Ursa Health

Blog Supplement

Fairness criteria

There is currently no consensus on the best criteria for determining fairness. Further, the proposed criteria can not all be satisfied simultaneously. At minimum, therefore, assessing fairness requires first making difficult decisions about what types of harm are of greatest concern (e.g., are false negatives worse than false positives?).

Another challenge is that differences in model performance are expected when groups have varying levels of underlying risk, a phenomenon known as the spectrum effect. Researchers and stakeholders then must determine whether these differences are justified or due to harmful discrimination.

An additional complication is that group-based metrics are susceptible to fairness gerrymandering, where an algorithm can appear fair for each individual group while simultaneously demonstrating serious violations of fairness criteria within subgroups constructed from combinations of the original groups.

Label choice bias criterion

Label choice bias is an observational criterion that compares health outcomes across groups conditional on their risk scores. Specifically, Obermeyer and coauthors compare health status Y for black patients B and white patients W conditional on risk score R. Under this criterion, the equality of $E[Y|R,B] = E[Y|R,W]$ indicates that there is no algorithmic bias (i.e., no harmful discrimination).

Structural causal model

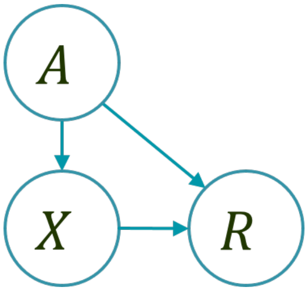
A causal graph implies a set of equations that forms the structural causal model. The variables from the two scenarios in the blog post include:

- $\mathbf{X}$: illness score
- $\mathbf{A}$: race $\{0 = \text{Fasites}, 1 = \text{Noffians}\}$
- $\mathbf{R}$: costs (either total or avoidable)
- $\mathbf{M_X}$: managed illness

To keep the example simple, we assume an additive model with linear effects and normally distributed disturbance terms.

Scenario 1: Total cost outcome

Recall the causal graph for the first scenario:



In scenario 1, Noffian group membership causes:

- Increased health need, and
- Reduced healthcare utilization.

Additional assumptions are that:

- Total cost increases linearly in illness.

The following equations show the structural causal model from the blog post:

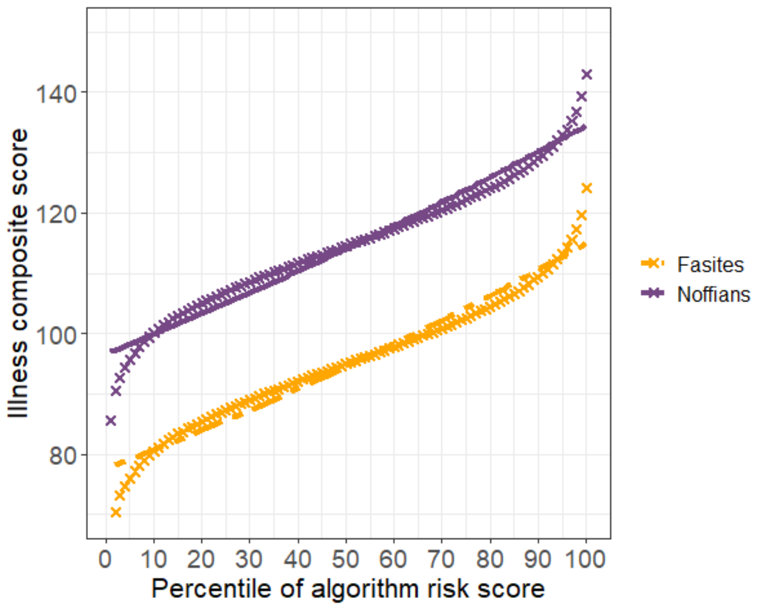
$$\begin{aligned} X &= 100 + 10A + \epsilon \\ R &= 500 + 100X - 2000A + \epsilon \end{aligned}$$

If total costs were based on differences in illness alone, Noffians would average $1,000 more in total costs. Due to barriers of access depressing costs, however, Noffians actually average $1,000 less. We simulated data from the above structural model and found risk scores using linear regression. The R code below shows the data simulation for Scenario 1:

```r
# parameters
N <- 10000
sigma <- 1000

# proportion of Noffians
prop <- 0.5
A <- rep(c(0,1), times=c((1-prop)*N, prop*N))

# additive shift in illness and cost due to race
set.seed(1)
X <-  100 + 10*A + rnorm(N,0,10)
R <- 500 + 100*X - 2000*A + rnorm(N,0,sigma)

# models
train_data <- data.frame(A, X, R)
m1 <- lm(R ~ X, data = train_data)
m2 <- lm(R ~ X + A, data = train_data)
```

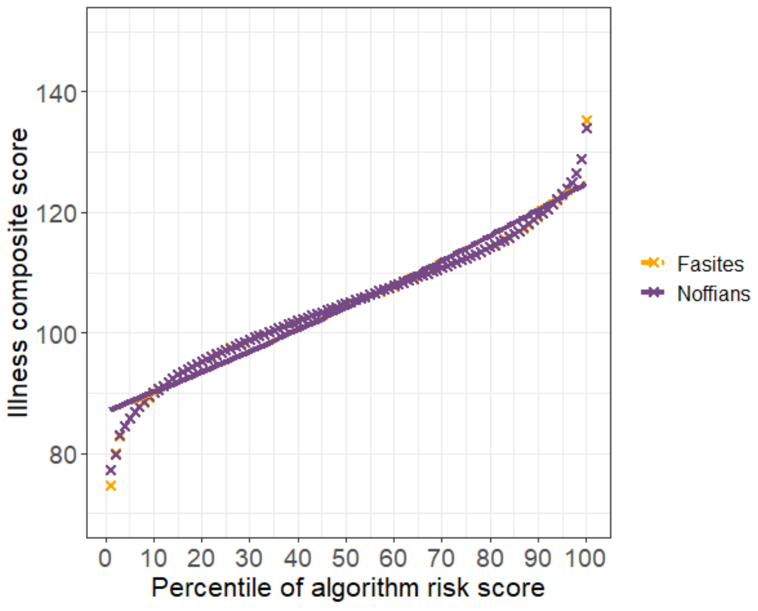
The label choice criterion for the scenario 1 model, which includes race, is shown below:

Scenario 1: Label choice criterion, including race



And the model that omits race gives:

Scenario 1: Label choice criterion, omitting race



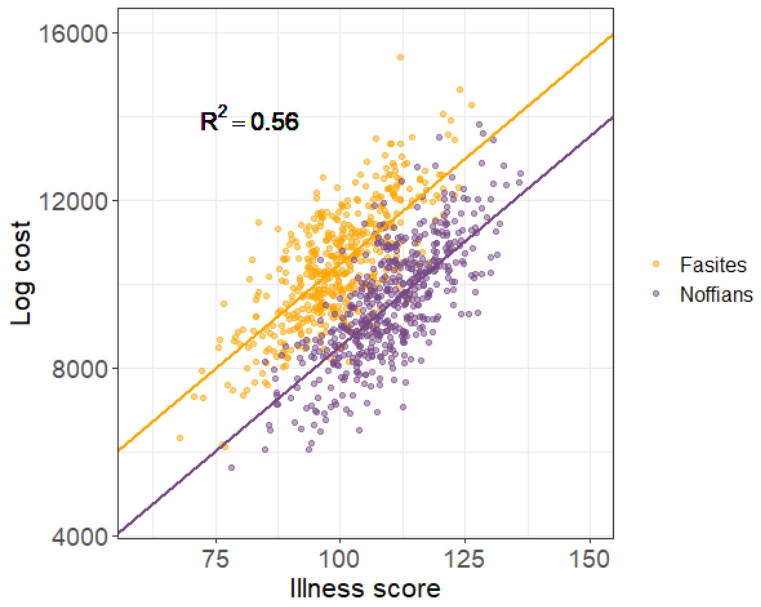
Thus, making the algorithm “race blind” enforces fairness. The reason this works in this simple example is transparent: If the only predictor is the patient illness score, then patients with the same illness levels will necessarily be assigned the same risk.

Excluding race is not the only strategy for addressing fairness. Some advocate for a bias-aware approach that includes the protected attribute during model training and then applies another procedure to measure and remove the influence of bias. In this simple example, the correction is straightforward, as one can simply add 2000 to each Noffian risk score.

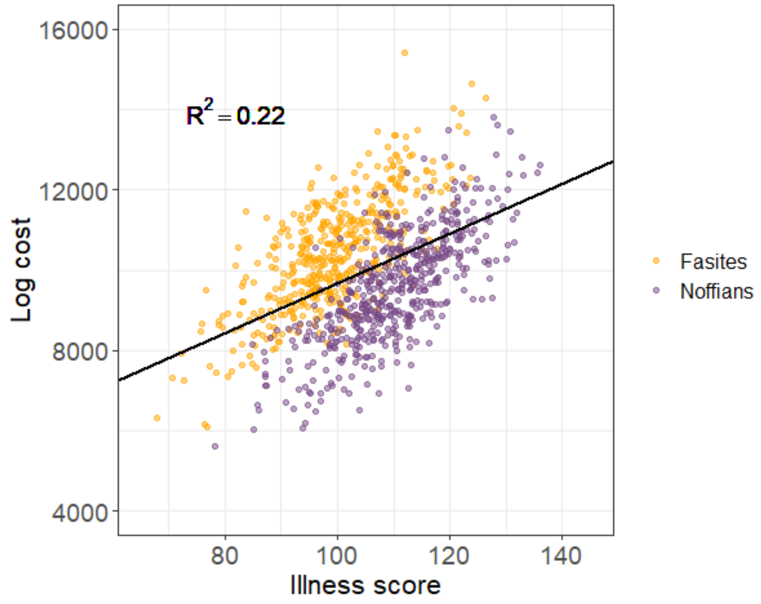
Forcing the model to ignore race is easy in our contrived example, but it is often quite difficult to achieve in practice. The problem is that protected attributes are likely to be correlated with other predictors. This allows for redundant encodings of those attributes.

An additional concern is that a trade-off between fairness and accuracy often must be made. Race is a confounding variable in our example, so omitting it will result in biased coefficient estimates. The accuracy trade-off between models that do and do not include race is shown in the plots below:

Total cost: Race included in the model



Total cost: Race omitted from the model

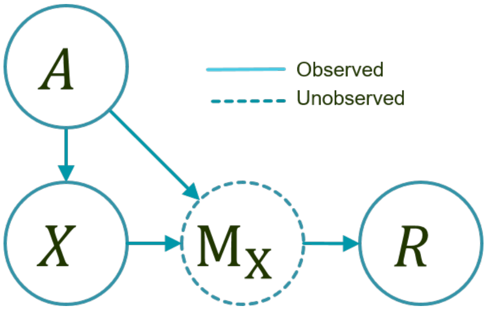


Some question whether fairness and accuracy are necessarily in tension with one another. In our example, the issue is that the labels are biased: Given their true health need, Noffians *should* receive more units of utilization.

It therefore seems dubious to strongly prioritize accuracy if there is a reasonable belief that the labels are partially corrupted.

Scenario 2: Avoidable costs outcome

The causal graph for the second scenario is:



In scenario 2, Noffian group membership causes:

- Increased health need, and
- Worse management of illness.

Additional assumptions are that:

- Avoidable costs increase linearly in managed illness.

The structural causal model is given by:

$$\begin{aligned} X &= 100 + 10A + \epsilon \\ M_X &= X - 10(1 - A) + \epsilon \\ R &= 500 + 50M_X + \epsilon \end{aligned}$$

Based on their higher managed illness levels, Noffians should average $1,000 more in avoidable costs. The R code below shows how data were generated for Scenario 2:

```r
# parameters
N <- 10000
sigma <- 500

# prop Noffian patients
prop <- 0.5
A <- rep(c(0,1), times=c((1-prop)*N, prop*N))

# greater reduction in managed illness for Fasites
set.seed(1)
X <- 100 + 10*A + rnorm(N,0,10)
Mx <- X - 10*(1-A) + rnorm(N,0,10)
R <- 500 + 50*Mx + rnorm(N,0,sigma)
# Model can also be expressed just in terms of X, but then it conceals that Mx is a resolving
variable

# models
train_data <- data.frame(A, X, Mx, R)
m1 <- lm(R ~ X, data = train_data)
m2 <- lm(R ~ X + A, data = train_data)
```