Sample checklist for reporting ML-based science (beta)

Standard version - Checklist for Obermeyer et al. (2019)

About the checklist

This 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 35 questions spread across 8 sections. 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.

This is a beta version of our checklist. We are soliciting feedback and will continue to update the template. For feedback or questions, contact: sayashk@princeton.edu. The checklist starts on the next page. After filling it out, save it starting from that page.

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

Section 1: Study design

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

Adult patients at a U.S. hospital whose risk for high medical costs is evaluated using a predictive model based on past insurance data.

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

To evaluate racial bias in a typical commercial algorithm used to determine the eligibility of patients for entry into high-risk medical programs.

1c. Motivation for the use of ML-methods in the study, as opposed to traditional statistical methods.

To replicate the algorithm with different output labels and check if choice of label impacts racial bias.

Section 2: Computational reproducibility

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

The data used in this analysis contain private health information, so they cannot be made publicly available. Instead, we provide a synthetic dataset and code based on the real world data, to enable replication. The GitLab repository contains a data dictionary.

URL: https://gitlab.com/labsysmed/dissecting-bias

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.

https://gitlab.com/labsysmed/dissecting-bias

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.

All software environment details and dependencies included in the GitLab repository: https://gitlab.com/labsysmed/dissecting-bias

Hardware infrastructure, operating system details, and time estimate **not reported**.

A **hypothetical** example of reporting these details:

- Intel Core i7-13700E Processor, no external GPU, 16 GB RAM, 256 GB disk space
- Windows 11 Build 22621.1413
- Time estimate to run results: 3 hours 30 minutes

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

Included in the GitLab, with instructions for generating the result.

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

Not included

As an example, Paluck et al. (2018) provide a CodeOcean repository alongside their paper.

Section 3: Data quality

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 of ground-truth annotations, and other data documentation.

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. For our outcome, we study several measures of health constructed from electronic health records, linked to algorithmic predictions. We

include a data dictionary is included in the GitLab repository: https://gitlab.com/labsysmed/dissecting-bias

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

We include data from all patients at a large academic hospital in the U.S. between 2013-2015.

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

Our dataset describes a typical commercial risk-prediction algorithm. It contains the algorithm's predictions, the ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. This allows us to quantify racial disparities in algorithms and isolate the mechanisms by which they arise.

3d. The outcome variable of the model, along with descriptive statistics (split by class for a categorical outcome variable) and its definition.

We develop three algorithms trained in the same way to predict in year t: Total cost, avoidable cost, and health. The sample was 71.2% enrolled in commercial insurance and 28.8% in Medicare; on average, 50.9 years old; and 63% female. Table 1 provides descriptive statistics.

3e. Number of samples in the dataset.

(i) 6079 patients who self-identified as Black and (ii) 43,539 patients who self-identified as White without another race or ethnicity.

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

Not reported.

A **hypothetical** example could be: 5% of our samples have missing data for at least one feature or the outcome variable. We exclude these samples from our analysis.

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.). If it is not representative, note among threats to external validity (8a.)

We use a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today, Optum ImpactPro. It contains both the algorithm's predictions as well as the data needed to understand its inner workings. The authors also focus on the pitfalls of using cost prediction as an outcome, which is something many commercial algorithms use.

Section 4: Data preprocessing

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

Not reported.

A **hypothetical** example could be: "We excluded patients who did not self-identify their race, noting that marginalized patients might have a higher tendency of concealing their race."

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

Not reported.

A **hypothetical** example could be: "We checked for impossible patient age by bounding the range (18 to 130), and removed samples with impossible values."

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. Preferably through a flow chart.

Not reported.

Obermeyer et al. example: Not reported.

A **hypothetical** example could be: "We calculated the mean biomarker values for all measurements for each patient in one year, and used each mean biomarker as a feature."

Section 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.

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

We randomly divide all patient-years into a ⅔ training set and a ⅓ holdout set. For each observation, we generate 149 features from year t-1. A complete list of features is included in the supplementary materials. Using these features, we train an L1-regularized regression (lasso) to deliver a risk score for year t.

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

Not reported.

A **hypothetical** example could be: "Since the industry-standard for risk prediction typically applies L1-regularized regression, we also used this model to replicate and study the results."

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

We train all models in a random ⅔ training set and show all results only from the ⅓ holdout set. *This is out-of-sample testing with a separate holdout set.*

5d. Method for selecting the final model(s) reported in the paper.

Not reported.

A **hypothetical** example would be: "To compare predictive performance between two models (e.g., random forests and lasso regression), we conducted paired t tests comparing the sizes of the standardized residuals of predictions from the models."

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.

The regularization penalty is tuned via ten-fold cross validation in the training set. We evaluate results only on the holdout set.

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

We train models with alternative labels to compare against models that are currently used for predicting high healthcare costs. This shows that label choice can have marked differences in the racial bias of healthcare algorithms.

Section 6: Data leakage

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

Only patient data in year t-1 is used to predict in year t. The train-test split is also performed at the patient level, i.e., no patient can appear in both the training and test set.

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).

Each patient can only appear in the train or test set.

6c. Justification for why each feature used in the model is legitimate for the task at hand and does not lead to leakage.

We use features that are used in typically available healthcare algorithms for predicting healthcare costs of patients in the following year. We only use information about these features from past years.

Section 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.

We compare different models based on the concentration of a given outcome of interest at or above the 97th percentile of predicted risk. We choose this threshold because only 3% of patients are admitted directly into the high-risk program.

7b. Uncertainty estimates (e.g., confidence intervals, standard deviations), details of how these are calculated, and a justification for why they are valid for each metric reported in the paper.

We report standard errors for the concentrations in (7a).

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

No statistical testing was performed for comparing model performance.

Section 8: Generalizability and limitations

8a. Threats to external validity.

Not reported.

Example: The authors' study only applies for a 2-year period from 2013-2015, and thus might not be applicable to later datasets or an updated commercial algorithm with similar labels.

8b. Contexts in which you don't expect the study's findings to hold.

The result is replicable in a larger national dataset of 3,695,943 commercially insured patients. Contexts in which the study doesn't hold weren't reported.

The authors implicitly noted that this result would likely apply to other US-based commercial algorithms that use cost prediction as the risk outcome, with commercial insurance patients.