rosenthal S, Maibach E. Inoculating the public against misinformation about climate change. Global Challenges. 2017;1(2):1600008.

[105] Ballew MT, Leiserowitz A, Roser-Renouf C, Rosenthal SA, Kotcher JE, Marlon JR, et al. Climate Change in the American Mind: Data, Tools, and Trends. Environment: Science and Policy for Sustainable Development. 2019;61(3):4–18.

[106] Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013;309(13):1351–1352.

[107] Pettorelli N, Safi K, Turner W, et al. Satellite remote sensing, biodiversity research and conservation of the future. Philosophical Transactions of the Royal Society B Biological Sciences. 2014;369(1643):20130190.

[108] Connelly R, Playford CJ, Gayle V, Dibben C. The role of administrative data in the big data revolution in social science research. Social Science Research. 2016;59:1–12.

[109] Meng XL, et al. Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics. 2018;12(2):685–726.

[110] Klau S, Martin-Magniette ML, Boulesteix AL, Hoffmann S. Sampling uncertainty versus method uncertainty: A general framework with applications to omics biomarker selection. Biometrical Journal, DOI=101002/bimj201800309. 2019;.

Acknowledgments

Competing interests

The authors declare no competing interests.