RESEARCH METHODS

REFORMS: Consensus-based Recommendations for Machine-learning-based Science

Sayash Kapoor[1,2]*, Emily M. Cantrell[3,4], Kenny Peng[5], Thanh Hien Pham[1,2], Christopher A. Bail[6,7,8], Odd Erik Gundersen[9,10], Jake M. Hofman[11], Jessica Hullman[12], Michael A. Lones[13], Momin M. Malik[14,15,16], Priyanka Nanayakkara[12,17], Russell A. Poldrack[18], Inioluwa Deborah Raji[19], Michael Roberts[20,21], Matthew J. Salganik[2,3,22], Marta Serra-Garcia[23], Brandon M. Stewart[2,3,22,24], Gilles Vandewiele[25], Arvind Narayanan[1,2]

Machine learning (ML) methods are proliferating in scientific research. However, the adoption of these methods has been accompanied by failures of validity, reproducibility, and generalizability. These failures can hinder scientific progress, lead to false consensus around invalid claims, and undermine the credibility of ML-based science. ML methods are often applied and fail in similar ways across disciplines. Motivated by this observation, our goal is to provide clear recommendations for conducting and reporting ML-based science. Drawing from an extensive review of past literature, we present the REFORMS checklist (recommendations for machine-learning-based science). It consists of 32 questions and a paired set of guidelines. REFORMS was developed on the basis of a consensus of 19 researchers across computer science, data science, mathematics, social sciences, and biomedical sciences. REFORMS can serve as a resource for researchers when designing and implementing a study, for referees when reviewing papers, and for journals when enforcing standards for transparency and reproducibility.

INTRODUCTION

Machine learning (ML) methods are being widely adopted for scientific research (1–11). Compared to older statistical methods, they offer increased predictive accuracy (1), the ability to process large amounts of data (12), and the ability to use different types of data for scientific research, such as text, images, and video (7). However, the rapid uptake of ML methods has been accompanied by concerns of validity, reproducibility, and generalizability (13–20). There are several reasons for concern. Performance evaluation is notoriously tricky in ML (21–24). ML code tends to be complex and as yet lacks standardization (25, 26), leading to a lack of computational reproducibility (27). Subtle pitfalls arise from the differences

[1]Department of Computer Science, Princeton University, Princeton, NJ 08544, USA. [2]Center for Information Technology Policy, Princeton University, Princeton, NJ 08544, USA. [3]Department of Sociology, Princeton University, Princeton, NJ 08544, USA. [4]School of Public and International Affairs, Princeton University, Princeton, NJ 08544, USA. [5]Department of Computer Science, Cornell University, Ithaca, NY 14850, USA. [6]Department of Sociology, Duke University, Durham, NC 27708, USA. [7]Department of Political Science, Duke University, Durham, NC 27708, USA. [8]Sanford School of Public Policy, Duke University, Durham, NC 27708, USA. [9]Department of Computer Science, Norwegian University of Science and Technology, Trondheim, Norway. [10]Aneo AS, Trondheim, Norway. [11]Microsoft Research, New York, NY 10012, USA. [12]Department of Computer Science, Northwestern University, Evanston, IL 60208, USA. [13]School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK. [14]Center for Digital Health, Mayo Clinic, Rochester, MN 55905, USA. [15]School of Social Policy & Practice, University of Pennsylvania, Philadelphia, PA 19104, USA. [16]Institute in Critical Quantitative, Computational, & Mixed Methodologies, Johns Hopkins University, Baltimore, MD 21218, USA. [17]Department of Communication Studies, Northwestern University, Evanston, IL 60208, USA. [18]Department of Psychology, Stanford University, Stanford, CA 94305, USA. [19]Department of Computer Science, University of California, Berkeley, Berkeley, CA 94720, USA. [20]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, UK. [21]Department of Medicine, University of Cambridge, Cambridge, UK. [22]Office of Population Research, Princeton University, Princeton, NJ 08544, USA. [23]Rady School of Management, University of California, San Diego, La Jolla, CA 92093, USA. [24]Department of Politics, Princeton University, Princeton, NJ 08544, USA. [25]Department of Information Technology, Ghent University, Ghent, Belgium.
*Corresponding author. Email: sayashk@princeton.edu

between explanatory and predictive modeling (28). The hype and overoptimism about commercial artificial intelligence (AI) may spill over into scientific research (29). In addition, publication biases that have led to past reproducibility crises (30) are also present in ML research (31, 32). If left unchecked, these flaws can lead to a feedback loop of overoptimism because nonreplicable findings are cited more than replicable ones (33). There is an urgent need to systematically address errors in ML-based science rather than finding errors in individual studies after publication.

Here, we focus on a specific subset of ML applications: ML-based science. We use this term to refer to research that makes a scientific claim using the performance of an ML model as evidence. For example, Salganik *et al.* (34) use ML models to investigate the predictability of life outcomes. This contrasts with ML methods research, which involves improving widely applicable ML methods instead of making scientific claims using ML models. In the next section ("The scope of our claims: ML-based science"), we clarify the distinctions between ML methods research and ML-based science and outline the scope of the paper in greater detail. Box 1 summarizes this discussion.

One promising way to detect and prevent errors in scientific research is by improving standards for conducting and reporting science (35–37). Clear expectations for using ML methods can allow researchers and referees to spot errors early. Despite the use of ML methods across disciplines, there are no widely applicable best practices for reporting the design, implementation, and evaluation of ML-based science. This leads to different, and often no fixed standards for conducting and reporting research in each field adopting ML methods. As a result, common failure modes in using ML methods recur across disciplines (38, 39).

Here, we introduce a checklist for ML-based science, with the goal of preventing known but common errors that occur when ML is used in scientific research. To that end, we review the literature on best practices and common errors in ML-based science. We introduce the REFORMS checklist for reporting ML-based science, which consists of 32 items across eight modules (Table 1). We accompany

> **Box 1.**
>
> **What is ML-based science?**
>
> ML-based science refers to scientific research that uses ML models to contribute to scientific knowledge. This includes making predictions, conducting measurements, or performing other tasks that help answer scientific questions.
>
> ML-based science could use various ML techniques such as supervised learning, unsupervised learning, and reinforcement learning. The research should be geared toward answering a scientific question of interest.
>
> **Exclusions.** Not all ML research and applications qualify. For example, ML methods research and predictive analytics fall outside the scope. Similarly, quantitative research not using ML methods, such as explanatory modeling and simulations, do not qualify as ML-based science.
>
> **Applicability.** Our checklist for ML-based science will be more useful for some types of research than others. For instance, research on the use of ML for predictions may benefit more than research on using ML for search tasks within vast and complex spaces.

REFORMS with a detailed set of guidelines to set expectations for each item (text S1).

Checklists have been adopted in many scientific fields (*35–37*, *40*), and they have been impactful in improving reporting practices (*41*, *42*). In 2014, the U.S. National Institutes of Health (NIH) created principles to improve reproducibility and rigor, endorsed by several journals. One item was the creation of guidelines and checklists for journals (*43*). The EQUATOR network collects reporting guidelines for health research and includes more than 500 checklists (*44*). Several checklists have been proposed in ML methods research (*32*, *45*, *46*).

REFORMS differs from this large body of past work in crucial ways. First, past checklists for scientific research are field or method specific. For example, the CLAIM checklist (*35*) provides best practices for reporting research on AI in medical imaging. As a result, many items in thechecklist do not apply to other scientific fields adopting ML methods. Because ML methods are being rapidly adopted across fields, and research that uses ML methods suffers from similar errors, we aimed to make the REFORMS checklist field agnostic. We selected items that broadly apply to fields that use ML methods. Second, past checklists for ML methods research focus on common errors in developing ML methods. However, these errors differ from the ones that arise in scientific research (see Box 1 for the distinction between ML methods research and ML-based science). For instance, the checklist provided by Pineau *et al.* (*32*) does not include questions about the distribution about which the claims are made, because ML methods research often focuses on improving a model's performance on a benchmark dataset (*47*). In contrast, clearly specifying the distribution of interest is core to a scientific claim. Still, past work in both scientific research and ML methods research has helped inform our checklist.

We present findings from an extensive review of past research on best practices and common shortcomings relevant to each of the eight checklist modules. Notably, the REFORMS checklist represents consensus-based recommendations for ML-based science. These recommendations subtly differ from reporting guidelines: While reporting guidelines only include items that can be addressed by authors after a study has been conducted, consensus-based recommendations, such as the REFORMS checklist, also inform readers of best practices and can take a stance on how certain research activities should be conducted.

Alongside the checklist, we release guidelines paired with each item in the checklist (text S1). Similar guidelines have been included in past efforts at establishing standards. For instance, the STROBE-RDS statement (*37*) included an "explanation and elaboration" document to clarify expectations for each item in the checklist. Our guidelines aim to increase the usability of the REFORMS checklist and to provide pointers to best practices for ML-based science.

The REFORMS checklist can help address common failure modes that lead to errors in ML-based science. To guard against corners being cut because of time and publication pressures, the standards provide a set of clear expectations. To aid researchers new to ML-based science, the guidelines for each item identify resources and relevant past literature. While no single document would be enough to familiarize researchers with all the nuances of ML-based science, our hope is that the guidelines can be one useful pedagogical resource. Finally, there are many steps involved in successfully reporting ML-based science. It is hard to keep all of these items in mind when writing up a study. REFORMS provides all 32 items in one document to prevent omission errors.

THE SCOPE OF OUR CLAIMS: ML-BASED SCIENCE

We define ML-based science as scientific research that uses the performance of an ML model as evidence for a scientific claim. This includes, but is not limited to, making predictions, conducting measurements, or performing other tasks that contribute to the body of scientific knowledge. This definition has two parts: First, the research should be geared toward answering a scientific question of interest. This means that other types of ML research and applications do not fit under the umbrella of ML-based science:

ML methods research

Research focusing on developing and refining ML methods, such as a typical NeurIPS paper, does not constitute ML-based science. While such work does contribute new ML methods, the main objective is not to solve specific scientific problems of interest about generalizable populations. (Still, elements of the REFORMS checklist could be valuable to ML methods research. This is particularly the case when these newly developed methods are evaluated on benchmarks that directly influence their application in scientific contexts despite not being representative of real problems.)

Predictive analytics

Many real-world applications of ML models emphasize predictive accuracy but are not conducted to gain scientific insights. For example, social media platforms use ML to predict if a user will click on ads millions of times a day. There are many differences between predictive analytics and ML-based science. In predictive analytics, the population or distribution of interest may not be clearly defined. The relationships found using ML models only need to hold within a company or organization and need not generalize. Relative accuracy is more important than absolute accuracy numbers because the decision to deploy a model has often already been made—so the only question being answered by the modeling activity is which of a given set of models should be deployed in production. In addition, verifying whether the predictions later come true is often easy. This feedback is a better test for whether an application works as intended (compared to using a checklist).

Table 1. The REFORMS checklist for ML-based science. See text S1 for guidelines on how to use the checklist. Alongside each item, authors should report the section or page number where the item is reported. Some items in the REFORMS checklist could be hard to report for specific studies. Instead of requiring strict adherence for each item, authors and referees should decide which items are relevant for a study and how details can be reported better. To that end, we hope that the checklist can offer a useful starting point for authors and referees working on ML-based science.

Module	Item
Study goals	1a. State the population or distribution about which the scientific claim is made.
	1b. Describe the motivation for choosing this population or distribution *(1a.)*.
	1c. Describe the motivation for the use of ML methods in the study.
Computational reproducibility	2a. Describe the dataset used for training and evaluating the model and provide a link or DOI to uniquely identify the dataset.
	2b. Provide details about the code used to train and evaluate the model and produce the results reported in the paper along with link or DOI to uniquely identify the version of the code used.
	2c. Describe the computing infrastructure used.
	2d. Provide a README file which contains instructions for generating the results using the provided dataset and code.
	2e. Provide a reproduction script to produce all results reported in the paper.
Data quality	3a. Describe source(s) of data, separately for the training and evaluation datasets (if applicable), along with the time when the dataset(s) are collected, the source and process of ground-truth annotations, and other data documentation.
	3b. State the distribution or set from which the dataset is sampled (i.e., the sampling frame).
	3c. Justify why the dataset is useful for the modeling task at hand.
	3d. State the outcome variable of the model, along with descriptive statistics (split by class for a categorical outcome variable) and its definition.
	3e. State the sample size and outcome frequencies.
	3f. State the percentage of missing data, split by class for a categorical outcome variable.
	3g. Justify why the distribution or set from which the dataset is drawn *(3b.)* is representative of the one about which the scientific claim is being made *(1a.)*.
Data preprocessing	4a. Describe whether any samples are excluded with a rationale for why they are excluded.
	4b. Describe how impossible or corrupt samples are dealt with.
	4c. Describe all transformations of the dataset from its raw form *(3a.)* to the form used in the model, for instance, treatment of missing data and normalization—preferably through a flow chart.
Modeling	5a. Describe, in detail, all models trained.
	5b. Justify the choice of model types implemented.
	5c. Describe the method for evaluating the model(s) reported in the paper, including details of train-test splits or cross-validation folds.
	5d. Describe the method for selecting the model(s) reported in the paper.
	5e. For the model(s) reported in the paper, specify details about the hyperparameter tuning.
	5f. Justify that model comparisons are against appropriate baselines.
Data leakage	6a. Justify that preprocessing (Module 4) and modeling (Module 5) steps only use information from the training dataset (and not the test dataset).
	6b. Describe methods used to address dependencies or duplicates between the training and test datasets (e.g. different samples from the same patients are kept in the same dataset partition).
	6c. Justify that each feature or input used in the model is legitimate for the task at hand and does not lead to leakage.
Metrics and uncertainty	7a. State all metrics used to assess and compare model performance (e.g., accuracy, AUROC etc.). Justify that the metric used to select the final model is suitable for the task.
	7b. State uncertainty estimates (e.g., confidence intervals, standard deviations), and give details of how these are calculated.
	7c. Justify the choice of statistical tests (if used) and a check for the assumptions of the statistical test.
Generalizability and limitations	8a. Describe evidence of external validity.
	8b. Describe contexts in which the authors do not expect the study's findings to hold.

The second criteria for ML-based science is that the research should use ML methods. By this, we refer to a variety of techniques including, but not limited to, supervised learning, unsupervised learning, and reinforcement learning. Consequently, other types of quantitative research not using these methods do not fall under the umbrella of ML-based science:

Explanatory modeling
Such research uses traditional statistical methods for improving our understanding of real-world phenomena rather than predicting outcomes (*48*). This is an example of quantitative scientific research that does not use ML methods. For more information about how explanatory modeling differs from ML methods, see the section on item 1c ("Motivation for the use of ML methods in the study"). While explanatory modeling suffers from many shortcomings (*13*), errors in this domain are subtly different compared to ML-based science, which means that much of our checklist does not apply to explanatory modeling research.

Simulation
Similarly, physics-based models and simulation methods are sometimes evaluated iteratively by output and by their fidelity with existing theory (*49*, *50*) and do not involve ML or fitting models to data—except potentially as a tool to understand and summarize the simulation results. As a result, many items in the REFORMS checklist do not apply. Note that some algorithms used in ML, like Markov chain Monte Carlo, are examples of numerical simulations; however, these are used only to fit models to data, which is different to simulating an underlying phenomenon of interest.

In sum, we call research at the intersection of ML methods and quantitative science "ML-based science." The work of Salganik *et al.* on the Fragile Families Challenge (*34*) is an example of this category. They used ML to predict children's life outcomes and answer scientific questions about the predictability of outcomes studied by sociologists. We discuss many other examples in our review.

The checklist is broadly applicable to research where the scientific claim is supported by the accuracy of an ML model on an out-of-sample test set. The specific ML method used is not important. As an example, research using logistic regression models would be within scope if the accuracy of the logistic regression model on an out-of-sample dataset is used to support the scientific claim of interest. See the section on item 1c ("Motivation for the use of ML methods in the study").

In some cases, our checklist may be applicable to scientific work with foundation models (*51*), which are a type of ML model, although our focus in this paper is more general. There are several challenges with using proprietary foundation models for scientific research. For example, they are nondeterministic and can change without adequate notice (*52*, *53*). This could lead to hard-to-resolve shortcomings in computational reproducibility. Foundation models are one example of a gray area in the definition of ML-based science. In such cases, we leave it up to the authors and referees to decide whether the REFORMS checklist is useful.

Note that our checklist will likely be more helpful for some types of ML-based science than others. For instance, it is likely to be more useful for predictive modeling compared to research that uses ML methods for search tasks in vast and complex spaces, such as the search for new materials or new phases of matter (*54*). In cases where verifying the result of an ML-based experiment is easy (for instance, verifying the properties of a new drug in a lab), our checklist might be less useful than such verification, though it could help ensure the validity of the experiments before verification.

METHODS
To develop the REFORMS checklist, we started with a focus on steps in a canonical ML pipeline, drew from previous checklists used in other domains, and went through a consensus process with all authors, involving multiple rounds of feedback and a virtual discussion. Table 2 lists the modules in our checklist and the corresponding stages in the ML pipeline. For each module, we focus on three goals: (i) establishing the scientific claim and its relationship to the ML modeling process; (ii) providing an overview of the best practices and common shortcomings in building ML models correctly; and (iii) enabling the verification of the results by an independent researcher. In other words, we aim to decrease the likelihood of errors of interpretation (goal 1) or execution (goal 2) and to make it easier for independent researchers to spot errors (goal 3). Box 2 outlines these goals in more detail. Box 3 summarizes how different stakeholders—authors, referees, and journals—can use the checklist.

To build on previous efforts at improving the reporting quality of research, we used three past checklists to ensure our coverage of important items in reporting an ML model. Pineau *et al.* (*32*) provided the checklist used alongside papers submitted to NeurIPS 2020, a prominent ML methods conference. Collins *et al.* (*40*) provided the TRIPOD checklist for prediction models in health research. Mongan *et al.* (*35*) introduced the CLAIM checklist for AI models in clinical imaging. We chose these checklists because they covered diverse modeling approaches and were applicable in different settings—ML methods research, models for individual diagnosis and prognosis, and ML for medical imaging, respectively.

Consensus process for developing the REFORMS checklist
Once we had an initial set of items, authors met virtually for a discussion. One of the main outcomes of the discussion was the need for a paired set of guidelines alongside each item to clarify how these items should be reported, which we discuss in more detail below. Then, the authors collaboratively edited the checklist to choose commonly applicable items across disciplines. We paid close attention to usability:

Table 2. Stages of ML-based science and corresponding checklist modules.

Stage of scientific study	Section of the checklist
Study design	Study goals (Module 1)
	Computational reproducibility (Module 2)
Data collection and preparation	Data quality (Module 3)
	Data preprocessing (Module 4)
Modeling	Modeling decisions (Module 5)
Evaluation	Data leakage (Module 6)
	Metrics and uncertainty quantification (Module 7)
Scope and limitations	Generalizability and limitations (Module 8)

To decrease the time and cognitive load that using the checklist would entail, we removed items that were too specific and would apply to a small subset of ML-based science. Last, the authors independently flagged unclear items to improve the quality of the checklist.

A recurring theme in our conversation was making the checklist easy to use. To that end, we developed a set of accompanying guidelines to help researchers understand the motivation for each section and clarify what is expected for each item in REFORMS. The guidelines (text S1) are based on our review of past literature for items in the checklist (this review is presented next). We include references to key prior work to help onboard researchers new to using ML methods. This includes a mix of peer-reviewed scientific research that details best practices, as well as resources from the ML methods community that outline common shortcomings and ways to address them. Crucially, we do not take prescriptive stances on matters of ongoing methodological debate. Instead, we present best practices to minimize and detect known types of errors in ML-based science.

Organization of the paper

In the remainder of the paper, we present the REFORMS checklist. They comprise eight modules based on the stages of an ML-based science study (see Table 2). For each module, we motivate why items in this module are important to address in ML-based science. For each item, we include expectations about what it means to address the item sufficiently. In our review, we draw from past literature on best practices and common errors in ML-based science. We also occasionally draw on literature about science with traditional statistical methods, as best practices and shortcomings are shared in many aspects of ML-based science and other quantitative science. In Table 1, we provide a template for the REFORMS checklist. In text S1, we distill the guidelines for filling out the checklist as a standalone document. In text S2, we present a table with additional details on the references used in our review.

MODULE 1: STUDY GOALS

This section focuses on stating a study's goals. This is motivated by recent research that shows that reporting study goals in adequate depth and clarity is not trivial or common (*55*). Studies that appear to ask the same research question may actually have subtle differences in their questions which lead to substantially different findings (*28*).

1a) Population or distribution about which the scientific claim is made

The population of interest is the group to which the researchers intend the findings of the study to generalize. In the parlance of ML methods research, population is analogous to a "distribution" of scientific interest. That is, scientific claims are often made about data sampled from a certain distribution (*56*), rather than a specific population of individuals. For example, the performance of an image classification tool might be made in the context of images taken from various distributions, such as satellite images of a certain region, instead of human populations. For brevity, we treat population and distribution interchangeably.

The population is typically broader than the sampling frame and sample, which are discussed in Module 3. Defining the population of interest is important because it shapes the article's conclusions, places boundaries on those conclusions, and provides the basis for metrics of significance and uncertainty that derive from the concept

of sampling from a population (*55, 57, 58*). Unfortunately, research articles do not always clearly state their population of interest (*55, 59, 60*).

1b) Motivation for choosing this population or distribution

The choice of a particular population of interest may be motivated by pure scientific interest or by a need for applied knowledge. Explaining the motivation for studying the population of interest helps the reader understand the importance of the study and contextualize its results.

We acknowledge that the motivation for choosing a particular population of interest may arise from what data is available. The development of the motivation depends on whether researchers followed a deductive, inductive, or iterative approach (*61*). In a deductive approach, researchers begin with a theory and a population of interest and then select or create a dataset based on the data's ability to test that theory for that population. In an inductive approach, researchers begin with a dataset and then determine what research questions and populations of interest that dataset can address. In an iterative approach, researchers iterate between data collection, data analysis, and theory until they develop a hypothesis, and then collect additional data to test that hypothesis (*61*). The deductive approach is the most widely accepted approach in the scientific community, but Grimmer *et al.* (*61*) argue that there is also great value in inductive and iterative approaches and that it is important to communicate which approach was used.

1c) Motivation for the use of ML methods in the study

The research questions asked in ML-based studies often differ from those asked with traditional statistical methods. Breiman (*12*) famously argued that there are "two cultures" in statistical modeling: the "data modeling" culture and the "algorithmic modeling" culture. In the data modeling culture, which is aligned with traditional statistical methods, researchers' focus is on estimating the parameters of a function that is meant to meaningfully represent the process by which input data produces output data in the world. This includes describing relationships between quantities of interest ("descriptive modeling") and estimating causal effects ("explanatory modeling") (*48*). For example, researchers might ask, "what is the relationship between household income and the likelihood of experiencing clinical depression?," and estimate the magnitude and direction of that relationship using a logistic regression model. In the "algorithmic modeling" culture, which aligns with much of ML-based research, the focus is on building a model that reliably maps input data to output data. In this culture, the parameters do not necessarily need to provide a faithful and interpretable description of patterns in the world; rather, the goal is to accurately predict output data when given a new sample of input data that is separate from the original data sample in which the model was trained. For example, researchers might ask, "given all the data available about a person in a particular dataset, how accurately can we predict whether that person will experience clinical depression?," and test a variety of models to find the one with the best predictive accuracy. (This could include predictions made with simple models like linear or logistic regression.) Breiman's proposition was that more scientific research should use algorithmic modeling (*12*). Several responses to Breiman have argued for the value of merging the two cultures or moving iteratively between them (*48, 62–66*).

Because non-ML (data modeling) methods are currently the standard approach in many scientific disciplines, explaining the motivation for using ML methods will help readers better understand a study's goals. Considering the differences and similarities between Breiman's two cultures may be useful to researchers when motivating the use of ML methods. In addition, many recent articles provide guidance on the value of ML methods for science (*1, 7, 48, 62, 67–71*).

MODULE 2: COMPUTATIONAL REPRODUCIBILITY

Computational reproducibility refers to the ability of an independent researcher to get the same results as reported in a paper or manuscript. It is an essential part of computational research (*72*).

Computational reproducibility can help independent researchers evaluate the findings in a paper and verify whether they hold up under scrutiny. The availability of reproducibility materials has led to several errors being spotted (*73–78*). Conversely, if the code and data for reproducing all results in a study are not available, then identifying the precise sources of errors in a study becomes hard (*79–81*).

Current computational reproducibility standards fall short

Some journals require authors to make their computational reproducibility materials available after publication without requiring these materials at the time of publication. However, such measures can miss the mark. Stodden *et al.* (*82*) attempted to contact the authors of 204 papers published in the journal *Science* to obtain reproducibility materials. Only 44% of authors responded. Similarly, Gabelica *et al.* (*83*) studied papers published in 333 open-access journals indexed on BioMed Central in January 2019. Of the 1792 papers that claimed they would share data upon request, 1669 did not share the data. That is, they were unable to get the data for 93% of the papers. This indicates the importance of requiring computational materials at the time of publication rather than at the authors' discretion later.

Vasilevsky *et al.* (*84*) studied the data-sharing policies at 318 biomedical journals. They found that almost a third of these journals had no data-sharing policies in place. Even the journals that did have data-sharing policies did not have clear guidelines for authors to comply with their policies.

ML methods research has also struggled to ensure computational reproducibility (*85, 86*). Gundersen and Kjensmo (*18*) systematically analyzed 400 papers that were published at leading conferences. In addition to code and dataset availability, they evaluated the documentation of methods in a paper's text, for instance, whether the experimental setup is described. They found that none of the 400 papers satisfied all of their reproducibility criteria, and in general, papers only satisfied 20 to 30% of the criteria.

Pineau *et al.* (*32*) found that only around half of the papers submitted to NeurIPS 2018, a leading ML conference, contained the code and data needed to reproduce results. To improve reproducibility, they introduced a checklist that was used in NeurIPS 2019. In the checklist, including reproducibility materials was optional but recommended. Still, after the checklist was introduced, more than 75% of the papers included reproducibility materials along with the submissions. Similar checklists have become the standard at several ML conferences (*87–89*). While ML methods research differs from ML-based science in its goals and practices, we can learn from these experiences to emphasize the importance of computational reproducibility in ML-based science.

Ensuring computational reproducibility in ML-based science is challenging

ML methods used in scientific research can be complex and often require numerous packages and dependencies. This makes computational reproducibility challenging (*27*). Liu and Salganik (*25*) describe their experiences ensuring computational reproducibility while editing a special issue in Socius on the Fragile Families Challenge. The Fragile Families Challenge was a prediction competition where multiple participants tried to predict children's life outcomes on the same dataset (*34*). The special issue published papers based on a few of the resulting models. Liu and Salganik wanted to ensure that the code and data alongside every publication were verified to ensure that they produced the same results as those presented in the paper. Despite spending 13 months working on achieving computational reproducibility and exchanging dozens of emails with the authors, they were unable to verify the computational reproducibility of all papers. Even preliminary steps, like installing the correct versions of each package, were nontrivial when a large number of packages were used in a study. They could eventually verify the computational reproducibility of 7 of the 12 papers. They published the rest with the code and data available at the time of publication.

Interventions adopted by journals

Journals have adopted several measures to improve computational reproducibility (*90–92*). The Transparency and Openness Promotion standards introduced by the Open Science Foundation (*93*) have a few sections that focus on computational reproducibility in scientific research. They are divided into three levels: Level 1 requires stating whether the computational reproducibility materials are available. If they are, authors should provide details about how to access them. Level 2 requires the materials to be available in a trusted repository at the time of publication. Level 3 requires the materials to be verified by the journal to ensure that they generate the results reported in the paper before publication.

Several social science journals use the Data and Code Availability Standards for computational reproducibility (*94*). Other journals have taken additional measures to verify whether the code is correct. For instance, *Nature Methods* conducts code reviews in addition to peer reviews for papers that provide computational artifacts (*95*).

In our checklist, we include the basic details that can enable independent researchers to verify the computational reproducibility of a result. Our checklist requires information about the dataset, code, computing environment, documentation for how to get the results in a study, and a reproduction script to automatically run the code and generate results. We acknowledge that computational reproducibility is hard and that some items in this module are more challenging compared to others. For instance, it is not always possible to release private datasets. When one aspect of computational reproducibility cannot be entirely met, researchers can still aim to satisfy the other aspects (e.g., provide a reproduction script even when data cannot be made available). Researchers can also address limitations through alternative means. We provide some such options below, such as providing a synthetic imitation of the data when the real data cannot be publicly released.

2a) Dataset

Peng *et al.* (*96*) and Nosek *et al.* (*93*) highlight the importance of citing datasets with permanent links to clarify which version of a dataset is used in a study. If an original dataset is provided alongside a study, documentation for the dataset is also important. For instance, authors can include data dictionaries (*97*) or datasheets (*98*). Such documentation should report basic details about the properties and format of the data.

Some datasets could contain sensitive information and cannot be publicly released. To address this, authors have previously released synthetic datasets when working with sensitive data (*99–102*). However, a limitation of synthetic data is that we can never know the patterns inherent to the original data to check if those patterns are preserved in the synthetic data (*103*, *104*). As a result, we cannot understand whether important relationships and properties of the original data have been preserved.

2b) Code

Similarly, it is important to report the exact version of code used for running the experiments and producing the results in a paper (*105*). Authors can accomplish this by providing a Digital Object Identifier (DOI), commit tag (for instance, from code repositories such as GitHub, GitLab, or BitBucket), or other documentation to precisely identify the version of the code used to train and evaluate the model and produce the results reported in the paper. Note that using archiving systems that provide permanent identifiers for the code used, like Dataverse, is likely to aid long-term reproducibility (*106*, *107*).

2c) Computing environment

Different computational experiments require different amounts of computing resources. To help readers understand the precise computing requirements for reproducing the study, authors should report details about the hardware (CPU, RAM, and disk space), software (operating system, programming language, and version number for each package used), and computing resources (time taken to generate the results) used to generate their results. Stodden and Miguez (*108*) provide best practices to document computing infrastructure.

2d) Documentation

Good documentation helps researchers unfamiliar with a project by walking them through the steps of setting up and running the code provided, starting from environment requirements and installation, to examples of usage and expected results (*109*, *110*).

2e) Reproduction script

A script to produce all results reported in the paper using the reproducibility materials can substantially reduce the time it takes for an independent researcher to reproduce the results reported in a study. Reproduction scripts can download packages with the version numbers needed to run the code, set the right dependencies, download and store datasets in the correct location, set up the computing environment, and run the code to produce the results reported in the paper.

Authors can implement such scripts in several ways, such as using a bash script (*111*) or using an online reproducibility platform such as CodeOcean (*112*). Note that this is a high bar for computational reproducibility. In some cases, it might not be possible to provide a script that would allow an independent researcher to reproduce all results—for instance, if the analysis is run on an academic high-performance computing cluster or if the dataset does not allow for programmatic download.

MODULE 3: DATA QUALITY

This module helps readers and referees understand and evaluate the quality of the data used in the study. Using poor-quality data or data that is not suitable for answering a research question can lead to results that are meaningless or misleading.

3a) Data source(s)

Describing a study's data source(s) allows readers to evaluate the data's strengths and weaknesses and to judge whether the data is appropriate for the study's goals. Most studies provide descriptions of their data source, but those descriptions sometimes lack important details (*113–115*). In addition, the quality of reporting about ground-truth annotation methods in ML-based science varies widely (*116*). To ensure a minimum level of information about data sources is reported, our checklist asks researchers to report when, where, and how data was collected, and how ground-truth annotations were performed on the dataset, if applicable.

3b) Sampling frame

A sampling frame is a list of people or units from which a sample is drawn. Because of practical limitations, the sampling frame in many studies does not include all members of the target population. It is important to describe a study's sampling frame so that readers understand the boundaries of the study's sample and how that sample relates to the target population (discussed further under checklist item 3g: *Dataset for evaluation is representative*).

Some research papers do not provide a clear description of their sampling frame or eligibility criteria for inclusion in the sample (*59, 113, 117*). Furthermore, one review article found that when papers stated information about their sampling frame or eligibility criteria was available in a prior publication, the prior publication was not always accessible, and the relevant information was often extremely difficult or impossible to find (*117*).

3c) Justification for why the dataset is useful for the modeling task

Our checklist asks researchers to justify why the dataset is useful for the modeling task because the appropriateness of a data source will depend on the research question. For example, while biased or incomplete data is inappropriate for some research questions, such data can work well for other research questions as long as the researcher understands how these shortcomings affect the analysis and communicates the limitations (*61*). A broad claim like "this is the best dataset available on this topic" does not help readers understand the strengths and weaknesses of the data for the study's research question; researchers should be specific about why the dataset is well suited to the question.

Modern ML-based research often relies on repurposed data sources, which are sometimes termed "big data:" for example, social media data, digital trace data, or digital administrative records (*118*). Salganik (*118*) describes 10 common characteristics of "big data" that result in differing strengths and weaknesses compared to traditional data sources. Researchers who are using ML with repurposed data can use these 10 characteristics as a guide when justifying why the data source is appropriate for their research goals and identifying shortcomings of the data.

3d) Outcome variable

Our checklist asks researchers to report how their outcome variable is defined. The outcome or target variable is the quantity that the model is used to predict, detect, classify, or estimate.

The outcome variable is typically an empirical proxy for an unobservable theoretical construct (*119*). For example, researchers might pose a question about the construct "academic performance" and use grade point average as measured in school administrative data as the empirical proxy for this construct. The outcome variable is usually not a perfect match for the theoretical construct it represents. Thus, to allow readers to evaluate a paper's claims, authors should describe precisely how their outcome variable is measured and note any ways in which this outcome variable might not align with the associated construct. This is especially important because mismatches between variables and the constructs they are purported to represent can create fairness issues (*120*).

Our checklist also asks for descriptive statistics about the outcome variable. Reviews of prior literature have found that descriptive statistics are not always sufficiently reported (*121, 122*). Descriptive statistics about the outcome variable help readers to understand the context being studied and to identify concerns related to rare values or skewed data.

3e) Number of samples in the dataset

Reporting sample size is important because a study must have sufficient sample size to achieve its objectives. Some research objectives can be achieved with small to moderate-sized samples: for example, detecting large differences between groups. Other objectives generally require large samples: for example, studying rare events, studying heterogeneity, or detecting small differences (*118*).

Note that there can be downsides to large sample sizes. When the sampling frame is unrepresentative, increasing the sample size can shrink confidence intervals without decreasing bias in estimates, thus giving false confidence (*123*). Furthermore, if a study exposes participants to any level of risk, larger samples may magnify harms (*118*).

Scientific literature is generally consistent about reporting total sample size (*59, 124*). However, reporting on sample size for subgroups or sample size after attrition in longitudinal research is less consistent (*59*). To ensure clear reporting, our checklist asks that in addition to the total sample size, researchers who are conducting a classification task report the number of samples in each class. This follows recent calls for more granular details about data, such as by Gebru *et al.* (*98*). We also ask researchers to distinguish between the number of individuals in the dataset and the number of rows in the dataset, in cases where an individual can appear in more than one row.

3f) Missingness

Missing data is highly prevalent in many research domains and can affect the results (*34, 125, 126*). It is important for researchers to report the prevalence of missing data in their dataset and to specify how they handled missingness (*125*). Missingness is particularly important to address carefully when it is nonrandom (*127*). Extensive literature across multiple fields has established that research articles frequently provide insufficient information about the presence and handling of missing data (*59, 113, 126, 128–132*).

Checklist item 3f focuses on reporting the prevalence of missing data. Reporting how missing data is handled is covered in item 4c: *Data transformations*.

3g) Dataset for evaluation is representative

This item asks researchers to justify that their sample is representative of the target population defined in Module 1. Representativeness is

important for the study's ability to generalize from the sample to the target population. Lack of representativeness in the sampling process is sometimes underreported; for example, one review of past literature found underreporting of information about selection bias (*59*).

Probability sampling is a common approach for achieving representativeness. However, probability sampling is not always necessary. Because of coverage errors and nonresponse in probability sampling, the differences between probability sampling and nonprobability sampling are not always as large as they first appear (*118*). For studies that use nonprobability sampling, researchers may be able to make a reasonable argument for representativeness by comparing sample characteristics with population characteristics (*57*). Researchers can also use statistical methods to adjust for nonprobability sampling (or to adjust for errors in probability sampling), such as post-stratification, sample matching, propensity score weighting, and calibration (*118*).

In some studies, the dataset will not be representative of the target population. Reasons a sample might fail to be representative of the target population depend on the type of data collected. For example, Salganik (*118*) describes three types of representation error in survey data, and Grimmer *et al.* (*61*) describe four sources of bias in sample selection for text analysis. A nonrepresentative sample is okay for certain research goals. For example, Salganik argues that research that aims to make out-of-sample generalizations generally requires representative data, but research that aims to make within-sample comparisons can be well suited to nonrepresentative data (*118*). Concerns about nonrepresentativeness should be noted under checklist item 8a: *Evidence of external validity.*

MODULE 4: DATA PREPROCESSING

Preprocessing is the series of steps taken to convert the dataset from its rawest available form into the final form used in the modeling process. This includes data cleaning and selection (i.e., selecting a set of samples from the dataset to be included in the modeling process) as well as other transformations of the data, such as imputing missing data and normalizing feature values.

Our checklist focuses on two broad components of preprocessing: first, the subset of data to consider (i.e., which rows of a dataset are considered), and second, the transformations that are subsequently applied to the data (i.e., how entries of a dataset might be altered). Each of these components has implications on the scope and validity of resulting scientific claims and are essential for ensuring the reproducibility of the results. As discussed in Module 3, preprocessing methods are often not specified in papers.

4a) Excluded data and rationale

Researchers might exclude some samples from the dataset—for instance, to remove outliers or to only focus on certain subsets. Thus, the resulting scientific claims should be made in relation to the particular subset of a dataset that is ultimately used. This type of preprocessing is closely related to our discussion on data quality (Module 3) and generalizability (Module 8). This item underscores the importance of reporting the specific subset of the dataset used, in addition to details about the overall data.

Hofman *et al.* (*28*) note how the choice of the subset can substantially affect the performance of the resulting model. A specific example they use is the prediction of the reach ("cascade size") of a social media post as a function of the poster's past success. They show that the threshold of popularity used to determine the subset of posts considered plays a large role in influencing the predictability of cascade size. The scientific claim—regarding the predictability of the success of a post—depends on what data are included and excluded. We ask authors to justify why the particular subset of data used was chosen.

4b) How impossible or corrupt samples are dealt with

A dataset may contain erroneous or undesirable data points. Data may be impossible (e.g., a person whose height is recorded as 10 feet) or corrupted (e.g., a survey response filled out by a bot). Different techniques have been developed to detect such cases (*133, 134*). Attempting to filter such data can be important to ensure the dataset represents the intended population of the study. Data detected as impossible or corrupt may be removed or transformed—our checklist asks authors to report such steps in items 4a and 4c, respectively.

4c) Data transformations

Once the set of data points to be used is decided upon, researchers could transform the data in various ways: for example, by normalizing or augmenting data, or imputing missing data. Both imputing missing data and under/oversampling from a subpopulation—if done improperly—can harm the validity of model performance in relation to the scientific claim made. For example, mean imputation and oversampling must be done separately on each fold of a dataset. Failing to do so can result in overoptimistic results (*78*).

Specifying preprocessing steps is also important for ensuring the reproducibility of the results. Choices in preprocessing technique can substantially affect the properties of the resulting model, including accuracy and interpretability (*135, 136*).

MODULE 5: MODELING

Researchers make several choices in creating an ML model. The exact specification of the model is important to consider with respect to the particular scientific claim being made. In addition, because of the large number of choices involved in creating an ML model, it is important to report exact details of how an ML model is created—otherwise, reproducibility by independent researchers could be hindered. Raff (*137*) attempted to reproduce ML results from 255 papers using only the paper's text (i.e., without using the code accompanying a paper), and found that 93 could not be reproduced.

Modeling choices are also closely related to choices involving evaluation. Model selection—the process of choosing the model(s) whose results are reported in a study—often depends on the evaluation setup. An improper approach to evaluation or model selection can result in exaggerated performance estimates.

Major ML conferences, such as NeurIPS and International Conference on Machine Learning, incorporate checklists that ask authors to verify that they have included training details within their paper, similar to the specific items we propose below (*88, 89*). Similarly, Mitchell *et al.* (*138*) introduced model cards to document details about the modeling and evaluation process, with a primary focus on natural language processing and computer vision.

A note on terminology: There is no broad agreement on the use of different terms related to the modeling process. For the purpose of the discussion in this module, we make one conceptual distinction. Items 5a and 5b focuses on aspects of the model that are specified

before training. This includes, for example, the type of model and the loss function. Meanwhile, items 5c, 5d, 5e, and 5f focus on fitted (trained) models.

5a) Model description

Specific details about all models trained are essential for ensuring the reproducibility of a paper. This includes specifying the input and output of a model, the type of model (e.g., Random Forest or Neural Network), and the loss function and algorithm used to train the model.

5b) Justification for the choice of model types implemented

With many different possible types of ML models to choose from, the choice of types to consider can be dependent on the intended scientific claim of a paper. For example, if a scientific claim aims to establish how predictive a set of features can be [e.g., (34, 139)], then it may be appropriate to consider a wide range of types—the most important criteria is the resulting accuracy of the fitted models. If the scientific claim aims to establish the potential usability of a model in practical settings, there may be additional desiderata. For example, a model that is used in high-stakes settings like clinical decision-making may need to be interpretable or explainable (140), so it may be more appropriate to choose a model type that is likelier to have these properties.

It is not always clear what the best type of model to use is for a given scientific question. For example, there is an active debate on how or if interpretability or explainability should be implemented in the context of different applications [see e.g., (141)]. Here, we ask authors to provide reasons for why the set of model types they implemented is appropriate.

The focus of this item is on the types of models considered, but some claims may depend on the choice of a specific fitted model (and its performance). The following items address the process of evaluating, selecting, and comparing specific fitted models.

5c) Model evaluation method

We ask authors to report details about the model evaluation procedure. We include evaluation in this "modeling" module because evaluation is often used as part of the model selection process (see item 5d). That is, it is common to consider many models and select the best-performing one on the basis of the evaluation setup.

Evaluation of ML models must be done on test data separate from training data and any data that was used in model selection (see item 5d for details). To ensure reproducibility and verify validity, it is necessary to report how data were split and used. We ask authors to report how models are evaluated: For example, using a holdout test set, an external validation test set, or (nested) cross validation. We also ask for the sample size within each split of the data, including the number of samples of each class for classification tasks.

5d) Model selection method

There are many possible fitted models that can arise depending on specific choices in the modeling process. Even holding the type of model fixed, differences can arise because of the specific choice of hyperparameters, for example. Researchers should report how the final model(s) reported in a paper are selected.

A common goal of model selection is to select the model with the best performance on a hold-out set, but improper model selection can result in misleading performance (142, 143).

Testing multiple models on the holdout set and choosing the one with the best performance on the holdout set can result in an overoptimistic estimate of performance. Neunhoeffer and Sternberg (76) consider this type of error in the context of political science, showing how a study on civil war prediction conducted improper model selection. Similarly, if cross-validation (CV) is used in the model selection process, testing on the same data will result in an overoptimistic estimate of performance (144). To avoid this bias, a separate test set can be maintained, or when there is less data, nested CV can provide an alternative (142).

Model selection is not limited to choosing the model with the highest accuracy. Multiple models can achieve the same accuracy as one another yet make different predictions for the same individuals and have different characteristics regarding issues like fairness and interpretability. This phenomenon is known as predictive multiplicity (145, 146). Black et al. (147) provide guidance on opportunities and concerns that arise from such issues.

5e) Hyperparameter selection

Training procedures are often dependent on the choice of model hyperparameters, such as regularization weight, the number of training epochs, or the learning rate for a model. Similarly, hyperparameters can also be used to define the model space, such as the choice of activation function or the width of layers in an artificial neural network. The choice of hyperparameters is part of the model selection process.

In the context of natural language processing research, Dodge et al. (148) show that details about hyperparameter search affect the resulting accuracy. More resources devoted to searching over hyperparameters can improve performance substantially. Islam et al. (149) show that because of variance in performance depending on hyperparameters, specific choices of hyperparameters can be misleading when interpreting results. One method may perform better than another with one set of hyperparameters while performing worse given a different set of hyperparameters. Incorrect estimates of model performance due to hyperparameter optimization are especially concerning when comparing different models (150–152).

Last, note that the degree to which hyperparameter optimization affects results can vary depending on the model type (153). This may be a reason to prefer one model type over another.

5f) Appropriate baselines

It can be important to compare the performance of a model to baselines, especially when the scientific claim argues that a particular ML method can outperform existing approaches on a task. To clearly establish the performance of an ML model, it is necessary to detail how baseline models were trained and how the baseline methods are optimized. For example, if the baselines were not chosen using the same model selection methods, this could result in the baselines being "weak." As a result, the apparent benefits of the new method may be misleading. Sculley et al. (24) and Lin (154) detail examples in which ML models were compared to weak baselines.

Lones (155) discusses how to compare models more broadly. See also item 7c: *Appropriate statistical tests* for a discussion on the use of statistical testing to compare models.

MODULE 6: DATA LEAKAGE

Leakage is a spurious relationship between the features and the target variable that arises as an artifact of the data collection, sampling,

preprocessing, or modeling steps. For example, normalizing features in the training and test data together leads to leakage because information about the test data features is included in the training data (*156*).

Data leakage is a common error in the use of ML methods. Epic, a U.S. healthcare technology company, released a sepsis prediction model in hospitals nationwide. However, one of the features used in the model was whether a patient had been prescribed antibiotics. This is an error because antibiotics would typically be prescribed after the diagnosis of sepsis, so they act as a proxy for the outcome variable. Consequently, the model's performance was inflated because of having access to information it would not have in a real-world scenario (*157*).

Leakage has caused widespread reproducibility errors in ML-based science. In a survey of leakage across ML-based science, Kapoor and Narayanan (*39*) found that leakage affects hundreds of papers across 17 fields.

We ask authors to justify that their study does not suffer from major sources of leakage, which we elaborate on in this section. Kapoor and Narayanan (*39*) offer model info sheets to help detect and prevent leakage before publication. Our checklist focuses on the three main types of leakage in ML-based science found in their survey.

6a) Train-test separation is maintained

When information from the test set is used during the training process, it leads to overoptimistic performance results due to data leakage.

Not using a held-out test set is a textbook error in ML (*158*). Still, it is widespread. For example, Poldrack *et al.* (*159*) find that of the 100 neuropsychiatry studies that claimed to predict patient outcomes, 45 only reported in-sample statistical fit as evidence for predictive accuracy.

There can be other, more subtle variations of this error. For example, if the train-test split occurs after any of the other preprocessing or modeling steps (Modules 4 and 5), this also results in leakage. Vandewiele *et al.* (*78*) found overoptimistic results in 21 papers that claimed to predict the risk of preterm births. These papers suffered from the same error: oversampling data before partitioning it into training and test sets. This resulted in the test set becoming artificially similar to the training set and led to exaggerated performance claims across the literature on preterm risk prediction.

6b) Dependencies or duplicates between datasets

In some cases, samples in the dataset might have dependencies. For example, a clinical dataset might have many samples from the same patient. Oner *et al.* (*160*) find that when image data from the same patient is present in training and test sets, it leads to overoptimistic results. Similarly, for time-series forecasting models, randomly splitting a time-series dataset into training and test sets is likely to lead to overoptimism (*161*), because the training data has information "from the future" (*162*).

In such cases, the train-test split or CV split should take these dependencies into account—for instance, by including all samples from each patient in the same train-test split or CV fold. There are several ways to avoid these dependencies. Bergmeir and Benítez (*163*) find that blocked CV for time-series evaluation deals with temporal autocorrelation. Hammerla and Plotz (*164*) demonstrate how "neighborhood bias" can affect data recordings close in time. They introduce "meta-segmented CV" to deal with such dependencies. Roberts *et al.* (*165*) describe block CV strategies for a number of structures with dependencies, including temporal, spatial, and hierarchical dependencies.

Duplicates in the datasets can also spread across training and test sets if the dataset is split randomly. This should be avoided, as it leaks information across the train-test split. Roberts *et al.* (*15*) outline this error with Frankenstein datasets: Datasets that combine multiple other sources of data can end up using the same data twice—for instance, if two datasets that rely on the same underlying data source are combined into a larger dataset.

6c) Feature legitimacy

If any of the features used in a model is a proxy for the outcome, this can result in leakage. Filho *et al.* (*166*) found that a prominent paper on hypertension prediction (*167*) suffered from data leakage due to illegitimate features. The model included the use of antihypertensive drugs as a feature to predict hypertension. However, this feature is not available when predicting whether a new patient suffers from hypertension in a clinical setting, so it artificially inflates the performance of the ML model. Similarly, Epic's sepsis prediction tool used "antibiotics" as a feature to predict whether someone would get sepsis (*157*).

This type of leakage is more likely when there are a large number of features, due to the increased likelihood of including one or more illegitimate features. The sheer volume of features can make it challenging to scrutinize each one for potential leakage.

MODULE 7: METRICS AND UNCERTAINTY QUANTIFICATION

The performance of ML models is key to the scientific claims of interest. Because authors can make many possible choices when choosing performance metrics, it is important to reason why the metrics used are appropriate for the task (*8*, *168*). In addition, communicating and reasoning about uncertainty is important, but uncertainty is currently under-reported (*169–173*). For example, Simmonds *et al.* (*38*) find that studies often do not report the various kinds of uncertainty in the modeling process.

We ask authors to report their performance metrics and uncertainty estimates for those metrics in enough detail to enable a judgment about whether they made valid choices for evaluating the model's performance. The checklist requires authors to detail how they evaluate the performance of their model on its own and in relation to baselines. In addition, in this section, we ask authors to reason about their measurement of model performance and uncertainty in relation to their scientific claim.

7a) Performance metrics used

There are many possible ways to measure the performance of an ML model [see table 4 in (*8*) for an overview]. Certain metrics can be misleading or inappropriate. For example, accuracy might not be suitable to measure the performance of an ML model in the presence of heavy class imbalance: When most data points have positive labels, it is easy to obtain high accuracy simply by predicting positive for all cases (*174*). Other common metrics, like area under the receiver operating characteristic (AUROC), also have limitations (*175*).

The proper choice of metric is often influenced by the particular application being studied. In some domains, certain errors may be more costly than others. For example, false positives are more costly than false negatives in email spam detection (*176*). We ask authors to justify the use of a particular performance metric in relation to the scientific claim.

7b) Uncertainty estimates

Uncertainty in the performance of ML models can arise in many ways. For example, it could arise from randomness in the training data, evaluation data, or in the training process itself. Uncertainty is important to capture when evaluating the strength of a scientific claim. Because a dataset represents only a finite sample from a population, there is uncertainty in this sampling.

Research using ML methods can look to other areas of statistical practice for existing methods for quantifying uncertainty. For example, diagnostic testing in medicine is directly analogous to supervised classification, and so biostatistical methods for such settings (177) are broadly applicable. Beyond binomial proportion confidence intervals, such methods include McNemar's test, as well as analytic approaches for sample size calculations (particularly relevant for seeing whether there is sufficient power to audit for differential performance among subpopulations). Where methods do not exist, bootstrapping within the test set can be used to generate confidence intervals for performance claims (142).

Uncertainty in the performance of ML models can also affect downstream use of these labels in scientific inquiry. Angelopoulos et al. (173) show how this uncertainty can be propagated to subsequent analyses.

In a systematic review of uncertainty quantification across seven scientific fields, Simmonds et al. (38) found that scientific fields differ in what kinds of uncertainty they report. They propose best practices and a checklist to help researchers account for uncertainty.

7c) Appropriate statistical tests

Comparing the performance of different models can be key to some scientific claims. For example, a scientific claim may argue that one ML method outperforms others. Statistical testing is one tool to evaluate such differences in performance. Raschka (142) gives an overview of statistical tests for ML models. Still, in some or even most cases, an appropriate and valid statistical test may not be known; this can be an area for further development. The reliance on statistical significance testing has also led to misinterpretations and false conclusions. As a result, reporting uncertainty is better than performing statistical tests alone (178). Still, if a statistical test must be performed, it should be appropriate for the comparison.

MODULE 8: GENERALIZABILITY AND LIMITATIONS

External validity (or "generalizability") refers to the extent to which the findings from a study's sample apply to the target population, as well as the extent to which the findings apply to other populations, outcomes, and contexts (179, 180).

ML-based science faces a number of threats to external validity (13, 181). Because studies that use ML methods are often unaccompanied by external (i.e., out-of-distribution) validation (182), it is important to reason about these threats. In addition, authors are best positioned to identify the boundaries of applicability of their claims to prevent misunderstandings about the claims made in their study.

External validity of three types of claims in ML-based science

We distinguish among three types of external validity, corresponding to three types of claims made in ML-based science.

External validity of claims about observed patterns

In some studies, researchers use ML methods to make a claim about the presence of a pattern in the world. For example, Mathur et al. use structural topic modeling to study manipulative tactics in a corpus of U.S. political campaign emails from the 2020 election cycle. One of their findings is that the median active sender of campaign emails uses "sensationalist clickbait" in 37% of those emails (183). A question about the external validity of this claim might be: Is the frequency of sensationalist clickbait in campaign emails similar in other U.S. election cycles?

External validity of claims about fitted models

A "model" is a function that an algorithm learns when the algorithm is applied to data (184). In some studies, researchers train an ML model and make a claim about the model's performance when deployed in real-world settings. For example, Dugas et al. build a model to forecast influenza outbreaks at the city level using Google Flu Trends and other readily accessible data, with the goal that the model can be deployed by medical centers to provide warning of upcoming outbreaks. They find that "the model, on the average, predicts weekly influenza cases during 7 out-of-sample outbreaks within 7 cases for 83% of estimates" (p. 1) (185). A question about the external validity of their claim, which they note in their discussion, is: Does the model achieve similar performance in other years? External validity of claims about fitted models is more widely known as "domain generalization" or "robustness to distribution shift" and is a well-studied phenomenon in ML (186, 187).

External validity of claims about learning algorithms

An "algorithm" is a procedure for learning from data (184). In some studies, researchers make a claim about the usefulness of ML algorithms in a particular context. For example, Bansak et al. (188) develop an ML algorithm to assign refugees to resettlement sites by "leverag[ing] synergies between refugee characteristics and resettlement sites." They test the algorithm in retrospective data from the United States and Switzerland and find that their "approach led to gains of roughly 40 to 70%, on average, in refugees' employment outcomes relative to current assignment practices" (p. 1). They claim that governments can use this algorithmic assignment approach to improve resettlement outcomes for refugees. However, unlike the influenza forecasting example above, they do not claim that the specific model they trained can be deployed directly into use. A question about external validity of their claim might be: Does algorithmic assignment of refugee resettlement locations work similarly well for other time periods?

Reporting on external validity falls short in past literature

Reviews of past literature have found that scientific papers sometimes lack discussion about external validity (59) or lack information about sample demographics that would help readers draw their own conclusions about external validity (189). Furthermore, many ML-based studies make claims about the ability of a model to deploy in real-world settings but do not report whether their evaluation sample matches the population in the real-world setting (189) and do not conduct external validation (189, 190).

8a) Evidence of external validity

In this checklist item, we ask researchers to discuss the ability to generalize their claims from the sample to the target population and to other populations, outcomes, or contexts. Researchers can use a mix of quantitative and theoretical approaches to make arguments regarding their findings' external validity. They can report quantitative evidence by testing their claims in out-of-distribution data. They can make theoretical arguments about their expectations of external

validity by referring to prior literature and reasoning about the level of similarity between contexts (*60*).

Several threats to the external validity of ML models have been documented in past literature. Finlayson *et al.* (*191*) outline external validity failures due to dataset shifts in clinical settings. Hullman *et al.* (*13*) discuss the threats to external validity that arise in different phases of an ML research project. Liao *et al.* (*182*) outline a taxonomy of evaluation failures in ML, including failures in external validity. Geirhos *et al.* (*56*) discuss the phenomenon of shortcut learning in ML models, a phenomenon where models rely on shortcuts (such as the background color in an image) instead of detecting patterns that actually relate to the phenomena of interest.

Researchers who make claims about fitted models should be aware that even if their model currently generalizes to their target population, that performance may degrade over time because of temporal distribution shift (or "temporal drift"). Causes of temporal drift include changes in technology, changes in population and setting, and changes in behavior (*191*). Concerns about the risk of drift should be communicated in the paper, when applicable.

8b) Contexts in which the authors do not expect the study's findings to hold

Clarifying the circumstances under which scientific conclusions or model performance are not expected to hold helps to set clear expectations and avoid unjustified hype. Raji *et al.* (*181*) find that flaws in ML models deployed in real-world settings stem in part from a lack of focus on identifying when models are not expected to work. Simons *et al.* (*60*) argue that making an explicit "constraints on generality" statement that identifies boundaries of the circumstances where findings are expected to hold has several benefits, including helping to ensure that a study's conclusions accurately reflect its evidence, increasing the likelihood of successful replication, and inspiring follow-up studies that test the findings in new populations.

CONCLUSION

ML methods present an exciting advance for scientific research. Done right, they can allow researchers to analyze complex data and work with modalities such as images and video. Yet, recent failures of ML-based science reveal the urgent necessity of improving standards of transparency across fields that use ML methods. Our paper provides a cross-disciplinary bar for conducting and reporting ML-based science.

Supplementary Materials
This PDF file includes:
Texts S1 and S2
References

REFERENCES AND NOTES

1. S. Athey, G. W. Imbens, Machine learning methods that economists should know about. *Annu. Rev. Econ.* **11**, 685–725 (2019).
2. D. R. Schrider, A. D. Kern, Supervised machine learning for population genetics: A new paradigm. *Trends Genet.* **34**, 301–312 (2018).
3. J. J. Valletta, C. Torney, M. Kings, A. Thornton, J. Madden, Applications of machine learning in animal behaviour studies. *Anim. Behav.* **124**, 203–220 (2017).
4. R. Iniesta, D. Stahl, P. McGuffin, Machine learning, statistical learning and the future of biological research in psychiatry. *Psychol. Med.* **46**, 2455–2465 (2016).
5. S. Tonidandel, E. B. King, J. M. Cortina, Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organ. Res. Methods* **21**, 525–547 (2018).
6. T. Yarkoni, J. Westfall, Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspect. Psychol. Sci.* **12**, 1100–1122 (2017).
7. J. Grimmer, M. E. Roberts, B. M. Stewart, Machine learning for social science: An agnostic approach. *Annu. Rev. Polit. Sci.* **24**, 395–419 (2021).
8. A. K. Leist, M. Klee, J. H. Kim, D. H. Rehkopf, S. P. A. Bordas, G. Muniz-Terrera, S. Wade, Mapping of machine learning approaches for description, prediction, and causal inference in the social and health sciences. *Sci. Adv.* **8**, eabk1942 (2022).
9. T. L. Wiemken, R. R. Kelley, Machine learning in epidemiology and health outcomes research. *Annu. Rev. Public Health* **41**, 21–36 (2020).
10. H. R. Varian, Big data: New tricks for econometrics. *J. Econ. Perspect.* **28**, 3–28 (2014).
11. S. Mullainathan, J. Spiess, Machine learning: An applied econometric approach. *J. Econ. Perspect.* **31**, 87–106 (2017).
12. L. Breiman, Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, 199–231 (2001).
13. J. Hullman, S. Kapoor, P. Nanayakkara, A. Gelman, A. Narayanan, "The worst of both worlds: A comparative analysis of errors in learning from data in psychology and machine learning" in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (2022), pp. 335–348.
14. A. L. Beam, A. K. Manrai, M. Ghassemi, Challenges to the reproducibility of machine learning models in health care. *JAMA* **323**, 305–306 (2020).
15. M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J. H. F. Rudd, E. Sala, C.-B. Schonlieb, Common pitfalls and recommendations for using ¨ machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat. Mach. Intell.* **3**, 199–217 (2021).
16. X. Bouthillier, C. Laurent, P. Vincent, "Unreproducible research is reproducible" in *International Conference on Machine Learning* (PMLR, 2019), pp. 725–734.
17. M. B. McDermott, S. Wang, N. Marinsek, R. Ranganath, L. Foschini, M. Ghassemi, Reproducibility in machine learning for health research: Still a ways to go. *Sci. Transl. Med.* **13**, eabb1655 (2021).
18. O. E. Gundersen, S. Kjensmo, State of the art: Reproducibility in artificial Intelligence. *Proc. Conf. Artif. Intell.* **32**, (2018).
19. G. Varoquaux, V. Cheplygina, Machine learning for medical imaging: Methodological failures and recommendations for the future. *NPJ Digit. Med.* **5**, 48 (2022).
20. L. Messeri, M. J. Crockett, Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
21. R. M. Schmidt, F. Schneider, P. Hennig, "Descending through a crowded valley - benchmarking deep learning optimizers" in *Proceedings of the 38th International Conference on Machine Learning* (PMLR, 2021), pp. 9367–9376.
22. X. Bouthillier, P. Delaunay, M. Bronzi, A. Trofimov, B. Nichyporuk, J. Szeto, N. Mohammadi Sepahvand, E. Raff, K. Madan, V. Voleti, S. E. Kahou, V. Michalski, T. Arbel, C. Pal, G. Varoquaux, P. Vincent, Accounting for variance in machine learning benchmarks. *Proc. Mach. Learn. Syst.* **3**, 747–769 (2021).
23. O. DeMasi, K. Kording, B. Recht, Meaningless comparisons lead to false optimism in medical machine learning. *PLOS ONE* **12**, e0184604 (2017).
24. D. Sculley, J. Snoek, A. Wiltschko, A. Rahimi, *Winner's curse? On pace, progress, and empirical rigor* (2018); https://openreview.net/forum?id=rJWF0Fywf.
25. D. M. Liu, M. J. Salganik, Successes and struggles with computational reproducibility: Lessons from the fragile families challenge. *Socius* **5**, 10.1177/2378023119849803 (2019).
26. S. Dittmer, M. Roberts, J. Gilbey, A. Biguri, AIX-COVNET Collaboration, J. Preller, J. H. F. Rudd, J. A. D. Aston, C.-B. Schonlieb, Navigating the development challenges in creating complex data systems. *Nat. Mach. Intell.* **5**, 681–686 (2023).
27. O. E. Gundersen, S. Shamsaliei, R. J. Isdahl, Do machine learning platforms provide out-of-the-box reproducibility? *Future Gener. Comput. Syst.* **126**, 34–47 (2022).
28. J. M. Hofman, A. Sharma, D. J. Watts, Prediction and explanation in social systems. *Science* **355**, 486–488 (2017).
29. J. Banja, AI hype and radiology: A plea for realism and accuracy. *Radiol. Artif. Intell.* **2**, e190223 (2020).
30. V. E. Johnson, R. D. Payne, T. Wang, A. Asher, S. Mandal, On the reproducibility of psychological science. *J. Am. Stat. Assoc.* **112**, 1–10 (2017).
31. S. J. Bell, O. P. Kampman, Perspectives on machine learning from psychology's reproducibility crisis. arXiv:2104.08878 [cs.LG] (18 April 2021).
32. J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Lariviere, A. Beygelzimer, F. d'Alche-Buc, E. Fox, H. Larochelle, Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *J. Mach. Learn. Res.* **22**, 7459–7478 (2022).
33. M. Serra-Garcia, U. Gneezy, Nonreplicable publications are cited more than replicable ones. *Sci. Adv.* **7**, eabd1705 (2021).
34. M. J. Salganik, I. Lundberg, A. T. Kindel, C. E. Ahearn, K. al-Ghoneim, A. Almaatouq, D. M. Altschul, J. E. Brand, N. B. Carnegie, R. J. Compton, D. Datta, T. Davidson, A. Filippova, C. Gilroy, B. J. Goode, E. Jahani, R. Kashyap, A. Kirchner, S. McKay,

A. C. Morgan, A. Pentland, K. Polimis, L. Raes, D. E. Rigobon, C. V. Roberts, D. M. Stanescu, Y. Suhara, A. Usmani, E. H. Wang, M. Adem, A. Alhajri, B. AlShebli, R. Amin, R. B. Amos, L. P. Argyle, L. Baer-Bositis, M. Büchi, B. R. Chung, W. Eggert, G. Faletto, Z. Fan, J. Freese, T. Gadgil, J. Gagné, Y. Gao, A. Halpern-Manners, S. P. Hashim, S. Hausen, G. He, K. Higuera, B. Hogan, I. M. Horwitz, L. M. Hummel, N. Jain, K. Jin, D. Jurgens, P. Kaminski, A. Karapetyan, E. H. Kim, B. Leizman, N. Liu, M. Möser, A. E. Mack, M. Mahajan, N. Mandell, H. Marahrens, D. Mercado-Garcia, V. Mocz, K. Mueller-Gastell, A. Musse, Q. Niu, W. Nowak, H. Omidvar, A. Or, K. Ouyang, K. M. Pinto, E. Porter, K. E. Porter, C. Qian, T. Rauf, A. Sargsyan, T. Schaffner, L. Schnabel, B. Schonfeld, B. Sender, J. D. Tang, E. Tsurkov, A. van Loon, O. Varol, X. Wang, Z. Wang, J. Wang, F. Wang, S. Weissman, K. Whitaker, M. K. Wolters, W. L. Woon, J. Wu, C. Wu, K. Yang, J. Yin, B. Zhao, C. Zhu, J. Brooks-Gunn, B. E. Engelhardt, M. Hardt, D. Knox, K. Levy, A. Narayanan, B. M. Stewart, D. J. Watts, S. McLanahan, Measuring the predictability of life outcomes with a scientific mass collaboration. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 8398–8403 (2020).

35. J. Mongan, L. Moy, C. E. Kahn, Checklist for artificial intelligence in medical imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2**, e200029 (2020).

36. P. M. Bossuyt, J. B. Reitsma, D. E. Bruns, C. A. Gatsonis, P. P. Glasziou, L. Irwig, J. G. Lijmer, D. Moher, D. Rennie, H. C. W. de Vet, H. Y. Kressel, N. Rifai, R. M. Golub, D. G. Altman, L. Hooft, D. A. Korevaar, J. F. Cohen, STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies. *BMJ Open* **351**, h5527 (2015).

37. R. G. White, A. J. Hakim, M. J. Salganik, M. W. Spiller, L. G. Johnston, L. Kerr, C. Kendall, A. Drake, D. Wilson, K. Orroth, M. Egger, W. Hladik, Strengthening the Reporting of Observational Studies in Epidemiology for respondent-driven sampling studies: STROBE-RDS statement. *J. Clin. Epidemiol.* **68**, 1463–1471 (2015).

38. E. G. Simmonds, K. P. Adjei, C. W. Andersen, Janne Cathrin Hetle Aspheim, C. Battistin, N. Bulso, H. Christensen, B. Cretois, R. Cubero, I. A. Davidovich, L. Dickel, B. Dunn, E. Dunn-Sigouin, K. Dyrstad, S. Einum, D. Giglio, H. Gjerløw, A. Godefroidt, R. Gonzalez-Gil, S. G. Cogno, F. Grosse, P. Halloran, M. F. Jensen, J. J. Kennedy, P. E. Langsaether, J. H. Laverick, D. Lederberger, C. Li, E. Mandeville, C. Mandeville, E. Moe, T. N. Schroder, D. Nunan, J. S. Parada, M. R. Simpson, E. S. Skarstein, C. Spensberger, R. Stevens, A. Subramanian, L. Svendsen, O. M. Theisen, C. Watret, R. B. OHara, How is model-related uncertainty quantified and reported in different disciplines? arXiv:2206.12179 [stat.AP] (24 June 2022).

39. S. Kapoor, A. Narayanan, Leakage and the reproducibility crisis in machine-learningbased science. *Patterns* **4**, 100804 (2023).

40. G. S. Collins, J. B. Reitsma, D. G. Altman, K. G. Moons, Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD statement. *BMJ* **350**, g7594 (2015).

41. A. C. Plint, D. Moher, A. Morrison, K. Schulz, D. G. Altman, C. Hill, I. Gaboury, Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Med. J. Aust.* **185**, 263–267 (2006).

42. S. Han, T. F. Olonisakin, J. P. Pribis, J. Zupetic, J. H. Yoon, K. M. Holleran, K. Jeong, N. Shaikh, D. M. Rubio, J. S. Lee, A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review. *PLOS ONE* **12**, e0183591 (2017).

43. Principles and guidelines for reporting preclinical research (2015); www.nih.gov/research-training/rigor-reproducibility/principles-guidelinesreporting-preclinical-research.

44. Reporting guidelines. The EQUATOR Network; www.equatornetwork.org/reporting-guidelines/.

45. A. Rogers, T. Baldwin, K. Leins, 'just what do you think you're doing, Dave?' A checklist for responsible data use in NLP, in Findings of the Association for Computational Linguistics: EMNLP 2021, M.-F. Moens, X. Huang, L. Specia, S. Wen-tau Yih, Eds. (Punta Cana, Dominican Republic), (Association for Computational Linguistics, 2021). pp. 4821–4833.

46. O. E. Gundersen, Y. Gil, D. W. Aha, On reproducible AI: Towards reproducible research, open science, and digital scholarship in AI publications. *AI Mag.* **39**, 56–68 (2018).

47. D. Donoho, 50 years of data science. *J. Comput. Graph. Stat.* **26**, 745–766 (2017).

48. J. M. Hofman, D. J. Watts, S. Athey, F. Garip, T. L. Griffiths, J. Kleinberg, H. Margetts, S. Mullainathan, M. J. Salganik, S. Vazire, A. Vespignani, T. Yarkoni, Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).

49. E. Winsberg, *Science in the Age of Computer Simulation* (University of Chicago Press, 2010).

50. J. Pfeffer M. M. Malik, "Simulating the dynamics of socio-economic systems" in *Networked Governance: New Research Perspectives*, B. Hollstein, W. Matiaske, K.-U. Schnapp, Eds. (Springer International Publishing, 2017), pp.. 143–161.

51. R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. S. Chatterji, A. S. Chen, K. A. Creel, J. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. D. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. F. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. S. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles,

H. Nilforoshan, J. F. Nyarko, G. Ogut, L. J. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. H. Roohani, C. Ruiz, J. Ryan, C. R'e, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. P. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramer, R. E. Wang, W. Wang, B. Wu, J. Wu, ` Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. A. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, P. Liang, On the opportunities and risks of foundation models. arXiv:2108.07258 [cs.LG] (16 August 2021).

52. S. Kapoor, A. Narayanan, OpenAI's policies hinder reproducible research on language models (2022); www.aisnakeoil.com/p/openais-policies-hinderreproducible.

53. L. Chen, M. Zaharia, J. Zou, How is ChatGPT's behavior changing over time? arXiv:2307.09009 [cs.CL] (18 July 2023).

54. J. Carrasquilla, R. G. Melko, Machine learning phases of matter. *Nat. Phys.* **13**, 431–434 (2017).

55. I. Lundberg, R. Johnson, B. M. Stewart, What is your estimand? Defining the target quantity connects statistical evidence to theory. *Am. Sociol. Review* **86**, 532–565 (2021).

56. R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, F. A. Wichmann, Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).

57. L. Wilkinson, Statistical methods in psychology journals: Guidelines and explanations. *Am. Psychol.* **54**, 594–604 (1999).

58. A. Casteel, N. Bridier, Describing populations and samples in doctoral student research. *Int. J. Dr. Stud.* **16**, 339–362 (2021).

59. L. Tooth, R. Ware, C. Bain, D. M. Purdie, A. Dobson, Quality of reporting of observational longitudinal research. *Am. J. Epidemiol.* **161**, 280–288 (2005).

60. D. J. Simons, Y. Shoda, D. S. Lindsay, Constraints on generality (COG): A proposed addition to all empirical papers. *Perspect. Psychol. Sci.* **12**, 1123–1128 (2017).

61. J. Grimmer, M. E. Roberts, B. Stewart, *Text as data: A New framework for Machine Learning and the Social Sciences.* (Princeton Univ. Press, 2022).

62. I. Lundberg, J. E. Brand, N. Jeon, Researcher reasoning meets computational capacity: Machine learning for social science. *Soc. Sci. Res.* **108**, 102807 (2022).

63. A. Neufeld, D. Witten, Discussion of Breiman's 'two cultures': From two cultures to one. *Obs. Stud.* **7**, 171–174 (2021).

64. G. Shmueli, Comment on Breiman's "Two Cultures" (2002): From two cultures to multicultural. *Obs. Stud.* **7**, 197–201 (2021).

65. M. Baiocchi, J. Rodu, Reasoning using data: Two old ways and one new. *Obs. Stud.* **7**, 3–12 (2021).

66. E. L. Ogburn, I. Shpitser, Causal modelling: The two cultures. *Obs. Stud.* **7**, 179–183 (2021).

67. M. Molina, F. Garip, Machine learning for sociology. *Annu. Rev. Sociol.* **45**, 27–45 (2019).

68. H. H. Rashidi, N. Tran, S. Albahra, L. T. Dang, Machine learning in health care and laboratory medicine: General overview of supervised learning and auto-ML. *Int. J. Lab. Hematol.* **43**, 15–22 (2021).

69. A. L. Beam, I. S. Kohane, Big data and machine learning in health care. *JAMA* **319**, 1317 (2018).

70. L. G. McCoy, C. T. Brenna, S. S. Chen, K. Vold, S. Das, Believing in black boxes: machine learning for healthcare does not need explainability to be evidence-based. *J. Clin. Epidemiol.* **142**, 252–257 (2022).

71. A. L. Tarca, V. J. Carey, X.-W. Chen, R. Romero, S. Draghici, Machine learning and its applications to Biology. *PLoS Comput. Biol.* **3**, e116 (2007).

72. V. Stodden, Reproducing statistical results. *Annu. Rev. Stat. Its Appl.* **2**, 1–19 (2015).

73. T. Herndon, M. Ash, R. Pollin, Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Camb. J. Econ.* **38**, 257–279 (2014).

74. R. Herzog, P. A. M. Mediano, F. E. Rosas, R. Carhart-Harris, Y. S. Perl, E. Tagliazucchi, R. Cofre, Retraction note: A mechanistic model of the neural entropy increase elicited by psychedelic drugs. *Sci. Rep.* **12**, 15500 (2022).

75. M. R. Berenbaum, Retraction for Shu et al., Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2115397118 (2021).

76. M. Neunhoeffer, S. Sternberg, How cross-validation can go wrong and what to do about it. *Polit. Anal.* **27**, 101–106 (2019).

77. J. M. Hofman, D. G. Goldstein, S. Sen, F. Poursabzi-Sangdeh, J. Allen, L. L. Dong, B. Fried, H. Gaur, A. Hoq, E. Mbazor, N. Moreira, C. Muso, E. Rapp, R. Terrero, Expanding the scope of reproducibility research through data analysis replications. *Organ. Behav. Hum. Decis. Process* **164**, 192–202 (2021).

78. G. Vandewiele, I. Dehaene, G. Kovacs, L. Sterckx, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, J. Decruyenaere, S. Van Hoecke, T. Demeester, Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. *Artif. Intell. Med.* **111**, 101987 (2021).

79. J. P. A. Ioannidis, D. B. Allison, C. A. Ball, I. Coulibaly, X. Cui, A. C. Culhane, M. Falchi, C. Furlanello, L. Game, G. Jurman, J. Mangion, T. Mehta, M. Nitzberg, G. P. Page, E. Petretto, V. van Noort, Repeatability of published microarray gene expression analyses. *Nat. Genet.* **41**, 149–155 (2009).

80. T. Verstynen, K. P. Kording, Overfitting to 'predict' suicidal ideation. *Nat. Hum. Behav.* **7**, 680–681 (2023).

81. B. Haibe-Kains, G. A. Adam, A. Hosny, F. Khodakarami, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, C. S. Greene, T. Broderick, M. M. Hoffman, J. T. Leek, K. Korthauer, W. Huber, A. Brazma, J. Pineau, R. Tibshirani, T. Hastie, J. P. A. Ioannidis, J. Quackenbush, H. J. W. L. Aerts, Transparency and reproducibility in artificial intelligence. *Nature* **586**, E14–E16 (2020).

82. V. Stodden, J. Seiler, Z. Ma, An empirical analysis of journal policy effectiveness for computational reproducibility. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 2584–2589 (2018).

83. M. Gabelica, R. Bojcic, L. Puljak, Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *J. Clin. Epidemiol.* **150**, 33–41 (2022).

84. N. A. Vasilevsky, J. Minnier, M. A. Haendel, R. E. Champieux, Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ* **5**, e3208 (2017).

85. P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep reinforcement learning that matters. *Proc. AAAI Conf. Artif. Intell.* **32**, 3207–3214 (2018).

86. K. Musgrave, S. Belongie, S.-N. Lim, "A metric learning reality check" in *Computer Vision–ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm, Eds. (Springer International Publishing, Cham, 2020), pp. 681–699.

87. AAAI, AAAI reproducibility checklist; https://aaai.org/conference/aaai/aaai23/reproducibility-checklist/.

88. NeurIPS, NeurIPS 2023 paper guidelines; https://neurips.cc/public/guides/PaperChecklist.

89. ICML, ICML 2023 paper guidelines; https://icml.cc/Conferences/2023/PaperGuidelines.

90. Nature, Reporting standards and availability of data, materials, code and protocols; www.nature.com/nature-portfolio/editorial-policies/reporting-standards.

91. Science, Science journals: Editorial policies; www.science.org/content/page/sciencejournals-editorial-policies.

92. The Journal of Politics, Guidelines for data replication; www.journals.uchicago.edu/journals/jop/data-replication.

93. B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse, M. Humphreys, J. Ishiyama, D. Karlan, A. Kraut, A. Lupia, P. Mabry, T. Madon, N. Malhotra, E. Mayo-Wilson, M. McNutt, E. Miguel, E. L. Paluck, U. Simonsohn, C. Soderberg, B. A. Spellman, J. Turitto, G. VandenBos, S. Vazire, E. J. Wagenmakers, R. Wilson, T. Yarkoni, Promoting an open research culture. *Science* **348**, 1422–1425 (2015).

94. Koren, Miklos, Connolly, Marie, Lull, Joan, Vilhuber, Lars, Data and code availability standard (2022); https://zenodo.org/record/7436134.

95. Reviewing computational methods. *Nat. Methods* **12**, 1099–1099 (2015).

96. K. Peng, A. Mathur, A. Narayanan, Mitigating dataset harms requires stewardship: Lessons from 1000 papers. *Proc. Neural Inf. Process. Syst. Track on Datasets Benchmarks*, vol. 1, 2021.

97. U.S. Geological Survey, Data dictionaries; www.usgs.gov/datamanagement/data-dictionaries.

98. T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, K. Crawford, Datasheets for datasets. *Commun. ACM* **64**, 86–92 (2021).

99. J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, S. McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inform. Assoc.* **25**, 230–238 (2018).

100. J. Jordon, L. Szpruch, F. Houssiau, M. Bottarelli, G. Cherubin, C. Maple, S. N. Cohen, A. Weller, Synthetic Data–what, why and how? arXiv:2205.03257 [cs.LG] (6 May 2022).

101. Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).

102. B. Nowok, G. M. Raab, C. Dibben, Synthpop: Bespoke creation of synthetic data inR. *J. Stat. Softw.* **74**, 1–26 (2016).

103. J. Chen, D. Chun, M. Patel, E. Chiang, J. James, The validity of synthetic clinical data: a validation study of a leading synthetic data generator (Synthea) using clinical quality measures. *BMC Med. Inform. Decis. Mak.* **19**, 44 (2019).

104. S. Hao, W. Han, T. Jiang, Y. Li, H. Wu, C. Zhong, Z. Zhou, H. Tang, Synthetic data in AI: Challenges, applications, and ethical implications. arXiv:2401.01629v1 [cs.LG] (3 January 2024).

105. G. K. Sandve, A. Nekrutenko, J. Taylor, E. Hovig, Ten simple rules for reproducible computational Research. *PLoS Comput. Biol.* **9**, e1003285 (2013).

106. T. H. Vines, A. Y. Albert, R. L. Andrew, F. Debarre, D. G. Bock, M. T. Franklin, K. J. Gilbert, J.-S. Moore, S. Renaut, D. J. Rennison, The availability of research data declines rapidly with article age. *Curr. Biol.* **24**, 94–97 (2014).

107. E. Gibney, R. Van Noorden, Scientists losing data at a rapid rate. *Nature* (2013). https://doi.org/10.1038/nature.2013.14416.

108. V. Stodden, S. Miguez, Best practices for computational science: Software infrastructure and environments for reproducible and extensible research. *J. Open Research Softw.* **2**, e21 (2014).

109. L. Vilhuber, M. Connolly, M. Koren, J. Llull, P. Morrow, A template README for social science replication packages (2020); https://zenodo.org/record/4319999.

110. M. Singers, Awesome README; https://github.com/matiassingers/awesome-readme.

111. Harbert, Bash scripting (2018); https://rsh249.github.io/bioinformatics/bash script.html.

112. A. Clyburne-Sherin, X. Fei, S. A. Green, Computational reproducibility via containers in psychology. *Meta Psychol.* **3**, (2019).

113. C. L. A. Navarro, J. A. A. Damen, T. Takada, S. W. J. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. M. Moons, L. Hooft, Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic review. *BMC Med. Res. Methodol.* **22**, 12 (2022).

114. M. Yusuf, I. Atal, J. Li, P. Smith, P. Ravaud, M. Fergie, M. Callaghan, J. Selfe, Reporting quality of studies using machine learning models for medical diagnosis: A systematic review. *BMJ Open* **10**, e034568 (2020).

115. Y. Kim, J. Huang, S. Emery, Garbage in, garbage out: Data collection, quality assessment and reporting standards for social media data use in health research, infodemiology and digital disease detection. *J. Med. Internet Res.* **18**, e41 (2016).

116. R. S. Geiger, K. Yu, Y. Yang, M. Dai, J. Qiu, R. Tang, J. Huang, "Garbage in, garbage out? do machine learning application papers in social computing report where humanlabeled training data comes from?" in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, (Association for Computing Machinery, 2020).

117. F. Porzsolt, F. Wiedemann, S. I. Becker, C. J. Rhoads, Inclusion and exclusion criteria and the problem of describing homogeneity of study populations in clinical trials. *BMJ Evid. Based Med.* **24**, 92–94 (2019).

118. M. Salganik, *Bit by Bit: Social Research in the Digital Age* (Princeton Univ. Press, 2019).

119. S. Barocas, M. Hardt, A. Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* (2019); www.fairmlbook.org.

120. A. Z. Jacobs, H. Wallach, "Measurement and fairness" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (Association for Computing Machinery, New York, NY, USA, 2021), pp. 375–385.

121. M. Crede, P. D. Harms, Three cheers for descriptive statistics—and five more reasons why they matter. *Ind. Organ. Psychol.* **14**, 486–488 (2021).

122. J. Larson-Hall, L. Plonsky, Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Lang. Learn.* **65**, 127–159 (2015).

123. V. C. Bradley, S. Kuriwaki, M. Isakov, D. Sejdinovic, X.-L. Meng, S. Flaxman, Unrepresentative big surveys significantly overestimated US vaccine uptake. *Nature* **600**, 695–700 (2021).

124. L. Plonsky, Study quality in SLA. *Stud. Second Lang. Acquis.* **35**, 655–687 (2013).

125. P. McKnight, K. McKnight, S. Sidani, A. J. Figueredo, *Missing Data: A Gentle Introduction* (Guilford Publications, 2007).

126. J. L. Peugh, C. K. Enders, Missing data in educational research: A review of reporting practices and suggestions for improvement. *Rev. Educ. Res.* **74**, 525–556 (2004).

127. C. Mack, Z. Su, D. Westreich, *Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User's Guide (Third edition)* (2018); www.ncbi.nlm.nih.gov/books/NBK493611/.

128. S. Nijman, A. Leeuwenberg, I. Beekers, I. Verkouter, J. Jacobs, M. Bots, F. Asselbergs, K. Moons, T. Debray, Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *J. Clin. Epidemiol.* **142**, 218–229 (2022).

129. T. D. Little, T. D. Jorgensen, K. M. Lang, E. W. G. Moore, On the joys of missing data. *J. Pediatr. Psychol.* **39**, 151–162 (2014).

130. J. S. Nicholson, P. R. Deboeck, W. Howard, Attrition in developmental psychology. *Int. J. Behav. Dev.* **41**, 143–153 (2017).

131. W. R. Sterner, What is missing in counseling research? reporting missing data. *J. Couns. Dev.* **89**, 56–62 (2011).

132. J. A. Hussain, M. Bland, D. Langan, M. J. Johnson, D. C. Currow, I. R. White, Quality of missing data reporting and handling in palliative care trials demonstrates that further development of the CONSORT statement is required: a systematic review. *J. Clin. Epidemiol.* **88**, 81–91 (2017).

133. X. Chu, I. F. Ilyas, S. Krishnan, J. Wang, "Data cleaning: Overview and emerging challenges" in *Proceedings of the 2016 International Conference on Management of Data* (2016); pp. 2201–2206.

134. E. M. Buchanan, J. E. Scofield, Methods to detect low quality data and its implication for psychological research. *Behav. Res. Methods* **50**, 2586–2596 (2018).

135. T. Shadbahr, M. Roberts, J. Stanczuk, J. Gilbey, P. Teare, S. Dittmer, M. Thorpe, R. V. Torne, E. Sala, P. Lio, M. Patel, AIX-COCNET Collaboration; J. H. F. Rudd, T. Mirtti, A. Rannikko, J. A. D. Aston, J. Tang, C.-B. Schönlieb, Classification of datasets with imputed missing values: Does imputation quality matter? arXiv:2206.08478 [cs.LG] (16 June 2022).

136. E. Gryska, I. Bjorkman-Burtscher, A. S. Jakola, T. Dunas, J. Schneiderman, R. A. Heckemann, Deep learning for automatic brain tumour segmentation on mri: Evaluation of recommended reporting criteria via a reproduction and replication study. *BMJ Open* **12**, e059000 (2022).

137. E. Raff, "A step toward quantifying independently reproducible machine learning research" in Advances in Neural Information Processing Systems, vol. 32 (2019), pp. 5485–5495.

138. M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, T. Gebru, "Model cards for model reporting" in Proceedings of the Conference on Fairness, Accountability, and Transparency (2019), pp. 220–229.

139. J. Kleinberg, A. Liang, S. Mullainathan, "The theory is predictive, but is it complete? An application to human perception of randomness" in Proceedings of the 2017 ACM Conference on Economics and Computation (2017), pp. 125–126.

140. R. Roscher, B. Bohn, M. F. Duarte, J. Garcke, Explainable machine learning for scientific insights and discoveries. IEEE Access 8, 42200–42216 (2020).

141. C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215 (2019).

142. S. Raschka, Model evaluation, model selection, and algorithm selection in machine learning. arXiv:1811.12808v3 [cs.LG] (11 November 2020).

143. G. C. Cawley, N. L. C. Talbot, On over-fitting in model selection and subsequent selection bias in performance evaluation. J. Mach. Learn. Res. 11, 2079–2107 (2010).

144. S. Wager, Cross-validation, risk estimation, and model selection: Comment on a paper by rosset and tibshirani. J. Am. Stat. Assoc. 115, 157–160 (2020).

145. C. Marx, F. Calmon, B. Ustun, "Predictive multiplicity in classification" in Proceedings of the 37th International Conference on Machine Learning (PMLR, 2020), pp. 6765–6774.

146. J. Watson-Daniels, D. C. Parkes, B. Ustun, Predictive multiplicity in probabilistic classification. Proc AAAI Conf. Artif. Intell. 37, 10306–10314 (2023).

147. E. Black, M. Raghavan, S. Barocas, Model multiplicity: Opportunities, concerns, and solutions, in 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, 2022 June 21 to 24, 2022.

148. J. Dodge, S. Gururangan, D. Card, R. Schwartz, N. A. Smith, Show your work: Improved reporting of experimental results. arXiv:1909.03004 [cs.LG] (6 September 2019).

149. R. Islam, P. Henderson, M. Gomrokchi, D. Precup, Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. arXiv:1708.04133 [cs.LG] (10 August 2017).

150. A. F. Cooper, Y. Lu, J. Forde, C. M. De Sa, "Hyperparameter optimization is deceiving us, and how to stop it" in Advances in Neural Information Processing Systems (Curran Associates, Inc. 2021), vol. 34, pp. 3081–3095.

151. P. T. Sivaprasad, F. Mai, T. Vogels, M. Jaggi, Fe. Fleuret, Optimizer benchmarking needs to account for hyperparameter tuning, in Proceedings of the 37th International Conference on Machine Learning (PMLR, 2020), pp. 9036–9045.

152. G. E. Dahl, F. Schneider, Z. Nado, N. Agarwal, C. S. Sastry, P. Hennig, S. Medapati, R. Eschenhagen, P. Kasimbeg, D. Suo, J. Bae, J. Gilmer, A. L. Peirson, B. Khan, R. Anil, M. Rabbat, S. Krishnan, D. Snider, E. Amid, K. Chen, C. J. Maddison, R. Vasudev, M. Badura, A. Garg, P. Mattson, Benchmarking neural network training algorithms. arXiv:2306.07179 [cs.LG] (12 June 2023).

153. P. Probst, A.-L. Boulesteix, B. Bischl, Tunability: Importance of hyperparameters of machine learning algorithms. J. Mach. Learn. Res. 20, 1934–1965 (2019).

154. J. Lin, The neural hype and comparisons against weak baselines. ACM SIGIR Forum 52, 40–51 (2019).

155. M. A. Lones, How to avoid machine learning pitfalls: A guide for academic researchers. arXiv:2108.02497v3 [cs.LG] (9 February 2023).

156. S. Kaufman, S. Rosset, C. Perlich, O. Stitelman, Leakage in data mining: Formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data 6, 1–21 (2012).

157. C. Ross, Epic's sepsis algorithm is going off the rails in the real world. The use of these variables may explain why, 2021; www.statnews.com/2021/09/27/epic-sepsisalgorithm-antibiotics-model/.

158. M. Kuhn, K. Johnson, Applied Predictive Modeling (Springer-Verlag, 2013).

159. R. A. Poldrack, G. Huckins, G. Varoquaux, Establishment of best practices for evidence for prediction: A review. JAMA Psychiatry 77, 534–540 (2020).

160. M. U. Oner, Y.-C. Cheng, H. K. Lee, W.-K. Sung, Training machine learning models on patient level data segregation is crucial in practical clinical applications (2020); www.medrxiv.org/content/10.1101/2020.04.23.20076406v1.

161. M. M. Malik, A hierarchy of limitations in machine learning. arXiv:2002.05193 [cs.CY] (12 February 2020).

162. M. Lachanski, S. Pav, Shy of the character limit: Twitter mood predicts the stock market revisited. Econ J. Watch 14, 302–345 (2017).

163. C. Bergmeir, J. M. Benitez, On the use of cross-validation for time series predictor evaluation. Inform. Sci. 191, 192–213 (2012).

164. N. Y. Hammerla, T. Plotz, "Let's (not) stick together: Pairwise similarity biases cross-validation in activity recognition" in Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Osaka, Japan (ACM, 2015), pp. 1041–1051.

165. D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schroder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, C. F. Dormann, Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929 (2017).

166. A. Chiavegatto Filho, A. F. D. M. Batista, H. G. Dos Santos, Data leakage in health outcomes prediction with machine learning. Comment on Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. J. Med. Internet Res. 23, e10969 (2021).

167. C. Ye, T. Fu, S. Hao, Y. Zhang, O. Wang, B. Jin, M. Xia, M. Liu, X. Zhou, Q. Wu, Y. Guo, C. Zhu, Y.-M. Li, D. S. Culver, S. T. Alfreds, F. Stearns, K. G. Sylvester, E. Widen, D. McElhinney, X. Ling, Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. J. Med. Internet Res. 20, e22 (2018).

168. J. Z. Forde, A. F. Cooper, K. Kwegyir-Aggrey, C. De Sa, M. Littman, Model selection's disparate impact in real-world deep learning applications. arXiv:2104.00606 [cs.LG] (1 April 2021).

169. U. Bhatt, J. Antoran, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melanc¸on, R. Krishnan, J. Stanley, O. Tickoo, L. Nachman, R. Chunara, M. Srikumar, A. Weller, A. Xiang, "Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty" in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, AIES '21 (Association for Computing Machinery, New York, NY, USA, 2021), pp. 401–413.

170. A. F. Cooper, K. Lee, M. Z. Choksi, S. Barocas, C. De Sa, J. Grimmelmann, J. Kleinberg, S. Sen, B. Zhang, Arbitrariness and Social Prediction: The Confounding Role of Variance in Fair Classification. Proc. AAAI Conf. Artif. Intell. 38, 22004–22012 (2024).

171. S. Qian, V. H. Pham, T. Lutellier, Z. Hu, J. Kim, L. Tan, Y. Yu, J. Chen, S. Shah, "Are my deep learning systems fair? An empirical study of fixed-seed training, in Advances in Neural Information Processing Systems (Curran Associates Inc. 2021), vol. 34, (pp. 30211–30227).

172. C. Young, Model Uncertainty and the Crisis in Science. Socius 4, 2378023117737206 (2018).

173. A. N. Angelopoulos, S. Bates, C. Fannjiang, M. I. Jordan, T. Zrnic, Prediction-powered inference. Science 382, 669–674 (2023).

174. M. C. Monard, G. Batista, Learning with skewed class distributions. Adv. Logic, Artif. Intell. Robot. 85, 173–180 (2002).

175. J. M. Lobo, A. Jimenez-Valverde, R. Real, Auc: A misleading measure of the performance of predictive distribution models. Glob. Ecol. Biogeogr. 17, 145–151 (2008).

176. A. Bhowmick, S. M. Hazarika, E-mail spam filtering: A review of techniques and trends. Adv. Electron. Commun. Comput. 443, 583–590 (2018).

177. X.-H. Zhou, D. K. McClish, N. A. Obuchowski, Statistical Methods in Diagnostic Medicine (John Wiley & Sons, ed. 2, 2011).

178. V. Amrhein, S. Greenland, B. McShane, Scientists rise up against statistical significance. Nature 567, 305–307 (2019).

179. W. R. Shadish, T. D. Cook, D. T. Campbell, Experimental and Quasi-Experimental Designs for Generalized Causal Inference (Houghton Mifflin, ed. 2, 2001).

180. N. Egami, E. Hartman, Elements of external validity: Framework, design, and analysis. Am. Polit. Sci. Rev. 117, 1070–1088 (2022).

181. I. D. Raji, I. E. Kumar, A. Horowitz, A. Selbst, "The fallacy of AI functionality" in 2022 ACM Conference on Fairness, Accountability, and Transparency (ACM, Seoul Republic of Korea, 2022), p. 959–972.

182. T. Liao, R. Taori, I. D. Raji, L. Schmidt, "Are we learning yet? a meta review of evaluation failures across machine learning" in Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2021).

183. A. Mathur, A. Wang, C. Schwemmer, M. Hamin, B. M. Stewart, A. Narayanan, Manipulative tactics are the norm in political emails: Evidence from 300k emails from the 2020 US election cycle. Big Data Soc. 10, 205395172211453 (2023).

184. J. Brownlee, Difference between algorithm and model in machine learning (2020); https://machinelearningmastery.com/difference-between-algorithm-and-model-inmachine-learning/.

185. A. F. Dugas, M. Jalalpour, Y. Gel, S. Levin, F. Torcaso, T. Igusa, R. E. Rothman, Influenza forecasting with Google Flu trends. PLOS ONE 8, e56176 (2013).

186. O. Wiles, S. Gowal, F. Stimberg, S. Alvise-Rebuffi, I. Ktena, K. Dvijotham, T. Cemgil, A fine-grained analysis on distribution shift. arXiv:2110.11328 [cs.LG] (21 October 2021).

187. P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao, T. Lee, E. David, I. Stavness, W. Guo, B. A. Earnshaw, I. S. Haque, S. Beery, J. Leskovec, A. Kundaje, E. Pierson, S. Levine, C. Finn, P. Liang, "Wilds: A benchmark of in-the-wild distribution shifts" in International Conference on Machine Learning (PMLR, 2021), pp. 5637–5664.

188. K. Bansak, J. Ferwerda, J. Hainmueller, A. Dillon, D. Hangartner, D. Lawrence, J. Weinstein, Improving refugee integration through data-driven algorithmic assignment. Science 359, 325–329 (2018).

189. S. Bozkurt, E. M. Cahan, M. G. Seneviratne, R. Sun, J. A. Lossio-Ventura, J. P. A. Ioannidis, T. Hernandez-Boussard, Reporting of demographic data and representativeness in machine learning models using electronic health records. J. Am. Med. Inform. Assoc. 27, 1878–1884 (2020).

190. C. L. A. Navarro, J. A. Damen, T. Takada, S. W. Nijman, P. Dhiman, J. Ma, G. S. Collins, R. Bajpai, R. D. Riley, K. G. Moons, L. Hooft, Systematic review finds spin practices and poor reporting standards in studies on machine learning-based prediction models. *J. Clin. Epidemiol.* **158**, 99–110 (2023).

191. S. G. Finlayson, A. Subbaswamy, K. Singh, J. Bowers, A. Kupke, J. Zittrain, I. S. Kohane, S. Saria, The clinician and dataset shift in artificial intelligence. *New Engl. J. Med.* **385**, 283–286 (2021).

192. M. Macleod, A. M. Collings, C. Graf, V. Kiermer, D. Mellor, S. Swaminathan, D. Sweet, V. Vinson, The MDAR (Materials Design Analysis Reporting) framework for transparent reporting in the life sciences. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2103238118 (2021).

193. Unofficial guidance on various topics by social science data editors; https://socialscience-data-editors.github.io/guidance/.

194. Requirements file format–pip documentation v23.0.1; https://pip.pypa.io/en/stable/reference/requirements-file-format/.

195. Nature research code and software submission checklist; www.nature.com/documents/nr-software-policy.pdf, 2017.

196. T. Comi, Using Codeocean for sharing reproducible research; https://rse.princeton.edu/2021/03/using-codeocean-for-sharing-reproducible-research/.

197. K. S. Chmielinski, S. Newman, M. Taylor, J. Joseph, K. Thomas, J. Yurkofsky, Y. C. Qiu, The Dataset Nutrition Label (2nd Gen): Leveraging context to mitigate harms in artificial intelligence. arXiv:2201.03954 [cs.LG] (10 January 2022).

198. *Brain imaging data structure* (2023); https://bids.neuroimaging.io/index.

199. H. Taherdoost, *Sampling methods in research methodology*; How to choose a sampling technique for research, 2016.

200. J. P. Vandenbroucke, E. Von Elm, D. G. Altman, P. C. Gøtzsche, C. D. Mulrow, S. J. Pocock, C. Poole, J. J. Schlesselman, M. Egger, STROBE Initiative, Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration, *Int. J. Surg.* **4**, e297 (2007).

201. 3.2. Tuning the hyper-parameters of an estimator; https://scikitlearn/stable/modules/grid search.html.

202. A. Vehtari, Cross-validation FAQ; https://avehtari.github.io/modelselection/CVFAQ.html.