PULMONARY HEALTH CASE STUDY: BIAS EXPLORATION

Exploring Fairness in Machine Learning

Amit Gandhi, Olusobomi Olubeko, Richard Fletcher

D-Lab , Massachusetts Institute of Technology

MITD-Lab USAID FROM THE AMERICAN PEOPLE CITE

Acknowledgements

Clinical Study Sponsorship:

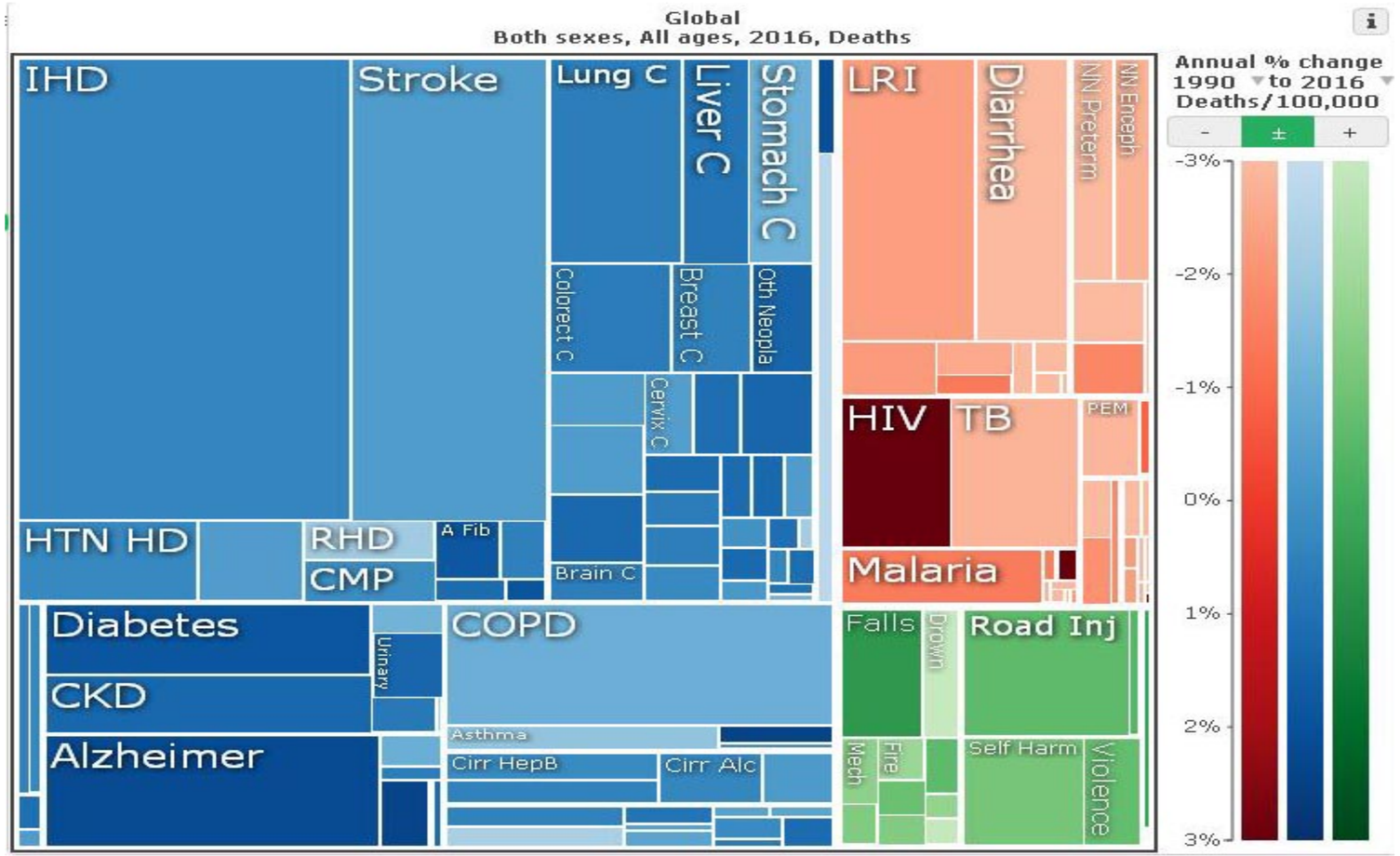
- **National Institutes of Health**
- **Vodafone Americas Foundation**
- **Tata Trust**

Clinical Partner:

- **Chest Research Foundation (Pune, India)**

Global Health Burden

Pulmonary Health



Study Overview

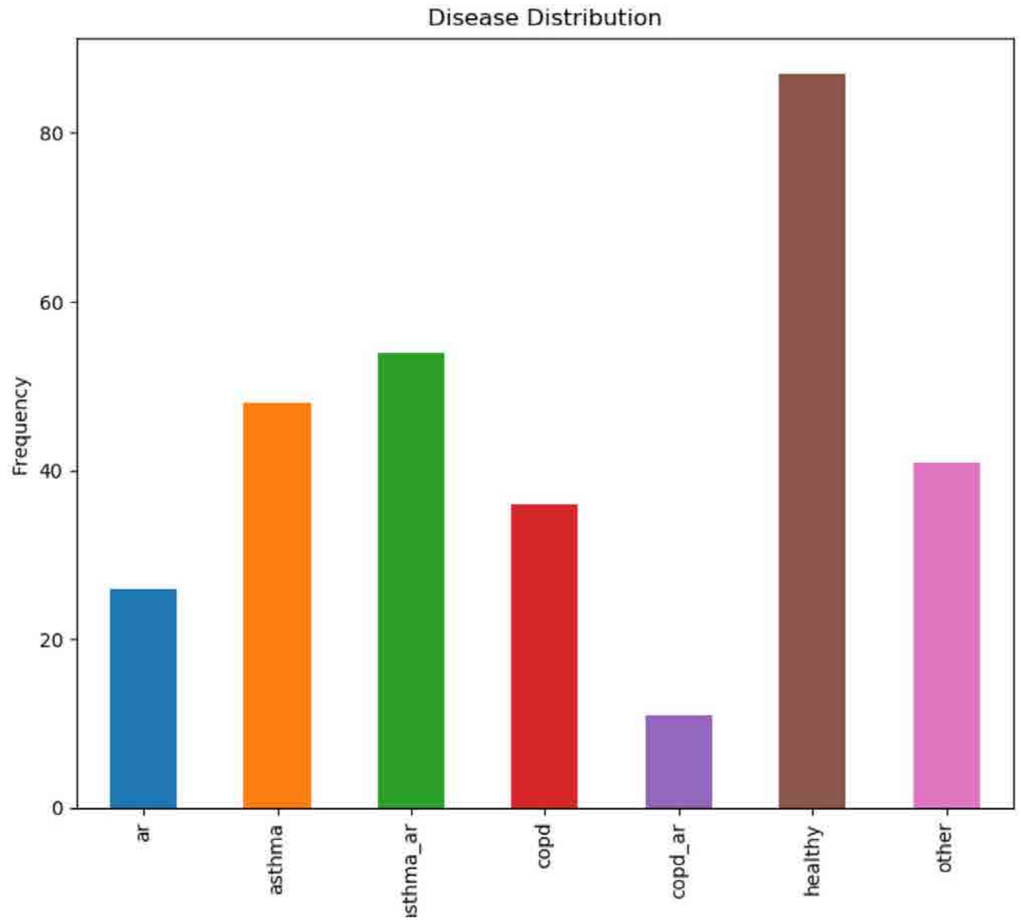
- 303 patients from Vodafone study
- Focused on diseases:

  -Allergic Rhinitis (AR)

  -Asthma

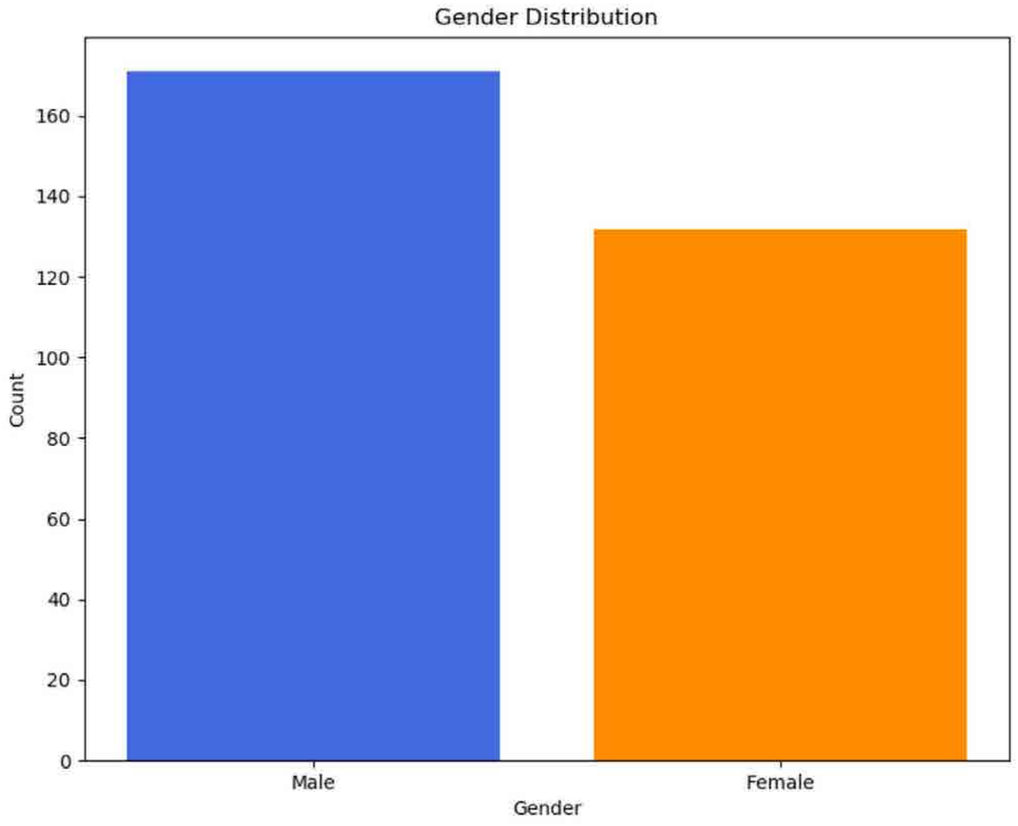
  -Chronic Obstructive Pulmonary Disease (COPD)
- Aim to explore effects of bias in gender and socioeconomic status

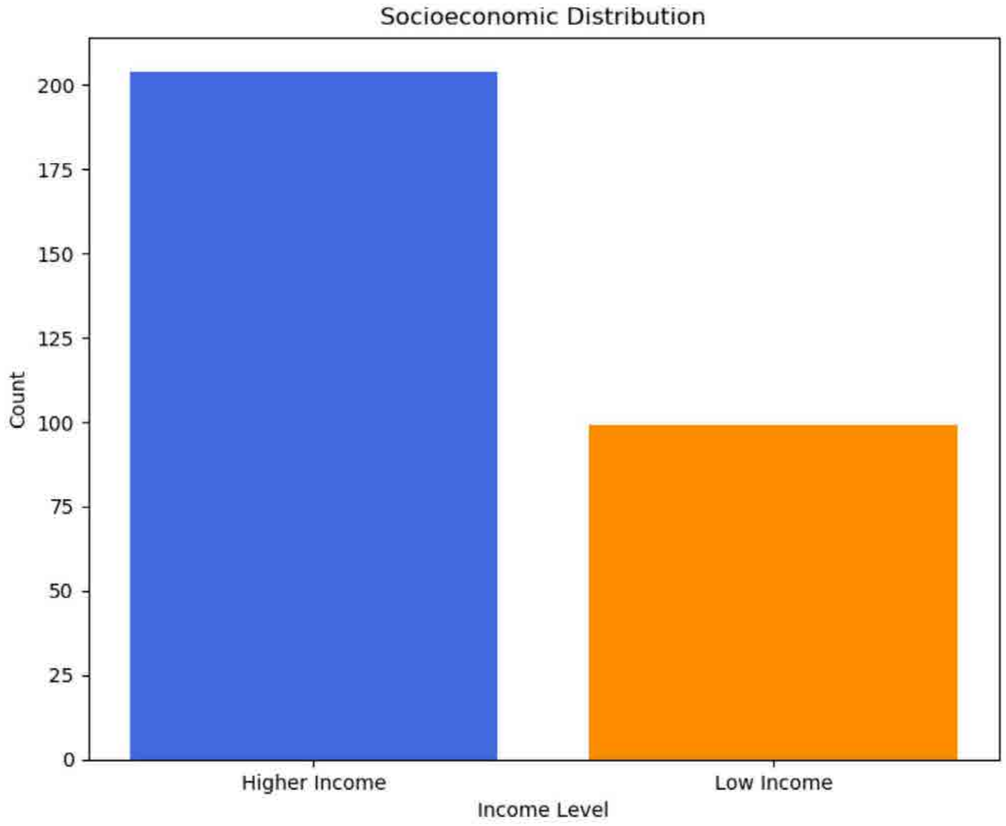
Gender and Income Distributions



- 26 AR Patients
- 48 Asthma Patients
- 54 Asthma + AR Patients
- 36 COPD Patients
- 11 COPD + AR Patients
- 87 Healthy Patients
- 41 Other Patients

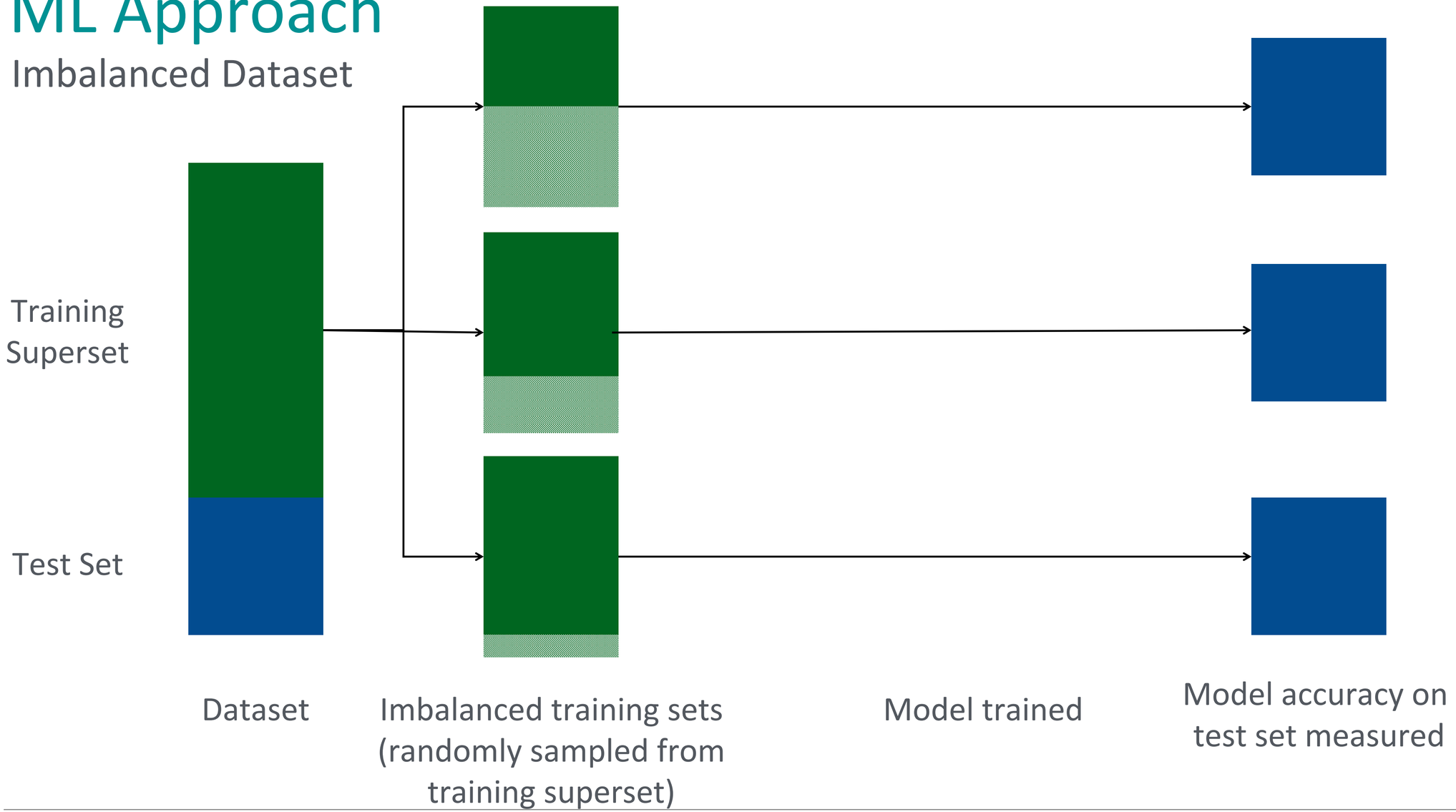
Overall Disease Distribution





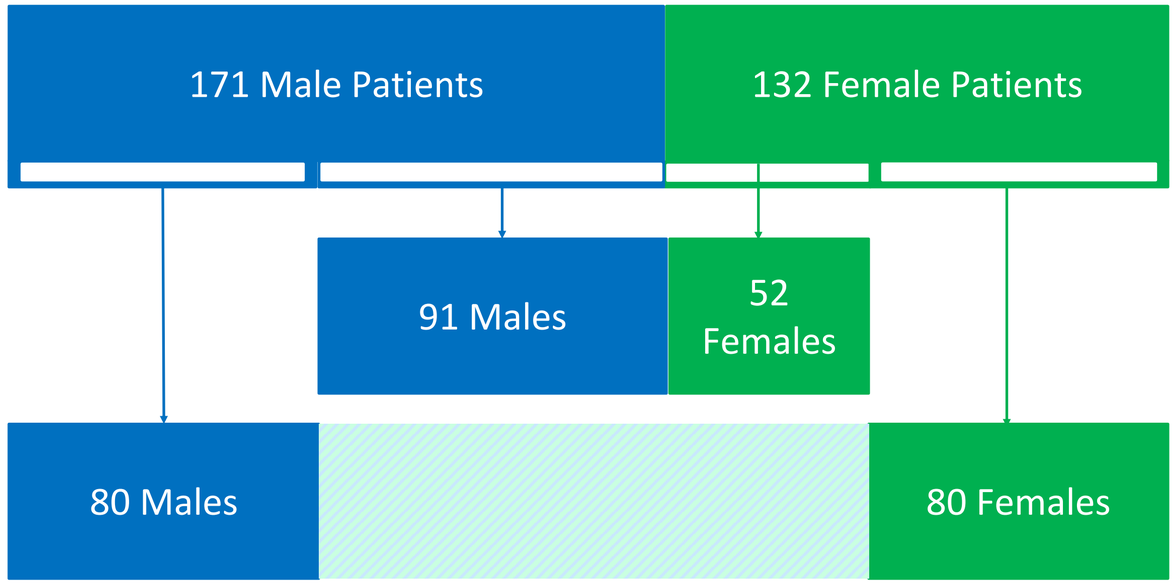
ML Approach

Imbalanced Dataset



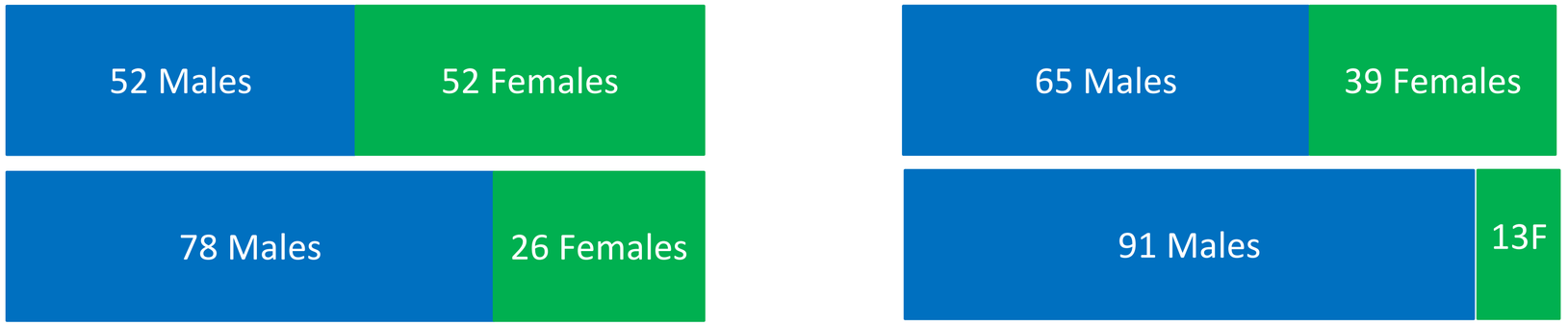
Gender Analysis

Design



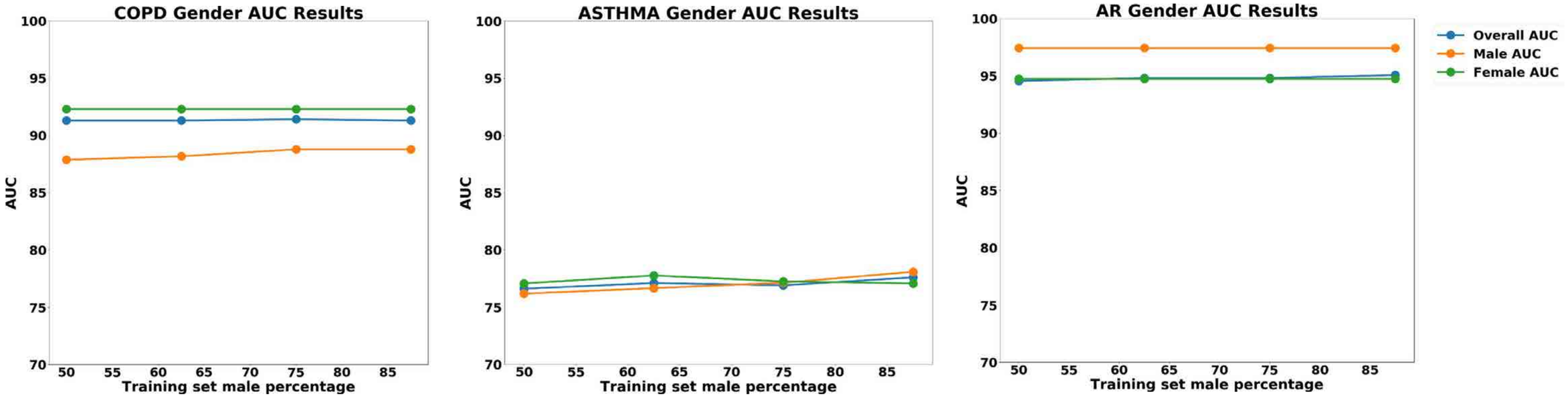
Randomly select 104 of 143 individuals from the training superset with ratios:

Repeat 1000x:



Gender Analysis

Results



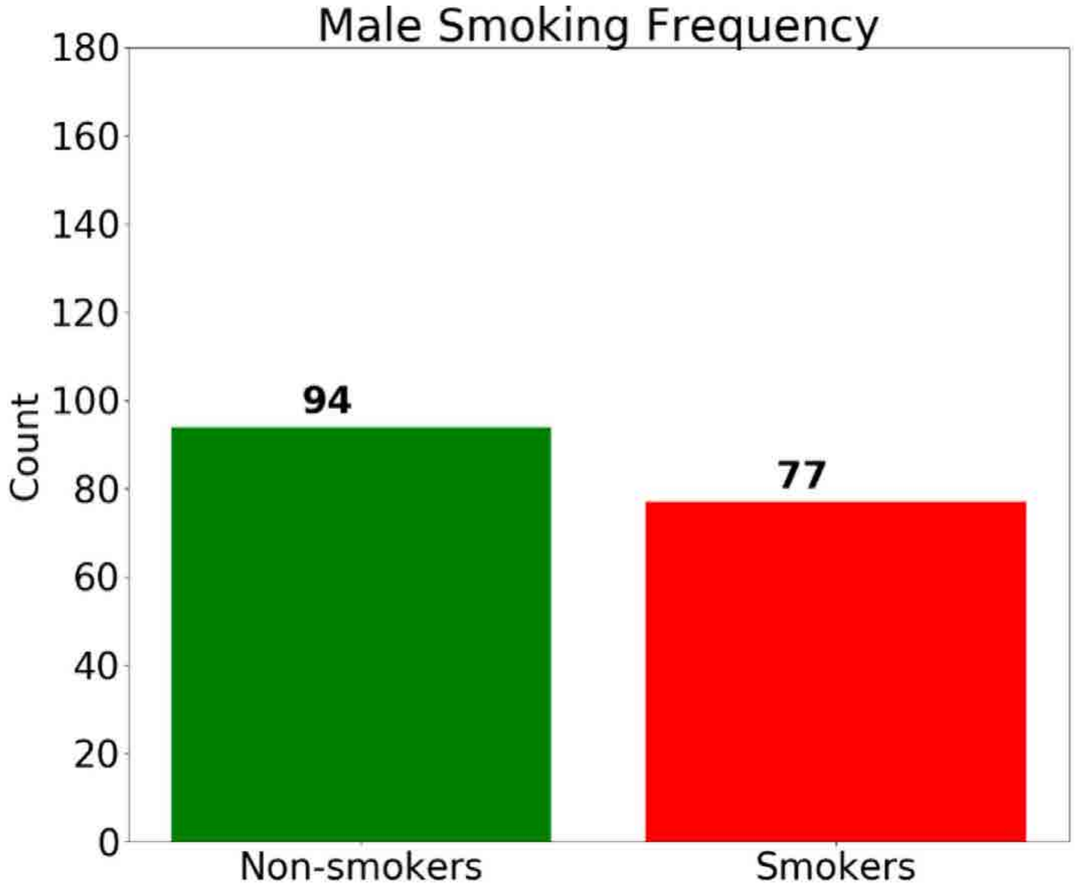
-COPD most sensitive to gender (6% variation)

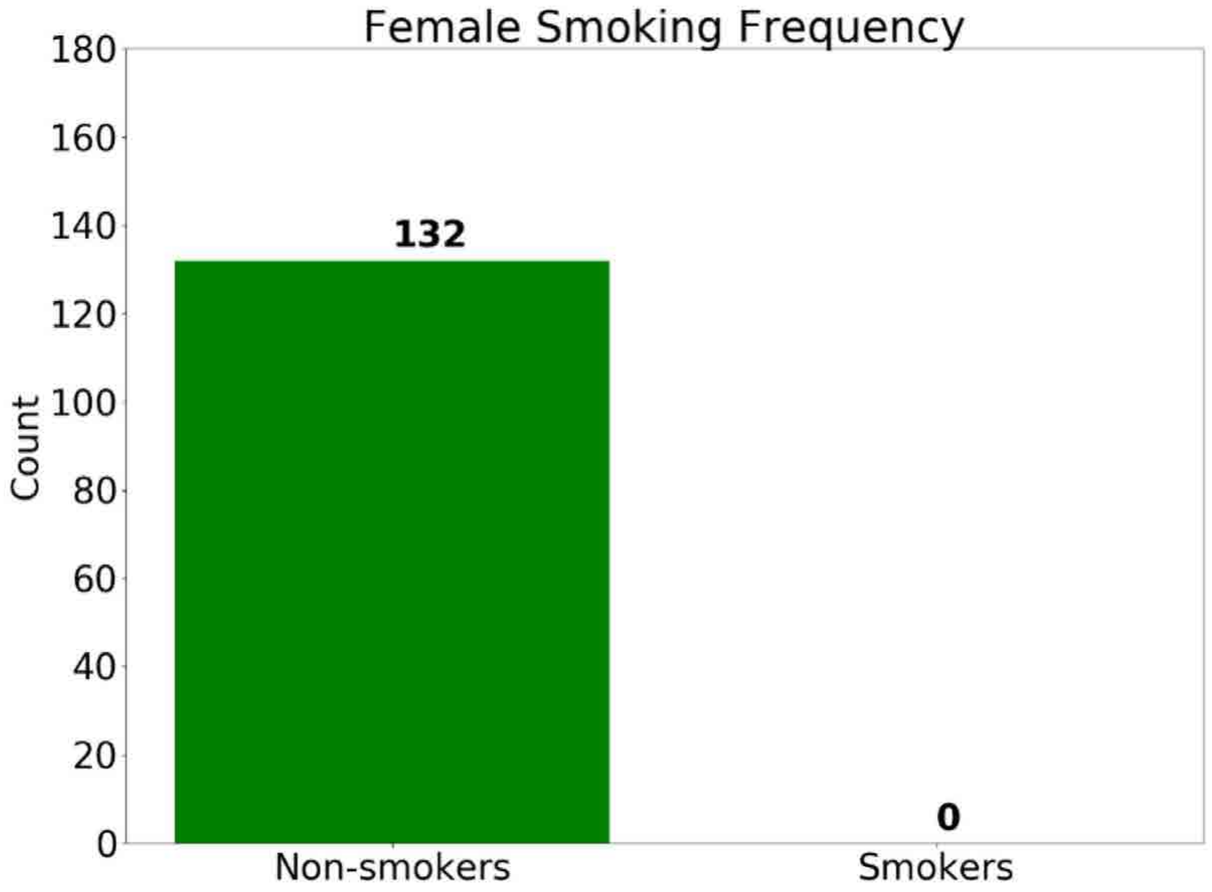
-Asthma has no gender differences

-Allergic Rhinitis has small sensitivity to gender, perhaps due to environmental exposure

Gender Analysis

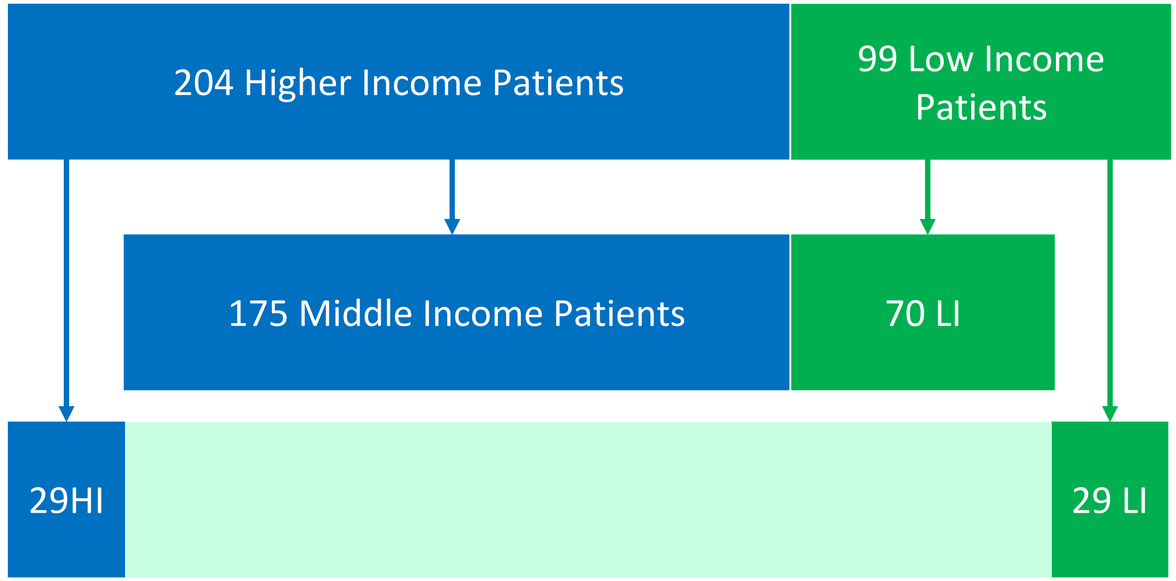
Smoking vs Gender





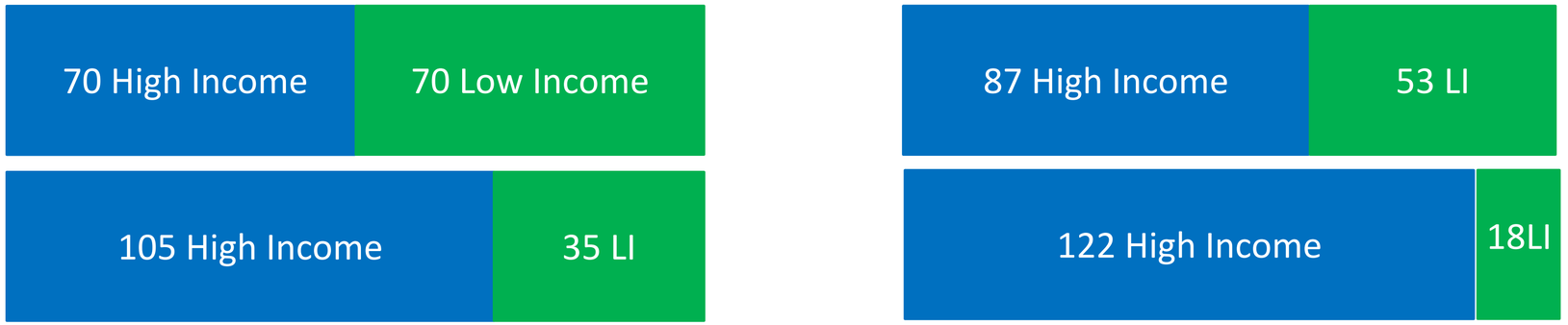
Income Analysis

Design



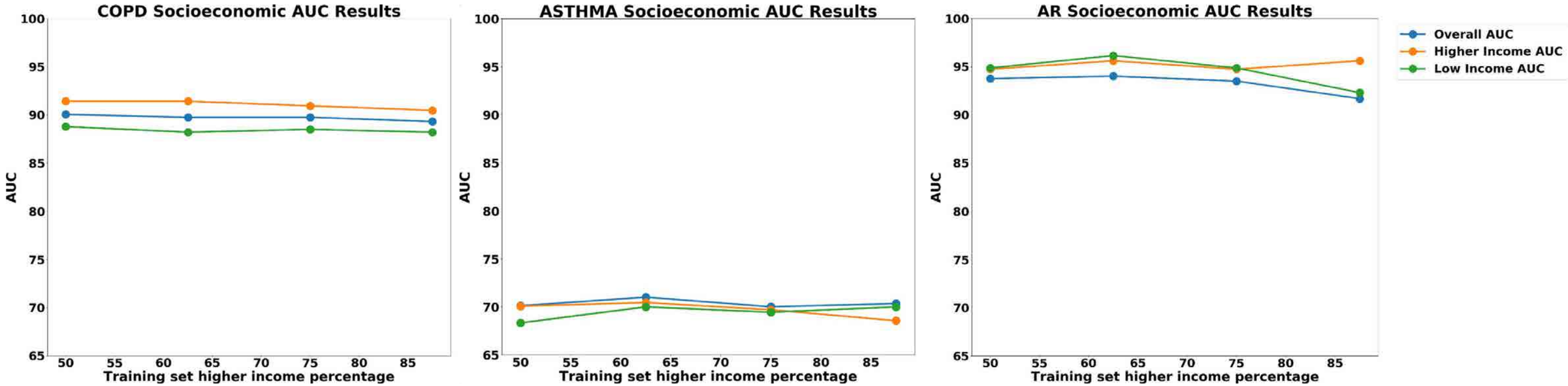
Randomly select 140 of 245 individuals from the training superset with ratios:

Repeat 1000x:



Income Analysis

Results



-COPD most sensitive to income (4% variation)

-Asthma and Allergic Rhinitis have no gender differences

Summary

- In this case study, balancing datasets across gender and socioeconomic status did not result in differences in model accuracy.
- Real-world datasets are often imbalanced – understanding the relative importance of balance for different protect variables will allow the analyst to make appropriate tradeoffs.
- Try to understand why variations in data/model accuracy exist: in this case we can most likely attribute them to smoking and environmental exposure.

Thank you

Amit Gandhi
Graduate Researcher, MIT
amitg@mit.edu

For health research questions:

Dr. Rich Fletcher
Research Scientist, MIT
fletcher@media.mit.edu

RES.EC-001 Exploring Fairness in Machine Learning

Spring 2019