Hello everybody. Welcome to join the presentation, I am Jade Zhang. My coauthors are Zejian Wu, Yan shi, and Chenhao You. We are second year graduate student studying Business Analytics at Clark University. In consultation of Professor Hamid , our research concentrates on investigation of Potential Bias and Discrimination In The Development of A Typical AI Platform For Heart Transplantation.

## INTRO

Heart failure is a global pandemic affecting an estimated 26 million patients worldwide. It has no cure. In this case, Heart transplantation is the most effective treatment for patients with end-stage heart failure. We want to support Decision makers in their decision for organ matching by demonstrating potential bias in predicting survival length of a patient with a potential heart organ.

Right now, we investigate the results of the latest developed tool in heart transplantation, for any evidence of bias in gender or region.

###

Our Dataset was obtained from National registry of U.S. heart transplants from 1987-2016. Survival of patients after one year from transplantation surgery is predicted. The best model with the highest AUC) is selected from all of ten thousand eight hundred combinations. The training algorithm was Logistic Regression.

###

The survival tool is available here. You are welcomed to check it out if you are interested,

## Definition of Discrimination

First, regarding discrimination, lemme give a brief introduction of the term we are very familiar with – discrimination. It is Initially originated from Latin for 'distinguishing’, refers to an unjustified treatment of people based on some properties of some groups. Human rights laws prohibit discrimination in many aspects, such as race, origin, color, gender…and so on.

## Slide 3: Discrimination in machine learning

It is a common – yet pretty upsetting – phenomenon in out society based on human perception; and sometimes, even “fair” and “logical” machines, can unintentionally generate some unfairness for different groups, which is sometimes called digital discrimination.

[example here?]

There are two categories of discrimination: direct and indirect. Direct discrimination means people that are similar in terms of non-protected characteristics do not receive similar predictions. Let’s say we have a male and a female who share exactly the same characteristics except for gender, and they didn’t get the same predictions. In this case, this is direct discrimination.

And the indirect discrimination, however, refers to differences in predictions across groups of people is not as large as justified by non-protected characteristics. What does it mean, for example, banks run algorithms to determine whether they should issue loans to clients or not. Even though race is not used as a decision criterion, if we say non-white population is much more than while population among the denied clients, then we say there exists indirect discrimination. Here people in the same neighborhood are treated in the same way, but since the algorithm lowers positive rate of the non-white dominated neighborhood, it still causes unfairness toward non-white people.

Our research focuses on the investigation of the existence of indirect discrimination in our dataset.

## protected groups & targets

We selected gender and region of the patients as protected groups. According to our dataset, gender includes male and female, and region includes southeast middle west nirtheast**.** The targets in prediction problem are survival status, which is either 1 representing survival and 0 means not; and survival possibility, which is a number between 0 and 1. We use statistical tests to investigate the existence of indirect discrimination

in predicted survival status and survival possibility among gender and region

### existence tests

The two basic statistical tests are used here. First is the regression slope test. Regression slop test checks if there is significant linear relationship between independent variable X and target Y. The null hypotheses is there is not a significant linear relationship.

Ans the second is mean differences test, it tests if there is difference between two population means, and for region which has three groups, we use ANOVA to check if there is difference between three population means.

### methodology:

The specific methodology is as follows:

* Gender
  1. Select the data generated by Logistic Regression
  2. Test if there is significant bias between male and female in the actual survival status (0 and 1)
  3. Test if there is significant bias between male and female in the predicted survival rate (0 and 1)
  4. Test if there is significant bias between male and female in the predicted survival possibility (0 to 1)
  5. Compare the test results
* Region
  1. Similar process
  2. Detect the bias between Southeast, Middle east, and northeast

### methodology

I know it sounds a little bit complicated…so I made a diagram here for conveniences. For each patient, there are his or her actual survival status, and predicted survival status and possibilities generated by three algorithms. Then we separate the data in groups in terms of gender and region. Here I specify three groups as regions; it also can be two if it’s gender. Then we run the regression slop test and mean differences test to see if there is significant differences among groups for actual survival status. And then we do the tests again for the predicted survival status and predicted survival possibilities. Finally, we compare the test results. For example, if there is no significant bias among protected variables in actual survival status, but then there does exist evidence of difference in predicted result, then we can somehow prove that there exists bias among the groups.

### Result:

Here comes our first test and result. As we can see here, we accept the null hypothesis that there is no significant linear relationship b/t gender and actual survival status over all the years, however, we have to reject the same hypothesis for predicted survival status and predicted survival possibilities in most years.

The mean differences test shows the similar result. It implies there is no significant difference between males’ and females’ actual survival status, but there exits differences between males’ and females’ predicted results. These two tests somehow proved that there exists bias between gender groups.

Then we go through the similar process for region. For all the three regions, we mostly accept the null hypothesis that there is no significant linear relationship b/t region and actual survival status, but then we can see the number of rejection increased obviously for predicted survival status and predicted survival possibilities.

The mean differences test for three groups – ANOVA, shows the similar result. The differences in means of the three groups is mostly not significant. Yet in the predicted results, it becomes extremely significant. These two tests somehow proved that there exists bias between gender groups.

### Conclusion

Here comes our conclusion, first, as I already mentioned, there exist bias in both gender and region groups. After we checked though the data, we conclude that The highest performing Model was in favor of female and northeast region. And, based on this produced tool, Any scoring algorithm is prone to bias and needed to be optimized to reduce bias

### future studies

This project is still going on and we have some ideas to be completed. Since we have proved the existence of bias, we can then measure it and try to find mathematical improvements in algorithms by developing cost function for bias

Also, to better detect the reasons behind the bias, we will run algorithm separately for different groups to investigate if there exists overfitting/underfitting problem for some specific groups

Then we will perform the sensitivity analysis of accuracy and bias to see if there is any relationship.

That’s all. Thank you!