Fair Phenotyping

Ryan King, Matthew Chang, Shima Salehi

Outline

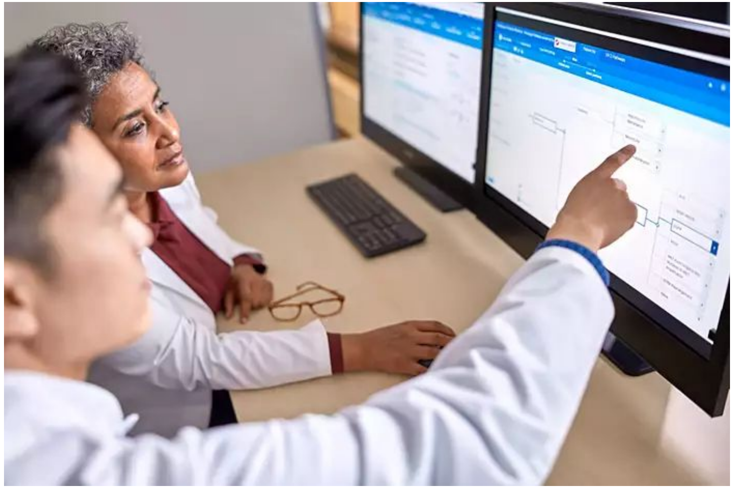
- ML bias in the medical field
- Introduction to Fairness
- Clustering with Fairlets
- Application for ICU dataset
- Methods
- Results

Griggs v Duke Power Co. (1971)

- Title VII of the Civil Rights Act of 1964
  - Prohibit discrimination based on race, color, religion, sex, and national origin
  - Explicit discrimination
- Disparate Impact
  - Decisions cannot affect protected classes disproportionately
  - Implicit discrimination

Fairness in Medical Data

- Protected groups: ethnicity, gender, socio-economic status
- No notable court cases of disparate impact
- Think about future impacts with growing use of ML

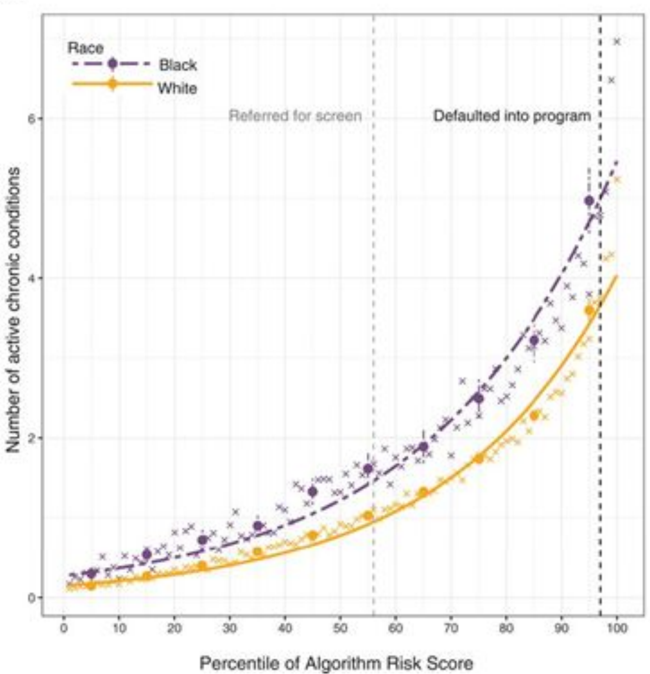


Obermeyer, et al 2019

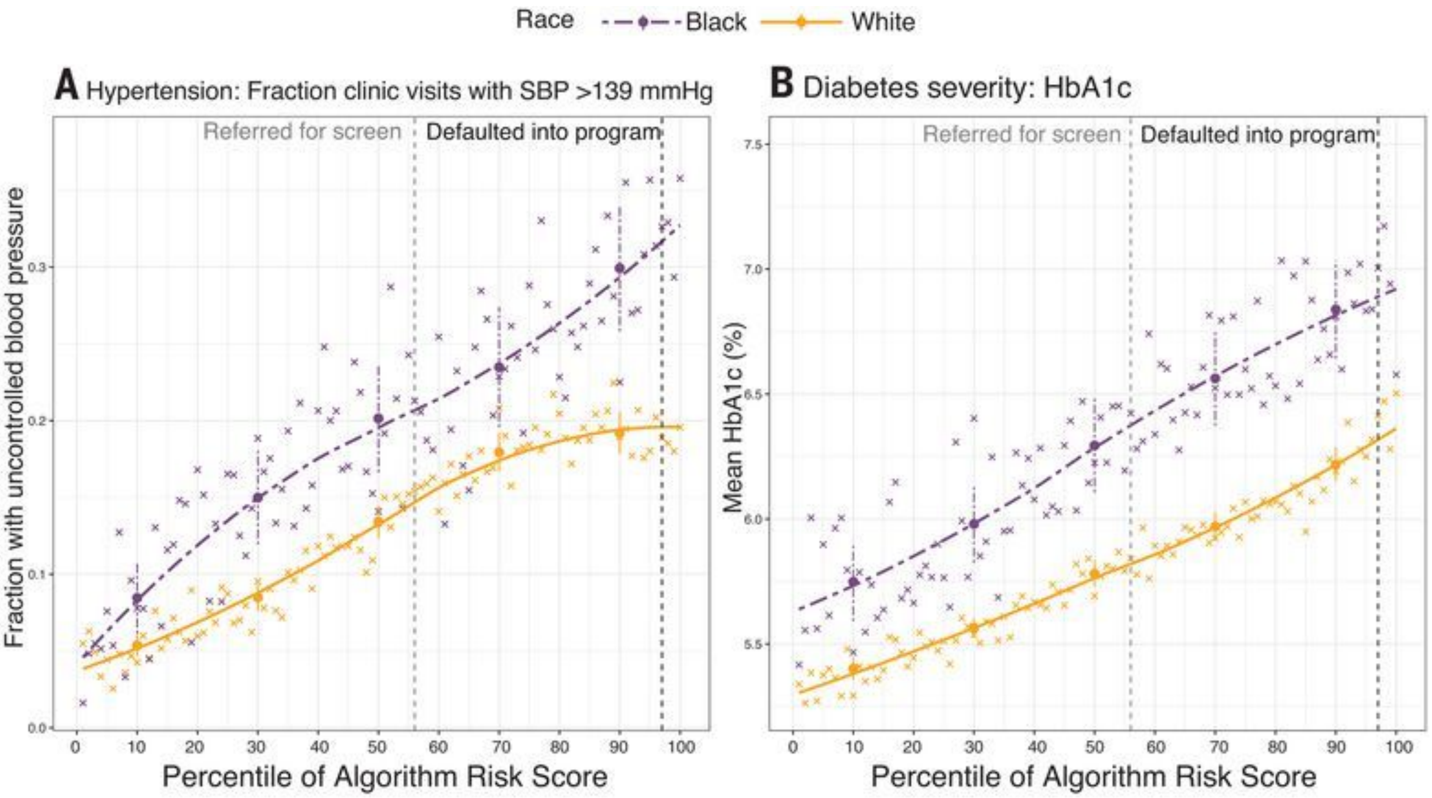
- Observed performance of widely used algorithm across industry, academia, governmental agencies, and non-profit hospitals
- Algorithm calculates a risk score which helps determine if individuals are qualified for a care coordination program

Obermeyer, et al 2019

- Less healthy blacks scored similarly to more healthy whites
- Patients at 97th percentile automatically identified for enrollment in care program



Obermeyer, et al 2019

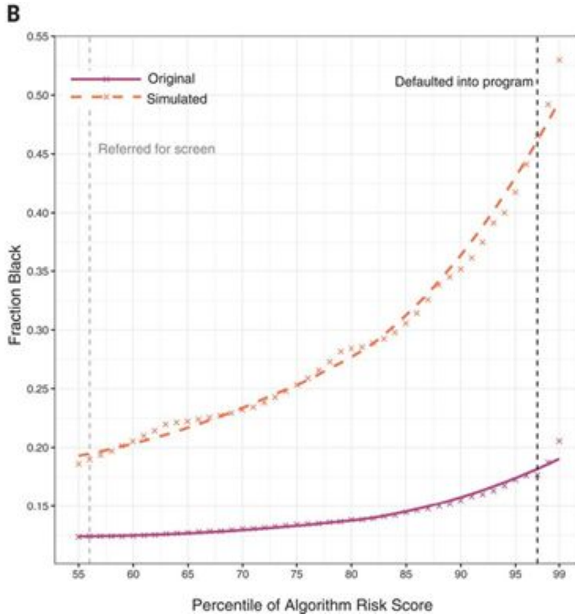


Obermeyer, et al 2019

- Issue: algorithm was trained on data from past healthcare spending
- Historically, blacks have spent less than whites
    - Spending more generally correlates to needing more care

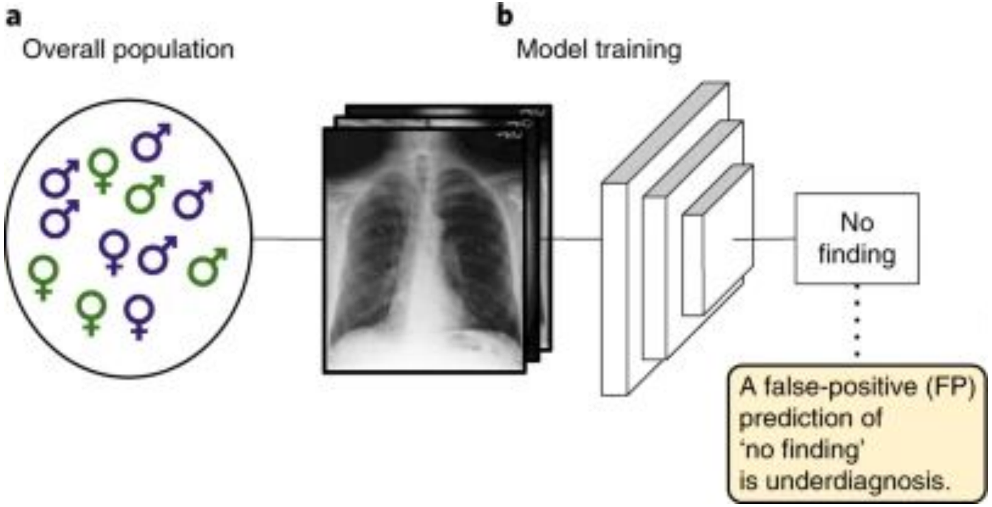
Obermeyer, et al 2019

- Algorithm has been corrected to be founded purely on health data
- Raises the question of how many other ML algorithms or datasets are biased and cause disparate impact between protected groups?



Seyyed-Kalantari, et al 2021

- Used X-ray dataset
- Algorithmic underdiagnosis in medical imaging

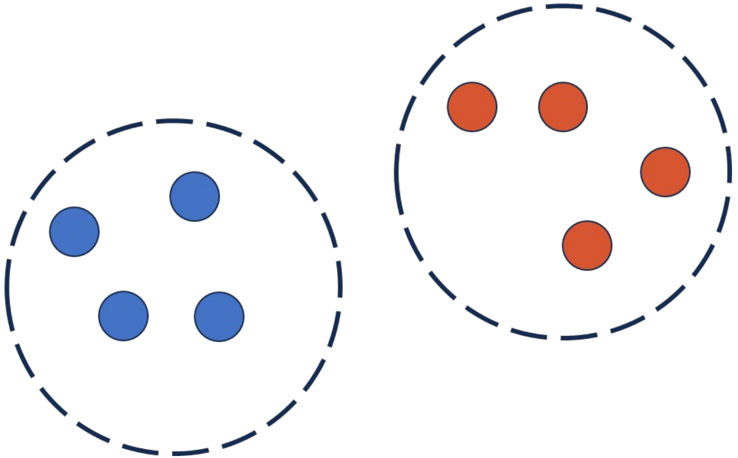


Seyyed-Kalantari, et al 2021

- Algorithm affects historically underserved populations
  - Females
  - Blacks
  - Hispanics
  - patients of low socioeconomic status
- Higher risk of being falsely flagged as healthy

Fairness in Clustering

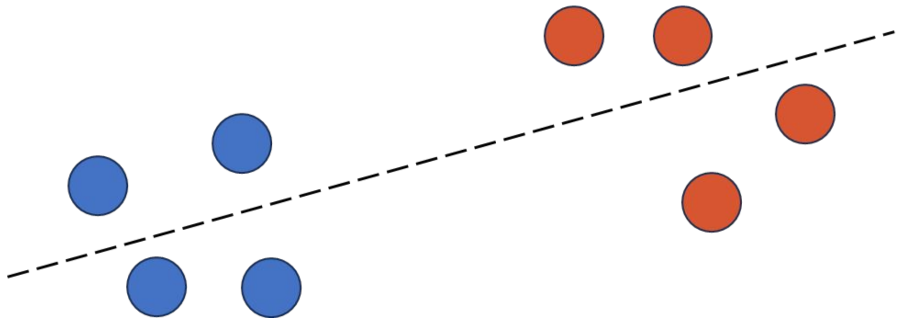
- Clustering Role: Enhancing basic learning methods via feature engineering.
- Objective: Ensuring fairness in generated features.
- Bias
- Unforeseen Results: Colorblind clustering can lead to unfair outcomes



Classical Partition "Color-blind"

Fairness in Medical Datasets

- Goal: Partitioning data into clusters while considering unprotected coordinates and 'color' attributes.
- Fair Representation: Aim for color balance within clusters.
- Complexity: Even in the two-color case, reveals significant intricacies.



“Fair” Partition

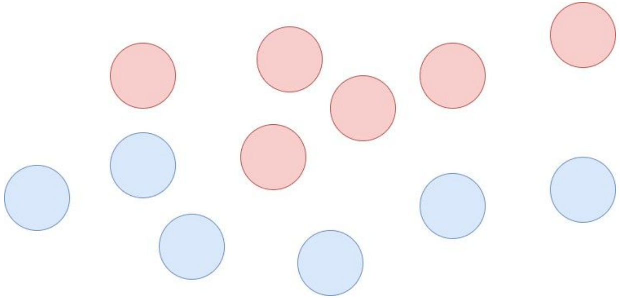
Fairness in Clustering Algorithms

- Deal with two kinds of clustering
  - Both are NP Hard

1. K-center
2. K-median

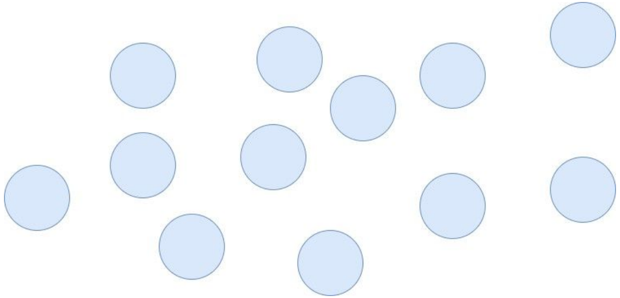
Balancing

$$balance(Y) = \min\left(\frac{\#RED(Y)}{\#BLUE(Y)}, \frac{\#BLUE(Y)}{\#RED(Y)}\right) \in [0,1]$$



Balance = 1

Perfectly Balanced



Balance = 0

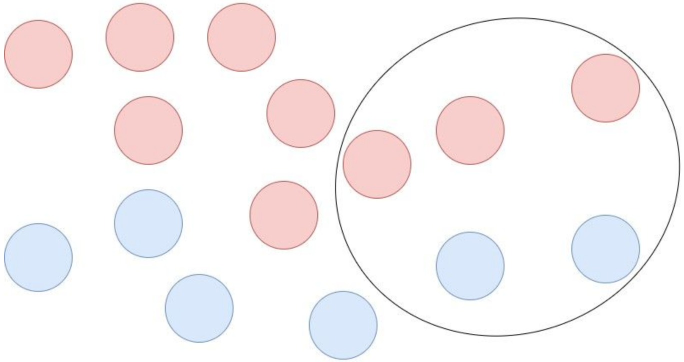
Monochromatic

Clustering with fairlet

Fairlets, subsets ensuring fairness within clusters

Reducing the clustering to the clustering of the fairlet

Specifically designed to maintain balance during the clustering, fairlets contribute to fairer clustering outcomes.



(R-B) = (9-6) ≥ (3-2) = (r-b)
|X| = (9 - 3) + (6 - 2) = 10
balance(X) = b/r = = 4/6 = 2/3

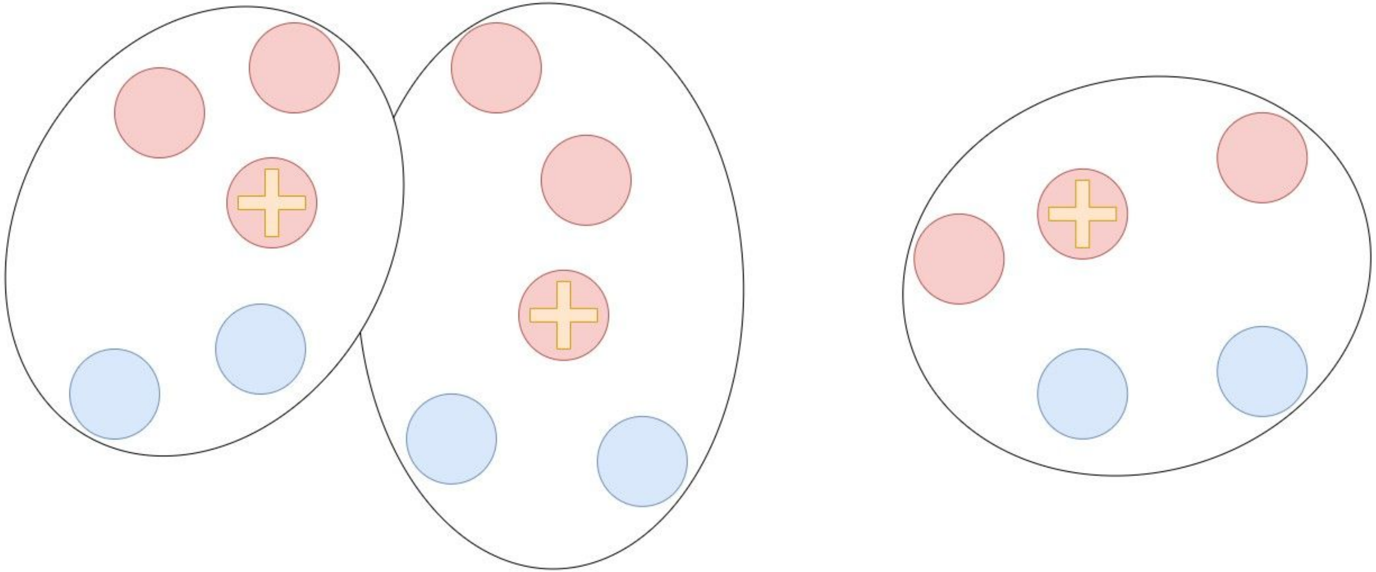
B=6, R=9

Balance = min(6/9,9/6)=0.667

Example

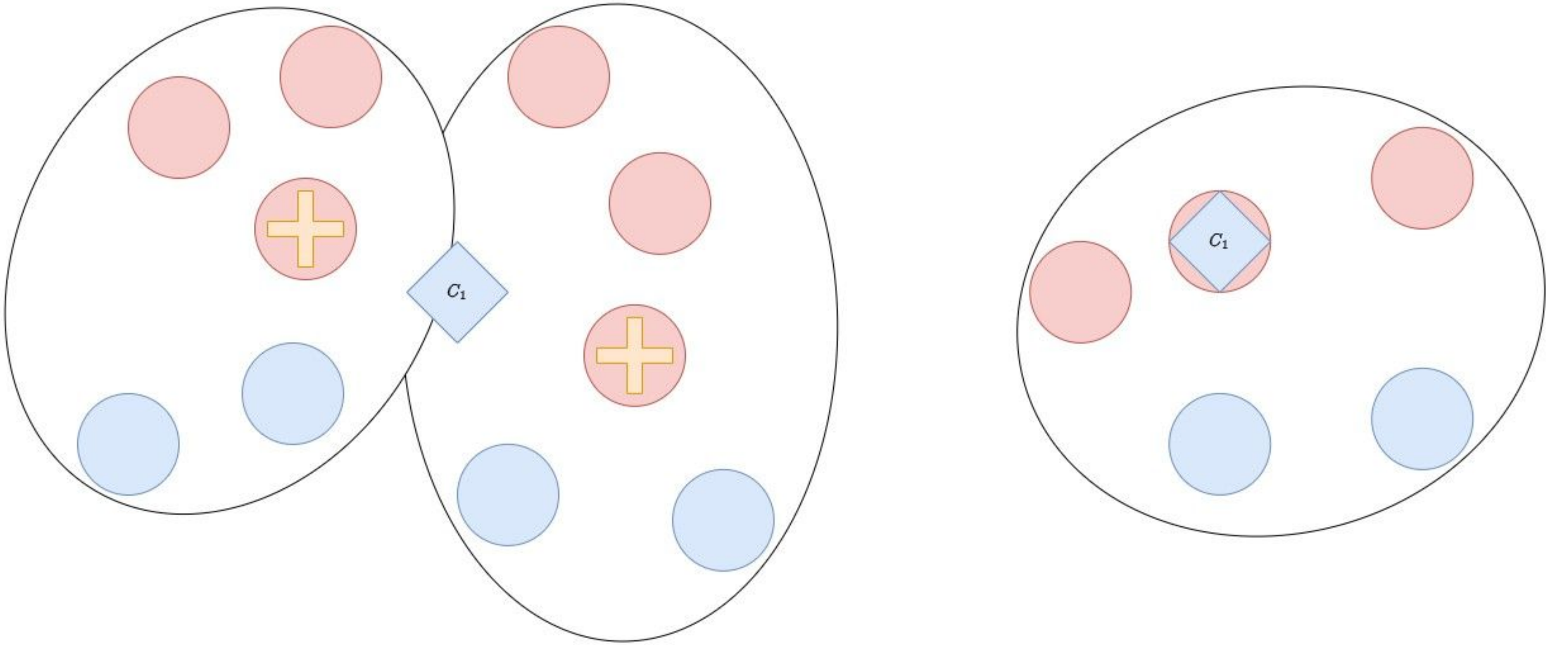
Continue Clustering

- Pick arbitrary centers



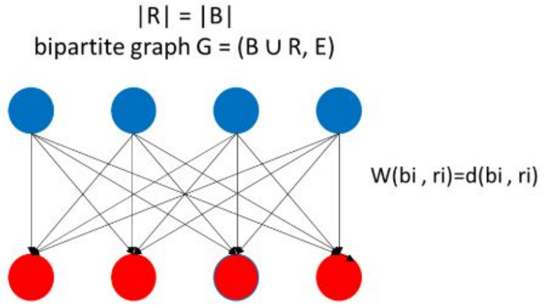
Cluster Your Clusters

Cluster your fairlet center (k-centers or k-medians)

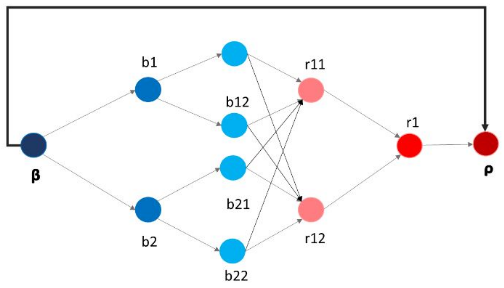


Finding fairlet decomposition

Graph covering : balance = 1 , binary search , polynomial



Using min cost flow : if balance < 1 , polynomial time



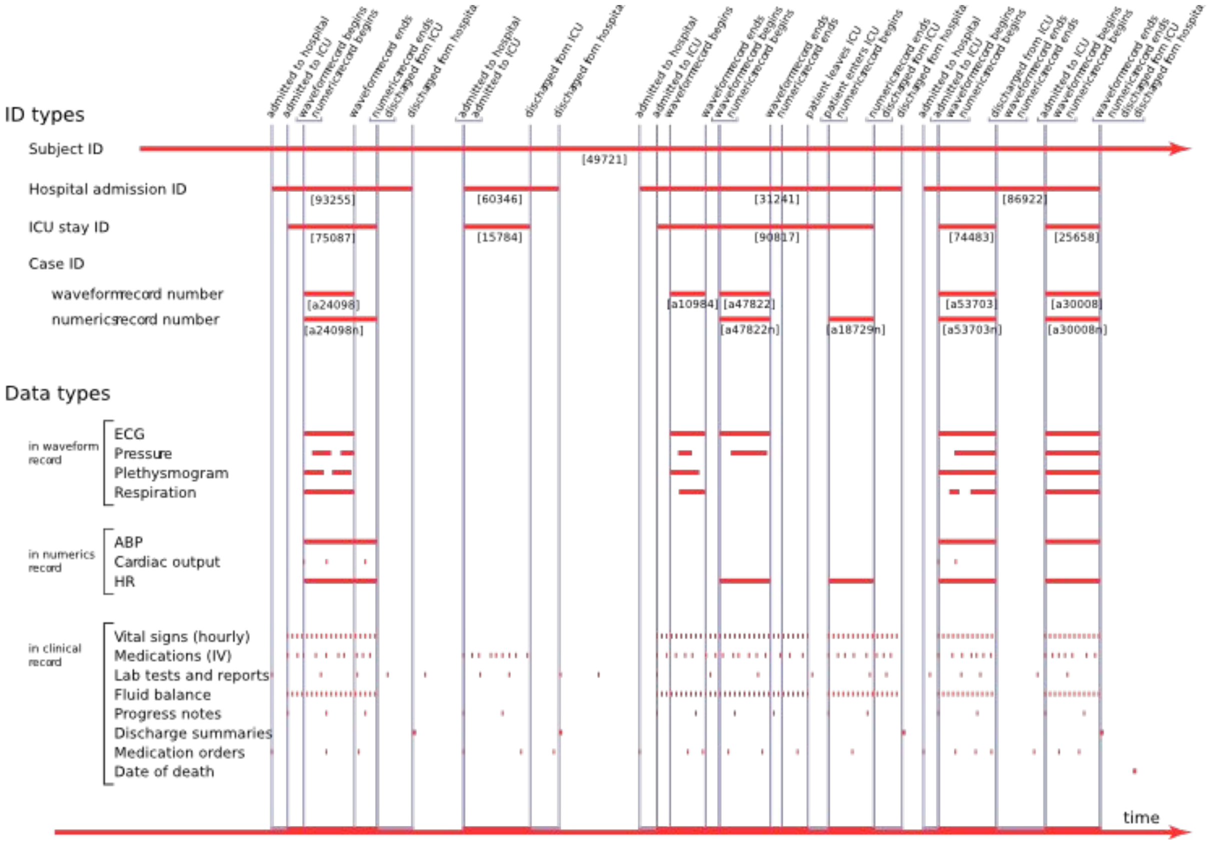
Application for ICU Dataset

- Avoid disparate impact on minority classes
- Our approach includes clustering with fairlets

MIMIC-III

- Medical Information Mart for Intensive Care III
- A publicly accessible critical care database
- The dataset contains de-identified health records of patients admitted to intensive care units (ICUs).

Patient Timeline



MIMIC-III Benchmark

Variable	MIMIC-III table	Impute value	Modeled as
Capillary refill rate	chartevents	0.0	categorical
Diastolic blood pressure	chartevents	59.0	continuous
Fraction inspired oxygen	chartevents	0.21	continuous
Glascow coma scale eye opening	chartevents	4 spontaneously	categorical
Glascow coma scale motor response	chartevents	6 obeys commands	categorical
Glascow coma scale total	chartevents	15	categorical
Glascow coma scale verbal response	chartevents	5 oriented	categorical
Glucose	chartevents, labevents	128.0	continuous
Heart Rate	chartevents	86	continuous
Height	chartevents	170.0	continuous
Mean blood pressure	chartevents	77.0	continuous
Oxygen saturation	chartevents, labevents	98.0	continuous
Respiratory rate	chartevents	19	continuous
Systolic blood pressure	chartevents	118.0	continuous
Temperature	chartevents	36.6	continuous
Weight	chartevents	81.0	continuous
pH	chartevents, labevents	7.4	continuous

MIMIC-III Benchmark

Phenotyping - Given a time series of clinical measurement from a patient's ICU stay, predict which phenotype

Train	Test
35621	6281

Phenotype	Type	Prevalence Train	Prevalence Test
Acute and unspecified renal failure	acute	0.214	0.212
Acute cerebrovascular disease	acute	0.075	0.066
Acute myocardial infarction	acute	0.103	0.108
Cardiac dysrhythmias	mixed	0.321	0.323
Chronic kidney disease	chronic	0.134	0.132
Chronic obstructive pulmonary disease	chronic	0.131	0.126
Complications of surgical/medical care	acute	0.207	0.213
Conduction disorders	mixed	0.072	0.071
Congestive heart failure; nonhypertensive	mixed	0.268	0.268
Coronary atherosclerosis and related	chronic	0.322	0.331
Diabetes mellitus with complications	mixed	0.095	0.094
Diabetes mellitus without complication	chronic	0.193	0.192
Disorders of lipid metabolism	chronic	0.291	0.289
Essential hypertension	chronic	0.419	0.423
Fluid and electrolyte disorders	acute	0.269	0.265
Gastrointestinal hemorrhage	acute	0.072	0.079
Hypertension with complications	chronic	0.133	0.130
Other liver diseases	mixed	0.089	0.089
Other lower respiratory disease	acute	0.051	0.057
Other upper respiratory disease	acute	0.040	0.043
Pleurisy; pneumothorax; pulmonary collapse	acute	0.087	0.091
Pneumonia	acute	0.139	0.135
Respiratory failure; insufficiency; arrest	acute	0.181	0.177
Septicemia (except in labor)	acute	0.143	0.139
Shock	acute	0.078	0.082
All acute diseases (macro-averaged)			
All mixed (macro-averaged)			
All chronic diseases (macro-averaged)			
All diseases (macro-averaged)			

Training

- Model: Bi-Directional LSTM
- Optimizer: Adam
- Epochs: 20
- Sampling: Uniform w/o replacement
- Objective Function: Binary Cross Entropy

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

Results of the Benchmark

Train a Bi-Directional LSTM on the phenotyping task

	Micro-AUC-ROC	Macro-AUC-ROC
Benchmark	81.43	76.22

Is this fair for all demographics?

MIMIC-III

- Frequency of top 10 ethnicities in MIMIC dataset

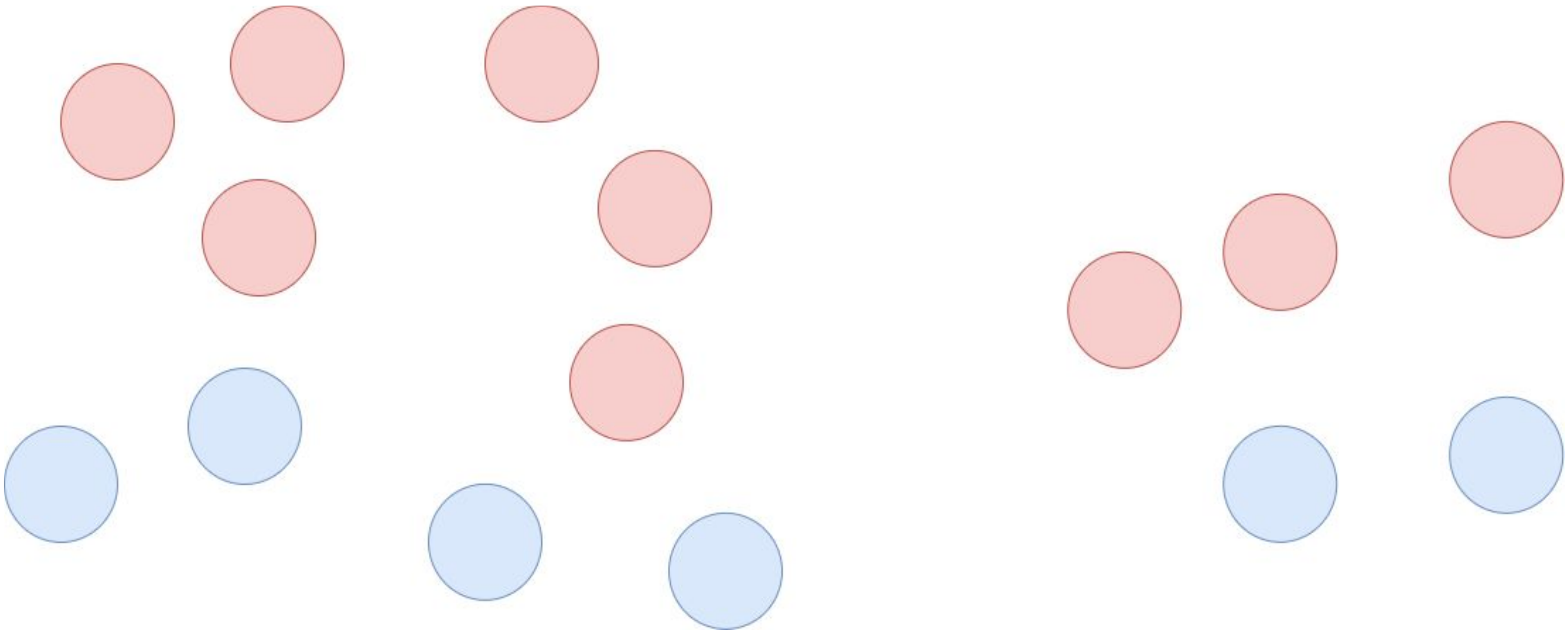
WHITE	76855
BLACK/AFRICAN AMERICAN	20304
UNKNOWN/NOT SPECIFIED	5683
HISPANIC OR LATINO	3138
OTHER	2518
ASIAN	2073
UNABLE TO OBTAIN	946
HISPANIC/LATINO - PUERTO RICAN	818
PATIENT DECLINED TO ANSWER	747
ASIAN - CHINESE	409

Results of the Benchmark

	Micro-AUC-ROC	Macro-AUC-ROC
Benchmark	81.43	76.22
African American	79.99	74.48
Other	81.57	76.29

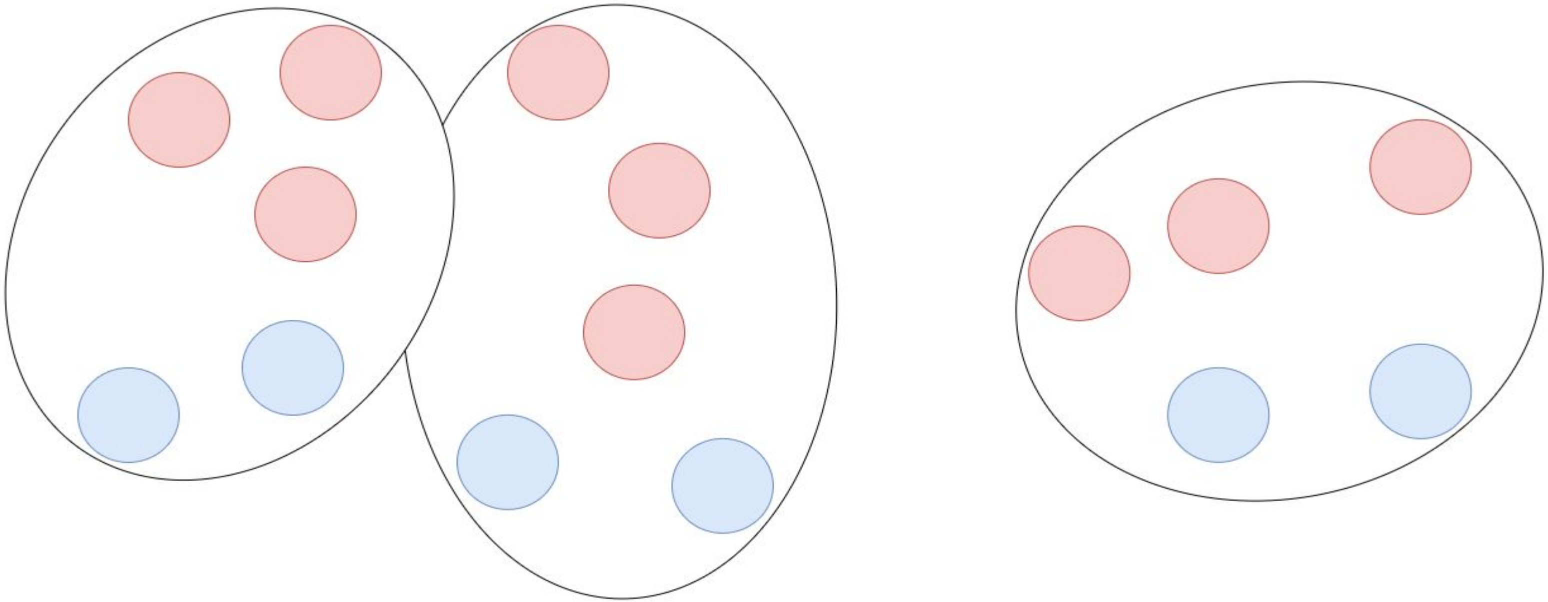
Fair Sampling

- Each fairlet has the same balance as the entire dataset
- Can we sample the fairlets during training



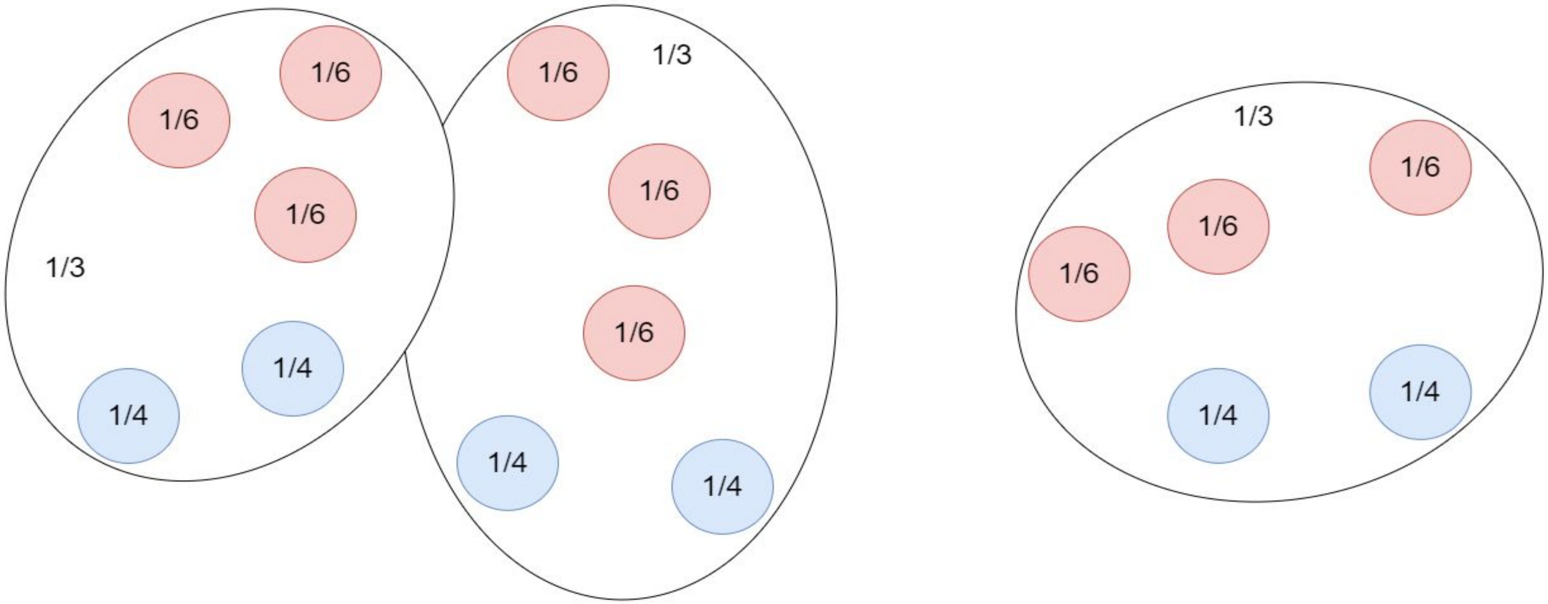
Fair Sampling

- Each fairlet has the same balance as the entire dataset
- Can we sample the fairlets during training



Fair Sampling

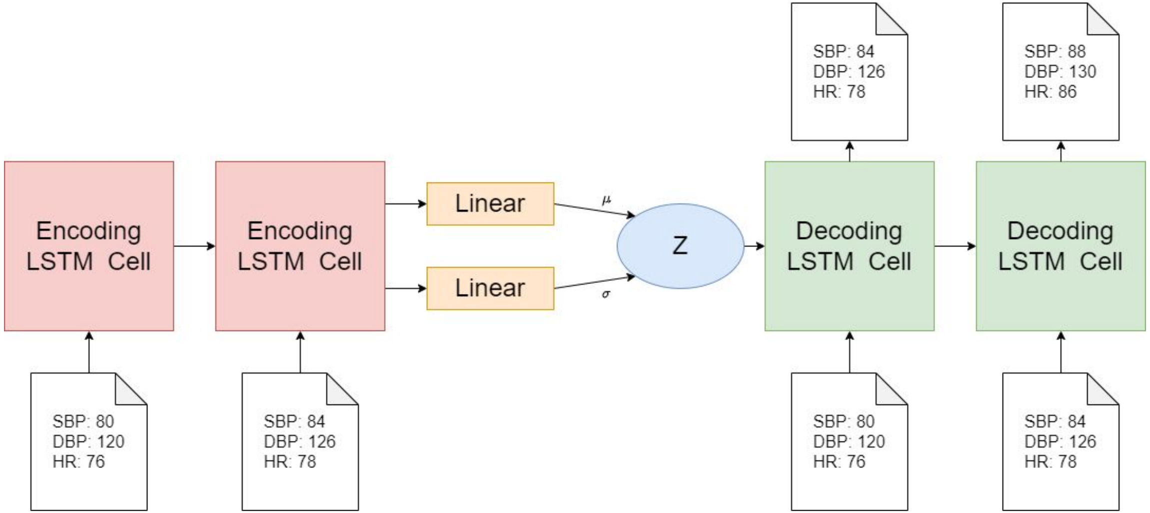
- Each fairlet has the same balance as the entire dataset
- Can we sample the fairlets during training



Creating Fairlets

We need a single vector representation for each ICU time series

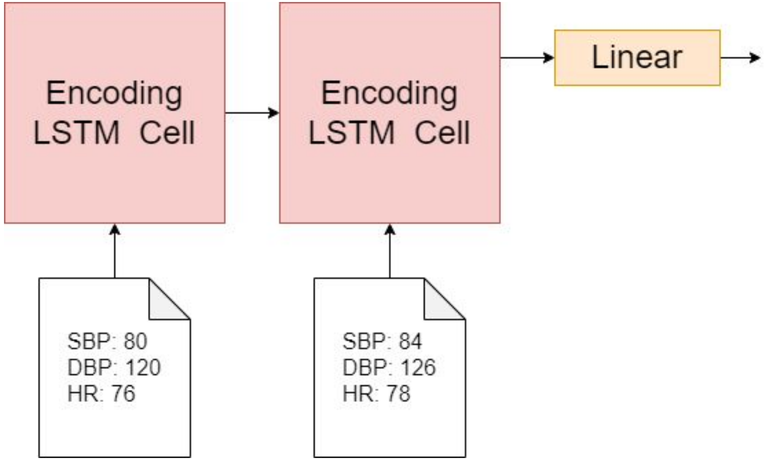
We start by training an LSTM-VAE [1] using the time series



[1] Bowman, Samuel R., et al. "Generating sentences from a continuous space." *arXiv preprint arXiv:1511.06349* (2015).

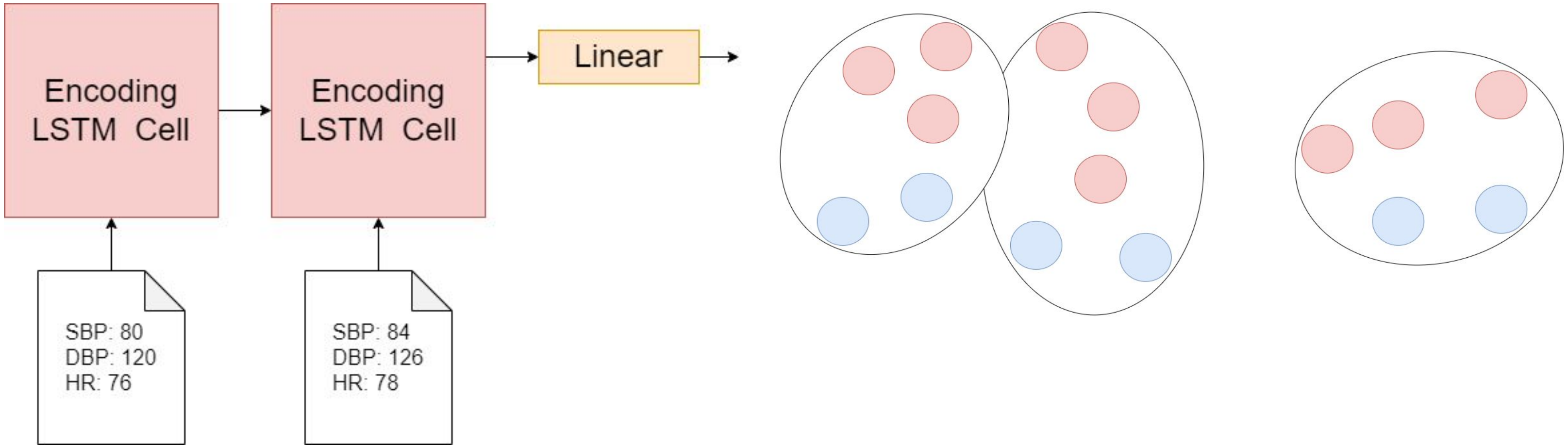
Creating Fairlets

We then use the encoder to transform our time series into a low-dimensional representation



Creating Fairlets

Create fairlets using the low-dimensional representation



Training

- Model: Bi-Directional LSTM
- Optimizer: Adam
- Epochs: 20
- **Sampling: Fairlet Sampling**
- Objective Function: Binary Cross Entropy

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$$

Results

	Sampling	Micro-AUC-ROC	Macro-AUC-ROC
Benchmark	Uniform	81.43	76.22
African American	Uniform	79.99	74.48
Other	Uniform	81.57	76.29
African American	Fairlet Sampling	79.97	74.17
Other	Fairlet Sampling	83.68	78.46

Questions