RAND ML Equity Tool: Tutorial

Inez Khan, Irineo Cabreros, Joshua Snoke

10/21/2021

1 Introduction

In recent years, there has been a growing awareness that Machine Learning (ML) algorithms can reinforce or exacerbate human biases. The RAND ML Equity Tool was developed to help identify and mitigate biases in algorithms that assist in decision-making processes. In particular, the tool helps users visualize tradeoffs, such as diminished predictive accuracy, that are inherent to enforcing equity. This tool was produced as part of a research effort for RAND, with the goal of assisting the Department of Defense (DoD) as they invest in the development of ML algorithms for a growing number applications. The companion report that further discusses this tool and a framework for reducing inequities are in Cabreros et al. (2021).

While ML algorithms are deployed in a wide variety of applications, this tool is specifically designed for a algorithms that assist in decision-making processes. In particular, this tool is useful when algorithmic output is used to influence binary decisions about individuals. A hypothetical example within this framework is an algorithm that produces individual-level employee performance scores, which are subsequently considered in promotional decisions.

1.1 Binary Classification Setup

Because this tool is currently designed for binary decision processes only, we briefly describe the binary classification setup in this section. Binary classification algorithms may be used to influence binary decisions, such as the decision to promote or not promote an individual. We refer to binary predictions as positive or negative, with positive corresponding to the assumed beneficial outcome, such as promotion. Often algorithmic output takes the form of predicted probabilities, which are then transformed to binary values based on a cutoff threshold. For example, a binary classification algorithm may predict promotion for a particular individual if the predicted probability of promotion is greater than 0.5 (or some other cutoff value).

Classification algorithms are trained on historical data for which the true outcome is observed. Performance properties of an algorithm are derived by comparing algorithm predictions for historical data to historical outcomes. In the case of binary classification, there are four possible outcomes for each prediction: *true positive*, *false positive*, *true negative*, and *false negative.* These possibilities are summarized in Figure 1, which is commonly referred to as the *confusion matrix. False positives* and *false negatives* correspond to prediction errors: the algorithmic prediction and true outcome are not the same. For instance, an individual who was in fact promoted but was not predicted to be promoted is a *false negative.* Conversely, an individual who was in fact not promoted but was predicted to be promoted is a *false positive.*

Algorithmic performance can be described by various different error rates. For instance, the *false positive rate* refers to the proportion of incorrect predictions made within the population of truly negative individuals. Likewise, the *false negative rate* refers to the proportion of incorrect predictions made within the population of truly positive individuals. The *overall error rate* of an algorithm refers to the proportion incorrect predictions made in the entire population.

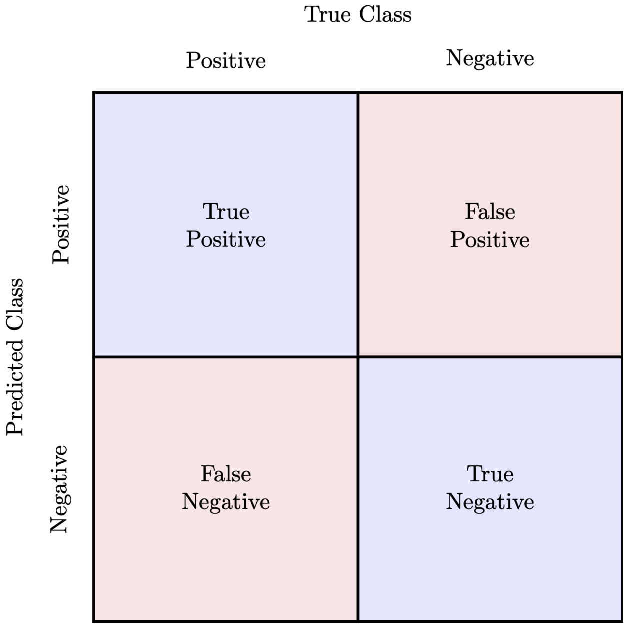


Figure 1: Confusion matrix for the binary classification problem. Red cells indicate incorrect predictions, while blue cells indicate correct predictions

1.2 Equity Definitions

Equity definitions concern properties of algorithmic predictions with respect to a *protected characteristic* (e.g., race) which has several *levels* (e.g., Black, white, and Hispanic). There are many different ways to define equity, and we will provide a very brief overview of several of the most important ones in this subsection.

A simple equity definition is that of *statistical parity.* This definition requires that positive predictions occur at equal rates for each level of the protected class. When statistical parity holds, the demographics of the population predicted to belong to the positive class match the demographics of the entire population. A variant on statistical parity is *conditional statistical parity*, in which statistical parity holds conditional on a subset of important covariates X. Many additional equity concepts are constructed by comparing different error rates across levels of a protected characteristic. For example, an important equity definition is that of *false negative rate balance* (false negative rate balance is equivalent to *true positive rate balance*) For the case of promotions, false negative rate balance may require that the algorithm does not. erroneously predict non-promotion for one race at a higher rate than another. *False positive rate balance* is defined analogously (false positive rate balance is equivalent to *true negative rate balance*). *Equalized odds* is attained when both false positive rate and false negative rate balance is attained by an algorithm. The Table below provides mathematical definitions many of the common equity concepts. In this table, Y is the binary outcome, $\hat{Y}$ is the algorithmic prediction, G is a protected characteristic, and X are additional covariates.

Table 1: Fairness definitions.

Name	Definition
Statistical parity	$P(\hat{Y}=1\|G=g)=P(\hat{Y}=1)$
Conditional statistical parity	$P(\hat{Y}=1\|G=g,X=x)=P(\hat{Y}=1\|X=x)$
FP balance	$P(\hat{Y}=1\|Y=0,G=g)=P(\hat{Y}=1\|Y=0)$
FN balance (Equal opportunity)	$P(\hat{Y}=0\|Y=1,G=g)=P(\hat{Y}=0\|Y=1)$
PPV balance	$P(Y=1\|\hat{Y}=1,G=g)=P(Y=1\|\hat{Y}=1)$
NPV balance	$P(Y=0\|\hat{Y}=0,G=g)=P(Y=0\|\hat{Y}=0)$

Name	Definition
Equalized Odds	FP balance & FN balance
Conditional use accuracy equality	PPV balance & NPV balance
Overall accuracy equality	$P(\hat{Y} = Y \| G = g) = P(\hat{Y} = Y)$

Another common equity definition is *fairness through unawareness.* This definition simply requires that protected characteristics are removed from any input data upstream of algorithmic training. With respect to race, fairness through unawareness intuitively captures the notion of a "race-blind" decision. Despite its simplicity, many have noted that fairness through unawareness is a problematic notion within the ML framework. This is because simply removing protected characteristics may not actually remove their influence from predictions. For example, even if race is removed from the input data, an individual's race may be predicted with high confidence from other variables that remain in the data, such as geography.

1.3 Approaches to Enforcing Equity

The fairness literature has developed many methods to enforce fairness. These methods differ both in the fairness criteria they enforce as well as the stage in the algorithm at which the method intervenes. Methodologies are typically categorized into three classes: pre-processing, in-processing, and post-processing (Berk et al., 2018). Pre-processing refers to any method that alters the input data before it is used by the algorithm to produce predictions. In-processing refers to any corrective measure that intervenes during the ML training process. Finally, post-processing methods are applied entirely downstream of training the ML algorithm, directly adjusting the algorithmic predictions themselves.

Figure 2 summarizes the distinction between these methods. The RAND ML Equity Tool implements both pre-processing and post-processing methods, but not in-processing methods. This is because pre-processing and post-processing methods are implemented independently from the development of the predictive ML algorithm and are therefore easier to apply in practice. By contrast, in-processing methods require more customized implementation. We provide further details of the pre- and post-processing methods implemented in the RAND ML Equity Tool below.

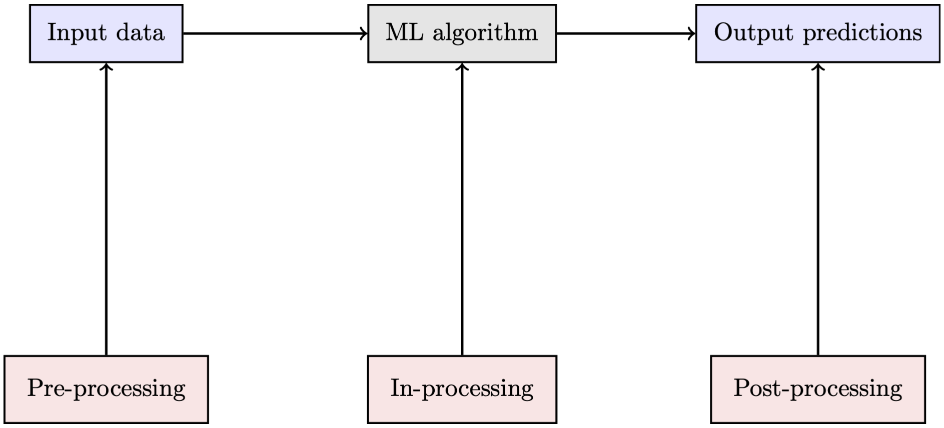


Figure 2: Methods for enforcing equity constraints

2 Tool Overview

The RAND ML Equity tool contains five tabs with features that assess the equity properties of a model and adjust model input/output to meet equity goals. The subsections below provide details for each of these tabs.

The tool works in common browsers such as Google Chrome, Firefox, Microsoft Edge, and Safari, but there may be minor differences in user layout depending on the browser. All data must be input as a .csv file and include a categorial variable named ‘G’; this is the protected characteristic (e.g., gender or race) with respect to which equity is being assessed. Different tabs have different additional data requirements.

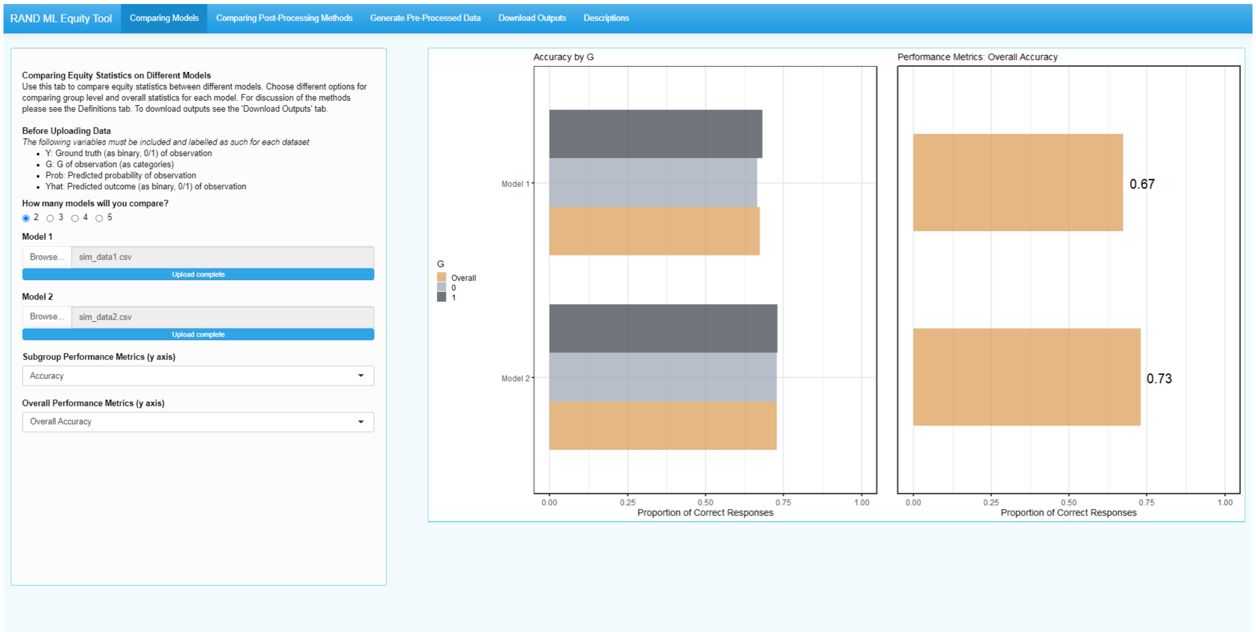
We note that in any real-world application of our tool, legal considerations and institutional policies may constrain the types of interventions possible when enforcing equity criteria as demonstrated in this tutorial. See Cabreros et al. (2021) for a more detailed discussion of this point.

2.1 Compare Models

The ‘Comparing Models’ tab allows you to compare equity performance across different models. One can use this tab to assess the tradeoffs across different candidate models or the same model with different pre-processing methods applied to the data (described below). This tab allows for the comparison of 2-5 models. A dataset for each model must be uploaded and contain the following three variables, named as follows:

1) **Y**: The ground truth of the observation. It must be a binary indicator. The tool currently does not allow continuous or categorial outcomes for this tab.
2) **G**: A categorial variable the identifies the protected characteristic of the observation and the variable you would want to increase equity over.
3) **Yhat**: The predicted outcome of the observation, also as a binary indicator.

Note that these variable names must match exactly in the input data. Additional variables that are included in the dataset will be ignored. This tab allows you to assess models on the following performance metrics: *Overall Accuracy*, *False Positive Rate*, *False Negative Rate*, *Positive Predictive Value*, and *Negative Predictive Value*. These performance metrics can be displayed by group level (as defined by the levels of `G`) as well as aggregated across the entire population. Additionally, the proportion of positive predictions for each group can be displayed by selecting the *Statistical Parity* option. Figure 3 shows an example of the tab.



Comparing Equity Statistics on Different Models
Use this tab to compare equity statistics between different models. Choose different options for comparing group level and overall statistics for each model. For discussion of the methods please see the Definitions tab. To download outputs see the 'Download Outputs' tab.

Before Uploading Data
The following variables must be included and labelled as such for each dataset
- Y: Ground truth (as binary, 0/1) of observation
- G: G of observation (as categories)
- Prob: Predicted probability of observation
- Yhat: Predicted outcome (as binary, 0/1) of observation

How many models will you compare?
● 2 ○ 3 ○ 4 ○ 5

Model 1
Browse... sim_data1.csv
Upload complete

Model 2
Browse... sim_data2.csv
Upload complete

Subgroup Performance Metrics (y axis)
Accuracy

Overall Performance Metrics (y axis)
Overall Accuracy

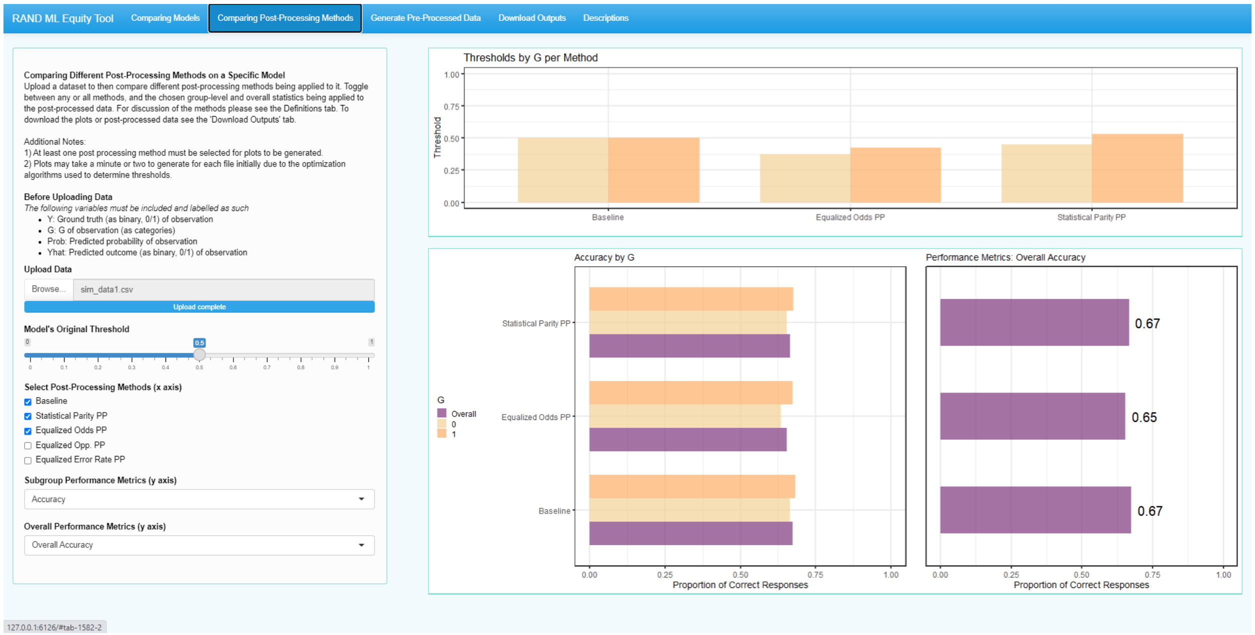
Accuracy by G
Model 1
Model 2
G
Overall
0
1
0.00
0.25
0.50
0.75
1.00
Proportion of Correct Responses
Performance Metrics: Overall Accuracy
0.67
0.73
0.00
0.25
0.50
0.75
1.00
Proportion of Correct Responses

Figure 3: Comparing Models Tab

2.2 Post-process Predictions

The ‘Comparing Post-Processing Methods’ tab allows you to apply various post-processing methods to a single specific model and visualize their effect on equity measures. The data configuration requirements are the same as for the ‘Comparing Models’ tab, except that an additional variable **Prob** is required. This variable specifies the probability of a positive ($Y = 1$) outcome for each individual. We note that, while the ‘Comparing Models’ tab is applicable to any classification algorithm, the ‘Comparing Post-Processing Methods’ tab requires output from algorithms with probabilistic predictions (e.g., logistic regressions, regression trees, etc.). While this limits the available algorithmic models amenable to the methods provided in this tab, we note that it is often the case that the output from predictive models without probabilistic output can be coerced into probabilistic output.

This tab also provides an additional plot for comparing the different thresholds applied to the data for each post-processing method broken down by race. Currently, the post-processing methods that can be applied are *Statistical Parity*, *Equalized Odds*, *Equalized Opportunity*, and *Equalized Error Rate*. One or more of these methods can then be compared to the unaltered model designated as *Baseline*. Figure 4 shows an example of the tab.



Comparing Different Post-Processing Methods on a Specific Model
Upload a dataset to then compare different post-processing methods being applied to it. Toggle between any or all methods, and the chosen group-level and overall statistics being applied to the post-processed data. For discussion of the methods please see the Definitions tab. To download the plots or post-processed data see the 'Download Outputs' tab.

Additional Notes:
1) At least one post processing method must be selected for plots to be generated.
2) Plots may take a minute or two to generate for each file initially due to the optimization algorithms used to determine thresholds.

Before Uploading Data
The following variables must be included and labelled as such

- Y: Ground truth (as binary, 0/1) of observation
- G: G of observation (as categories)
- Prob: Predicted probability of observation
- Yhat: Predicted outcome (as binary, 0/1) of observation

Upload Data
Browse... sim_data1.csv
Upload complete

Model's Original Threshold
0.5

Select Post-Processing Methods (x axis)
☑ Baseline
☑ Statistical Parity PP
☑ Equalized Odds PP
☐ Equalized Opp. PP
☐ Equalized Error Rate PP

Subgroup Performance Metrics (y axis)
Accuracy

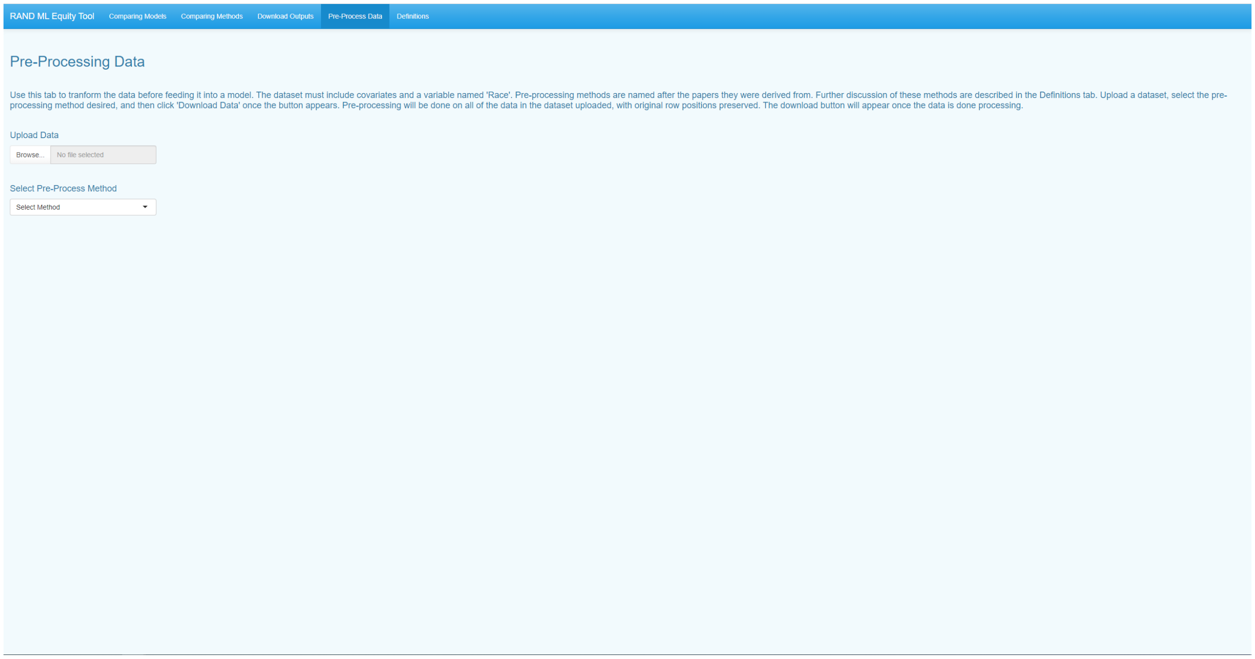
Overall Performance Metrics (y axis)
Overall Accuracy

127.0.0.1:6126/#tab-1582-2

Figure 4: Comparing Post-Processing Methods Tab

2.3 Generate Pre-Processed Data

The ‘Generate Pre-Processed Data’ tab provides a pre-processing method to apply to data before feeding it into a model. The input dataset must include all covariates used in the predictive model in addition to the protected variable ‘G’. Currently, there is one implemented pre-processing method to apply, which is the simple pairwise independence method described in ‘Johndrow and Lum (2019)’. This method ensures pairwise independence between each predictive covariate and G, but does not ensure full mutual independence. (‘Johndrow and Lum (2019)’ describe an approach for mutual independence, but we do not implement this in the present version of the tool.) The pre-processed data will be available to download as a .csv file in the ‘Download Outputs’ tab (described below). This dataset will be of the same dimensions as the input dataset, but with each variable pairwise-independent of `G`. Figure 5 shows an example of this tab.



Pre-Processing Data

Use this tab to tranform the data before feeding it into a model. The dataset must include covariates and a variable named 'Race'. Pre-processing methods are named after the papers they were derived from. Further discussion of these methods are described in the Definitions tab. Upload a dataset, select the pre-processing method desired, and then click 'Download Data' once the button appears. Pre-processing will be done on all of the data in the dataset uploaded, with original row positions preserved. The download button will appear once the data is done processing.

Figure 5: Generate Pre-Processed Data Tab

2.4 Download Outputs

The ‘Download Outputs’ tab allows you to download the data and/or plots produced by the three analytic tabs described in this section (i.e., the ‘Comparing Models’, ‘Comparing Post-Processing Methods’, and ‘Generate Pre-Processed Data’ tabs). The output available for download on this tab depends on the inputs selected from each of the other tabs. After output has been generated by one of the analytic tabs, download buttons will appear in the ‘Download Outputs’ tab in addition to a summary of the methods/metrics that have been selected. Data produced by each analytic tab will be output as a .xlsx file in the ‘Download Outputs’ tab. Each sheet in the excel file will contain data for each model or method applied, with the last tab providing threshold information for each model or method selected. All plots produced by a single analytic tab will be output into one .png file.

2.5 Descriptions

The last tab provides definitions and citations for all of the terms mentioned in the first four tabs. There is also brief discussion of implementation of the methods.

2.6 Usage notes

The tool can be run remotely on RAND’s RShiny server: INSERT LINK HERE. Alternatively, the tool’s source code can be downloaded and run locally in. Using the local version may be preferable, for instance, a particular dataset cannot be uploaded onto the RShiny server due to privacy considerations. When using the tool locally, it will be neccessary to install the following packages and their dependencies: `shiny`, `shinyWidgets`, `grid`, `gridExtra`, `openxlsx`, `dplyr`, and `ggplot2`. General instructions for running `RShiny` apps can be found here: https://shiny.rstudio.com/articles/running.html.

3 Examples

In this section, we provide three examples illustrating the use of the tool. The first example shows how one can use the tool to compare alternative model fits from an equity perspective. The second example demonstrates how to take an existing dataset, apply a pre-processing algorithm to it, and then visualize differences in performance over the unaltered data. The final example demonstrates the post-processing methods.

In each of the examples below, we will use the simulated dataset `data.csv`, which is available in the tutorial folder.

```
library(readr)
data <- read_csv(file = "./data.csv")
```

This dataset consists of the following variables:

1. G: A binary indicator of a protected characteristic (e.g., race/ethnicity).
2. Y: The true binary outcome of interest.
3. X and L: Covariates used to predict Y.

3.1: Comparing models

In the first example, we will compare the equity characteristics of two alternative methods for predicting an outcome Y. The first method uses just X to predict Y, while the second method uses X and L to predict Y. This type of setting may be useful when there is a concern that adding an additional covariate into a prediction model may have negative equity repercussions. X may be deemed a legitimate predictor of Y, while L is a variable that is contested.

In the code below, we fit two logistic regression models that we use to predict Y. The predicted probability is used to derive binary predictions of the outcome by assigning individuals with predicted probabilities greater than 0.5 to the outcome 1 and individuals with predicted probabilities less than 0.5 to the outcome 0. Note that we can use any alternative classification algorithm within the `Comparing Models` tab.

```
  # Predictions from model 1
  Prob1 <- predict(glm(Y ~ X,  data = data, family = "binomial"), type = "response")
  data$Prob1 <- Prob1
  data <- mutate(data, Yhat1 = ifelse(Prob1 <= 0.5, 0, 1))

  # Predictions from model 2
  Prob2 <- predict(glm(Y ~ X + L,  data = data, family = "binomial"), type = "response")
  data$Prob2 <- Prob2
  data <- mutate(data, Yhat2 = ifelse(Prob2 <= 0.5, 0, 1))
```

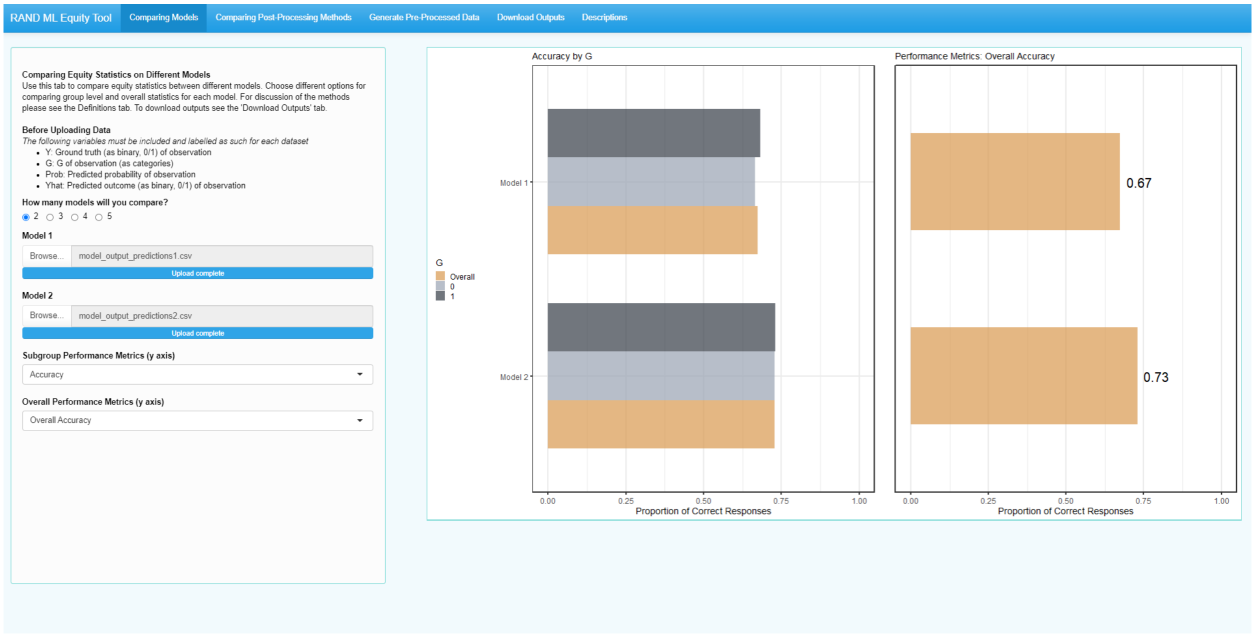
Next, we will write our files in the format expected by the RAND Equity Tool. For the 'Comparing Models' tab, the RAND Equity Tool requires an input dataset containing the following three variables: 1) The protected characteristic `G`, 2) The true outcome `Y`, and 3) The predicted outcome `Yhat`. Other variables may be included, but are unnecessary and will be ignored by computations in this tab. In other tabs, the probabilities of classification 'Prob' will also be used. We include these in our output below to facillitate later examples.

```
data1 <- dplyr::select(data, Prob1, Yhat1, Y, G)
names(data1) <- c("Prob", "Yhat", "Y", "G")
write_csv(data1, path = "./model_output_predictions1.csv")
```

```
## Warning: The `path` argument of `write_csv()` is deprecated as of readr 1.4.0.
## Please use the `file` argument instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
```

```r
data2 <- dplyr::select(data, Prob2, Yhat2, Y, G)
names(data2) <- c("Prob", "Yhat", "Y", "G")
write_csv(data2, path = "./model_output_predictions2.csv")
```

Once we have written these files, we can now use the ‘Comparing Models’ page to compare the equity characteristics of these models. A screen similar to Figure 6 should appear to compare the different methods. The scroll down menus allow the user to choose a metric to compare methods both in the whole population (using the “Overall Performance Metrics” menu) and across levels of the protected characteristic G (using the “Subgroup Performance Metrics” menu). In Figure 6, we show the accuracy results for both the whole population and stratified by G. We notice the second method produces higher overall accuracy. Examining the performance at each level of the protected class G, we notice that not only does the second method produce higher accuracy overall, but that discrepencies in performance are smaller across G as well.


RAND ML Equity Tool
Comparing Models
Comparing Post-Processing Methods
Generate Pre-Processed Data
Download Outputs
Descriptions
Accuracy by G
Performance Metrics: Overall Accuracy
Model 1
Model 2
0.67
0.73
G
Overall
0
1
0.00
0.25
0.50
0.75
1.00
Proportion of Correct Responses

Comparing Equity Statistics on Different Models
Use this tab to compare equity statistics between different models. Choose different options for comparing group level and overall statistics for each model. For discussion of the methods please see the Definitions tab. To download outputs see the 'Download Outputs' tab.

Before Uploading Data
The following variables must be included and labelled as such for each dataset

- Y: Ground truth (as binary, 0/1) of observation
- G: G of observation (as categories)
- Prob: Predicted probability of observation
- Yhat: Predicted outcome (as binary, 0/1) of observation

How many models will you compare?
● 2 ○ 3 ○ 4 ○ 5

Model 1
Browse... model_output_predictions1.csv
Upload complete

Model 2
Browse... model_output_predictions2.csv
Upload complete

Subgroup Performance Metrics (y axis)
Accuracy

Overall Performance Metrics (y axis)
Overall Accuracy

Figure 6: Comparing Two Different Models

Example 3.2: Pre-processing

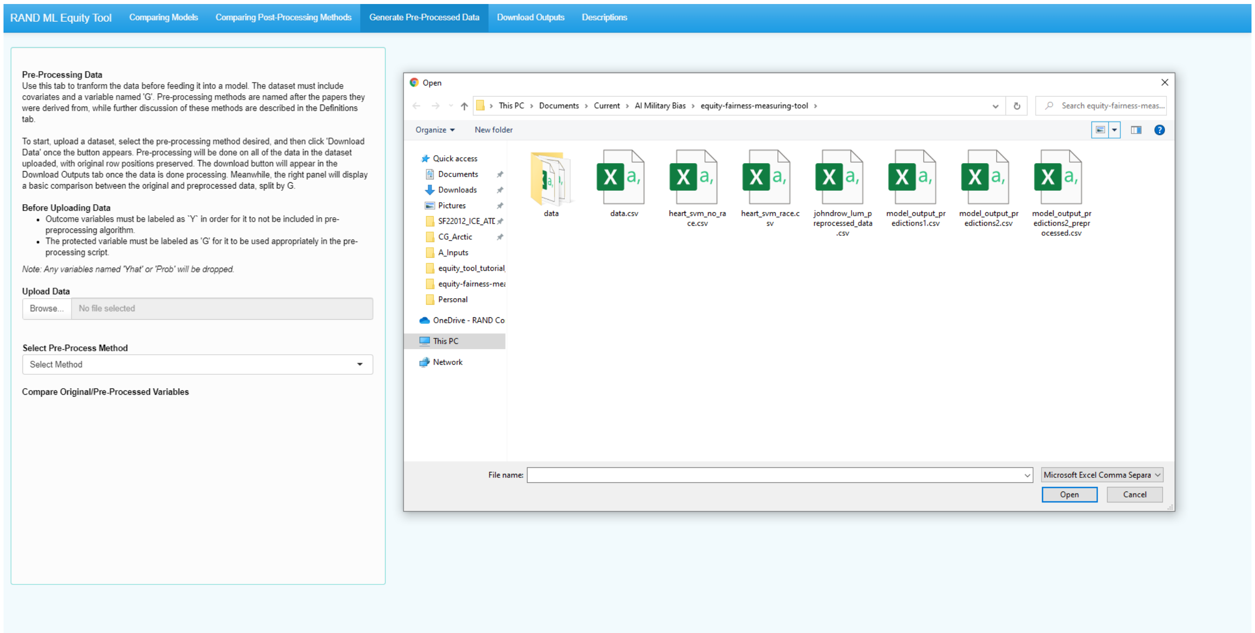
In this example, we apply a pre-processing method to remove the effects of G from a dataset in order to mitigate potential correlated effects it may have with other covariates.

First, we go to the ‘Generate Pre-Processed Data’ tab and click ‘Browse’ under the Upload Data section. A window will appear to select the file we wish to process, such as in Figure 7. We select the file `data.csv`, and then select ‘Johndrow and Lum (2019)’ under the select ‘Select Pre-Process Method’ option. After the pre-processing algorithm completes, the pre-processed dataset can be downloaded in the ‘Download Data’ tab by clicking the ‘Download Data’ button under the ‘Generate Pre-Processed Data’ section.

Now that we have obtained our pre-processed dataset, we will now re-fit our model usign the pre-processed data. Below, we fit a logistic regression analogous to model 2 in the section above, but using the pre-processed data rather than the original data. Again, we save the resulting output.

```r
  data_preprocess <- read_csv(file = "johndrow_lum_preprocessed_data.csv")
```

```
## Rows: 10000 Columns: 4
```



Pre-Processing Data
Use this tab to tranform the data before feeding it into a model. The dataset must include covariates and a variable named 'G'. Pre-processing methods are named after the papers they were derived from, while further discussion of these methods are described in the Definitions tab.

To start, upload a dataset, select the pre-processing method desired, and then click 'Download Data' once the button appears. Pre-processing will be done on all of the data in the dataset uploaded, with original row positions preserved. The download button will appear in the Download Outputs tab once the data is done processing. Meanwhile, the right panel will display a basic comparison between the original and preprocessed data, split by G.

Before Uploading Data
- Outcome variables must be labeled as `Y` in order for it to not be included in pre-preprocessing algorithm.
- The protected variable must be labeled as 'G' for it to be used appropriately in the pre-processing script.

Note: Any variables named 'Yhat' or 'Prob' will be dropped.

Upload Data
Browse... No file selected

Select Pre-Process Method
Select Method

Compare Original/Pre-Processed Variables

Open
This PC › Documents › Current › AI Military Bias › equity-fairness-measuring-tool ›
Search equity-fairness-meas...
Organize
New folder
Quick access
Documents
Downloads
Pictures
SF22012_ICE_ATC
CG_Arctic
A_Inputs
equity_tool_tutorial
equity-fairness-mea
Personal
OneDrive - RAND Co
This PC
Network
data
data.csv
heart_svm_no_race.csv
heart_svm_race.csv
johndrow_lum_preprocessed_data.csv
model_output_predictions1.csv
model_output_predictions2.csv
model_output_predictions2_preprocessed.csv
File name:
Microsoft Excel Comma Separa
Open
Cancel

Figure 7: Generate Pre-Processed Data File Upload

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (4): X, L, G, Y

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

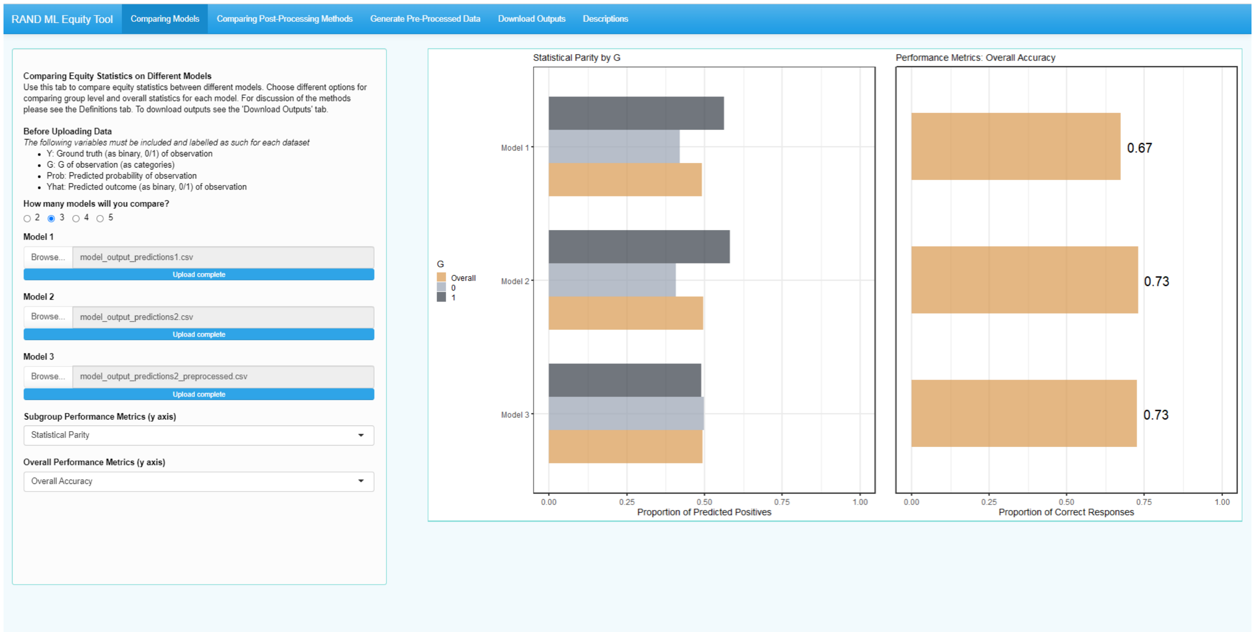
```
# Predictions from pre-processed data
rhat <- predict(glm(Y ~ X + L,  data = data_preprocess, family = "binomial"), type = "response")
data_preprocess$rhat <- rhat
data_preprocess <- mutate(data_preprocess, Yhat = ifelse(rhat <= 0.5, 0, 1))

data_preprocess <- dplyr::select(data_preprocess, rhat, Yhat, Y, G)
names(data_preprocess) <- c("Prob", "Yhat", "Y", "G")
write_csv(data_preprocess, path = "./model_output_predictions2_preprocessed.csv")
```

Now that we have saved our predictions from the model trained on the pre-processed data, we can use the RAND Equity Tool to understand the consequent equity impact. To do this, we go to the ‘Comparing Models’ tab. We select the option to allow us to compare three models. Under Model 1, we select the original file `model_output_predictions1.csv`, for Model 2 we select `model_output_predictions2.csv`, and under Model 3 we select the pre-processed file we obtained from the re-processing tab (`model_output_predictions2_preprocessed.csv`). In Figure 8 we view overall accuracy for the whole population and statistical parity for the subgroup analyses. As seen in Figure 8, we notice applying the pre-processing method increases statistical parity by producing the same proportion of predicted positives for each group. (It is a result shown by Johndrow and Lum (2019) that this form of pre-processing implicitly induces statistical parity.)

Next, we compare the False Negative Rate between the three models as in Figure 9. Under the Subgroup Performance Metrics, we select ‘False Negative Rate’. We notice the rate of false negatives appear to decrease in the pre-processed data.

The plots produced can be downloaded from the ‘Download Outputs’ tab. Under the section titled ‘Comparing Models,’ there is a button to download the plots as well as the metrics we have selected from the ‘Comparing Models’ tab, as in Figure 10. You can optionally name the plot file using the ‘Title of Saved Plots’ text box before downloading.



Comparing Equity Statistics on Different Models
Use this tab to compare equity statistics between different models. Choose different options for comparing group level and overall statistics for each model. For discussion of the methods please see the Definitions tab. To download outputs see the 'Download Outputs' tab.

Before Uploading Data
The following variables must be included and labelled as such for each dataset

- Y: Ground truth (as binary, 0/1) of observation
- G: G of observation (as categories)
- Prob: Predicted probability of observation
- Yhat: Predicted outcome (as binary, 0/1) of observation

How many models will you compare?
○ 2 ● 3 ○ 4 ○ 5

Model 1
Browse... model_output_predictions1.csv
Upload complete

Model 2
Browse... model_output_predictions2.csv
Upload complete

Model 3
Browse... model_output_predictions2_preprocessed.csv
Upload complete

Subgroup Performance Metrics (y axis)
Statistical Parity

Overall Performance Metrics (y axis)
Overall Accuracy

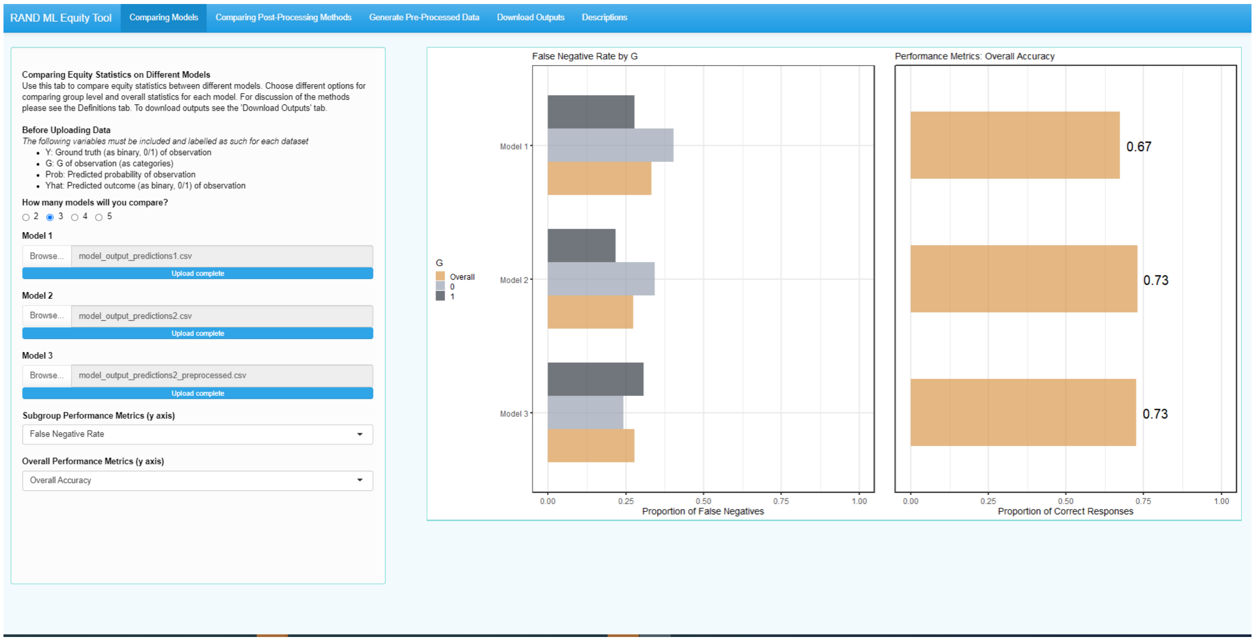
Figure 8: Comparing Preprocessed Data to Original for Statistical Parity

Example 3.3 Post Processing

Several post-processing methods can be applied to the outputs of any single model by using the ‘Comparing Post-Processing Methods’ tab of the RAND Equity Tool. However, this tab requires a variable called `Prob` in addition to the three variables required by the ‘Comparing Models’ tab. The variable `Prob` is the predicted probability that the binary outcome Y equals 1 for each individual.

As an example of the ‘Comparing Post-Processing Methods’ tab, we upload `model_output_predictions2.csv`. Note that it may take a couple minutes for the plots to generate as the tool processes all of the methods beforehand. Once the plots have appeared onto the screen, we select all of the Post-Processing Methods available, producing Figure 11. The Subgroup Performance Metrics menu allows the user to select an equity metric to compare post-processing methods on. First, we select Statistical Parity under the Subgroup Performance Metrics option. We leave the Overall Performance Metrics as Overall Accuracy. Unsurprisingly, we notice the Statistical Parity post-processing method produces the most equitable outcomes in terms of postive labels for both races. We also notice that none of the post-processing methods affect accuracy too much, which could simplify our decision for determining which post-processing method to use.

Next we compare the False Negative Rate by selecting that option under Subgroup Performance Metrics, which then gives us the screen in Figure 12. We notice Equalized Odds and Equalized Opportunity produces the lowest false negative rates across both of the races, while Equalized Error Rate produces the highest. Here, the context of the problem we are attempting to address is important where considering which post processing method to use. If our problem needs to minimize the false negative rate, we would want to choose a method that would minimize it. If our problem needs to produce equitable false negative rates, we would want to select the method that has the smallest difference between the races. However, if we do not have a need to consider false negative rates, we may want to focus on overall accuracy instead. In this case, the accuracy does not differ.



Comparing Equity Statistics on Different Models
Use this tab to compare equity statistics between different models. Choose different options for comparing group level and overall statistics for each model. For discussion of the methods please see the Definitions tab. To download outputs see the 'Download Outputs' tab.

Before Uploading Data
The following variables must be included and labelled as such for each dataset

- Y: Ground truth (as binary, 0/1) of observation
- G: G of observation (as categories)
- Prob: Predicted probability of observation
- Yhat: Predicted outcome (as binary, 0/1) of observation

How many models will you compare?
○ 2 ● 3 ○ 4 ○ 5

Model 1
Browse... model_output_predictions1.csv
Upload complete

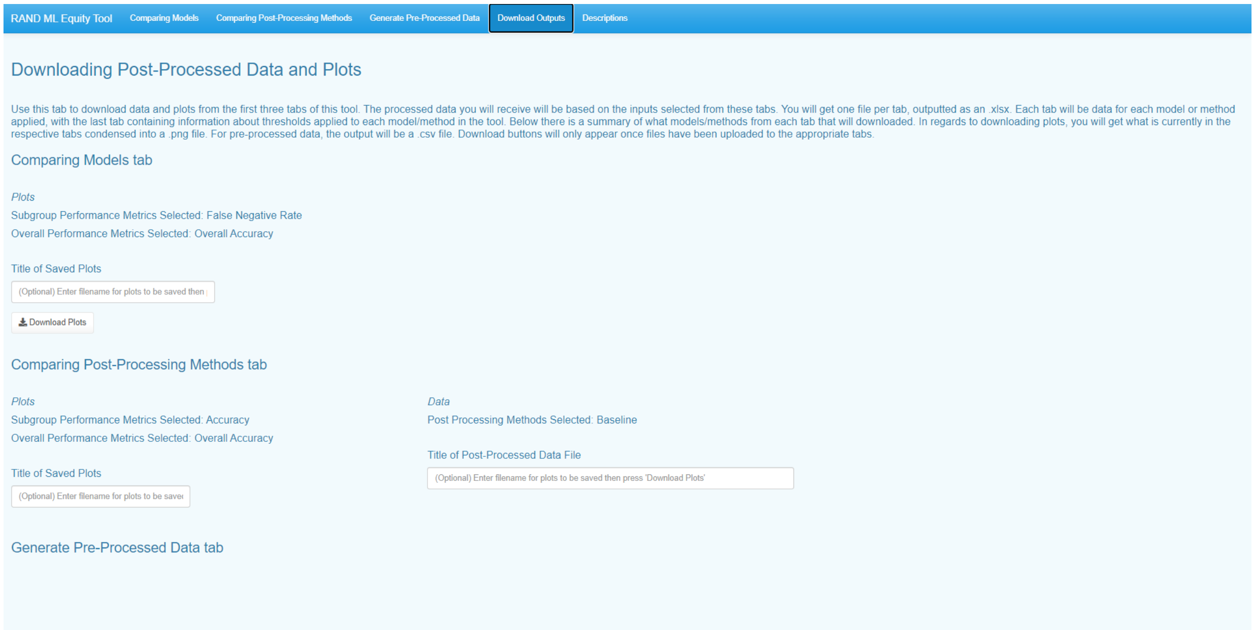
Model 2
Browse... model_output_predictions2.csv
Upload complete

Model 3
Browse... model_output_predictions2_preprocessed.csv
Upload complete

Subgroup Performance Metrics (y axis)
False Negative Rate

Overall Performance Metrics (y axis)
Overall Accuracy

Figure 9: Comparing Preprocessed Data to Original for False Negative Rates



Downloading Post-Processed Data and Plots

Use this tab to download data and plots from the first three tabs of this tool. The processed data you will receive will be based on the inputs selected from these tabs. You will get one file per tab, outputted as an .xlsx. Each tab will be data for each model or method applied, with the last tab containing information about thresholds applied to each model/method in the tool. Below there is a summary of what models/methods from each tab that will downloaded. In regards to downloading plots, you will get what is currently in the respective tabs condensed into a .png file. For pre-processed data, the output will be a .csv file. Download buttons will only appear once files have been uploaded to the appropriate tabs.

Comparing Models tab

Plots

Subgroup Performance Metrics Selected: False Negative Rate

Overall Performance Metrics Selected: Overall Accuracy

Title of Saved Plots

(Optional) Enter filename for plots to be saved then

Comparing Post-Processing Methods tab

Plots

Subgroup Performance Metrics Selected: Accuracy

Overall Performance Metrics Selected: Overall Accuracy

Title of Saved Plots

(Optional) Enter filename for plots to be saved

Data

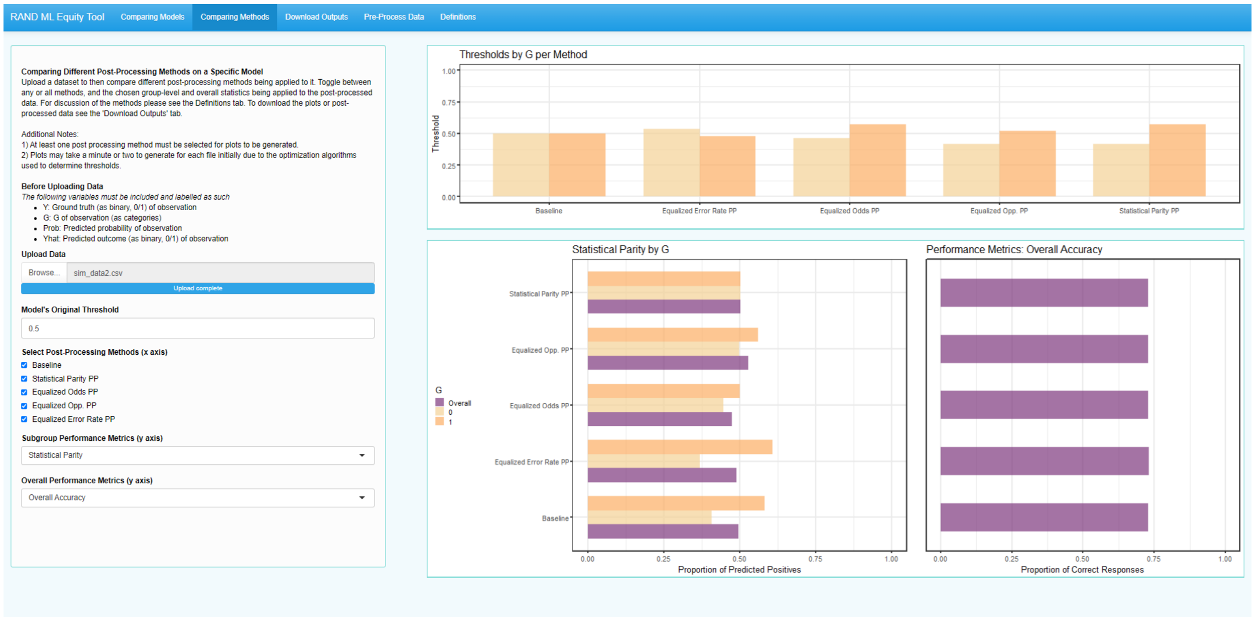
Post Processing Methods Selected: Baseline

Title of Post-Processed Data File

(Optional) Enter filename for plots to be saved then press 'Download Plots'

Generate Pre-Processed Data tab

Figure 10: Downloading Output from 'Comparing Models' Tab



Comparing Different Post-Processing Methods on a Specific Model
Upload a dataset to then compare different post-processing methods being applied to it. Toggle between any or all methods, and the chosen group-level and overall statistics being applied to the post-processed data. For discussion of the methods please see the Definitions tab. To download the plots or post-processed data see the 'Download Outputs' tab.

Additional Notes:
1) At least one post processing method must be selected for plots to be generated.
2) Plots may take a minute or two to generate for each file initially due to the optimization algorithms used to determine thresholds.

Before Uploading Data
The following variables must be included and labelled as such

- Y: Ground truth (as binary, 0/1) of observation
- G: G of observation (as categories)
- Prob: Predicted probability of observation
- Yhat: Predicted outcome (as binary, 0/1) of observation

Upload Data
Browse... sim_data2.csv
Upload complete

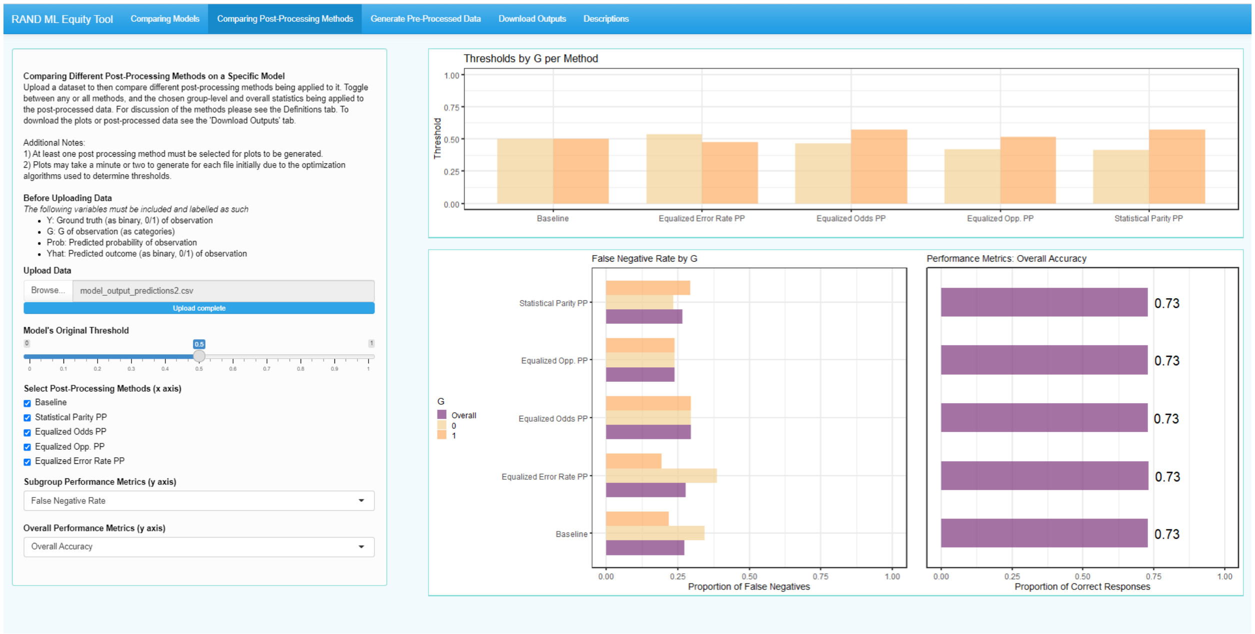
Model's Original Threshold
0.5

Select Post-Processing Methods (x axis)
☑ Baseline
☑ Statistical Parity PP
☑ Equalized Odds PP
☑ Equalized Opp. PP
☑ Equalized Error Rate PP

Subgroup Performance Metrics (y axis)
Statistical Parity

Overall Performance Metrics (y axis)
Overall Accuracy

Figure 11: Comparing Statistical Parity Among Post Processed Data



Comparing Different Post-Processing Methods on a Specific Model
Upload a dataset to then compare different post-processing methods being applied to it. Toggle between any or all methods, and the chosen group-level and overall statistics being applied to the post-processed data. For discussion of the methods please see the Definitions tab. To download the plots or post-processed data see the 'Download Outputs' tab.

Additional Notes:
1) At least one post processing method must be selected for plots to be generated.
2) Plots may take a minute or two to generate for each file initially due to the optimization algorithms used to determine thresholds.

Before Uploading Data
The following variables must be included and labelled as such

- Y: Ground truth (as binary, 0/1) of observation
- G: G of observation (as categories)
- Prob: Predicted probability of observation
- Yhat: Predicted outcome (as binary, 0/1) of observation

Upload Data
Browse... model_output_predictions2.csv
Upload complete

Model's Original Threshold
0.5

Select Post-Processing Methods (x axis)
☑ Baseline
☑ Statistical Parity PP
☑ Equalized Odds PP
☑ Equalized Opp. PP
☑ Equalized Error Rate PP

Subgroup Performance Metrics (y axis)
False Negative Rate

Overall Performance Metrics (y axis)
Overall Accuracy

Figure 12: Comparing False Negative Rate Among Post Processed Data

The produced plots can be downloaded from the ‘Download Outputs’ tab under the section titled ‘Comparing Post-Processing Methods’. We also see the options we have selected from this tab. If we wish to download either of the plots or the data, we can click on the appropriate download button.