The preregistration revolution

Brian A. Nosek[a,b,1], Charles R. Ebersole[b], Alexander C. DeHaven[a], and David T. Mellor[a]

[a]Center for Open Science, Charlottesville, VA 22903; and [b]Department of Psychology, University of Virginia, Charlottesville, VA 22904

Edited by Richard M. Shiffrin, Indiana University, Bloomington, IN, and approved August 28, 2017 (received for review June 15, 2017)

Progress in science relies in part on generating hypotheses with existing observations and testing hypotheses with new observations. This distinction between postdiction and prediction is appreciated conceptually but is not respected in practice. Mistaking generation of postdictions with testing of predictions reduces the credibility of research findings. However, ordinary biases in human reasoning, such as hindsight bias, make it hard to avoid this mistake. An effective solution is to define the research questions and analysis plan before observing the research outcomes—a process called preregistration. Preregistration distinguishes analyses and outcomes that result from predictions from those that result from postdictions. A variety of practical strategies are available to make the best possible use of preregistration in circumstances that fall short of the ideal application, such as when the data are preexisting. Services are now available for preregistration across all disciplines, facilitating a rapid increase in the practice. Widespread adoption of preregistration will increase distinctiveness between hypothesis generation and hypothesis testing and will improve the credibility of research findings.

methodology | open science | confirmatory analysis | exploratory analysis | preregistration

Progress in science is marked by reducing uncertainty about nature. Scientists generate models that may explain prior observations and predict future observations. Those models are approximations and simplifications of reality. Models are iteratively improved and replaced by reducing the amount of prediction error. As prediction error decreases, certainty about what will occur in the future increases. This view of research progress is captured by George Box's aphorism: "All models are wrong but some are useful" (1, 2).

Scientists improve models by generating hypotheses based on existing observations and testing those hypotheses by obtaining new observations. These distinct modes of research are discussed by philosophers and methodologists as hypothesis-generating versus hypothesis-testing, the context of discovery versus the context of justification, data-independent versus data-contingent analysis, and exploratory versus confirmatory research (e.g., refs. 3–6). We use the more general terms—postdiction and prediction—to capture this important distinction.

A common thread among epistemologies of science is that postdiction is characterized by the use of data to generate hypotheses about why something occurred, and prediction is characterized by the acquisition of data to test ideas about what will occur. In prediction, data are used to confront the possibility that the prediction is wrong. In postdiction, the data are already known and the postdiction is generated to explain why they occurred.

Testing predictions is vital for establishing diagnostic evidence for explanatory claims. Testing predictions assesses the uncertainty of scientific models by observing how well the predictions account for new data. Generating postdictions is vital for discovery of possibilities not yet considered. In many cases, researchers have very little basis to generate predictions, or evidence can reveal that initial expectations were wrong. Progress in science often proceeds via unexpected discovery—a study reveals an inexplicable pattern of results that sends the investigation on a new trajectory.

Why does the distinction between prediction and postdiction matter? Failing to appreciate the difference can lead to overconfidence in post hoc explanations (postdictions) and inflate the likelihood of believing that there is evidence for a finding when there is not. Presenting postdictions as predictions can increase the attractiveness and publishability of findings by falsely reducing uncertainty. Ultimately, this decreases reproducibility (6–11).

Mental Constraints on Distinguishing Predictions and Postdictions

It is common for researchers to alternate between postdiction and prediction. Ideas are generated, and observed data modify those ideas. Over time and iteration, researchers develop understanding of the phenomenon under study. That understanding might result in a model, hypothesis, or theory. The dynamism of the research enterprise and limits of human reasoning make it easy to mistake postdiction as prediction. The problem with this is understood as post hoc theorizing or hypothesizing after the results are known (12). It is an example of circular reasoning—generating a hypothesis based on observing data, and then evaluating the validity of the hypothesis based on the same data.

Hindsight bias, also known as the I-knew-it-all-along effect, is the tendency to see outcomes as more predictable after the fact compared with before they were observed (13, 14). With hindsight bias, the observer uses the data to generate an explanation, a postdiction, and simultaneously perceives that they would have anticipated that explanation in advance, a prediction. A common case is when the researcher's prediction is vague so that many possible outcomes can be rationalized after the fact as supporting the prediction. For example, a biomedical researcher might predict that a treatment will improve health and postdictively identify the one of five health outcomes that showed a positive benefit as the one most relevant for testing the prediction. A political scientist might arrive at a model using a collection of covariates and exclusion criteria that can be rationalized after the fact but would not have been anticipated as relevant beforehand. A chemist may have random variation occurring across a number of results and nevertheless be able to construct a narrative post facto that imbues meaning in the randomness. To an audience of historians (15), Amos Tversky provided a cogent explanation of the power of hindsight for considering evidence:

> All too often, we find ourselves unable to predict what will happen; yet after the fact we explain what did happen with a great deal of confidence. This "ability" to explain that which we cannot predict, even in the absence of any additional information, represents an important, though subtle, flaw in our reasoning. It leads us to believe that there is a less uncertain world than there actually is....

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Reproducibility of Research: Issues and Proposed Remedies," held March 8–10, 2017, at the National Academy of Sciences in Washington, DC. The complete program and video recordings of most presentations are available on the NAS website at www.nasonline.org/Reproducibility.

Author contributions: B.A.N. designed research; B.A.N. performed research; and B.A.N., C.R.E., A.C.D., and D.T.M. wrote the paper.

Conflict of interest statement: B.A.N., A.C.D., and D.T.M. are employed by the nonprofit Center for Open Science that has as its mission to increase openness, integrity, and reproducibility of research.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. Email: nosek@virginia.edu.

Published online March 12, 2018.

Mistaking postdiction as prediction underestimates the uncertainty of outcomes and can produce psychological overconfidence in the resulting findings.

The values of impartiality and objectivity are pervasive (16), particularly for scientists, but human reasoning is not reliably impartial or objective (17, 18). Scientists are motivated to advance knowledge; scientists are also motivated to obtain job security, awards, publications, and grants. In the present research culture, these rewards are more likely to be secured by obtaining certain kinds of research outcomes over others. Novel results are rewarded more than redundant or incremental additions to existing knowledge. Positive results—finding a relationship between variables or an effect of treatments on outcomes—are rewarded more than negative results—failing to find a relationship or effect; clean results that provide a strong narrative are rewarded more than outcomes that show uncertainty or exceptions to the favored narrative (9, 19–21). Novel, positive, clean results are better results both for reward and for launching science into new domains of inquiry. However, achieving novel, positive, clean results is a rare event. Progress in research is halting, messy, and uncertain. The incentives for such results combined with their infrequency create a potential conflict of interest for the researcher. If certain kinds of results are more rewarded than others, then researchers are motivated to obtain results that are more likely to be rewarded regardless of the accuracy of those results.

Lack of clarity between postdiction and prediction provides the opportunity to select, rationalize, and report tests that maximize reward over accuracy. Moreover, good intentions are not sufficient to overcome the fallibility of memory, motivated reasoning, and cognitive biases that can occur outside of conscious awareness or control (22–26). Researchers may design a study to investigate one question and, upon observing the outcomes, misremember the original purposes as more aligned with what was observed. Researchers may genuinely believe that they would have predicted, or even that they did predict, the outcomes as observed (22). Researchers may employ confirmation bias by seeking evidence consistent with their expectations and finding fault or ignoring evidence that is inconsistent with their expectations (24). These reasoning challenges are exacerbated by the misuse of common tools of statistical inference to provide false comfort about the reliability of evidence.

Standard Tools of Statistical Inference Assume Prediction

Null hypothesis significance testing (NHST) is designed for prediction—testing hypotheses—not for postdiction—generating hypotheses (6, 27). The pervasiveness in many disciplines of NHST and its primary statistic, the *P* value, implies either that most research is prediction or that postdiction is frequently mistaken as prediction with errant application of NHST. [This paper focuses on NHST because of its pervasive use (e.g., refs. 28 and 29). The opportunities and challenges discussed are somewhat different with other statistical approaches, such as Bayesian methods. However, no statistical method on its own avoids researcher opportunity for flexibility in analytical decisions, such as exclusion criteria or the creation of variables (30).]

In NHST, one usually compares a null hypothesis of no relationship among the variables and an alternate hypothesis in which the variables are related. Data are then observed that lead to rejection or not of the null hypothesis. Rejection of the null hypothesis at $P < 0.05$ is a claim about the likelihood that data as extreme or more extreme than the observed data would have occurred if the null hypothesis were true. It is underappreciated that the presence of "hypothesis testing" in the name of NHST is consequential for constraining its appropriate use to testing predictions. The diagnosticity of a *P* value is partly contingent on knowing how many tests were performed (27). Deciding that a given $P < 0.05$ result is unlikely, and therefore evidence against the null hypothesis, is very different if it was the only test conducted versus one of 20, 200, or 2,000 tests. [Notably, *P* values near 0.05 are not actually very unlikely in typical research practices (31), leading some researchers to recommend 0.005 as a more stringent criterion for claiming "significance" (32).]

If there were only one inference test to perform and only one way to conduct that test, then the *P* value is diagnostic about its intended likelihood. It is not hyperbole to say that this almost never occurs. Even in the simplest studies, there is more than one way to perform the statistical inference test. For example, researchers must decide whether any observations should be excluded from the analysis, whether any measures should be transformed or combined, and whether any other variables should be included in the model as covariates.

Correcting the diagnosticity of *P* values for the number of tests that were actually conducted is relatively straightforward (33, 34), although inconsistently—even rarely—applied in practice (35, 36). However, counting the literal performance of statistical tests is not sufficient to account for how observing the data can influence the selection of tests to conduct. Gelman and Loken (37) refer to the problem as the garden of forking paths. There are a vast number of choices for analyzing data that could be made. If those choices are made during analysis, observing the data may make selecting some paths more likely and others less likely. By the end, it may be impossible to estimate the paths that could have been selected if the data had looked different or if analytic decisions were influenced by hindsight, confirmation, and outcome biases. This leaves the observed *P* values with unknown diagnosticity, rendering them uninterpretable. In other words, NHST cannot be used with confidence for postdiction.

In prediction, the problem of forking paths is avoided because the analytic pipeline is specified before observing the data. As such, with correction for the number of tests conducted, *P* values retain their diagnosticity. In postdiction, analytic decisions are influenced by the observed data, creating the forking paths. The researcher is exploring the data to discover what is possible. The data help generate, not test, new questions and hypotheses.

The problem of failing to distinguish between postdiction and prediction is vastly underestimated in practice. Researchers may conduct lots of studies and test many possible relationships. Even if there are no relationships to find, some of those tests will elicit apparent evidence—positive results—by chance (27). If researchers selectively report positive results more frequently than negative results, then the likelihood of false positives will increase (38–40). Moreover, researchers have substantial degrees of freedom to conduct many different tests, and selection of those that yield positive results over those that yield negative results will increase the likelihood of attractive results at the expense of accuracy (30, 41, 42).

If researchers are clear about when they are in prediction and postdiction modes of research, then the benefits (and limits) of statistical inference will be preserved. However, with means, motive, and opportunity to misperceive postdiction as prediction and to selectively rationalize and report a biased subset of outcomes, researchers are prone to false confidence in evidence. Preregistration is a solution that helps researchers maintain clarity between prediction and postdiction and preserve accurate calibration of evidence.

Preregistration Distinguishes Prediction and Postdiction

Preregistration of an analysis plan is committing to analytic steps without advance knowledge of the research outcomes. That commitment is usually accomplished by posting the analysis plan to an independent registry such as https://clinicaltrials.gov/ or https://osf.io/. The registry preserves the preregistration and makes it discoverable, sometimes after an embargo period. With preregistration, prediction is achieved because selection of tests is not influenced by the observed data, and all conducted tests

are knowable. The analysis plan provides constraint to specify how the data will be used to confront the research questions.

In principle, inferences from preregistered analyses will be more reproducible than NHST analyses that were not preregistered because the relation between the analysis choices and findings cannot be influenced by motivation, memory, or reasoning biases. We say "in principle" because the case for preregistration is theoretically strong as a matter of inductive inference and empirically bolstered by some correlational evidence. However, there is not yet sufficient experimental evidence establishing its superiority for reproducibility. Correlational evidence suggests that hypothesizing in advance relates to increased replicability (11). Further, preregistration is correlated with outcomes that suggest reduced publication or reporting biases. For example, Kaplan and Irvin (43) observed a dramatic drop in the rate of positive results following the requirement to preregister primary outcomes in a sample of clinical trials. The benefits of preregistration are lost if researchers do not follow the preregistrations (44, 45). However, there is evidence that preregistration makes it possible to detect and possibly correct selection and reporting biases (e.g., compare-trials.org/). Franco et al. (38) observed that 40% of published papers in their sample of preregistered studies failed to report one or more of the experimental manipulations (treatment conditions), and 70% of published papers failed to report one or more of the outcome variables. Moreover, there was substantial selection bias in outcomes that were reported in papers included in the study (96% of consistently significant findings included in published articles) versus those that were left out (65% of null effects not included in published articles).

Formally speaking, analyses conducted on the data that are not part of the preregistration inform postdiction. In principle, preregistration can establish a bright line between prediction and postdiction. This preserves the diagnosticity of NHST inference for predictions and clarifies the role of postdiction for generating possible explanations to test as predictions in the future. In practice, there are challenges for implementing preregistration and maintaining a clear distinction between prediction and postdiction. Nevertheless, there are opportunities to benefit from preregistration even when the idealistic bright line cannot be achieved.

Preregistration in Practice

Preregistration does not favor prediction over postdiction; its purpose is to make clear which is which. There are practical challenges for effective integration of preregistration in many areas of research. We first describe the ideal preregistration and then address some of the practical challenges.

The Ideal. The idealized scenario for preregistration follows the simplified model of research taught in elementary school. A scientist makes observations in the world and generates a research question or hypothesis from those observations. A study design and analysis plan are created to evaluate that question. Then data are collected according to the design and analyzed according to the analysis plan. This confronts the hypothesis by testing whether it predicts the outcomes of the experiment. Following that, the researcher might explore the data for potential discoveries that generate hypotheses or potential explanations after the fact. The most interesting postdictions are then converted into predictions for designing the next study and the cycle repeats. In this idealized model, preregistration adds very little burden—the researcher just posts the study design and analysis plan to an independent registry before observing the data and then reports the outcomes of the analysis according to that plan. However, the idealized model is a simplification of how most research actually occurs.

Challenge 1: Changes to Procedure During Study Administration. Sometimes the best laid plans are difficult to achieve. Jolene preregisters an experimental design using human infants as participants. She plans to collect 100 observations. Data collection is difficult. She can only get 60 parents to bring their infants to her laboratory. She also discovers that some infants fall asleep during study administration. She had not thought of this in advance; the preregistered analysis plan does not exclude sleeping babies.

Deviations from data collection and analysis plans are common, even in the most predictable investigations. Deviations do not necessarily rule out testing predictions effectively. If the outcomes have not yet been observed, Jolene can document the changes to her preregistration without undermining diagnosticity. However, even if the data have been observed, preregistration provides substantial benefit. Jolene can transparently report changes that were made and why. Most of the design and analysis plan is still preserved, and deviations are reported transparently, making it possible to assess their impact. Compared with the situation in which Jolene did not preregister at all, preregistration with reported deviations provides substantially greater confidence in the resulting statistical inferences.

There is certainly increased risk of bias with deviations from analysis plans after observing the data, even when changes are reported transparently. For example, under NHST, if Jolene uses the observed results to help decide whether to continue data collection, the likelihood of misleading results may increase (30, 46). With transparent reporting, observers can assess the deviations and their rationale. The only way to achieve that transparency is with preregistration.

Challenge 2: Discovery of Assumption Violations During Analysis. During analysis, Courtney discovers that the distribution of one of her variables has a ceiling effect and another is not normally distributed. These violate the assumptions of her preregistered tests. Violations like these cannot be identified until observing the data. Nevertheless, multiple strategies are available to address contingencies in data analytic methods without undermining diagnosticity of statistical inference.

For some kinds of analysis, it is possible to define stages and preregister incrementally. For example, a researcher could define a preregistration that evaluates distributional forms of variables to determine data exclusions, transformations, and appropriate model assumptions that do not reveal anything about the research outcomes. After that, the researcher preregisters the model most appropriate for testing the outcomes of interest. Effective application of sequential preregistration is difficult in many research applications. If an earlier stage reveals information about outcomes to be tested at a subsequent stage, then the preregistration is compromised.

A more robust option is to blind the dataset by scrambling some of the observations so that distributional forms are still retained, but there is no way to know the actual outcomes until the dataset is unblinded (47, 48). Researchers can then address outliers and modeling assumptions without revealing the outcomes. Blinding can be difficult to achieve in practice, depending on the nature of the dataset and outcomes of interest.

Another method is to preregister a decision tree. The decision tree defines the sequence of tests and decision rules at each stage of the sequence. For example, the decision tree might specify testing a normality assumption and, depending on the outcome, selection of either a parametric or nonparametric test. A decision tree is particularly useful when the range of possible analyses is easily described. However, it is possible to preregister biases into decision trees. For example, one could preregister testing a sequence of exclusion rules and stopping when one achieves $P < 0.05$. On the positive side, this misbehavior is highly detectable; on the negative side, it invalidates the diagnosticity of statistical inference. Preregistration does not eliminate the possibility of poor statistical practices, but it does make them detectable.