UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques i Informàtica
M.Sc. Data Science
Final Master's Thesis

A comparative study of fairness methods for clinical predictions using MIMIC-IV database

Author:
Iris VUKOVIC

Supervisor:
Laura IGUAL MUÑOZ

Introduction

- All decisions made by humans can be biased
- Using **fairness methods**, we try to mitigate existing bias in **machine learning models** for **clinical decision-making**

Clinical Tasks

Task 1:

hospital admission after emergency department (ED) stay

Task 2:

invasive mechanical ventilation (IMV) occurrence within the first 48 hours of a patient's stay in the ICU

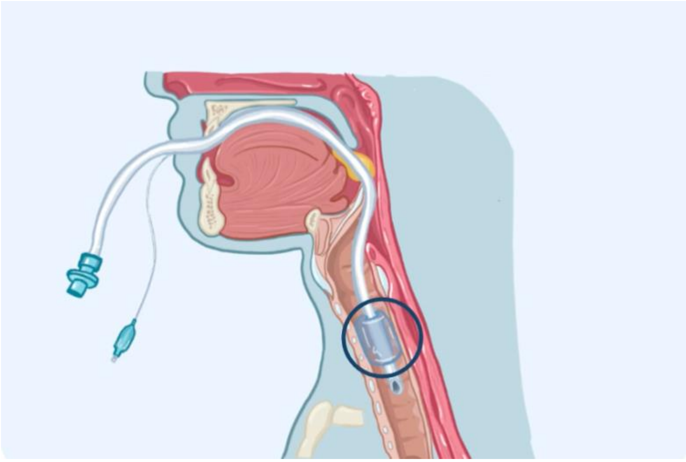
Clinical Task 2: What is IMV?

Task 2:

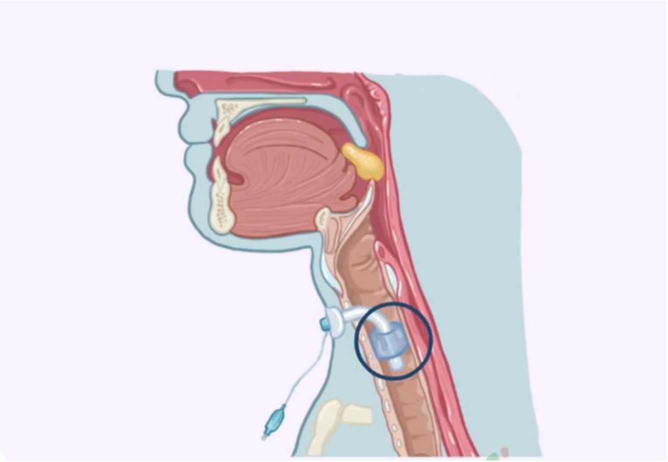
invasive mechanical ventilation (IMV) occurrence within the first 48 hours of a patient's stay in the ICU

IMV:

Standard



Tracheostomy



Non-IMV:

Face mask

Nasal plug

Helmet

Clinical Task 2: Motivation

Task 2:

invasive mechanical ventilation (IMV) occurrence within the **first 48 hours** of a patient's stay in the ICU

- Inspired by study (Abdelmalek et al., 2024) which found that in **MIMIC-IV** database, there were **different rates of IMV** for patients with respiratory failure based on race:

Lower IMV rates
- Black
- Asian
- Hispanic

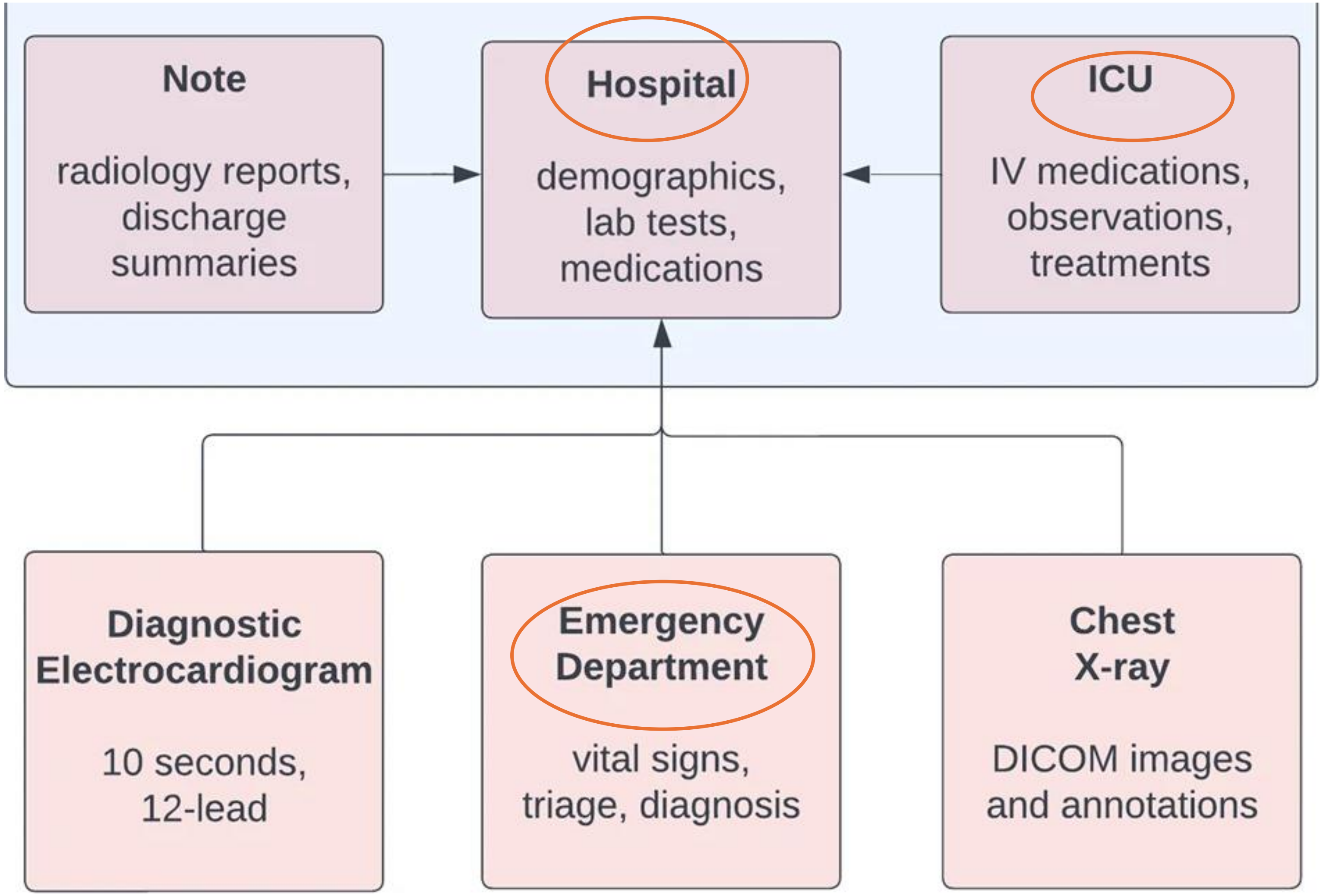
Higher IMV rates
- White

- Time constraint was chosen to mirror the **urgency of real-world decision-making timeframes** in clinical settings

Abdelmalek *et al.*, "Association between patient race and ethnicity and use of invasive ventilation in the United States," Feb. 2024.

MIMIC-IV Database (2024)

Medical Information Mart for Intensive Care (MIMIC)



- Collection of **electronic health records** from patients between 2008 and 2022 who had a stay in:
  - ***intensive care unit (ICU)***
  - ***emergency department (ED)***

Massachusetts Institute of Technology

Sensitive Attributes

Intersectionality: the way that social categorizations interact to create **unique biases** different than those stemming from individual sensitive attributes

Race

- **@White**
- **Black**
- **Asian**
- **Hispanic**
- **Other**

Sex

- **@Male**
- **Female**

Intersections

- **Sex x Race**

- **Female x Black**
- **Female x Hispanic**
- **@Male x Other**

.
.
.
.

Sensitive Attributes

Race

- **@White**
- **Black**
- **Asian**
- **Hispanic**
- **Other**

Sex

- **@Male**
- **Female**

The most privileged subgroups are marked with an **@** symbol in the results tables as the reference groups

Reference Group Encoding:

- **No categorical variable representation is created for them.**
- The model learns how being Female, Black, Hispanic, Asian, or Other affects outcomes **relative to being Male or White**

FAIM: Fairness-Aware Interpretable Modeling

- Traditional fairness methods have **trade-off** between :

• **model performance** • **transparency**	vs.	• **bias mitigation**

- **Fairness-aware interpretable modeling (FAIM):**
  - fairness method that **mitigates demographic bias** AND maintains **performance** and **transparency**

FAIM Framework

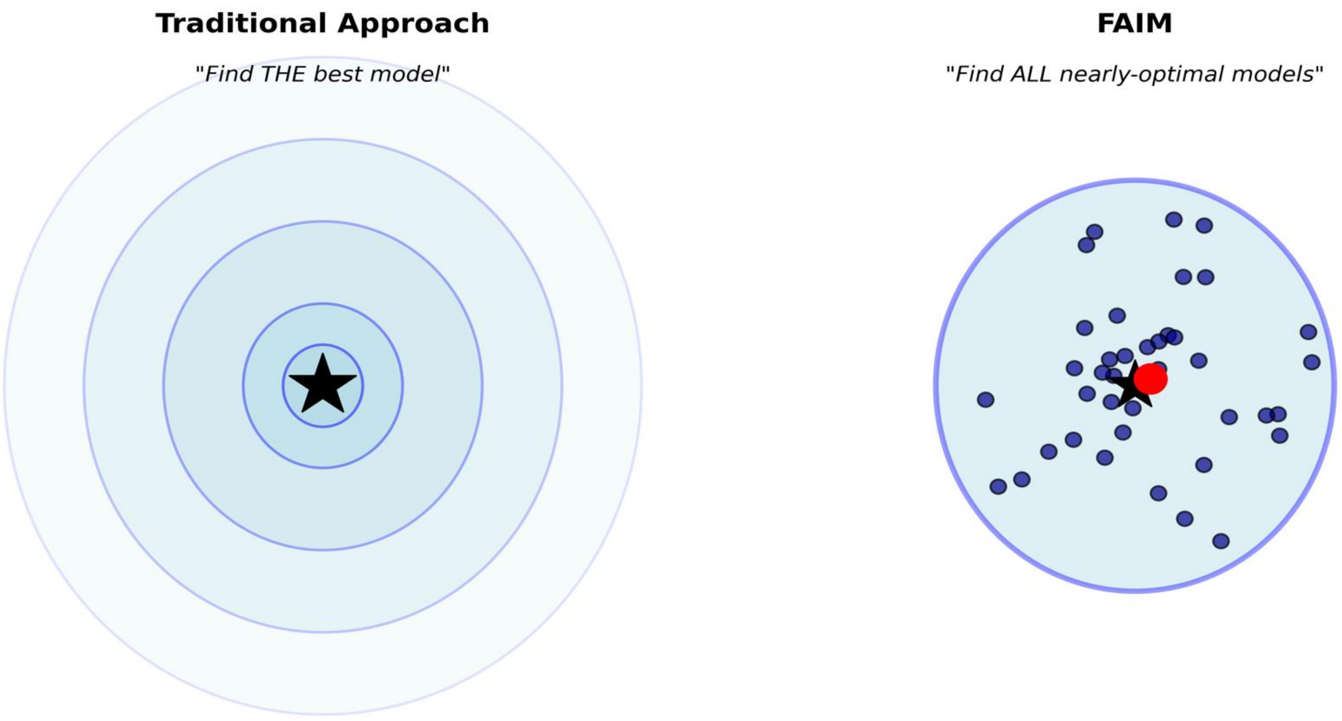
- FAIM generates a set of **nearly-optimal models** using the **ShapleyVIC** algorithm for each of four attribute-exclusion scenarios:
  - **no exclusion**
  - **sex exclusion**
  - **race exclusion**
  - **sex and race exclusion**

- Based on **Shapley value/SHAP** and **Rashomon Effect, (**Liu et al., 2024) expanded SHAP framework into **Shapley Variable Importance Cloud (VIC)**

Liu *et al.*, "FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare," *Patterns*, vol. 5, Oct. 2024.

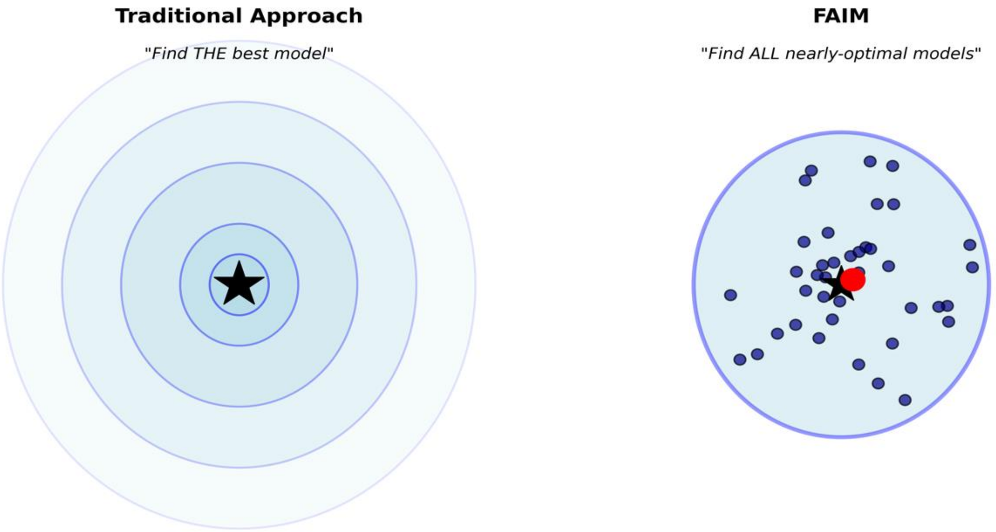
Nearly-optimal Models

- Generates set of models that fall within a threshold of up to 5% degradation from optimal area under the curve (AUC)



Nearly-optimal Models

- Generates set of models that fall within a threshold of up to 5% degradation from optimal area under the curve (AUC)



- **Fairness metrics** are evaluated on the validation set for the selection of a final fairness-aware model

Fairness Metric

- **independence-based** metrics only ensure equal prediction rates **regardless of actual patient outcomes**, potentially sacrificing diagnostic accuracy

- FAIM ranks models based on three **separation-based** fairness metrics for binary classification problems:
  - evaluate whether the model **performs equally well** across demographic subgroups

Equalized Odds

- balance **TPR** and **FPR** across subgroups

Equal Opportunity

- balance **TPR** across subgroups

Balanced Error Rate (BER) Equality

- balance **FPR** and **FNR** across subgroups

Fairness Metrics

Equalized Odds

- balance **TPR** and **FPR** across subgroups

Equal Opportunity

- balance **TPR** across subgroups

Balanced Error Rate (BER) Equality

- balance **FPR** and **FNR** across subgroups

Fairness Ranking Index (FRI)

- a measurement developed by (Liu et al., 2024) that aggregates those fairness metrics into one conclusive score

Liu *et al.*, "FAIM: Fairness-aware interpretable modeling for trustworthy machine learning in healthcare," *Patterns*, vol. 5, Oct. 2024.

Fairness Metrics: FRI

Fairness Ranking Index (FRI)

- a measurement developed by (Liu et al., 2023) that aggregates those fairness metrics into one conclusive score

$$FRI = \frac{1}{\Sigma\, metric_i \times metric_j + \varepsilon} \quad where\ i, j \ \in \{EqOdds, EqOpp, BEREq\}$$

- a **higher FRI** score indicates a **fairer model**
- when two metrics are large, indicating **significant bias**, their product is even larger, therefore more drastically **shrinking the overall FRI**

Results: Dataset sizes

Task 1: Hospital Admission

Split	**N**
Overall	**418,025**
Training (70%)	292,617
Validation (10%)	41,802
Test (20%)	83,606

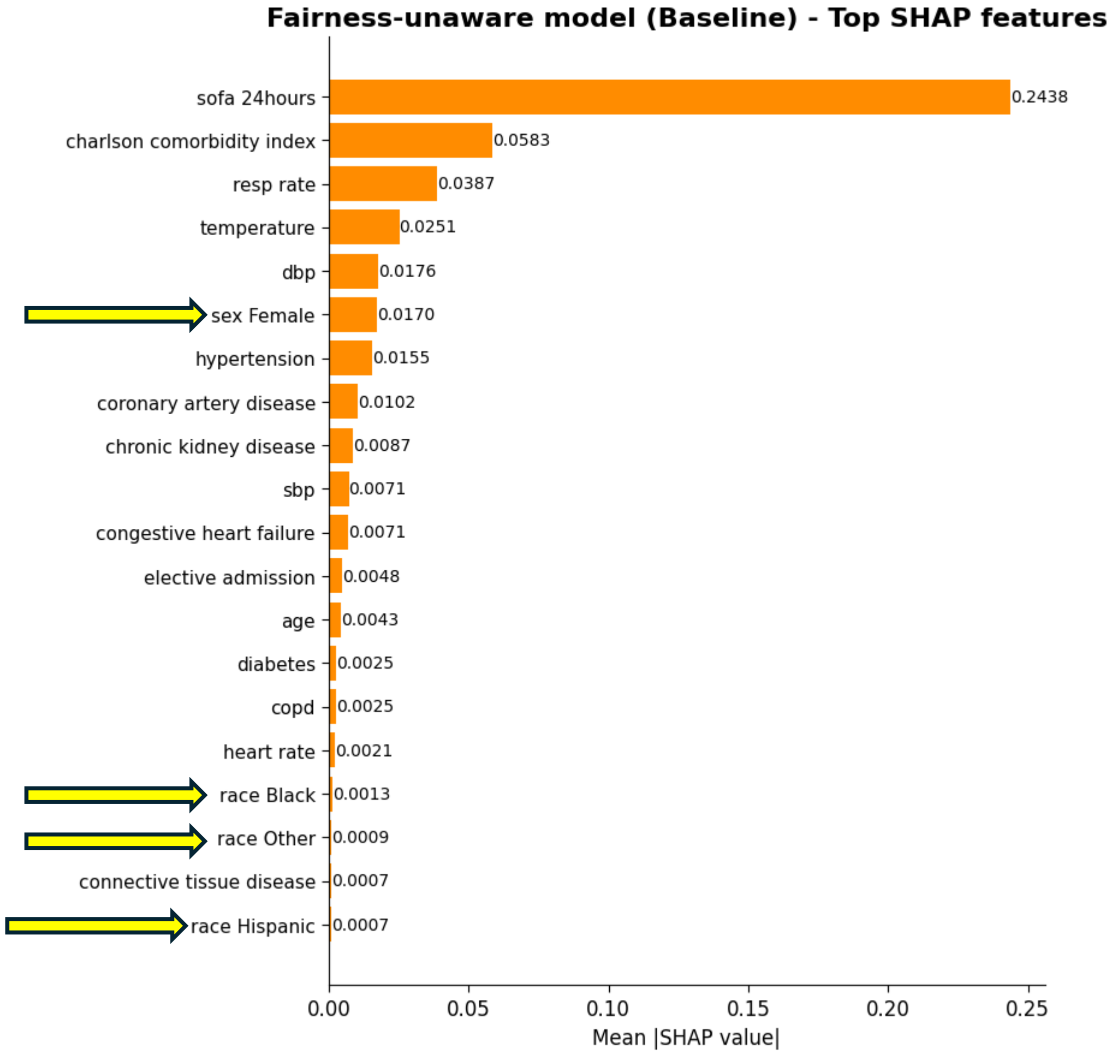
Task 2: IMV

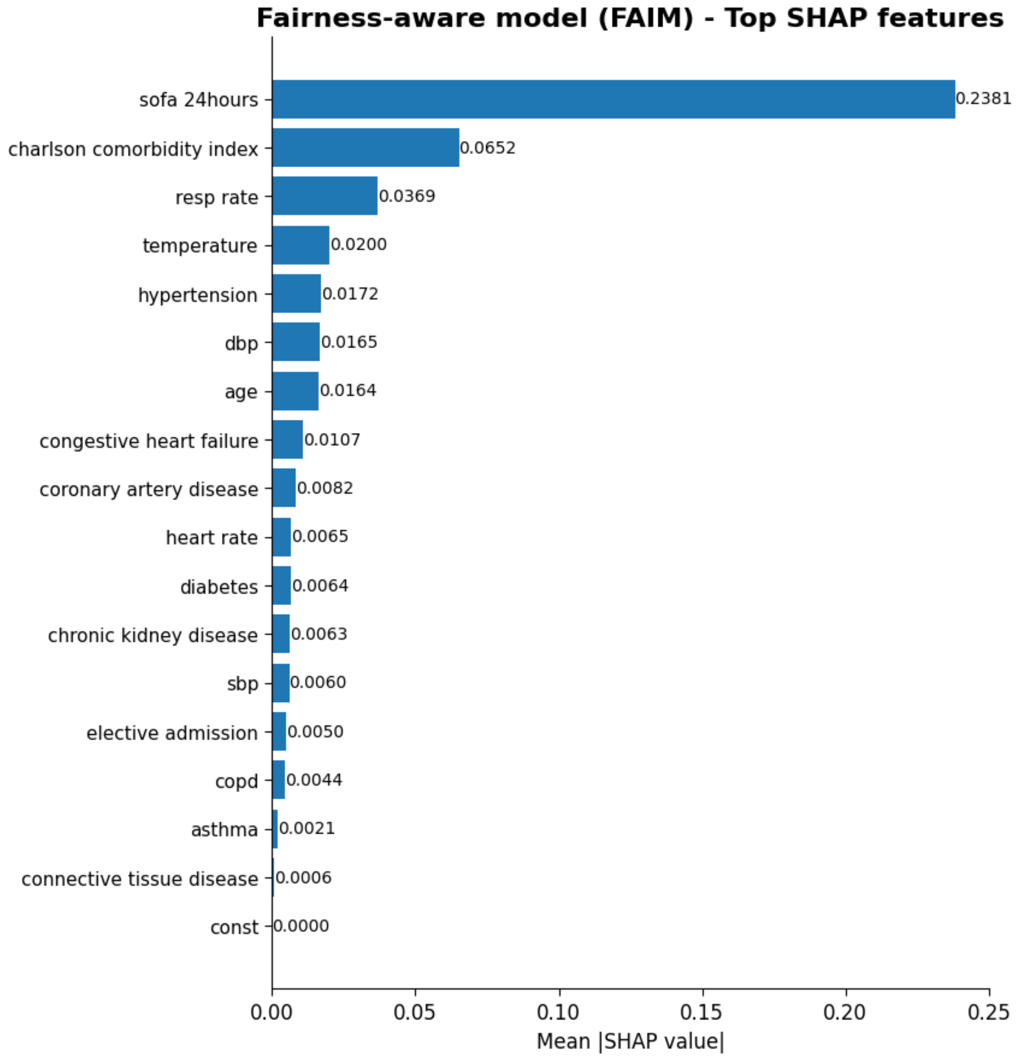
Split	**N**
Overall	**56,150**
Training (70%)	39,305
Validation (10%)	5,615
Test (20%)	11,230

Results: FRI scores

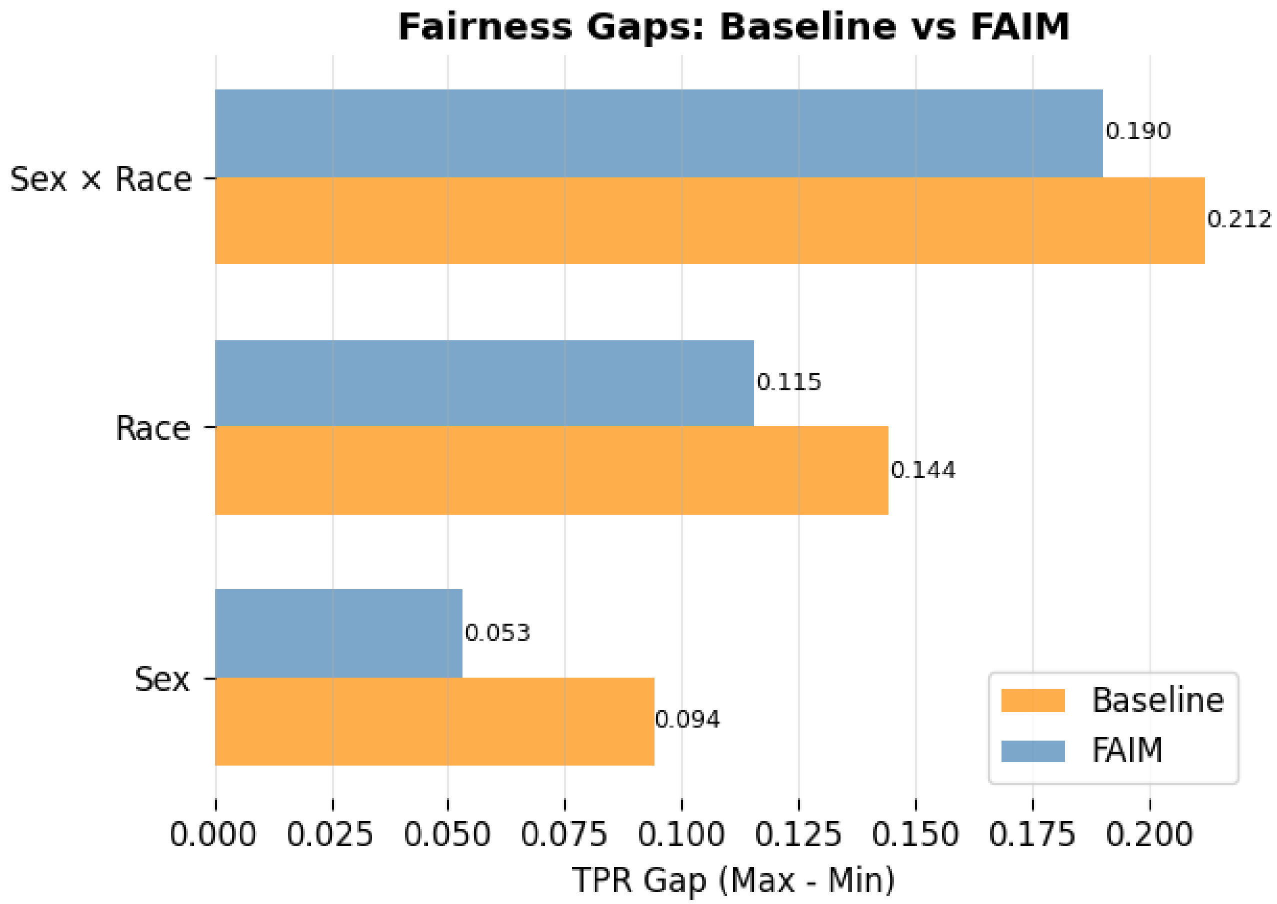
- From 800 models (200 per exclusion scenario), **360** were nearly optimal
- Final fairness-aware model came from **excluding sex and race**
  - **FAIM FRI: 22.11**
  - Mean FRI: 9.94
  - **Baseline (Logistic Regression) Model FRI: 5.05**

Results: SHAP comparison between Baseline and FAIM



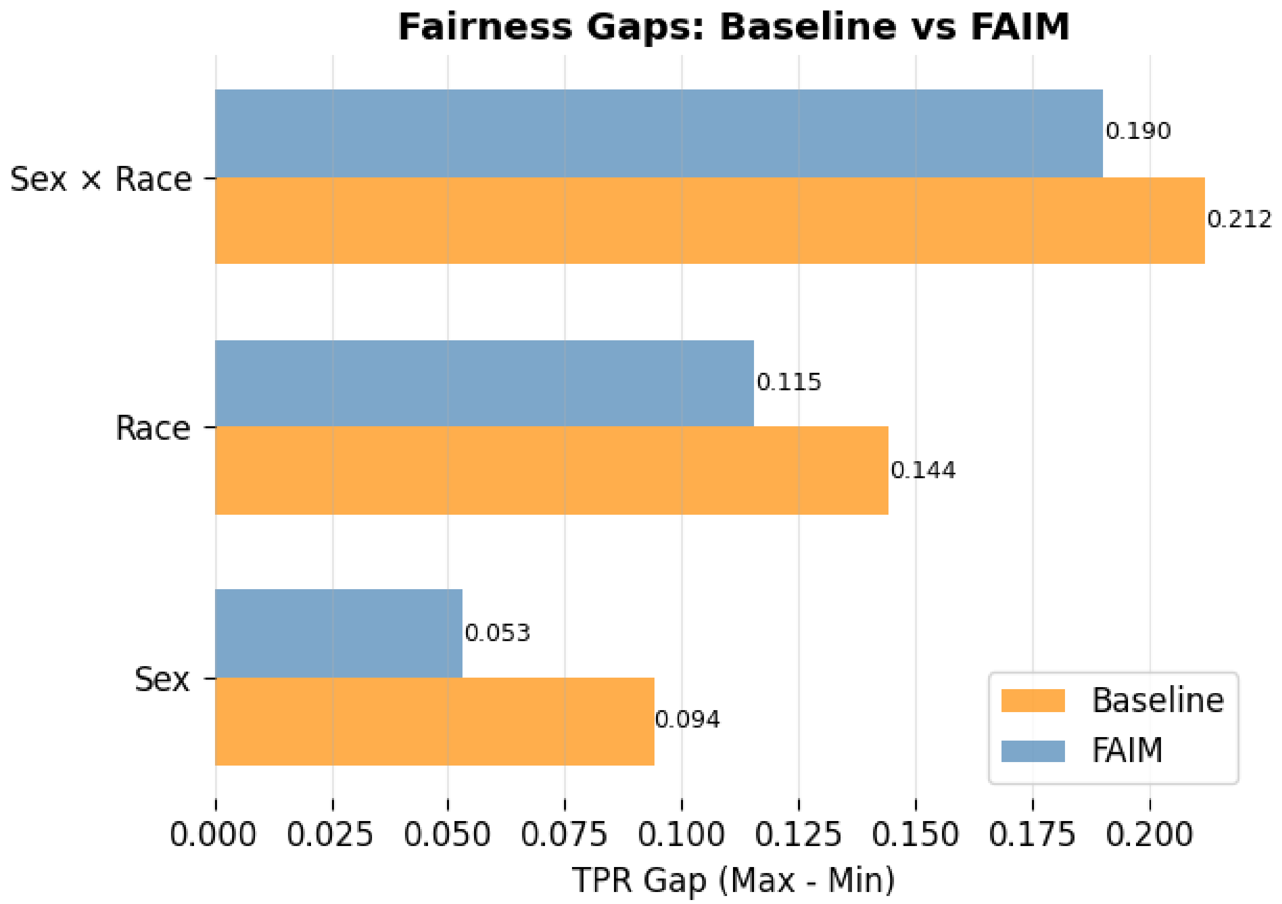


Results: TPR Gap Decrease between Baseline and FAIM



Sensitive Attribute	Gap Reduction %
Sex	43.6
Race	19.9
Sex × Race	10.2

Results: TPR Gap Decrease between Baseline and FAIM



Sensitive Attribute	**Gap Reduction %**
Sex	43.6
Race	19.9
Sex × Race	10.2

Results: TPR Gaps between Intersectional Subgroups

Model	**Metric**	**Intersection with Minimum TPR Value**	**Intersection with Maximum TPR Value**	**Gap**	**@Male_ @White vs. Minimum Intersection Gap**
Baseline	Equal Opportunity (TPR)	Female_ Black	@Male_ Asian	0.2118	0.1749
FAIM	Equal Opportunity (TPR)	Female_ Black	Female_ Asian	0.1902	0.1285

Results: TPR Gaps between Intersectional Subgroups

Model	Metric	Intersection with Minimum TPR Value	Intersection with Maximum TPR Value	Gap	@Male_ @White vs. Minimum Intersection Gap
Baseline	**Equal Opportunity (TPR)**	**Female_ Black**	**@Male_ Asian**	**0.2118**	**0.1749**
FAIM	**Equal Opportunity (TPR)**	**Female_ Black**	**Female_ Asian**	**0.1902**	**0.1285**

Results: TPR Gaps between Intersectional Subgroups

N: 4,795

Model	Metric	Intersection with Minimum TPR Value	Intersection with Maximum TPR Value	Gap	@Male_ @White vs. Minimum Intersection Gap
Baseline	**Equal Opportunity (TPR)**	**Female_ Black**	**@Male_ Asian**	**0.2118**	**0.1749**
FAIM	**Equal Opportunity (TPR)**	**Female_ Black**	**Female_ Asian**	**0.1902**	**0.1285**

N: 227

N: 184

Test Set: 11,230

Results: Fairness Metric Comparison across Methods

		Separation-based metrics						Independence-based metrics	
		Equal Opportunity	**Equalized Odds**	**BER Equality**	**Sensitivity (TPR)**	**Specificity (TNR)**	**AUC**	**Statistical Parity**	**Accuracy Equality**
	Baseline	0.211805	0.211805	0.074427	0.811897	0.704988	0.830100	0.239369	0.097973
In-processing	**FAIM**	0.190151	0.190151	0.064302	0.787513	0.725794	0.827863	0.145732	0.074915
In-processing	**Adversarial Learning**	0.256410	0.435675	0.073466	0.771575	0.764642	0.847577	0.250004	0.079148
In-processing	**Reductions**	0.241569	0.241569	0.114205	0.500268	0.896772	0.698457	0.113930	0.111166
Pre-processing	**Unawareness**	0.195433	0.195433	0.062793	0.806806	0.710723	0.828955	0.151995	0.082708
Pre-processing	**Reweighing**	0.169955	0.179479	0.081232	0.808146	0.706989	0.828427	0.191367	0.097049
Post-processing	**Equalized Odds**	0.853659	0.853659	0.267729	0.367095	0.860630	0.614064	0.467391	0.151583
Post-processing	**Calibrated Equalized Odds**	0.779633	0.779633	0.258896	0.416667	0.878901	0.647954	0.455892	0.202744
Post-processing	**Reject Option Classifier**	0.187976	0.187976	0.069573	0.784566	0.728728	0.756595	0.189684	0.082880

Results: Fairness Metric Comparison across Methods

		Separation-based metrics						Independence-based metrics	
		Equal Opportunity	Equalized Odds	BER Equality	Sensitivity (TPR)	Specificity (TNR)	AUC	Statistical Parity	Accuracy Equality
	Baseline	0.211805	0.211805	0.074427	0.811897	0.704988	0.830100	**0.23936**	0.097973
In-processing	**FAIM**	0.190151	0.190151	0.064302	0.787513	0.725794	0.827863	**0.14573**	**0.074915**
In-processing	**Adversarial Learning**	0.256410	0.435675	0.073466	0.771575	0.764642	0.847577	0.250004	0.079148
In-processing	**Reductions**	0.241569	0.241569	0.114205	0.500268	0.896772	0.698457	0.113930	**0.111166**
Pre-processing	**Unawareness**	0.195433	0.195433	0.062793	0.806806	0.710723	0.828955	0.151995	0.082708
Pre-processing	**Reweighing**	0.169955	0.179479	0.081232	0.808146	0.706989	0.828427	0.191367	0.097049
Post-processing	**Equalized Odds**	0.853659	0.853659	0.267729	0.367095	0.860630	0.614064	0.467391	**0.151583**
Post-processing	**Calibrated Equalized Odds**	0.779633	0.779633	0.258896	0.416667	0.878901	0.647954	0.455892	**0.202744**
Post-processing	**Reject Option Classifier**	0.187976	0.187976	0.069573	0.784566	0.728728	0.756595	0.189684	0.082880

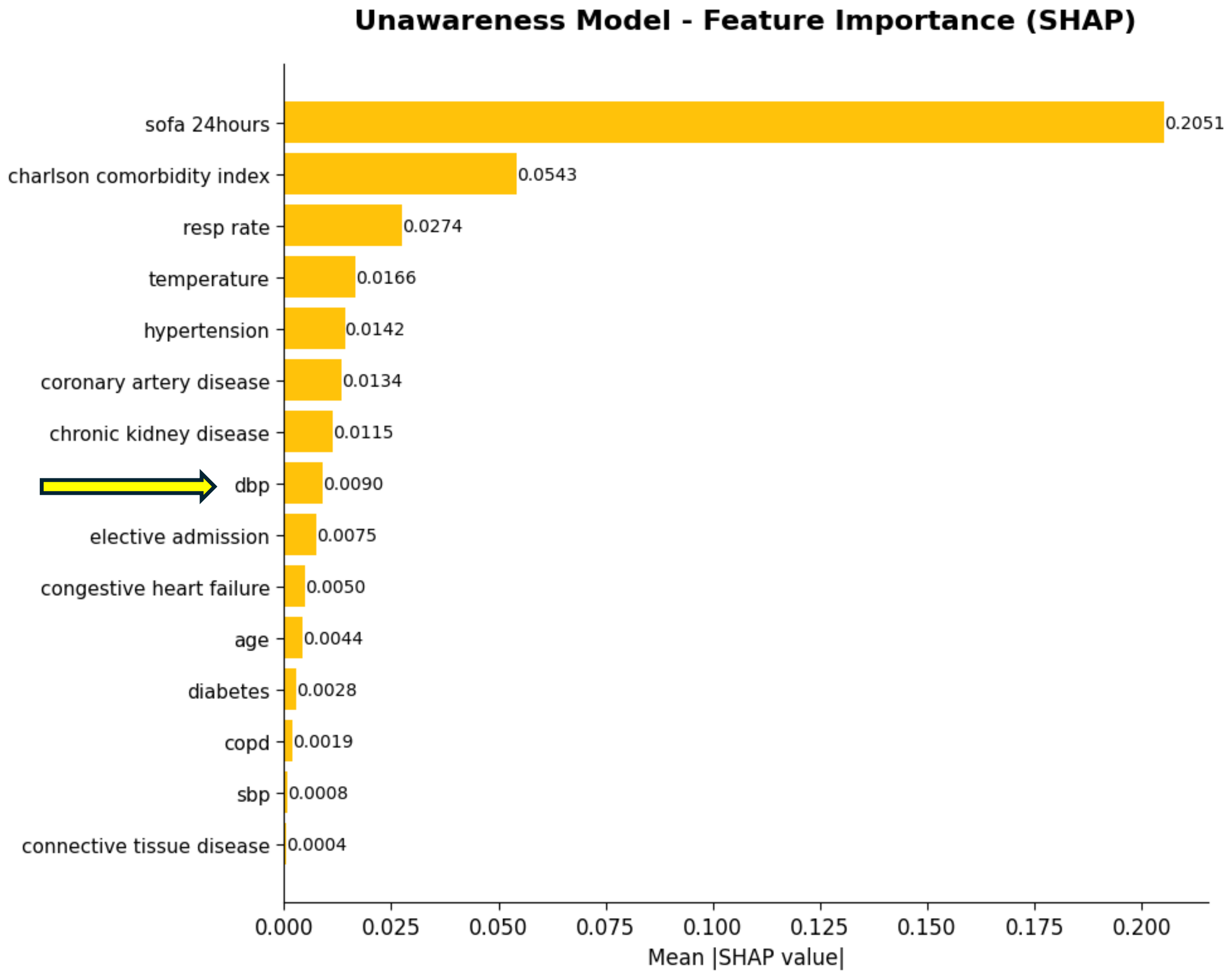
Results: Fairness Metric Comparison across Methods

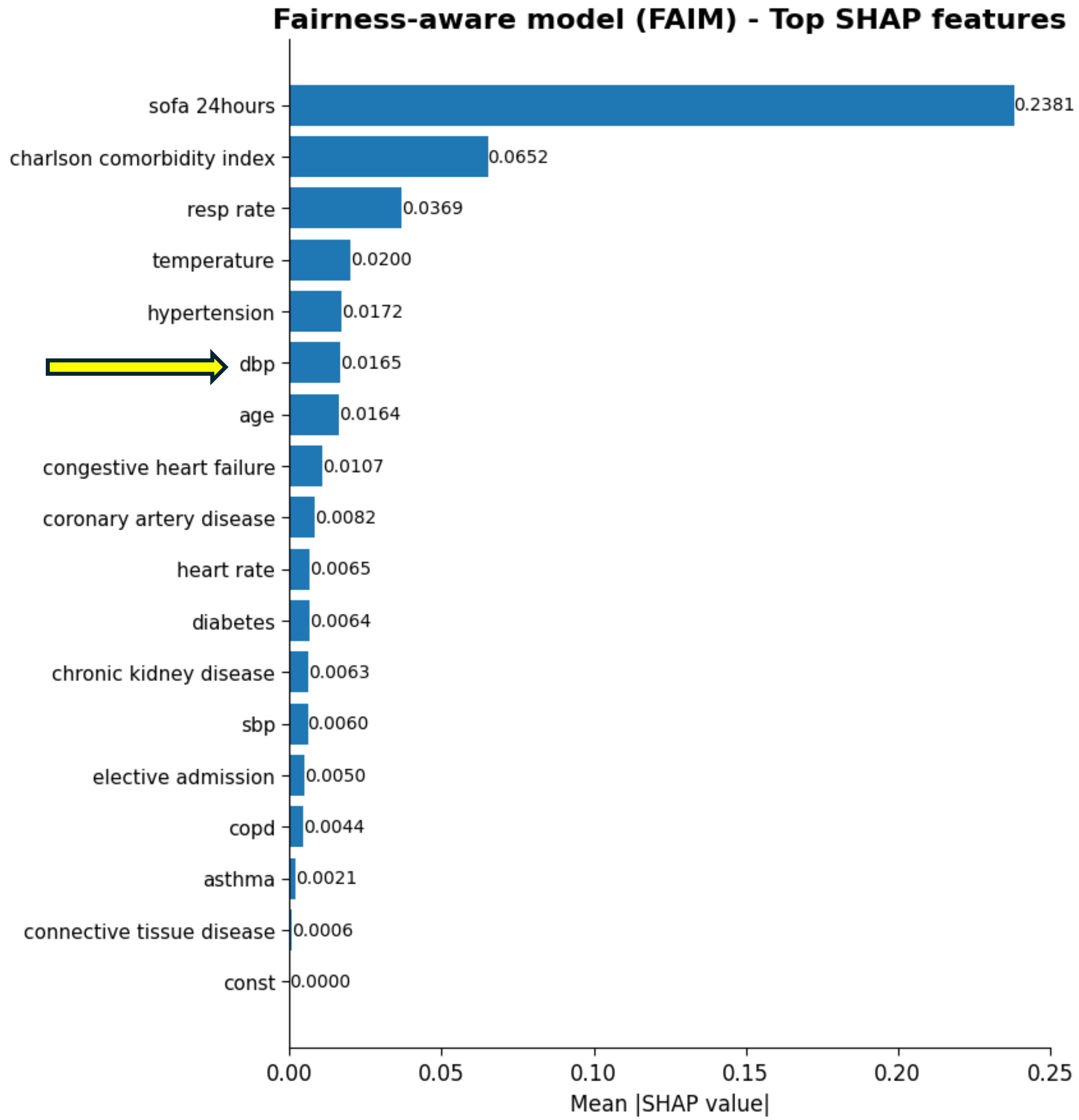
		Separation-based metrics						Independence-based metrics	
		Equal Opportunity	Equalized Odds	BER Equality	Sensitivity (TPR)	Specificity (TNR)	AUC	Statistical Parity	Accuracy Equality
	Baseline	0.211805	0.211805	0.074427	0.811897	0.704988	0.830100	0.239369	0.097973
In-processing	**FAIM**	0.190151	**0.1901**	0.06430	0.787513	**0.725794**	**0.82786**	0.14573	0.074915
In-processing	**Adversarial Learning**	0.256410	**0.43567**	0.07346	0.771575	**0.764642**	**0.84757**	0.25000	0.079148
In-processing	**Reductions**	0.241569	0.241569	0.114205	0.500268	0.896772	0.698457	0.113930	0.111166
Pre-processing	**Unawareness**	0.195433	0.195433	0.062793	0.806806	0.710723	0.828955	0.151995	0.082708
Pre-processing	**Reweighing**	0.169955	0.179479	0.081232	0.808146	0.706989	0.828427	0.191367	0.097049
Post-processing	**Equalized Odds**	0.853659	0.853659	0.267729	0.367095	0.860630	0.614064	0.467391	0.151583
Post-processing	**Calibrated Equalized Odds**	0.779633	0.779633	0.258896	0.416667	0.878901	0.647954	0.455892	0.202744
Post-processing	**Reject Option Classifier**	0.187976	0.187976	0.069573	0.784566	0.728728	0.756595	0.189684	0.082880

Results: Fairness Metric Comparison across Methods

		Separation-based metrics						Independence-based metrics	
		Equal Opportunity	**Equalized Odds**	**BER Equality**	**Sensitivity (TPR)**	**Specificity (TNR)**	**AUC**	**Statistical Parity**	**Accuracy Equality**
	Baseline	0.211805	0.211805	0.074427	0.811897	0.704988	0.830100	0.239369	0.097973
In-processing	**FAIM**	**0.190151** ☆	0.19015	0.06430	0.787513	0.725794	0.82786	0.14573	0.074915
In-processing	**Adversarial Learning**	0.256410	0.435675	0.073466	0.771575	0.764642	0.847577	0.250004	0.079148
In-processing	**Reductions**	0.241569	0.241569	0.114205	0.500268	0.896772	0.698457	0.113930	0.111166
Pre-processing	**Unawareness**	0.195433	0.19543	0.06279	**0.806806** ☆	0.710723	0.82895	0.15199	0.082708
Pre-processing	**Reweighing**	0.169955	0.179479	0.081232	0.808146	0.706989	0.828427	0.191367	0.097049
Post-processing	**Equalized Odds**	0.853659	0.853659	0.267729	0.367095	0.860630	0.614064	0.467391	0.151583
Post-processing	**Calibrated Equalized Odds**	0.779633	0.779633	0.258896	0.416667	0.878901	0.647954	0.455892	0.202744
Post-processing	**Reject Option Classifier**	0.187976	0.187976	0.069573	0.784566	0.728728	0.756595	0.189684	0.082880

Results: SHAP comparison between Unawareness and FAIM





Conclusions

- FAIM mitigates bias in machine learning models for clinical decision-making

Limitations

- Dataset scope
- Dataset size

Future Work

- **Dataset expansion**
- Include **more sensitive attributes** (marital status, insurance, etc.)
- Consider less inherently transparent model as baseline, such as **neural network**

UNIVERSITAT DE BARCELONA

Facultat de Matemàtiques i Informàtica
M.Sc. Data Science
Final Master’s Thesis

A comparative study of fairness methods for clinical predictions using MIMIC-IV database

Author:
Iris VUKOVIC

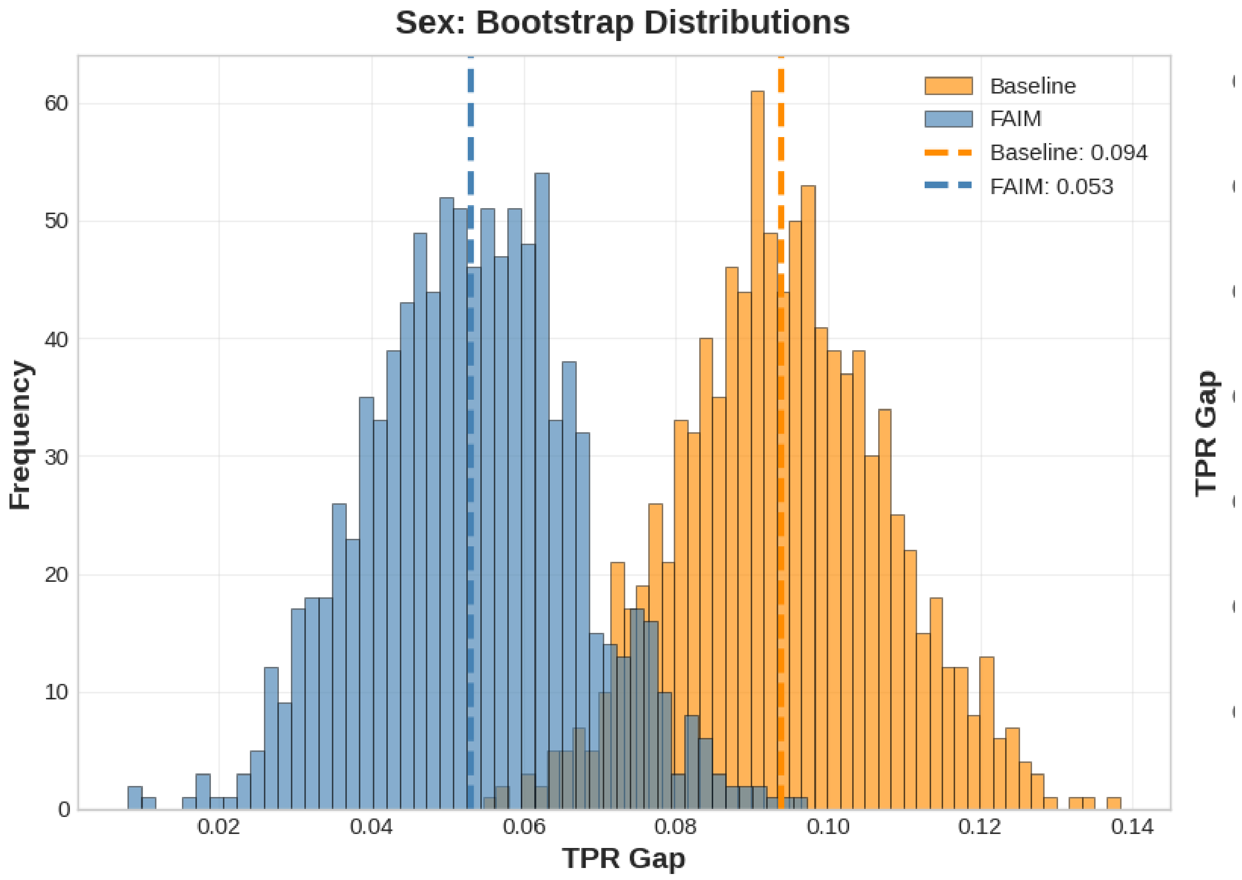
Supervisor:
Laura IGUAL MUÑOZ

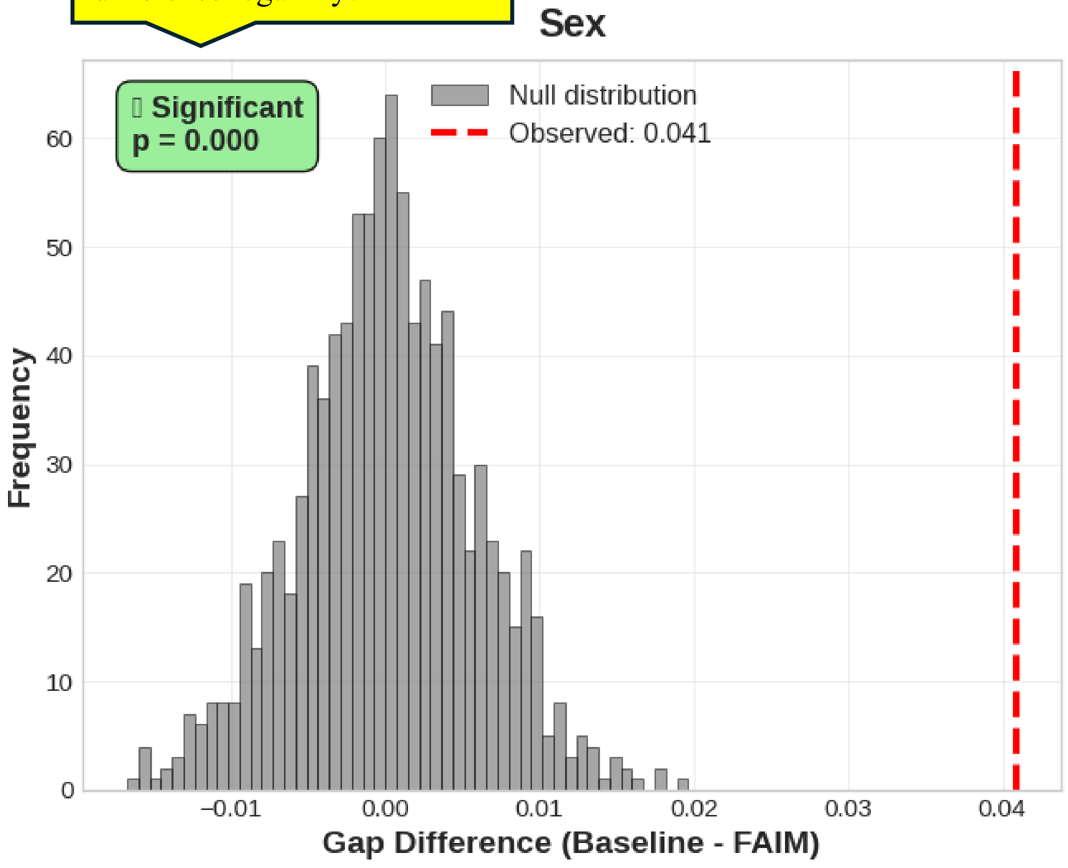
Thank you for your time!

Results: Bootstrapping and Permutation Testing (Task 2)

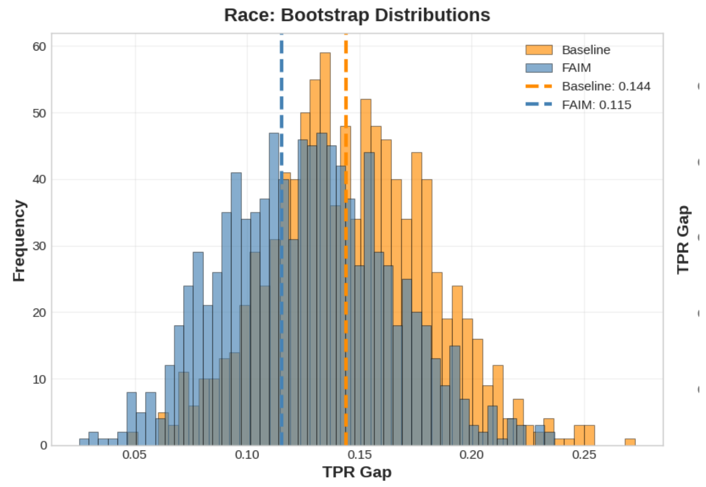
If I had 1000 different test sets, what range of TPR gaps would I see?

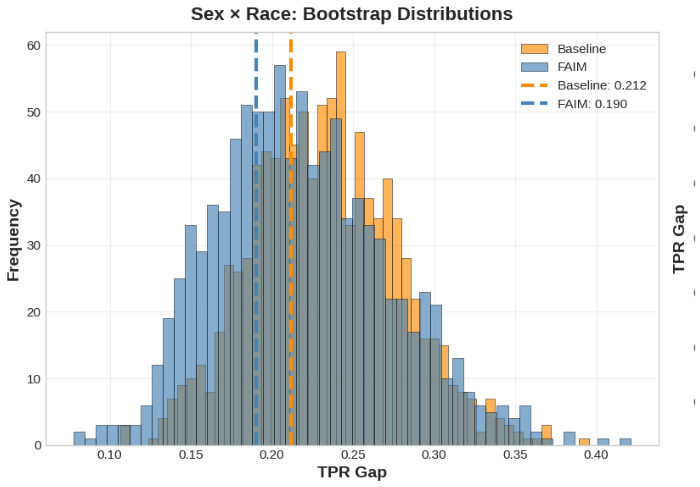
Does randomly scrambling the labels produces differences as large as the observed difference regularly?

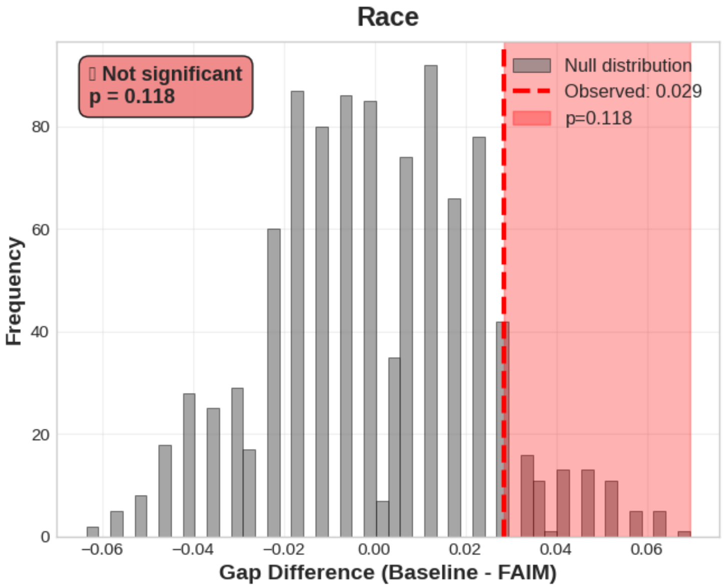


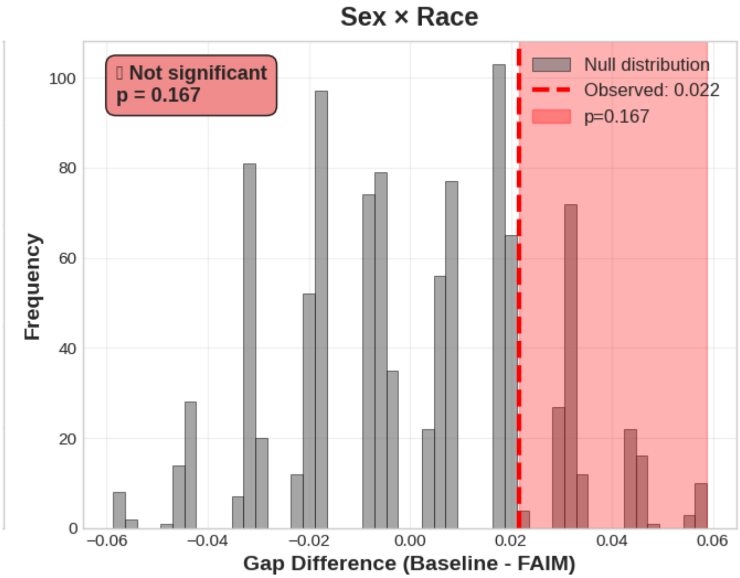


Results: Bootstrapping and Permutation Testing (Task 2)

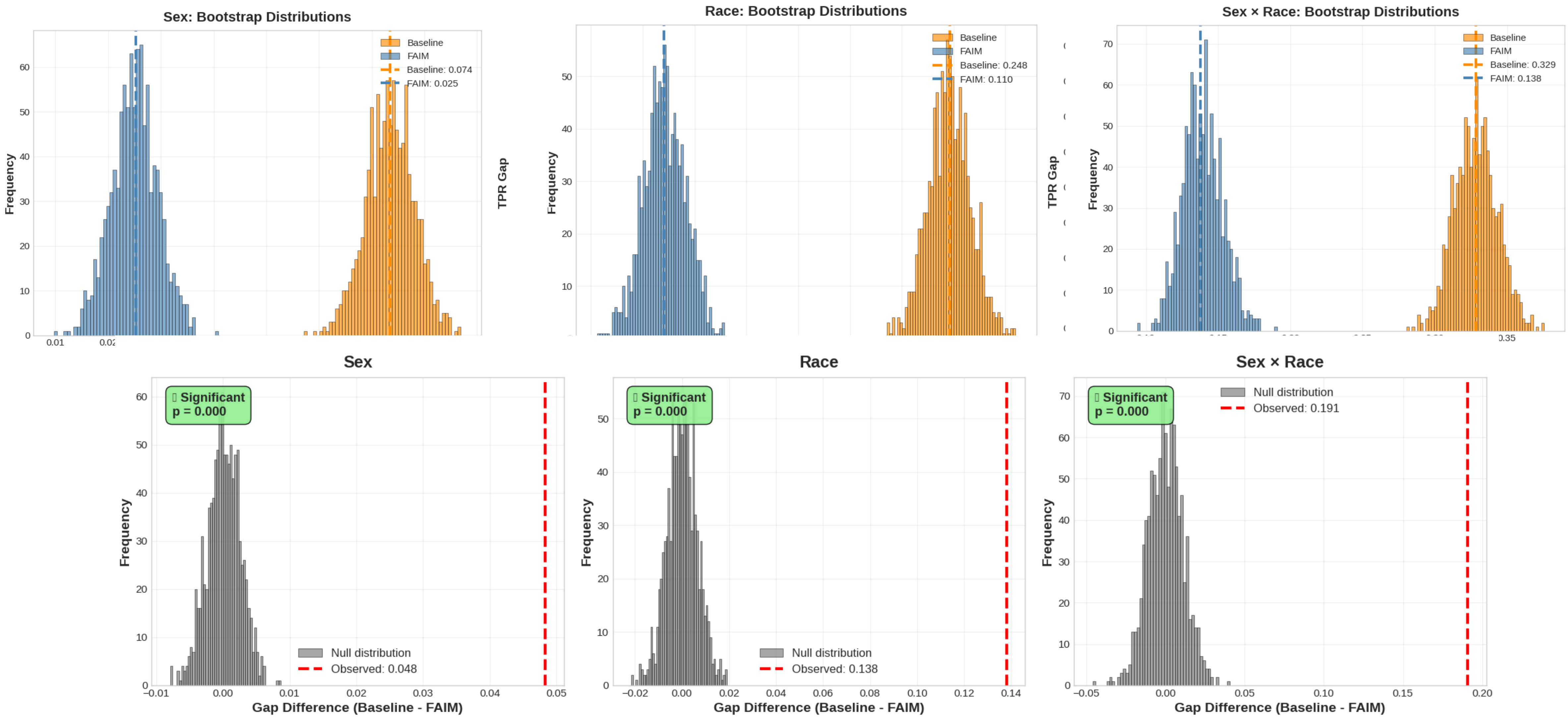




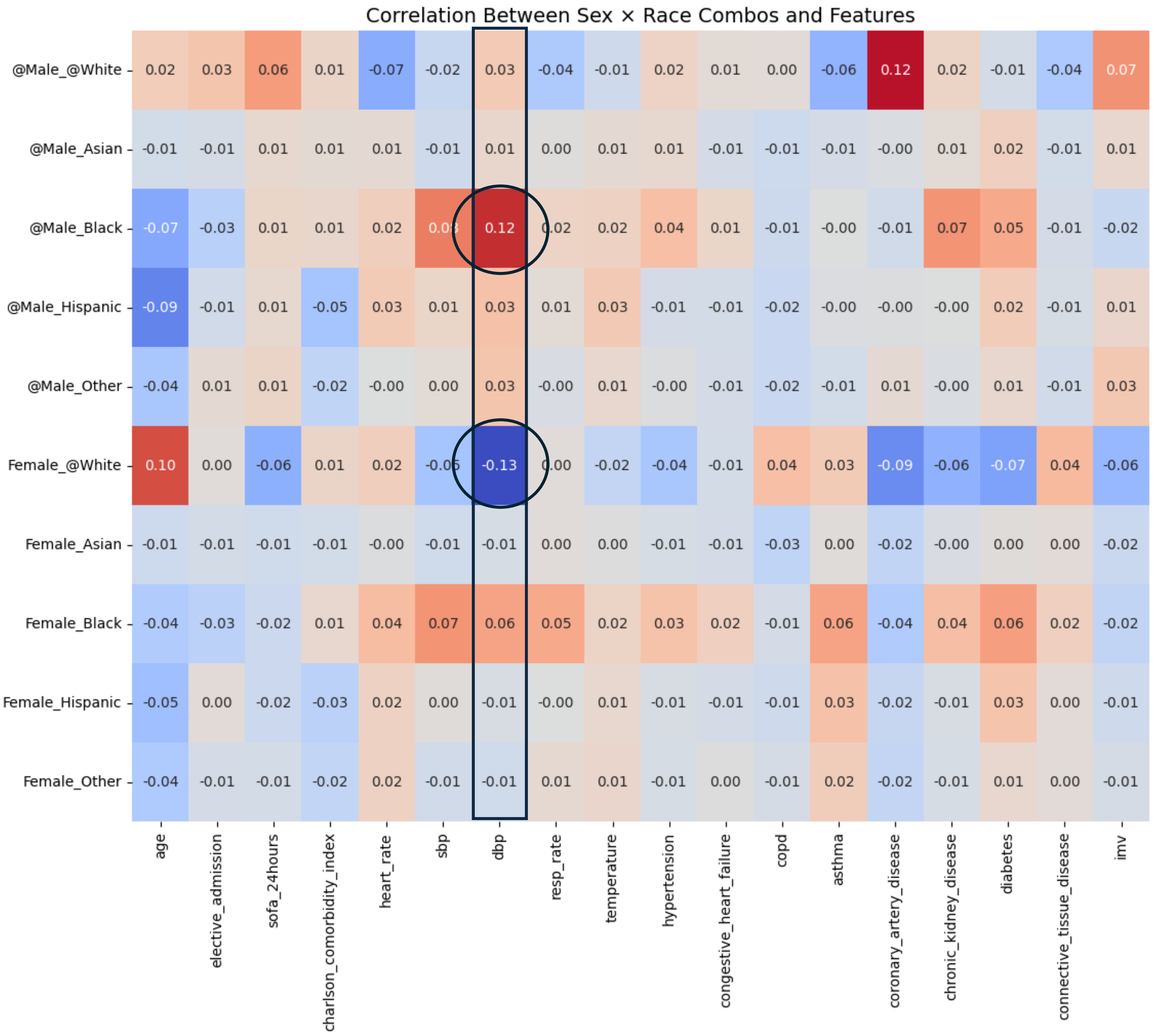




Results: Bootstrapping and Permutation Testing (Task 1)



Results


Correlation Between Sex × Race Combos and Features
@Male_@White 0.02 0.03 0.06 0.01 -0.07 -0.02 0.03 -0.04 -0.01 0.02 0.01 0.00 -0.06 0.12 0.02 -0.01 -0.04 0.07
@Male_Asian -0.01 -0.01 0.01 0.01 0.01 -0.01 0.01 0.00 0.01 0.01 -0.01 -0.01 -0.01 -0.00 0.01 0.02 -0.01 0.01
@Male_Black -0.07 -0.03 0.01 0.01 0.02 0.08 0.12 0.02 0.02 0.04 0.01 -0.01 -0.00 -0.01 0.07 0.05 -0.01 -0.02
@Male_Hispanic -0.09 -0.01 0.01 -0.05 0.03 0.01 0.03 0.01 0.03 -0.01 -0.01 -0.02 -0.00 -0.00 -0.00 0.02 -0.01 0.01
@Male_Other -0.04 0.01 0.01 -0.02 -0.00 0.00 0.03 -0.00 0.01 -0.00 -0.01 -0.02 -0.01 0.01 -0.00 0.01 -0.01 0.03
Female_@White 0.10 0.00 -0.06 0.01 0.02 -0.05 -0.13 0.00 -0.02 -0.04 -0.01 0.04 0.03 -0.09 -0.06 -0.07 0.04 -0.06
Female_Asian -0.01 -0.01 -0.01 -0.01 -0.00 -0.01 -0.01 0.00 0.00 -0.01 -0.01 -0.03 0.00 -0.02 -0.00 0.00 0.00 -0.02
Female_Black -0.04 -0.03 -0.02 0.01 0.04 0.07 0.06 0.05 0.02 0.03 0.02 -0.01 0.06 -0.04 0.04 0.06 0.02 -0.02
Female_Hispanic -0.05 0.00 -0.02 -0.03 0.02 0.00 -0.01 -0.00 0.01 -0.01 -0.01 -0.01 0.03 -0.02 -0.01 0.03 0.00 -0.01
Female_Other -0.04 -0.01 -0.01 -0.02 0.02 -0.01 -0.01 0.01 0.01 -0.01 0.00 -0.01 0.02 -0.02 -0.01 0.01 0.00 -0.01
age
elective_admission
sofa_24hours
charlson_comorbidity_index
heart_rate
sbp
dbp
resp_rate
temperature
hypertension
congestive_heart_failure
copd
asthma
coronary_artery_disease
chronic_kidney_disease
diabetes
connective_tissue_disease
imv

Results: TPR Gaps between Intersectional Subgroups

TPR -4.7%

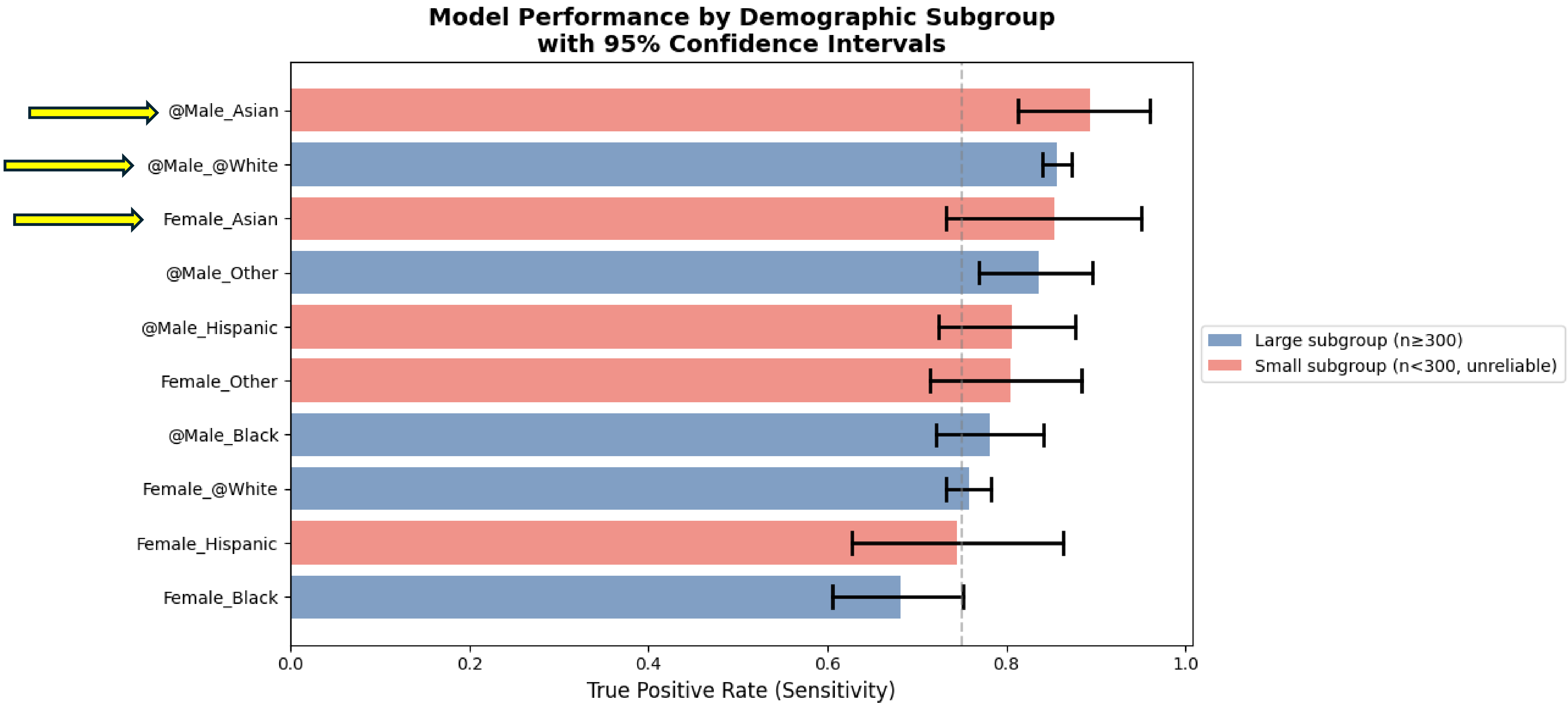
Model	Metric	Intersection with Minimum TPR Value	Intersection with Maximum TPR Value	Gap	@Male_ @White vs. Minimum Intersection Gap
Baseline	**Equal Opportunity (TPR)**	**Female_ Black**	**@Male_ Asian**	**0.2118**	**0.1749**
FAIM	**Equal Opportunity (TPR)**	**Female_ Black**	**Female_ Asian**	**0.1902**	**0.1285**

TPR -7.5%

TPR +0.9%

TPR +2.9%

Results: Confidence Intervals per Intersectional Subgroups



Results: Intersections Subgroups TPR Changes

Intersection	TPR Baseline	TPR FAIM	TPR Change	TPR Change (%)
Female_ Black	0.6815	0.6879	0.0064	0.9
Female_ Hispanic	0.7451	0.7255	-0.0196	-2.6
Female_ @White	0.7578	0.7596	0.0018	0.2
@Male_ Black	0.7814	0.765	-0.0164	-2.1
Female_ Other	0.8052	0.7792	-0.026	-3.2
@Male_ Hispanic	0.8061	0.7653	-0.0408	-5.1
@Male_ Other	0.8358	0.7761	-0.0597	-7.1
Female_ Asian	0.8537	0.878	0.0244	2.9
@Male_ @White	0.8564	0.8164	-0.0401	-4.7
@Male_ Asian	0.8933	0.8267	-0.0667	-7.5