Guidelines for reporting ML-based science

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 with answers based on Obermeyer et al. (2019). This sample checklist can be found here.

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.

Section 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 maximize predictive accuracy. “ML methods allow early detection of disease and can therefore be used to increase quality of life” or “ML methods could outperform traditional methods for disease identification.”

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

Section 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.

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, 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.

Section 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, split by class for a categorical outcome variable. Report how the outcome variable is defined.

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.).

The sampling frame could be unrepresentative by being a convenience sample, under-representing minorities, or constituting a too small or too large sample size (Hullman et al., 2022). If it is not representative, note this among threats to external validity (8a.).

Section 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, the height of a human is zero feet or 100 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 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.

Section 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

4. Coefficients where possible (e.g., linear regression or decision trees with a small number of features).

5b. Justification for choice of model

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.

Lones (2022) describes how to compare models against an appropriate baseline.

Section 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.

Section 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 metrics, it is important to reason about why the metrics used are appropriate for the task at hand. 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).

Section 8: Generalizability and limitations

8a. Threats to external validity

External validity refers to the applicability of a scientific claim beyond the specific dataset based on which it is made. 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.

Several threats to the external validity of ML models have been documented in past literature. Finlayson et al. (2021) outline external validity failures due to dataset shifts in clinical settings. Hullman et al. (2022) discuss the threats to validity that arise in different phases of an ML research project. Liao et al. (2021) outline a taxonomy of evaluation failures in ML, including failures in external validity.

Report threats to external validity which apply to the study.

8b. Contexts in which the authors don't 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 claim 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.

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] 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

[8] 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

[9] 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

[10] Harbert. 2018. Bash Scripting. Retrieved March 16, 2023 from https://rsh249.github.io/bioinformatics/bash_script.html

[11] 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

[12] 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

[13] 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

[14] 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

[15] 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

[16] 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

[17] 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

[18] 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

[19] Momin M. Malik. 2020. A Hierarchy of Limitations in Machine Learning. Retrieved March 16, 2023 from http://arxiv.org/abs/2002.05193

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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

[26] 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

[27] Sebastian Raschka. 2020. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. Retrieved March 16, 2023 from http://arxiv.org/abs/1811.12808

[28] 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

[29] 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

[30] 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

[31] 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

[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, 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

[33] Matias Singers. Awesome README. *GitHub*. Retrieved from https://github.com/matiassingers/awesome-readme

[34] 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

[35] Hamed Taherdoost. 2016. Sampling Methods in Research Methodology; How to Choose a Sampling Technique for Research. DOI:https://doi.org/10.2139/ssrn.3205035

[36] 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

[37] 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

[38] Aki Vehtari. Cross-validation FAQ. *Model selection tutorials and talks*. Retrieved from https://avehtari.github.io/modelselection/CV-FAQ.html

[39] 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

[40] 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

[41] 2017. Nature Research | Code and Software Submission Checklist. Retrieved March 16, 2023 from https://www.nature.com/documents/nr-software-policy.pdf

[42] 2023. About Brain Imaging Data Structure. *Brain Imaging Data Structure*. Retrieved March 16, 2023 from https://bids.neuroimaging.io/index

[43] 3.1. Cross-Validation: Evaluating Estimator Performance. *scikit-learn*. Retrieved March 16, 2023 from https://scikit-learn/stable/modules/cross_validation.html

[44] 3.2. Tuning the Hyper-Parameters of an Estimator. *scikit-learn*. Retrieved March 16, 2023 from https://scikit-learn/stable/modules/grid_search.html

[45] Data Dictionaries | U.S. Geological Survey. *USGS*. Retrieved from https://www.usgs.gov/data-management/data-dictionaries

[46] 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

[47] Requirements File Format - pip Documentation v23.0.1. Retrieved March 16, 2023 from https://pip.pypa.io/en/stable/reference/requirements-file-format/

[48] 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/