Checklist for reporting ML-based science: Obermeyer et al. (2019) example

This is an example of how the REFORMS checklist could be applied to a paper (Obermeyer et al. 2019, "Dissecting racial bias in an algorithm used to manage the health of populations"). This example was filled out by authors of the checklist, based on the information available in Obermeyer et al.'s paper and supplemental materials.

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

Module 1: Study goals

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

Black and White adult patients in U.S. health systems that use risk prediction tools.

Obermeyer et al. do not state this population explicitly, but they note that the algorithm they study "is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year" (p.1). The study focuses on Black and White patients.

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

This population is subject to health care risk prediction algorithms like the algorithm this paper studies (as described on p.1).

Regarding the decision to focus only on Black and White patients: "This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial and ethnic identities" (p.1-2).

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

To replicate a predictive algorithm that is used in a real-world application and determine how the choice of output label impacts racial bias in the algorithm's decisions.

Module 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 (p.7)

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.

URL: https://gitlab.com/labsysmed/dissecting-bias (p.7)

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 (p.7)

[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. (p.7)

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

[Not reported.]

[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.

As an example, Paluck et al. (2018) provide a CodeOcean repository alongside their paper. (URL: https://codeocean.com/capsule/8235972/tree/v7)

Module 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 and process 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." The data included algorithmic risk scores from a risk prediction algorithm, which could be linked to electronic health records, "including all diagnoses (in the form of International Classification of Diseases codes) as well as key quantitative laboratory studies and vital signs capturing the severity of chronic illnesses" and "insurance claims data on utilization, including outpatient and emergency visits, hospitalizations, and health care costs" (p.2). Additional detail about the data source is provided in the supplementary materials.

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

The sample consists of "all primary care patients enrolled in risk-based contracts from 2013 to 2015 [at a large academic hospital]" (p.1). There was no probability sampling; the hospital records constituted the full sample.

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

"Our dataset describes [a] typical [health care risk prediction] algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise" (p.1).

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).

The commercial algorithm that this paper examines predicts "total medical expenditures… in year *t*" (p.3).

The “experiments on label choice” section uses three outcome labels:

- To mimic the commercial algorithm’s prediction of medical expenditures: “total cost in year t (this tailors cost predictions to our own dataset rather than the national training set)” (p.5)
- As an alternative outcome label: “avoidable cost in year t (due to emergency visits and hospitalizations)” (p.5)
- As another alternative outcome label: “health in year t (measured by the number of chronic conditions that flare up in that year)” (p.5)

Descriptive statistics related to the outcomes are presented in Table 1. Additional information on outcomes is available in the supplemental materials.

3e. Number of samples in the dataset.

“Our main sample… consisted of (i) 6079 patients who self-identified as Black and (ii) 43,539 patients who self-identified as White without another race or ethnicity, whom we observed over 11,929 and 88,080 patient-years, respectively (1 patient-year represents data collected for an individual patient in a calendar year)” (p.2).

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

[Not reported.]

A **hypothetical** example could be: X% of patients have a missing value for the commercial algorithm’s risk score, X% of patient-years have missing data for the “total cost” outcome, X% of patient-years have missing data for the “avoidable cost” outcome, and X% of patient-years have missing data for the “health” outcome.

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.).

[Not reported.]

A **hypothetical** example could be: Descriptive statistics for age, gender, income, chronic health conditions, and health care spending for Black and White members of our sample are similar to the descriptive statistics for Black and White patients in large U.S. health systems nationally, as shown in Table X.

Module 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.

[Not reported.]

A **hypothetical** example could be: We manually selected a set of features from year *t-1* that resembles the features used in the commercial algorithm, which includes data on demographics (omitting race/ethnicity), active chronic conditions, costs, and biomarkers. We then implemented the following steps, in this order, prior to training our models:

- Dropped patient-years that had missing data in any of the three outcomes (for example, if a patient-year had missing data for the "total costs" outcome, it was also omitted from the "avoidable costs" outcome and "health" outcome analyses)
- Calculated the outcome values as follows:
    - "Total costs" is the sum of all billed costs for a patient in year *t*
    - "Avoidable costs" is the sum of all billed costs associated with emergency visits and hospitalizations for a patient in year *t*
    - "Health" is the sum of the number of active chronic conditions for a patient in year *t*, i.e., the number of chronic conditions for which the patient utilized health care
- Split the data into a ⅔ train set and ⅓ test set
- Used multiple imputation for missing feature data
- Calculated the mean biomarker values for all measurements for each patient in one year, and used each mean biomarker as a feature
- Applied one-hot encoding to categorical features
- Normalized features with min-max scaling

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.

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." (Appendix p.6)

A complete list of features is included in the GitLab. "Using these features, we train an L1-regularized regression (lasso)" to deliver a risk score for year t. (Appendix p.6)

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 model(s) reported in the paper, 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. (p.5, Appendix p.6)

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

The regularization penalty is tuned via ten-fold cross validation in the training set. (Appendix p.6)

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. (Appendix p.6)

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

Each of the models we compare were trained and tuned using the same procedure. (Appendix p.6)

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).

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. (Appendix p.6)

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. (Appendix p.6)

6c. Justification that each feature or input 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. (Appendix p.6)

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.

We compare different models based on the concentration of a given outcome of interest at or above the 97th percentile of predicted risk. We report the fraction of Black patients above this threshold to assess the biases of the different models. We choose this threshold because only 3% of patients are admitted directly into the high-risk program. (p.5-6)

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

We report standard errors for the concentrations in (7a). (Table 2)

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.

Module 8: Generalizability and limitations

8a. Evidence of external validity.

External validity for the actual population where the commercial algorithm is used was confirmed: “We contacted the algorithm manufacturer for an initial discussion of our results. In response, the manufacturer independently replicated our analyses on its national dataset of 3,695,943 commercially insured patients. This effort confirmed our results” (p.7).

The authors implicitly note that their results likely apply to other U.S.-based commercial health care algorithms that use cost prediction as the risk outcome, due to similarities between algorithms.

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

[Not reported.]

A **hypothetical** example could be: This study only applies for a 2-year period from 2013-2015; the racial bias associated with the use of health care costs as an outcome label might increase or decrease for future time periods, as racial differences in health care access and utilization shift.