Appendix A: REFORMS checklist template

Visit reforms.cs.princeton.edu for the latest version.

About. The REFORMS checklist lists items that should be reported in a scientific study that uses machine learning (ML) methods. It is intended to accompany the paper or report that introduces an ML model: for instance, as an appendix or supplemental material. The checklist consists of 32 questions spread across 8 modules. For each item, either list the section name, section number, or page number in the paper where the item is reported, or justify why a given item is not filled out. Note that not all of these items need to be reported in the main text of the paper; they could be reported in an appendix or supplementary files.

Some items in the checklist could be hard to report for specific studies. For instance, including a reproduction script to computationally reproduce all results (2e.) might not be possible for studies performed on academic computing clusters or those which use private data that cannot be released. Instead of requiring strict adherence for each item, we suggest authors and referees decide which items are relevant for a study and where details can be reported better. The items in our reporting standards could be a helpful starting point.

Use the accompanying Guidelines for reporting ML-based science to see how each item can be filled out. We also provide a sample checklist based on Obermeyer et al. (2019) (URL: https://reforms.cs.princeton.edu/obermeyer-sample.pdf)

This is a beta version of our checklist. We are soliciting feedback and will continue to update the template (visit reforms.cs.princeton.edu for the latest version). For feedback or questions, contact: sayashk@princeton.edu. The checklist starts on the page after the author list. After filling it out, save it starting from that page.

Authors

Sayash Kapoor
Emily Cantrell
Kenny Peng
Thanh Hien Pham
Christopher A. Bail
Odd Erik Gundersen
Jake M. Hofman
Jessica Hullman
Michael A. Lones
Momin M. Malik
Priyanka Nanayakkara
Russell A. Poldrack

Inioluwa Deborah Raji
Michael Roberts
Matthew J. Salganik
Marta Serra-Garcia
Brandon M. Stewart
Gilles Vandewiele
Arvind Narayanan

Checklist for reporting ML-based science

Module 1: Study goals

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

1b. Motivation for choosing this population or distribution (1a.).

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

Module 2: Computational reproducibility

2a. Dataset used for training and evaluating the model along with link or DOI to uniquely identify the dataset.

2b. 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. Description of the computing infrastructure used.

- Hardware infrastructure: CPU, GPU, RAM, disk space etc.
- Operating system.
- Software environment: Programming language and version, documentation of all packages used along with versions and dependencies (e.g., through a requirements.txt file).
- An estimate of the time taken to generate the results.

2d. README file which contains instructions for generating the results using the provided dataset and code.

2e. Reproduction script to produce all results reported in the paper[1].

Module 3: Data quality

[1] Note that this is a high bar for computational reproducibility. It might not be possible to provide such a script—for instance, if the analysis is run on an academic computing cluster, or if the dataset does not allow for programmatic download.

3a. 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. Distribution or set from which the dataset is sampled (i.e., the sampling frame).

3c. Justification for why the dataset is useful for the modeling task at hand.

3d. The definition of the outcome variable of the model along with descriptive statistics, if applicable.

(The outcome variable is also known as the dependent variable, the target variable, the output variable or the predicted variable).

3e. Number of samples in the dataset.

3f. Percentage of missing data, split by class for a categorical outcome variable.

3g. Justification for 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.).

Module 4: Data preprocessing

4a. Identification of whether any samples are excluded with a rationale for why they are excluded.

4b. How impossible or corrupt samples are dealt with.

4c. 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.

Module 5: Modeling

5a. Detailed descriptions of all models trained, including:

- All features used in the model (including any feature selection).
- Types of models implemented (e.g., Random Forests, Neural Networks).
- Loss function used.

5b. Justification for the choice of model types implemented.

5c. Method for evaluating the model(s) reported in the paper, including details of train-test splits or cross-validation folds.

5d. 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:

- Range of hyper-parameters used and a justification for why this range is reasonable.
- Method to select the best hyper-parameter configuration.
- Specification of all hyper-parameters used to generate results reported in the paper.

5f. Justification that model comparisons are against appropriate baselines.

Module 6: Data leakage

6a. Justification that pre-processing (Section 4) and modeling (Section 5) steps only use information from the training dataset (and not the test dataset).

6b. Methods 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. Justification that each feature or input used in the model is legitimate for the task at hand and does not lead to leakage.

Module 7: Metrics and uncertainty

7a. 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. Uncertainty estimates (e.g., confidence intervals, standard deviations), and details of how these are calculated.

7c. Justification for the choice of statistical tests (if used) and a check for the assumptions of the statistical test.

Module 8: Generalizability and limitations

8a. Evidence of external validity.

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

Appendix B: Guidelines for filling out the REFORMS checklist

Visit reforms.cs.princeton.edu for the latest version.

These guidelines provide documentation for each item in the Reporting standards for ML-based science. We elaborate on why researchers should consider reporting the item, link to additional helpful resources to accomplish each item and add references to articles that describe the issues in depth.

We also provide a sample checklist based on Obermeyer et al. (2019) (URL: https://reforms.cs.princeton.edu/obermeyer-sample.pdf).

As noted in our paper, some of the items in our reporting standards could be hard to report for specific studies. For instance, including a reproduction script to computationally reproduce all results (2e.) might not be possible for studies performed on academic computing clusters or those which use private data that cannot be released.

Instead of requiring strict adherence for each item, we suggest authors and referees decide which items are relevant for a study and how details can be reported better.

Module 1: Study design

The items in this section help communicate the purpose and goals of the study and how various decisions in the study design were arrived at. Details about the design of the study are important to clarify the applicability of the scientific claims of the study. They also help communicate the motivation behind researchers' various degrees of freedom, i.e., decisions researchers make throughout the research and analysis process that influence their findings.

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

Researchers make scientific claims about a given distribution or population that they are interested in studying. Note that this is the population of interest, and not the sample, which can be specified later in (3b.)

To communicate the applicability of the claims, explicitly report the distribution or population about which you expect the scientific claims to hold. For example, "US children aged between 12 and 18" or "people engaging in online debates on climate change."

1b. Motivation for choosing this population or distribution (1a).

Justify why the researchers chose this population or distribution. For example: “We aimed to determine whether existing vaccines for COVID-19 are effective in children aged between 12 and 18. There are no prior studies on vaccine efficacy in this population.”

A valid motivation is having access to a dataset that inspired a research question, and thus the population or distribution of interest is limited by the dataset. For example, studying CDC data for all U.S. counties would limit the population of interest to US counties.

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

Report the reasons for using ML methods and consider comparing it with alternative or traditional methods that could be used for similar aims.

For example, if the goal of the research is to make a prediction, i.e., if explanation is not a goal of the study, ML methods can help improve predictive accuracy.

See Hofman et al. (2021) for an overview of the different types of modeling and their aims.

Module 2: Computational reproducibility

Computational reproducibility refers to the ability of a researcher to get the same figures and results that are reported in a paper or manuscript without making any changes to the code, data, or computing environment. This is important for ensuring the scientific validity of a study: errors can be uncovered quickly, independent researchers can verify the findings in a study, and researchers can easily build on a study’s results. Several journals currently require computational reproducibility and have specific guidelines. If you’re already using a discipline or journal-specific checklist, specify that here.

See Liu and Salganik (2019) for a discussion on the importance and challenges of ensuring computational reproducibility.

Sandve et al. (2013) discuss high-level imperatives and research practices that can enable computational reproducibility.

See the Social Science Data Editors’ guidance on computational reproducibility.

Include as many of the items below as possible, in supplementary documents alongside a paper or pre-print that describes the study. Ideally, upload them to an established repository that provides a persistent identifier for the resources (such as Harvard Dataverse or Zenodo). Since code, data, and computational environments can have different versions over time, include the precise version that you use to generate the results reported in a study.

For some domains, sharing the code and dataset is not possible due to the presence of sensitive data. Specify below if such a restriction applies.

2a. Dataset

Report a permanent link or DOI to the specific version of the dataset used for training and evaluating the model. For a discussion of the importance of DOIs, see Peng, Mathur, Narayanan (2021).

If an original dataset was used, also include the data dictionary for the dataset. A data dictionary describes metadata about the dataset, and familiarizes a reader to the properties and format of the data. The US Geological Survey has a detailed guide to data dictionaries, complete with examples and instructions.

If the dataset contains sensitive information and cannot be publicly released, consider releasing a synthetic dataset, or releasing the data per request or application. There are packages that support generation of a synthetic dataset such as synthpop for R.

2b. Code

Provide a commit tag (for instance, on Github, GitLab, or BitBucket), a DOI, or equivalent documentation to precisely identify the version of the code used to train and evaluate the model and produce the exact results reported in the paper.

In the code, include comments with explanations of variables and operations to sufficiently mark different stages of the analysis for an unfamiliar reader. The documentation in (2d) can refer to these comments for greater clarity.

2c. Computing infrastructure

To help readers understand the precise computing requirements for reproducing your study, whenever possible, report the following details on the infrastructure used to generate the results:

1. Hardware infrastructure: CPU, GPU, RAM, disk space.
2. Operating system and its version.
3. Software environment: Programming language and version, documentation of all packages used along with versions and dependencies (e.g., through a requirements.txt file).
4. An estimate of the time taken to generate the results.

Computing infrastructure is always changing, and thus could make it difficult or impossible to replicate a study with a slightly different environment. Having the exact details is crucial for replication.

See Requirements File Format from Python's pip installer for an example of how to document package versions.

See Stodden and Miguez (2014) for more detailed best practices to document computing infrastructure.

2d. README

Report the exact steps that should be taken by independent researchers to reproduce the results in your study, given access to the code, dataset, and computing environment specified in 2a-c.

A good README helps someone unfamiliar with the 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.

Consider using Nature's README for software submission. See also the README template for social science replication packages.

The "Awesome README" repository compiles examples, templates, and best practices for writing README files.

2e. Reproduction script

A script to produce all results reported in the paper using the code and dataset can significantly reduce the time it takes for an independent researcher to reproduce the results reported in a study.

The script should go through all steps involved in producing the results. For example, the script should download the packages, set the right dependencies, download and store the dataset in the correct location, set up the computational environment, pre-process the data, and run the code to produce exactly the same results as reported in the paper.

One option is a bash script which carries out each of the steps you list in (2d). Another way is to use an online reproducibility platform such as CodeOcean, which allows researchers to share the required materials in 2a-c along with a reproduction script.

Note that this is a high bar for computational reproducibility, and in some cases, it might not be possible to provide such a script—for instance, if the analysis is run on an academic high-performance computing cluster, or if the dataset does not allow for programmatic download. It could also be challenging to set up, and resources listed here might help. In case you are not able to share a reproduction script, specify why.

Comi (2021) introduces CodeOcean for reproducible research, and shares how to create a CodeOcean capsule from Git.

Module 3: Data quality

This section is focused on reporting details about how the data used for developing and evaluating the model is collected. A good quality dataset is key to making valid scientific claims using ML models. The items in this section help readers understand and evaluate the quality of the data used in the modeling process.

3a. Data source(s)

Report details about the source of the dataset, separately for the training and validation data sets (if applicable). For instance, if re-using the dataset from a previous study, cite the study and explain what the source of the data collection was.

If collecting a new dataset, report the data collection process, who annotated the dataset, and how the annotations were carried out. Report the time-period and geographic locations of data collection.

You can also follow discipline-specific best-practices when releasing or using datasets. Examples include Datasheets for Datasets (Gebru et al., 2021), Dataset Nutrition Labels (Chmielinski et al., 2022), or the Brain Imaging Data Structure for Neuroimaging. If available, include such supplementary documents as supplementary materials along with the paper.

3b. Sampling frame

The sampling frame is the source from which a sample is drawn (using a sampling method.) The unit of the sampling frame is typically also the unit of the sample.

Report the sampling frame, which is the distribution or set from which the dataset is sampled. Include the sampling method (e.g., simple random, stratified, cluster sampling, etc.) Include any details about the distribution or population that pertains to the study (1a.).

Taherdoost, (2016) compiled a short guide to sampling in research.

3c. Justification for why the dataset is useful for the modeling task at hand

Report the rationale for why the dataset is useful for modeling and making the scientific claim reported in the study. Justifications could describe why the dataset is relevant to the modeling task, such as quantifying the population of interest well, or including novel insight that would be discovered through modeling.

3d. Details about the outcome variable

The outcome or target variable of the ML model is the quantity that the model is used to predict, detect, classify, or estimate. In other words, it is the variable of interest in the modeling process.

Report the outcome variable of the ML model. Provide descriptive statistics (e.g., mean, median, and variance) for the outcome variable, if applicable. For tasks with a continuous outcome variable (i.e., regression tasks), consider providing a plot of the outcome's distribution, such as a histogram.

3e. Number of samples in the dataset

Report the total number of samples (for a tabular dataset, this is the total number of rows in the dataset) as well as the number of samples in each class for a classification task.

If there are individuals or entities with multiple observations, report both the number of distinct individuals, as well as overall rows or units of data. For example, if you have a dataset with 10,000 rows with data on 5,000 unique patients, report both of these numbers. See also (6b.)

3f. Percentage of missing data, split by class for a categorical outcome variable

Datasets often have missing samples. An estimate of missingness can give readers an idea of how important the methods for dealing with missing data are in a given study.

Report the number or percentage of missing samples for each feature, when possible. Alternatively, provide summary statistics for the proportion of missing data.

See also (4c.) for methods for handling missing data.

3g. Dataset for evaluation is representative

Justify why the distribution or set from which the dataset is drawn (3b.) is representative of the population about which the scientific claim is being made (1a.).

There are many reasons the sampling frame could be unrepresentative: for example, if it is a convenience sample, if it under-represents minorities, or constitutes a too small sample size (Hullman et al., 2022). If the sample is unrepresentative of the target population, note this as a concern in the section on external validity (8a.).

Module 4: Data preprocessing

Pre-processing is the series of steps taken to convert the dataset used from its raw form into the final form used in the modeling process. This includes data 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.

Since pre-processing steps can influence the scientific claims made based on ML models (Hofman et al. 2017), it is important to specify the exact steps used in a study.

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. Report the criteria for selecting a subset of rows from the initial dataset (if any).

4b. How impossible or corrupt samples are dealt with

Some datasets might have feature values that are impossible (for instance, if the height of a human is recorded as greater than 10 feet). Some samples might have corrupt data.

Report the checks made for impossible or corrupt data. In case you find impossible or corrupt data, report mitigation strategies, such as methods used for detecting or removing outliers.

4c. Data transformations

Researchers often perform several transformations on a dataset before using it in an ML model. For example, they might impute missing data in a dataset using mean imputation or over-sample data from the minority class.

Report the precise sequence of all transformations of data from its raw form to the final form used in the model (e.g., missing data imputation, feature or outcome normalization, data augmentation using oversampling), preferably through a flow-chart, like a STROBE flow diagram.

Specify if each transformation is data-dependent (e.g., mean imputation) or data-independent (e.g., log transformation). Note that data-dependent transformations must be done within splits. For example, when using 5-fold cross-validation, perform mean imputation within each of the folds instead of performing it on the entire data together to avoid leaking information between the training and test data. See also 6a.

Shadbahr et al. (2022) discuss how poorly imputed data can lead to poor interpretability of the final model.

Module 5: Modeling

There are many steps involved in creating an ML model. This makes it hard to report the exact details of how an ML model is created, and can hinder replication by independent researchers. Specify the main steps in the modeling process, including feature selection, the types of models considered, and evaluation.

5a. Model description

To help readers determine how the models were trained, provide a detailed description of all models trained over the course of the study. For each model, include:

1. Inputs (including any feature selection steps and a description of the set of features used) and outputs
2. Types of models implemented (e.g., Random Forests, Neural Networks)
3. Loss function used

5b. Justification for choice of model types implemented

Describe why the types of models used are relevant for the study. Examples are “using a standard method for this field such as regularized regressions”, or “using decision trees for high explainability.”

Leist et al. (2022) describe various ML models that are suitable for different modeling tasks.

5c. Model evaluation method

Evaluating ML models requires testing them on data that they were not trained on, for instance by using a held-out test set or cross-validation (CV).

Report how the dataset is split for evaluating the ML model(s), for instance:

1. Cross-validation or nested CV
2. Held-out test set (internal validation set)
3. True out-of-sample set (external validation set; where the data comes from a different set compared to training data)

For the model evaluation method used, report details such as the number of samples in each train-test split or CV fold, as well as the number of samples of each class in each split (for a classification task).

Documentation from the Python package scikit learn elaborates why and how to do a train-validation-test split.

Vehtari (2020) describes various scenarios where using CV is appropriate.

Neunhoeffer and Sternberg (2018) highlight a common failure mode: using CV for *both* model selection and evaluation. Using nested CV helps address this issue.

Cawley and Talbot (2010) explore this issue in more detail and describe procedures for nested CV (section 5.1).

5d. Model selection method

Several ML models might be fit using the training set.

Report the criteria for choosing the final model(s) reported in the study. For instance, report if model performance on the training set, internal cross-validation fold (for nested cross-validation) or a separate validation set was used to select the final model(s) reported in the paper.

Raschka (2018) gives an overview of model selection techniques.

5e. Hyper-parameter selection

ML models often have hyperparameters. For example, Lasso regression has an additional penalty term (lambda or λ) that can be tuned. Tuning hyperparameters—trying different values and picking the one that works best—can help find the optimal performance for a given model and dataset.

Report the method used to compare the performance of different hyperparameter values. This should include details of what values for each parameter are considered, why these values are reasonable, how various hyperparameters are selected (for example, grid search or random search), and which hyperparameters are used in the final model(s) reported in the paper.

5f. Appropriate baselines

If comparing model performance against baselines, justify how the baselines are tuned appropriately and the model comparison is fair if applicable. (Note that this does not apply to comparisons against non-model based performance, such as comparing ML methods with human performance.)

Sculley et al. (2018) highlight several results in ML research that compare against weak baselines.

Lin (2019) highlights that comparisons against weak baselines can make results seem significant.

Module 6: Data leakage

Data leakage is a spurious relationship between the independent variables and the target variable that arises as an artifact of the data collection, sampling, pre-processing or modeling steps. Since the spurious relationship won’t be present in the distribution about which scientific claims are made, leakage usually leads to inflated estimates of model performance. Items in this section help detect and prevent leakage in the models developed and evaluated in a study.

Kapoor and Narayanan (2022) discuss the prevalence of leakage and provide “Model Info Sheets” to detect and prevent leakage in ML-based science.

6a. Train-test separation is maintained

When information from the test set is used during the training process, it leads to overly optimistic performance and results in data leakage.

Justify how all pre-processing (Section 4) and modeling (Section 5) steps only use information from the training data and not the entire dataset (e.g., they were performed after the data splits or cross-validation splits).

Vandewiele et al. (2020) show how oversampling before partitioning the training data and test data can cause errors in models, with several studies incorrectly reporting near-perfect accuracy.

6b. Dependencies or duplicates between training and test sets

In some cases, samples in the dataset might have dependencies. For example, a clinical dataset might have many samples from the same patient. In such cases, the train-test split or cross-validation (CV) split should take these dependencies into account—for instance, by including all samples from each patient in the same CV fold or train-test split.

Similarly, 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.

Report if the dataset used has dependencies or duplicates. If it does, detail how these are addressed (for example, by using block CV or removing duplicate rows of data).

Malik (2020) outlines alternatives for CV that helps reduce dependencies.

Bergmeir & Benítez (2012) find that blocked CV for time series evaluation deals with temporal autocorrelation.

Hammerla and Plotz (2015) demonstrate how neighborhood bias can affect data recordings close in time and introduce “meta-segmented CV” to deal with such dependencies.

Roberts et al. (2016) describe block CV strategies for a number of structures with dependencies, including temporal, spatial, and hierarchical dependencies.

6c. Feature legitimacy

Leakage can result from any of the features used in a model being a proxy for the outcome. For example, Filho et al. (2021) found that a prominent paper on hypertension prediction (Ye et al., 2018) suffered from data leakage due to illegitimate features. The model included the use of anti-hypertensive drugs as a feature in a clinical model used to predict hypertension.

Justify why each of the features used in the model is legitimate for the task at hand. Note that you do not necessarily need to list each feature individually; instead, you can provide arguments for a set of features together in case the same argument applies to all of them.

Module 7: Metrics and uncertainty

The performance of ML models is key to the scientific claims of interest. Since there are many possible choices that authors can make when choosing performance metrics, it is important to reason about why the metrics used are appropriate for the task at hand. Additionally, communicating and reasoning about uncertainty is important to discourage readers from ignoring the uncertainty in the final results.

7a. Performance metrics used

Several metrics are often used to assess how well an ML model performs and to compare the performance of different ML models. In some cases, these metrics are reported as part of a paper's final results, while in others, they are used to make intermediate decisions such as identifying which models to include in the study or to decide which hyperparameters should be used.

Report all metrics used to assess and compare model performance (e.g., Accuracy, AUC-ROC etc.). Include metrics that are used to make decisions about which model(s) are reported as well as the metrics used to evaluate the reported model(s).

Some metrics are unsuitable for certain problems. For example, accuracy might not be suitable to measure the performance of an ML model in the presence of heavy class imbalance (see Leist et al. (2022), Table 4). Justify the choice of metric(s) used for the scientific claim being made based on the ML model.

7b. Uncertainty estimates

For each performance metric reported in a paper, report an estimate of uncertainty such as standard deviations or confidence intervals. This could be part of graphs or tables in the paper.

Note that applying a bootstrap on the validation set is one way to get uncertainty estimates for a population mean based on a sample from that population.

Report the uncertainty estimate. Also report how the uncertainty estimate is calculated and justify why the method used for uncertainty estimation is valid.

Simmonds et al. (2022) outline the different sources of uncertainty that should be quantified in a study.

Raschka (2018) walks through bootstrapping to obtain an uncertainty estimate.

7c. Appropriate statistical tests

Statistical tests used for comparing model performance come with several assumptions.

Report the type of statistical test used in the paper (if any) for comparing model performance. Report the assumptions of the statistical test and justify why these assumptions are satisfied.

If using bootstrapped confidence intervals for performance metrics, one statistical test is to see if the interval contains a baseline value. Raschka (2018) outlines various statistical tests for comparing supervised learning algorithms. Note that reliance on statistical significance testing has led to misinterpretations and false conclusions (Amrhein, 2019).

Module 8: Generalizability and limitations

8a. Evidence of external validity

External validity (or "generalizability") refers to the applicability of a scientific claim beyond the specific dataset based on which it is made. This includes 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 (Egami and Hartman, 2021). For example, evaluating an ML model on a different dataset or a new clinical setting that it was not trained on is a test of its external validity.

Researchers can use a mix of quantitative and theoretical approaches to make arguments regarding their findings' ability to generalize to other populations, outcomes, and contexts. 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 (Simons et al., 2017).

Report evidence regarding the external validity of the study's findings.

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

Explicit boundaries around the applicability of a scientific claim can help clarify which settings we should expect the scientific claims to hold in. Authors are in the best position to understand limits to the applicability of their claims.

Report examples of settings or domains where the scientific claims made in the study do not hold.

Raji et al. (2022) discuss issues with ML models used in real-world settings. These issues stem in part from a lack of focus on identifying when models are not expected to work.

Guidelines references

1. Valentin Amrhein, Sander Greenland, and Blake McShane. 2019. Scientists Rise Up Against Statistical Significance. *Nature* 567, 7748 (March 2019), 305–307. DOI:https://doi.org/10.1038/d41586-019-00857-9
2. Christoph Bergmeir and José M. Benítez. 2012. On the Use of Cross-Validation for Time Series Predictor Evaluation. *Information Sciences* 191, (May 2012), 192–213. DOI:https://doi.org/10.1016/j.ins.2011.12.028
3. Gavin C. Cawley and Nicola L. C. Talbot. 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research* 11, 70 (2010), 2079–2107. Retrieved March 16, 2023 from http://jmlr.org/papers/v11/cawley10a.html
4. Alexandre Chiavegatto Filho, André Filipe De Moraes Batista, and Hellen Geremias Dos Santos. 2021. 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, 2 (February 2021), e10969. DOI:https://doi.org/10.2196/10969
5. Kasia S. Chmielinski, Sarah Newman, Matt Taylor, Josh Joseph, Kemi Thomas, Jessica Yurkofsky, and Yue Chelsea Qiu. 2022. The Dataset Nutrition Label (2nd Gen): Leveraging Context to Mitigate Harms in Artificial Intelligence. DOI:https://doi.org/10.48550/arXiv.2201.03954
6. Troy Comi. Using Codeocean for Sharing Reproducible Research | The Princeton Research Software Engineering Group Blog. Retrieved March 16, 2023 from https://rse.princeton.edu/2021/03/using-codeocean-for-sharing-reproducible-research/
7. Naoki Egami and Erin Hartman. Elements of external validity: Framework, design, and analysis. *American Political Science Review, 117*(3):1070–1088, October 2022. Retrieved July 31, 2023 from https://www.cambridge.org/core/journals/american-political-science-review/article/elements-of-external-validity-framework-design-and-analysis/2D0914404C84B3F169732FF1D5E39420
8. Samuel G. Finlayson, Adarsh Subbaswamy, Karandeep Singh, John Bowers, Annabel Kupke, Jonathan Zittrain, Isaac S. Kohane, and Suchi Saria. 2021. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med* 385, 3 (July 2021), 283–286. DOI:https://doi.org/10.1056/NEJMc2104626

9. Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (December 2021), 86–92. DOI:https://doi.org/10.1145/3458723
10. Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge & Felix A. Wichmann. 2020. Shortcut learning in deep neural networks. In *Nature Machine Intelligence* 2, 665–673 (2020). DOI: https://doi.org/10.1038/s42256-020-00257-z
11. Nils Y. Hammerla and Thomas Plötz. 2015. 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*, ACM, Osaka Japan, 1041–1051. DOI:https://doi.org/10.1145/2750858.2807551
12. Harbert. 2018. Bash Scripting. Retrieved March 16, 2023 from https://rsh249.github.io/bioinformatics/bash_script.html
13. Jake M. Hofman, Amit Sharma, and Duncan J. Watts. 2017. Prediction and Explanation in Social Systems. *Science* 355, 6324 (February 2017), 486–488. DOI:https://doi.org/10.1126/science.aal3856
14. Jessica Hullman, Sayash Kapoor, Priyanka Nanayakkara, Andrew Gelman, and Arvind Narayanan. 2022. 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*, ACM, Oxford United Kingdom, 335–348. DOI:https://doi.org/10.1145/3514094.3534196
15. Sayash Kapoor and Arvind Narayanan. Model Info Sheets for Addressing Leakage. *Leakage and the Reproducibility Crisis in ML-based Science*. Retrieved from https://reproducible.cs.princeton.edu/#model-info-sheets
16. Anja K. Leist, Matthias Klee, Jung Hyun Kim, David H. Rehkopf, Stéphane P. A. Bordas, Graciela Muniz-Terrera, and Sara Wade. 2022. Mapping of Machine Learning Approaches for Description, Prediction, and Causal Inference in the Social and Health Sciences. *Sci. Adv.* 8, 42 (October 2022), eabk1942. DOI:https://doi.org/10.1126/sciadv.abk1942
17. Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. 2021. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1, (December 2021). Retrieved March 16, 2023 from

https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/757b505cfd34c64c85ca5b5690ee5293-Abstract-round2.html

18. Jimmy Lin. 2019. The Neural Hype and Comparisons Against Weak Baselines. *SIGIR Forum* 52, 2 (January 2019), 40–51. DOI:https://doi.org/10.1145/3308774.3308781
19. David M. Liu and Matthew J. Salganik. 2019. Successes and Struggles with Computational Reproducibility: Lessons from the Fragile Families Challenge. *Socius* 5, (January 2019), 237802311984980. DOI:https://doi.org/10.1177/2378023119849803
20. Michael A. Lones. 2023. How To Avoid Machine Learning Pitfalls: A Guide for Academic Researchers. Retrieved March 16, 2023 from http://arxiv.org/abs/2108.02497
21. Momin M. Malik. 2020. A Hierarchy of Limitations in Machine Learning. Retrieved March 16, 2023 from http://arxiv.org/abs/2002.05193
22. Marcel Neunhoeffer and Sebastian Sternberg. 2019. How Cross-Validation Can Go Wrong and What to Do About It. *Polit. Anal.* 27, 1 (January 2019), 101–106. DOI:https://doi.org/10.1017/pan.2018.39
23. Beata Nowok, Gillian M. Raab, and Chris Dibben. 2016. synthpop : Bespoke Creation of Synthetic Data in R. *J. Stat. Soft.* 74, 11 (2016). DOI:https://doi.org/10.18637/jss.v074.i11
24. Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm Used To Manage the Health of Populations. *Science* 366, 6464 (October 2019), 447–453. DOI:https://doi.org/10.1126/science.aax2342
25. Betsy Levy Paluck, Seth Ariel Green, and Donald P. Green. 2021. The Contact Hypothesis Re-Evaluated: Code and Data. DOI:https://doi.org/10.24433/CO.4024382.V7
26. Elizabeth Levy Paluck, Seth A. Green, and Donald P. Green. 2019. The Contact Hypothesis Re-Evaluated. *Behav. Public Policy* 3, 02 (November 2019), 129–158. DOI:https://doi.org/10.1017/bpp.2018.25
27. Kenneth Peng, Arunesh Mathur, and Arvind Narayanan. 2021. Mitigating Dataset Harms Requires Stewardship: Lessons From 1000 Papers. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks* 1, (December 2021). Retrieved March 16, 2023 from https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/077e29b11be80ab57e1a2ecabb7da330-Abstract-round2.html

28. Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The Fallacy of AI Functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, ACM, Seoul Republic of Korea, 959–972. DOI:https://doi.org/10.1145/3531146.3533158
29. Sebastian Raschka. 2020. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. Retrieved March 16, 2023 from http://arxiv.org/abs/1811.12808
30. David R. Roberts, Volker Bahn, Simone Ciuti, Mark S. Boyce, Jane Elith, Gurutzeta Guillera-Arroita, Severin Hauenstein, José J. Lahoz-Monfort, Boris Schröder, Wilfried Thuiller, David I. Warton, Brendan A. Wintle, Florian Hartig, and Carsten F. Dormann. 2017. Cross-Validation Strategies for Data With Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* 40, 8 (August 2017), 913–929. DOI:https://doi.org/10.1111/ecog.02881
31. Geir Kjetil Sandve, Anton Nekrutenko, James Taylor, and Eivind Hovig. 2013. Ten Simple Rules for Reproducible Computational Research. *PLoS Comput Biol* 9, 10 (October 2013), e1003285. DOI:https://doi.org/10.1371/journal.pcbi.1003285
32. D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. 2018. Winner's Curse? On Pace, Progress, and Empirical Rigor. (June 2018). Retrieved March 16, 2023 from https://openreview.net/forum?id=rJWF0Fywf
33. Tolou Shadbahr, Michael Roberts, Jan Stanczuk, Julian Gilbey, Philip Teare, Sören Dittmer, Matthew Thorpe, Ramon Vinas Torne, Evis Sala, Pietro Lio, Mishal Patel, AIX-COVNET Collaboration, James H. F. Rudd, Tuomas Mirtti, Antti Rannikko, John A. D. Aston, Jing Tang, and Carola-Bibiane Schönlieb. 2022. Classification of Datasets With Imputed Missing Values: Does Imputation Quality Matter? (2022). DOI:https://doi.org/10.48550/ARXIV.2206.08478
34. Emily G. Simmonds, Kwaku Peprah Adjei, Christoffer Wold Andersen, Janne Cathrin Hetle Aspheim, Claudia Battistin, Nicola Bulso, Hannah Christensen, Benjamin Cretois, Ryan Cubero, Ivan A. Davidovich, Lisa Dickel, Benjamin Dunn, Etienne Dunn-Sigouin, Karin Dyrstad, Sigurd Einum, Donata Giglio, Haakon Gjerlow, Amelie Godefroidt, Ricardo Gonzalez-Gil, Soledad Gonzalo Cogno, Fabian Grosse, Paul Halloran, Mari F. Jensen, John James Kennedy, Peter Egge Langsaether, Jack H. Laverick, Debora Lederberger, Camille Li, Elizabeth Mandeville, Caitlin Mandeville, Espen Moe, Tobias Navarro Schroder, David Nunan, Jorge Sicacha Parada, Melanie Rae Simpson, Emma Sofie Skarstein, Clemens Spensberger, Richard Stevens, Aneesh Subramanian, Lea Svendsen, Ole Magnus Theisen, Connor Watret, and Robert B. OHara. 2022. How Is

Model-Related Uncertainty Quantified and Reported in Different Disciplines? (2022). DOI:https://doi.org/10.48550/ARXIV.2206.12179

35. Daniel J. Simons, Yuichi Shoda, and D. Stephen Lindsay. Constraints on generality (COG): A proposed addition to all empirical papers. Perspectives on Psychological Science, 12(6):1123–1128, August 2017. https://doi.org/10.1177/1745691617708630
36. Matias Singers. Awesome README. *GitHub*. Retrieved from https://github.com/matiassingers/awesome-readme
37. Victoria Stodden and Sheila Miguez. 2014. Best Practices for Computational Science: Software Infrastructure and Environments for Reproducible and Extensible Research. 2, 1 (July 2014), e21. DOI:https://doi.org/10.5334/jors.ay
38. Hamed Taherdoost. 2016. Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. DOI:https://doi.org/10.2139/ssrn.3205035
39. Jan P Vandenbroucke, Erik Von Elm, Douglas G Altman, Peter C Gøtzsche, Cynthia D Mulrow, Stuart J Pocock, Charles Poole, James J Schlesselman, Matthias Egger, and for the STROBE Initiative. 2007. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLoS Med* 4, 10 (October 2007), e297. DOI:https://doi.org/10.1371/journal.pmed.0040297
40. Gilles Vandewiele, Isabelle Dehaene, György Kovács, Lucas Sterckx, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Johan Decruyenaere, Sofie Van Hoecke, and Thomas Demeester. 2021. Overly Optimistic Prediction Results on Imbalanced Data: A Case Study of Flaws and Benefits When Applying Over-Sampling. *Artificial Intelligence in Medicine* 111, (January 2021), 101987. DOI:https://doi.org/10.1016/j.artmed.2020.101987
41. Aki Vehtari. Cross-validation FAQ. *Model selection tutorials and talks*. Retrieved from https://avehtari.github.io/modelselection/CV-FAQ.html
42. Vilhuber, Lars, Connolly, Marie, Koren, Miklós, Llull, Joan, and Morrow, Peter. 2020. A Template README for Social Science Replication Packages. (December 2020). DOI:https://doi.org/10.5281/ZENODO.4319999
43. Chengyin Ye, Tianyun Fu, Shiying Hao, Yan Zhang, Oliver Wang, Bo Jin, Minjie Xia, Modi Liu, Xin Zhou, Qian Wu, Yanting Guo, Chunqing Zhu, Yu-Ming Li, Devore S Culver, Shaun T Alfreds, Frank Stearns, Karl G Sylvester, Eric Widen, Doff McElhinney, and Xuefeng Ling. 2018. Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning. *J Med Internet Res* 20, 1 (January 2018), e22. DOI:https://doi.org/10.2196/jmir.9268

44. 2017. Nature Research | Code and Software Submission Checklist. Retrieved March 16, 2023 from https://www.nature.com/documents/nr-software-policy.pdf
45. 2023. About Brain Imaging Data Structure. *Brain Imaging Data Structure*. Retrieved March 16, 2023 from https://bids.neuroimaging.io/index
46. 3.1. Cross-Validation: Evaluating Estimator Performance. *scikit-learn*. Retrieved March 16, 2023 from https://scikit-learn/stable/modules/cross_validation.html
47. 3.2. Tuning the Hyper-Parameters of an Estimator. *scikit-learn*. Retrieved March 16, 2023 from https://scikit-learn/stable/modules/grid_search.html
48. Data Dictionaries | U.S. Geological Survey. *USGS*. Retrieved from https://www.usgs.gov/data-management/data-dictionaries
49. How To Write (and Set) a Run Script | Help | Code Ocean. Retrieved March 16, 2023 from https://help.codeocean.com/en/articles/2465281-how-to-write-and-set-a-run-script
50. Requirements File Format - pip Documentation v23.0.1. Retrieved March 16, 2023 from https://pip.pypa.io/en/stable/reference/requirements-file-format/
51. Unofficial Guidance on Various Topics by Social Science Data Editors. *Data and Code Guidance by Data Editors*. Retrieved March 16, 2023 from https://social-science-data-editors.github.io/guidance/

REFORMS: Reporting Standards for Machine Learning Based Science

Appendix C: Table of References on Reporting Quality & Problems in Scientific Literature

This appendix provides additional details on some of the citations from the main text. We include references from the main text that address: (1) the quality of reporting in past scientific literature, or (2) examples of problems that occurred in past scientific literature. This appendix does not constitute a comprehensive list of all published references on these topics. The table has 44 entries with details about their relevance to our review.

The citations are listed in order of appearance in the main text, with section headings corresponding to the headings from the text. Some sections from the main text are omitted because they do not contain references that match our criteria for inclusion in the table. Some citations are included in the table more than once because they appear in multiple sections. Many of the references focus specifically on machine learning (ML)-based science, but we also include references about science with traditional statistical methods because some of the best practices and shortcomings are shared in ML-based science and other quantitative sciences.

Reference	**Findings about reporting quality in past literature or problems in past literature**	**Discipline**	**Literature examined**	**ML-Focused?**
MODULE 1: STUDY GOALS				
Introduction				
Hofman et al., 2017, "Prediction and explanation in social systems" [1]	The authors re-evaluate data from a prior paper to demonstrate how different (but equally reasonable) choices in research design can lead to different results from the same data. This includes an example of how slight differences in the definition of a research question can lead to substantially different results.	Computational social science	Re-evaluation of data from 1 prior paper on prediction of information cascade size on Twitter	Yes
1a) Population or distribution about which the scientific claim is made				
Lundberg et al., 2021, "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory" [2]	Only 9 out of 32 papers papers (28%) provided sufficient information for a reader to "confidently" identify the target population about which the scientific claim is made (p. 553).	Sociology	32 quantitative papers in 2018 volume of a top sociology journal	No
Tooth et al., 2005, "Quality of Reporting of Observational Longitudinal Research" [3]	33 out of 49 papers (67%) define a target population.	Epidemiology & medicine	49 longitudinal studies on strokes in six journals, 1999-2003	No
MODULE 2: COMPUTATIONAL REPRODUCIBILITY				
Introduction				
Verstynen and Kording, 2023, "Overfitting to 'predict' suicidal ideation" [4]	The code for the feature selection step in a flawed prior paper was not released, so Verstynen and Kording could not pinpoint the exact source of errors.	Psychology, neuroscience, and biomedical engineering	1 paper on prediction of suicidal ideation	Yes
Current computational reproducibility standards fall short				

Stodden et al., 2018, “An empirical analysis of journal policy effectiveness for computational reproducibility” [5]	Stodden et al. attempted to contact the authors of 204 papers published in the journal Science to obtain reproducibility materials. Only 44% of authors responded.	Multi-disciplinary	204 quantitative papers in Science	No
Gabelica et al., 2022, “Many researchers were not compliant with their published data sharing statement: A mixed-methods study” [6]	Gabelica et al. examined 333 open-access journals indexed on BioMed Central in January 2019 and found that out of the 1,792 papers that pledged to share data upon request, 1,669 did not do so, resulting in a 93% data unavailability rate.	Biology, health sciences and medicine	1,792 papers published in 333 BioMed Central open-access journals in January 2019	No
Vasilevsky et al., 2017, “Reproducible and reusable research: Are journal data sharing policies meeting the mark?” [7]	Vasilevsky et al. examined the data-sharing policies of 318 biomedical journals and discovered that almost one-third lacked any such policies, and those that did often lacked clear guidelines for author compliance.	Biology, health sciences and medicine	318 biomedical journals (Biochemistry and Molecular Biology, Biology, Cell Biology, Crystallography, Developmental Biology, Biomedical Engineering, Immunology, Medical Informatics, Microbiology, Microscopy, Multidisciplinary Sciences, and Neurosciences)	No
Computational reproducibility allows independent researchers to find errors in original papers				
Hofman et al., 2021, “Expanding the scope of reproducibility research through data analysis replications” [8]	Hofman et al. analyze 11 papers and find various shortcomings in this body of literature.	Multi-disciplinary	11 computational social science papers	No
Vandewiele et al., 2021, “Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling” [9]	Vandewiele et al. analyze 24 papers on pre-term birth prediction and find 21 of these papers suffer from leakage.	Medicine	24 papers on pre-term risk prediction	Yes
MODULE 3: DATA QUALITY				
3a) Data source(s)				

Navarro et al., 2022, "Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review" [10]	98% of articles adhered to the guidelines for reporting data source from the TRIPOD statement.	Epidemiology & medicine	152 articles on diagnostic or prognostic prediction models across medical fields, published 2018-2019	Yes
Yusuf et al., 2020, "Reporting quality of studies using machine learning models for medical diagnosis: a systematic review" [11]	24 out of 28 papers (86%) reported information about their data source, defined as "Where and when potentially eligible participants were identified (setting, location and dates)" (p. 3).	Medicine	28 "medical research studies that used ML methods to aid clinical diagnosis," published July 2015-July 2018	Yes
Kim et al., 2016, "Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection" [12]	Studies that utilize social media data frequently omit important information about their data collection process, such as details about the development and assessment of search filters. This paper provides a framework for reporting this information.	Health media	Studies that use social media data (this is not a formal review paper, but it provides several examples)	No
Geiger et al., 2020, "Garbage In, Garbage Out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?" [13]	There was "wide divergence" in whether papers followed best practices for reporting the data annotation process, such as reporting: "who the labelers were, what their qualifications were, whether they independently labeled the same items, whether inter-rater reliability metrics were disclosed, what level of training and/or instructions were given to labelers, whether compensation for crowdworkers is disclosed, and if the training data is publicly available" (p. 325).	Multi-disciplinary: "the papers represented political science, public health, NLP, sentiment analysis, cybersecurity, content moderation, hate speech, information quality, demographic profiling, and more" (p. 328)	164 "machine learning application papers... that classified tweets from Twitter" (p. 326)	Yes
3b) Sampling frame				

Navarro et al., 2022, "Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review" [10]	105 out of 152 studies (69%) reported their eligibility criteria.	Epidemiology & medicine	152 articles on diagnostic or prognostic prediction models across medical fields, published 2018-2019	Yes
Tooth et al., 2005, "Quality of Reporting of Observational Longitudinal Research" [3]	41 out of 49 papers (84%) reported their sampling frame, and 32 out of 49 papers (65%) reported their eligibility criteria.	Epidemiology & medicine	49 longitudinal studies on strokes in six journals, 1999-2003	No
Porzsolt et al., 2019, "Inclusion and exclusion criteria and the problem of describing homogeneity of study populations in clinical trials" [14]	75 out of 100 studies (75%) reported inclusion criteria. 6 of those 75 studies (8%) also reported exclusion criteria.	Medicine	100 publications on "quality of life" assessments	No
3d) Outcome variable				
Credé and Harms, 2021, "Three cheers for descriptive statistics—and five more reasons why they matter" [15]	In a review of literature that was still a work-in-progress at the time Credé and Harms published this commentary, "Among the articles coded to date, less than half report the ethnicity of the participants or the types of jobs held by the participants and only 56% report data on the industry in which the data were collected. Other interesting—and to meta-analysts potentially important—information is also remarkably often unreported" (p. 486). (Note: This commentary discusses descriptive statistics broadly, not just descriptive statistics for outcome variables.)	Industrial and organizational psychology	Articles from four top journals in industrial and organizational psychology (number of articles is not reported)	No
Larson-Hall and Plonsky, 2015, "Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field" [16]	Meta-analyses frequently had to omit large numbers of primary articles from their analyses due to insufficient descriptive statistics in the primary articles. (Note: This article discusses descriptive statistics broadly, not just descriptive statistics for outcome variables.)	Second language acquisition	Approximately 90 meta-analyses in second language acquisition	No
3e) Sample size				
Plonsky, 2013, "Study Quality in SLA: An Assessment of Designs, Analyses, and Reporting Practices in Quantitative L2 Research" [17]	99% of studies reported sample size.	Second language acquisition	606 studies in second language acquisition journals, published 1990-2010	No
Tooth et al., 2005, "Quality of Reporting of Observational Longitudinal Research" [3]	100% of 49 longitudinal studies reported the total number of participants from the first wave of their study. However, only 25 out of 49 (51%) reported the number of participants after attrition at each subsequent wave.	Epidemiology & medicine	49 longitudinal studies on strokes in six journals, 1999-2003	No
3f) Missingness				

McKnight et al., 2007, "Missing Data: A Gentle Introduction" [18]	Around 90% of articles had missing data, and the average amount of missing data per study was over 30%. Furthermore, "few of the articles included explicit mention of missing data, and even fewer indicated that the authors attended to missing data, either by performing statistical procedures or by making disclaimers regarding the studies in the results and conclusions" (p. 3).	Psychology	Over 300 publications from a prominent psychology journal	No
Peugh and Enders, 2004, "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement" [19]	Among the articles Peugh and Enders reviewed, "[d]etails concerning missing data were seldom reported" and "[t]he methods used to handle missing data were, in many cases, difficult to ascertain because explicit descriptions of missing-data procedures were rare" (p. 537). However, Peugh and Enders were able to infer the amount of missingness in some studies by examining the "discrepancy between the reported degrees of freedom for a given analysis and the degrees of freedom that one would expect on the basis of the stated sample size and design characteristics" (p. 537). In articles published in 1999, they detected missing data in 16% of studies, but they write that this is likely a "gross underestimate" of the actual prevalence of missing data. Among articles published in 2003, they were able to detect missing data in 42% of articles, which is higher than in 1999 due to changes in reporting practices following a recommendation by an American Psychological Association task force.	Educational research	989 studies published in 1999 and 545 studies published in 2003 in 23 applied educational research journals	No
Salganik et al., 2020, Supplementary information for "Measuring the predictability of life outcomes using a scientific mass collaboration" [20]	There are many reasons for missing data in survey data, including a respondent not participating in a given wave of a longitudinal survey, respondents refusing to answer some questions, skip patterns in the survey design, and redaction for privacy. In a modified version of a well-known, high-quality social survey dataset, 73% of possible data entries were missing, and the largest source of missingness was survey skip patterns. This high level of missingness emphasizes the importance of careful attention to handling missing data.	Sociology	1 study with a well-known social survey data set	Yes
Nijman et al., 2022, "Missing data is poorly handled and reported in prediction model studies using machine learning: a literature review" [21]	"A total of 56 (37%) prediction model studies did not report on missing data and could not be analyzed further. We included 96 (63%) studies which reported on the handling of missing data. Across the 96 studies, 46 (48%) did not include information on the amount or nature of the missing data" (p. 220).	Medicine	152 ML-based clinical prediction model studies, published 2018-2019	Yes

Navarro et al., 2022, "Completeness of reporting of clinical prediction models developed using supervised machine learning: a systematic review" [10]	"Forty-four studies reported how missing data were handled (28.9%, 95% CI 22.3 to 36.6). The missing data item consists of four sub-items of which three were rarely addressed in included studies. Within 28 studies that reported handling of missing data: three studies reported the software used (10.7%, CI 3.7 to 27.2), four studies reported the variables included in the procedure (14.3%, CI 5.7 to 31.5) and no study reported the number of imputations (0%, CI 0.0 to 39.0)" (pp. 6-7).	Epidemiology & medicine	152 articles on diagnostic or prognostic prediction models across medical fields, published 2018-2019	Yes
Little et al., 2013, "On the Joys of Missing Data" [22]	"Among the 80 reviewed studies, only 45 (56.25%) mentioned missing data explicitly in the text or a table of descriptive statistics. Of those 45, only three mentioned testing whether the missingness was related to other variables, justifying their [missingness at random] assumption" (p. 156).	Pediatric psychology	80 empirical studies in the 2012 issues of a pediatric psychology journal	No
Nicholson et al., 2016, "Attrition in developmental psychology" [23]	Among 541 longitudinal studies, only 253 (47%) discussed missingness due to attrition, and only 99 (18%) explicitly discussed whether missingness due to attrition was "missing at random," "missing completely at random," or "missing not at random."	Developmental psychology	541 longitudinal studies in major developmental journals, published 2009 and 2012	No
Sterner, 2011, "What Is Missing in Counseling Research? Reporting Missing Data" [24]	In the first journal, "14 of 66 (21%) articles referenced missing data on some level. Of these 14 articles, 11 mentioned missing data specifically... In the remaining 52 JCD articles, no information was provided on whether missing data existed." In the second journal, "one of 28 (4%) empirically based research articles made reference to screening for missing data; however, no mention was made of missing data in the remaining articles" (p. 56).	Counseling	94 empirical research articles in two top counseling journals, published 2004 to 2008	No
Tooth et al., 2005, "Quality of Reporting of Observational Longitudinal Research" [3]	Only 19 out of 49 articles (39%) reported on missing data items at each longitudinal wave, and only 2 out of 42 articles (5%) that had missing data in their analyses described imputation, weighting, or sensitivity analyses for handling missing data.	Epidemiology & medicine	49 longitudinal studies on strokes in six journals, 1999-2003	No
Hussain et al., 2017, "Quality of missing data reporting and handling in palliative care trials demonstrates that further development of the CONSORT statement is required: a systematic review" [25]	101 out of 108 studies (94%) reported the number of participants who were missing in the primary outcome analysis; however, reporting rates were lower for other details about missing data and for methods of handling missing data.	Epidemiology	108 articles on palliative care randomized controlled trials, published 2009-2014	No
3g) Dataset for evaluation is representative				

Tooth et al., 2005, “Quality of Reporting of Observational Longitudinal Research” [3]	Among several reporting criteria this review examined, “the criteria in the checklist representing selection bias were the least frequently reported overall” (p. 285). Specifically, selection-in biases were discussed in 14 out of 49 articles (28%), comparison of consenters with non-consenters was discussed in 1 out of 47 applicable articles (2%), and loss to follow-up was accounted for in the analyses of 1/41 applicable articles (5%). Additionally, 37 out of 49 articles (75%) discuss how their results relate to the target population.	Epidemiology & medicine	49 longitudinal studies on strokes in six journals, 1999-2003	No
MODULE 4: DATA PREPROCESSING				
4c) Data transformations				
Vandewiele et al., 2021, “Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling” [9]	Vandewiele et al. analyze 24 papers on pre-term birth prediction and find 11 of these papers improperly transform data (by oversampling before splitting into train and test sets).	Medicine	24 papers on pre-term risk prediction	Yes
MODULE 5: MODELING				
5d) Model selection method				
Neunhoeffer and Sternberg, 2019, “How Cross-Validation Can Go Wrong and What to Do About It.” [26]	Neunhoeffer and Sternberg demonstrate that the main findings of a prominent political science paper fail to reproduce due to improper model selection. In particular, model selection was done on the same data that was used for evaluation.	Political Science	1 prominent political science paper	Yes
5e) Hyper-parameter selection				
Dodge et al., 2019, “Show Your Work: Improved Reporting of Experimental Results” [27]	Dodge et al. find that among 50 random papers from a prominent natural language processing conference, while 74% of papers reported at least some information about the best performing hyperparameters, 10% of fewer reported more specific details about hyperparameter search or the effect of hyperparameters on performance.	Natural language processing	50 random papers from a prominent natural language processing conference in 2018	Yes
5f) Appropriate baselines				
Sculley et al., 2018, “Winner’s curse? On pace, progress, and empirical rigor” [28]	Sculley et al. discuss five papers that provide evidence of improper comparison with baselines in different areas of ML, suggesting that empirical progress in the field can be misleading.	ML	5 papers identifying poor performance compared to baselines in different areas of ML	Yes
MODULE 6: DATA LEAKAGE				
Introduction				

Kapoor and Narayanan, 2022, "Leakage and the reproducibility crisis in ML-based science" [29]	Kapoor and Narayanan found that leakage affects hundreds of papers across 17 fields.	Multi-disciplinary	A survey of leakage issues across 17 fields	Yes
Train-test separation is maintained				
Poldrack et al., 2020, "Establishment of best practices for evidence for prediction: A review" [30]	Poldrack et al. 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.	Neuropsychiatry	100 published studies between December 24, 2017 and October 30, 2018 in PubMed using search terms "fMRI prediction" and "fMRI predict"	Yes
Dependencies or duplicates between datasets				
Roberts et al., 2021, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans" [31]	Roberts et al. discuss the issue of "Frankenstein" datasets: datasets that combine multiple other sources of data and can end up using the same data twice—for instance, if two datasets rely on the same underlying data source are combined into a larger dataset.	Medicine	62 studies that claimed to diagnose or prognose Covid-19 using chest x-rays	Yes
MODULE 7: METRICS AND UNCERTAINTY				
7b) Uncertainty estimates				
Simmonds et al., 2022, "How is model-related uncertainty quantified and reported in different disciplines?" [32]	Simmonds et al. show that across seven fields, no fields consistently reported complete model uncertainties, and that the type of uncertainties reported varied by field.	Multi-disciplinary	496 studies across 7 fields that included statistical models	No
MODULE 8: GENERALIZABILITY AND LIMITATIONS				
Introduction				
Raji et al., 2022, "The Fallacy of AI Functionality" [33]	Raji et al. review real-world applications of technologies that claim to use ML and cateogorize several ways in which such technology frequently failed, including "lack of robustness to changing external conditions" (p. 9).	Computer science and law (real-world ML applications)	283 cases of failures of technology that claimed to be AI, ML or data-driven between 2012 to 2021	Yes

Liao et al., 2021, “Are We Learning Yet? A Meta-Review of Evaluation Failures Across Machine Learning” [34]	Liao et al. find that the same types of evaluation failures occur across a wide range of ML tasks and algorithms. They provide a taxonomy of common internal and external validity failures.	Computer science	107 “survey papers from computer vision, natural language processing, recommender systems, reinforcement learning, graph processing, metric learning, and more”	Yes
Reporting on external validity falls short in past literature				
Tooth et al., 2005, “Quality of Reporting of Observational Longitudinal Research” [3]	37 out of 49 papers (75%) discuss how the findings from their sample generalize to their target population, and 26 out of 49 papers (53%) discuss generalizability beyond the target population.	Epidemiology & medicine	49 longitudinal studies on strokes in six journals, 1999-2003	No
Bozkurt et al., 2020, “Reporting of demographic data and representativeness in machine learning models using electronic health records” [35]	The authors argue that descriptive statistics about the study sample should be provided in order to be transparent about representativeness of the target population. They find that of 164 studies that trained ML models with electronic health records data, “Race/ethnicity was not reported in 64%; gender and age were not reported in 24% and 21% of studies, respectively. Socioeconomic status of the population was not reported in 92% of studies.” They also find, “Few models (12%) were validated using external populations” (p. 1878).	Medicine	164 studies that trained ML models with electronic health records data	Yes
Navarro et al., 2023, “Systematic review finds ‘spin’ practices and poor reporting standards in studies on machine learning-based prediction models” [36]	“In the main text, 86/152 (56.6% [95% CI 48.6 - 64.2]) studies made recommendations to use the model in clinical practice, however, 74/86 (86% [95% CI 77.2 - 91.8]) lacked external validation in the same article. Out of the 13/152 (8.6% [95% CI 5.1 - 14.1]) studies that recommended the use of the model in a different setting or population, 11/ 13 (84.6% [95% CI 57.8 - 95.7]) studies lacked external validation” (p. 104).	Epidemiology & medicine	152 articles on diagnostic or prognostic prediction models across medical fields, published 2018-2019	Yes

References

[1] Jake M. Hofman, Amit Sharma, and Duncan J. Watts. Prediction and explanation in social systems. *Science*, 355(6324):486–488, February 2017.

[2] Ian Lundberg, Rebecca Johnson, and Brandon M. Stewart. What is your estimand? Defining the target quantity connects statistical evidence to theory. *American Sociological Review*, 86(3):532–565, June 2021.

[3] Leigh Tooth, Robert Ware, Chris Bain, David M. Purdie, and Annette Dobson. Quality of reporting of observational longitudinal research. *American Journal of Epidemiology*, 161(3):280–288, February 2005.

[4] Timothy Verstynen and Konrad Paul Kording. Overfitting to ‘predict’ suicidal ideation. *Nature Human Behaviour*, pages 1–2, April 2023.

[5] Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academy of Sciences*, 115(11):2584–2589, March 2018.

[6] Mirko Gabelica, Ružica Bojčić, and Livia Puljak. Many researchers were not compliant with their published data sharing statement: A mixed-methods study. *Journal of Clinical Epidemiology*, 150:33–41, October 2022.

[7] Nicole A. Vasilevsky, Jessica Minnier, Melissa A. Haendel, and Robin E. Champieux. Reproducible and reusable research: Are journal data sharing policies meeting the mark? *PeerJ*, 5:e3208, April 2017.

[8] Jake M. Hofman, Daniel G. Goldstein, Siddhartha Sen, Forough Poursabzi-Sangdeh, Jennifer Allen, Ling Liang Dong, Brenda Fried, Harpreet Gaur, Adnan Hoq, Emeka Mbazor, Naomi Moreira, Cindy Muso, Etta Rapp, and Roymil Terrero. Expanding the scope of reproducibility research through data analysis replications. *Organizational Behavior and Human Decision Processes*, 164:192–202, May 2021.

[9] Gilles Vandewiele, Isabelle Dehaene, György Kovács, Lucas Sterckx, Olivier Janssens, Femke Ongenae, Femke De Backere, Filip De Turck, Kristien Roelens, Johan Decruyenaere, Sofie Van Hoecke, and Thomas Demeester. Overly optimistic prediction results on imbalanced data: A case study of flaws and benefits when applying over-sampling. *Artificial Intelligence in Medicine*, 111:101987, January 2021.

[10] Constanza L. Andaur Navarro, Johanna A. A. Damen, Toshihiko Takada, Steven W. J. Nijman, Paula Dhiman, Jie Ma, Gary S. Collins, Ram Bajpai, Richard D. Riley, Karel G. M. Moons, and Lotty Hooft. Completeness of reporting of clinical prediction models developed using supervised machine learning: A systematic review. *BMC Medical Research Methodology*, 22(1), January 2022.

[11] Mohamed Yusuf, Ignacio Atal, Jacques Li, Philip Smith, Philippe Ravaud, Martin Fergie, Michael Callaghan, and James Selfe. Reporting quality of studies using machine learning models for medical diagnosis: A systematic review. *BMJ Open*, 10(3), 2020.

[12] Yoonsang Kim, Jidong Huang, and Sherry 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. *Journal of Medical Internet Research*, 18(2):e41, February 2016.

[13] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. Garbage in, garbage out? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, January 2020.

[14] Franz Porzsolt, Felicitas Wiedemann, Susanne I. Becker, and C. J. Rhoads. Inclusion and exclusion criteria and the problem of describing homogeneity of study populations in clinical trials. *BMJ Evidence - Based Medicine*, 24(3):92, 06 2019.

[15] Marcus Credé and P. D. Harms. Three cheers for descriptive statistics—and five more reasons why they matter. *Industrial and Organizational Psychology*, 14(4):486–488, 2021.

[16] Jenifer Larson-Hall and Luke Plonsky. Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, 65(S1):127–159, May 2015.

[17] Luke Plonsky. Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, 35(4):655–687, 2013.

[18] Patrick McKnight, Katherine McKnight, Souraya Sidani, and Aurelio José Figueredo. *Missing Data : A Gentle Introduction*. Guilford Publications, 2007.

[19] James L. Peugh and Craig K. Enders. Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, 74(4):525–556, 2004.

[20] Matthew J. Salganik, Ian Lundberg, Alexander T. Kindel, Caitlin E. Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M. Altschul, Jennie E. Brand, Nicole Bohme Carnegie, Ryan James Compton, Debanjan Datta, Thomas Davidson, Anna Filippova, Connor Gilroy, Brian J. Goode, Eaman Jahani, Ridhi Kashyap, Antje Kirchner, Stephen McKay, Allison C. Morgan, Alex Pentland, Kivan Polimis, Louis Raes, Daniel E. Rigobon, Claudia V. Roberts, Diana M. Stanescu, Yoshihiko Suhara, Adaner Usmani, Erik H. Wang, Muna Adem, Abdulla Alhajri, Bedoor AlShebli, Redwane Amin, Ryan B. Amos, Lisa P. Argyle, Livia Baer-Bositis, Moritz Büchi, Bo-Ryehn Chung, William Eggert, Gregory Faletto, Zhilin Fan, Jeremy Freese, Tejomay Gadgil, Josh Gagné, Yue Gao, Andrew Halpern-Manners, Sonia P. Hashim, Sonia Hausen, Guanhua He, Kimberly Higuera, Bernie Hogan, Ilana M. Horwitz, Lisa M. Hummel, Naman Jain, Kun Jin, David Jurgens, Patrick Kaminski, Areg Karapetyan, E. H. Kim, Ben Leizman, Naijia Liu, Malte Möser, Andrew E. Mack, Mayank Mahajan, Noah Mandell, Helge Marahrens, Diana Mercado-Garcia, Viola Mocz, Katariina Mueller-Gastell, Ahmed Musse, Qiankun Niu, William Nowak, Hamidreza Omidvar, Andrew Or, Karen Ouyang, Katy M. Pinto, Ethan Porter, Kristin E. Porter, Crystal Qian, Tamkinat Rauf, Anahit Sargsyan, Thomas Schaffner, Landon Schnabel, Bryan Schonfeld, Ben Sender, Jonathan D. Tang, Emma Tsurkov, Austin van Loon, Onur Varol, Xiafei Wang, Zhi Wang, Julia Wang, Flora Wang, Samantha Weissman, Kirstie Whitaker, Maria K. Wolters, Wei Lee Woon, James Wu, Catherine Wu, Kengran Yang, Jingwen Yin, Bingyu Zhao, Chenyun Zhu, Jeanne Brooks-Gunn, Barbara E. Engelhardt, Moritz Hardt, Dean Knox, Karen Levy, Arvind Narayanan, Brandon M. Stewart, Duncan J. Watts, and Sara McLanahan. Supplementary information for: Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, 117(15):8398–8403, March 2020.

[21] SWJ Nijman, AM Leeuwenberg, I Beekers, I Verkouter, JJL Jacobs, ML Bots, FW Asselbergs, KGM Moons, and TPA Debray. Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review. *Journal of Clinical Epidemiology*, 142:218–229, February 2022.

[22] Todd D. Little, Terrence D. Jorgensen, Kyle M. Lang, and E. Whitney G. Moore. On the joys of missing data. *Journal of Pediatric Psychology*, 39(2):151–162, July 2013.

[23] Jody S. Nicholson, Pascal R. Deboeck, and Waylon Howard. Attrition in developmental psychology. *International Journal of Behavioral Development*, 41(1):143–153, July 2016.

[24] William R. Sterner. What is missing in counseling research? reporting missing data. *Journal of Counseling and Development*, 89(1):56–62, January 2011.

[25] Jamilla A. Hussain, Martin Bland, Dean Langan, Miriam J. Johnson, David C. Currow, and Ian 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. *Journal of Clinical Epidemiology*, 88:81–91, August 2017.

[26] Marcel Neunhoeffer and Sebastian Sternberg. How cross-validation can go wrong and what to do about it. *Political Analysis*, 27(1):101–106, January 2019.

[27] Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A Smith. Show your work: Improved reporting of experimental results. *arXiv preprint arXiv:1909.03004*, 2019.

[28] D. Sculley, Jasper Snoek, Alex Wiltschko, and Ali Rahimi. Winner's curse? on pace, progress, and empirical rigor. June 2018.

[29] Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in ML-based science, July 2022. arXiv:2207.07048 [cs, stat].

[30] Russell A. Poldrack, Grace Huckins, and Gael Varoquaux. Establishment of best practices for evidence for prediction: A review. *JAMA psychiatry*, 77(5):534–540, May 2020.

[31] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I. Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, Jonathan R. Weir-McCall, Zhongzhao Teng, Effrossyni Gkrania-Klotsas, James H. F. Rudd, Evis Sala, and Carola-Bibiane Schönlieb. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, March 2021. Number: 3 Publisher: Nature Publishing Group.

[32] Emily G Simmonds, Kwaku Peprah Adjei, Christoffer Wold Andersen, Janne Cathrin Hetle Aspheim, Claudia Battistin, Nicola Bulso, Hannah Christensen, Benjamin Cretois, Ryan Cubero, Ivan A Davidovich, et al. How is model-related uncertainty quantified and reported in different disciplines? *arXiv preprint arXiv:2206.12179*, 2022.

[33] Inioluwa Deborah Raji, I. Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. The fallacy of AI functionality. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 959–972, Seoul Republic of Korea, June 2022. ACM.

[34] Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. Are we learning yet? A meta review of evaluation failures across machine learning. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1, December 2021.

[35] Selen Bozkurt, Eli M Cahan, Martin G Seneviratne, Ran Sun, Juan A Lossio-Ventura, John P A Ioannidis, and Tina Hernandez-Boussard. Reporting of demographic data and representativeness in machine learning models using electronic health records. *Journal of the American Medical Informatics Association*, 27(12):1878–1884, September 2020.

[36] Constanza L. Andaur Navarro, Johanna A.A. Damen, Toshihiko Takada, Steven W.J. Nijman, Paula Dhiman, Jie Ma, Gary S. Collins, Ram Bajpai, Richard D. Riley, Karel G.M. Moons, and Lotty Hooft. Systematic review finds “spin” practices and poor reporting standards in studies on machine learning-based prediction models. *Journal of Clinical Epidemiology*, 158:99–110, June 2023.