Breakout Session 1: Track B

Measuring and Mitigating the Impact of Biases in Laboratory Testing on Machine Learning Models

Mr. Trenton Chang

Ph.D. Candidate, University of Michigan

UNIVERSITY OF MICHIGAN

Measuring and Mitigating the Impact of Biases in Laboratory Testing on AI Models

Trenton Chang, MS

ctrenton@umich.edu

PhD Candidate in Computer Science & Engineering, University of Michigan

PIs: Jenna Wiens, PhD & Michael W. Sjoding, MD

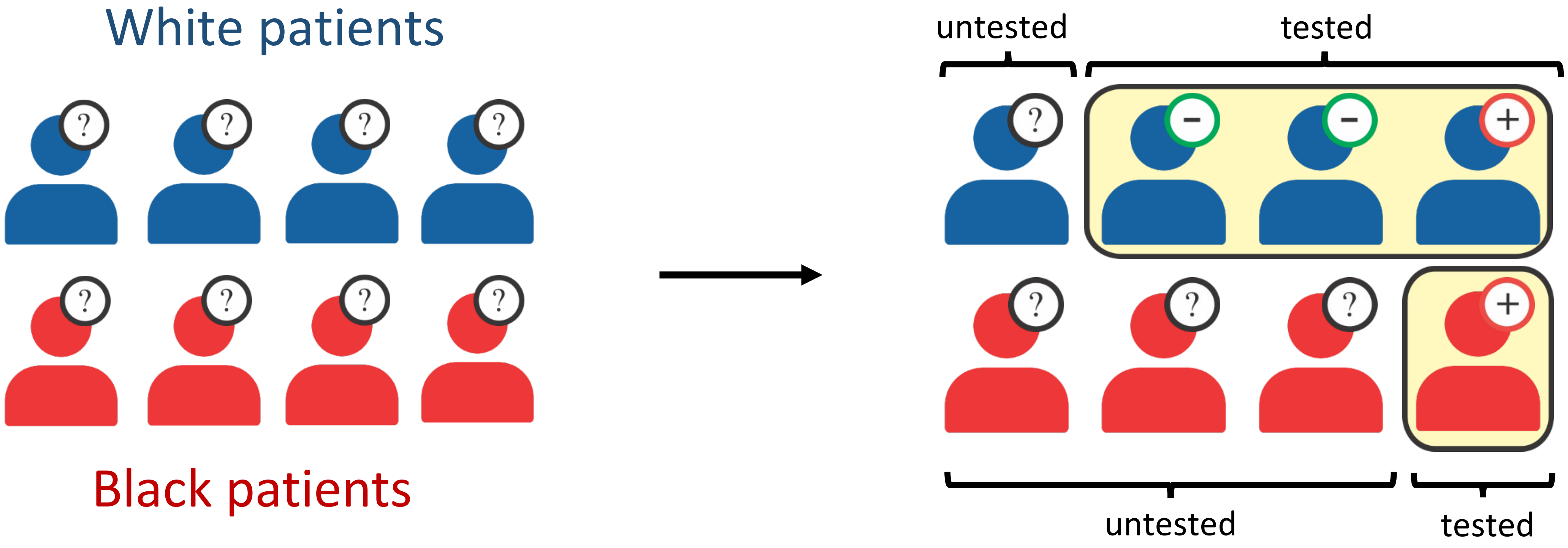
Project Summary & Goals

- Artificial intelligence (AI) tools can potentially assist in diagnostic decision making
- However, AI tools are susceptible to biases, resulting in poor generalization
- We aim to develop techniques and tools for *understanding* and *mitigating* potential biases

Highlights of our work:

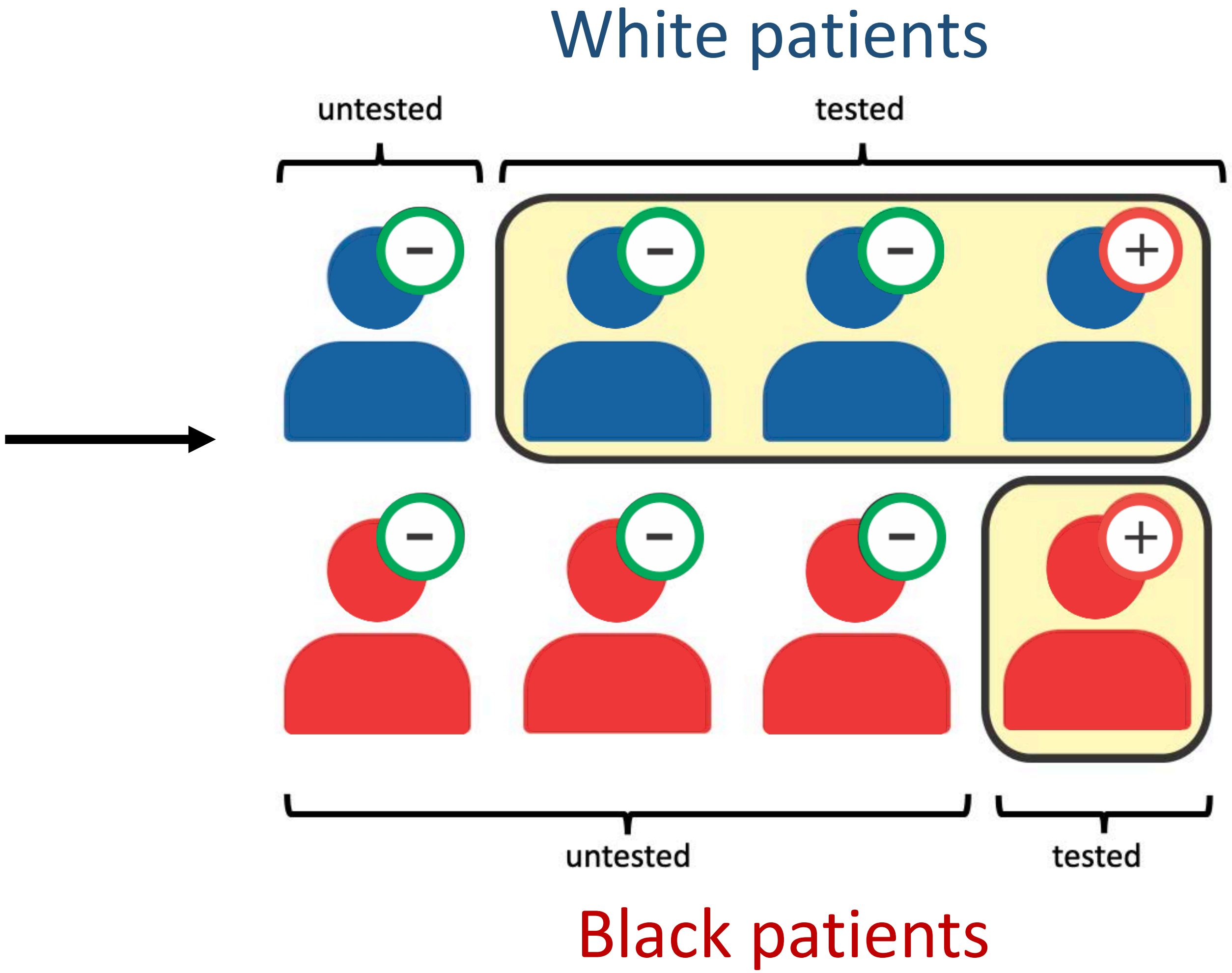
- A large-scale observational study of bias in laboratory testing (*under review*)
- A method for mitigating the impact of laboratory testing bias on AI models (*under review*)

Laboratory testing as a source of bias



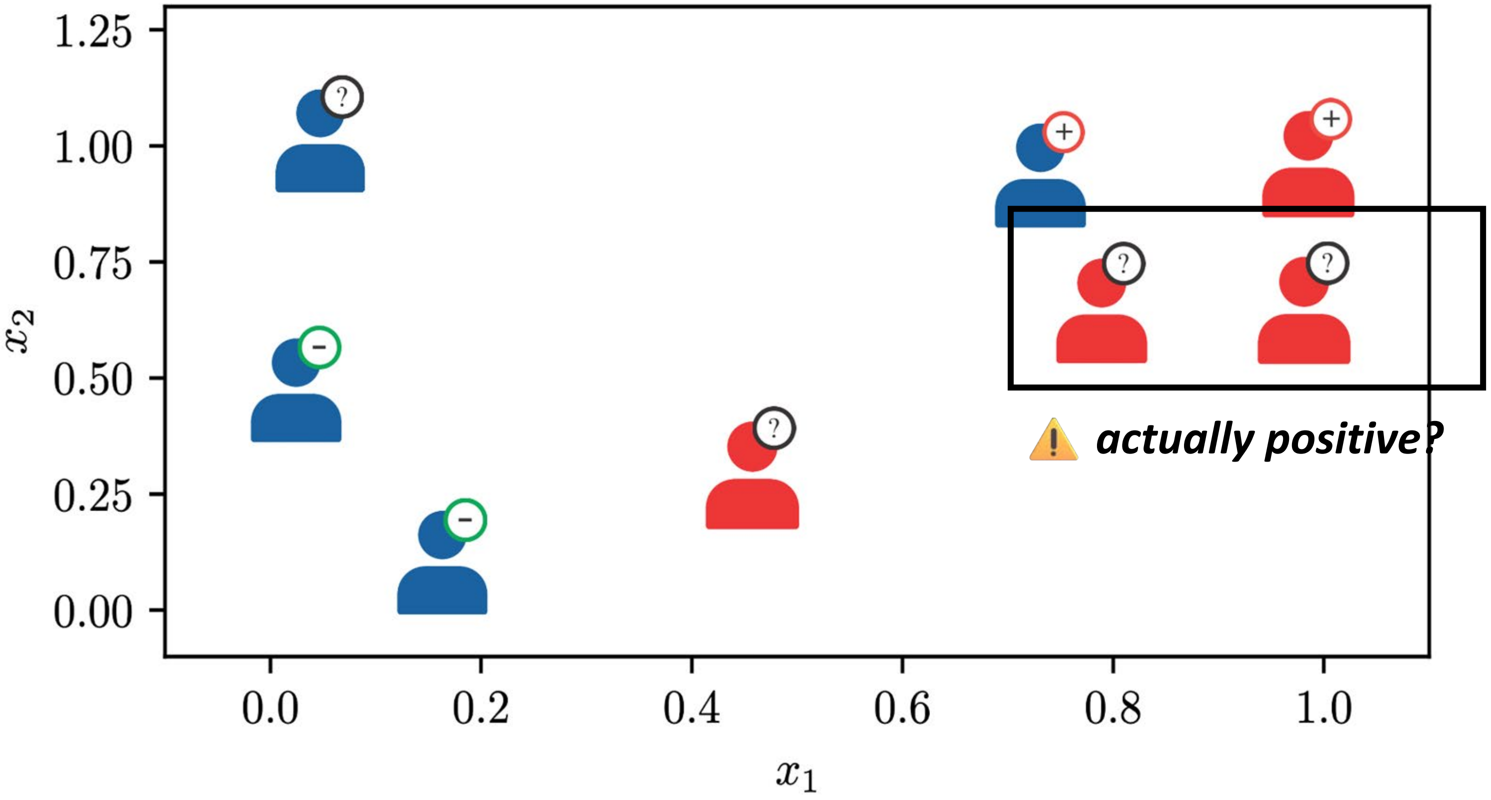
Untested = negative: the default assumption

Many works in practice assumed untested patients are negative:



Impact of testing bias on AI

An AI model might “see” training data as shown below:



⚠️ ***actually positive?***

In this example, an AI model trained on such data may *underpredict* the risk in Black patients.

Is there evidence of such undertesting?

- We conducted a retrospective matched cohort study of 235,830 emergency department (ED) visits
- **Question:** were there significant differences in laboratory testing rates between White vs. Black patients?
- **Cohorts:** All adult ED visits by White and Black patients at Michigan Medicine (U-M), 2015-2022 & Beth Israel Deaconness Medical Center (BIDMC), 2011-2019
- **Race:** as collected during patient registration
- **Main outcomes:** Testing rate difference (% White - % Black) for complete blood count, metabolic panel, arterial blood gas, blood culture, troponin, BNP, and d-dimer. *Secondary outcome:* hospital admission rate.
- **Matching:** exact 1:1 matching on age, biological sex, chief complaint (text), and ED triage score (1 to 5).

Cohort inclusion/exclusion summary

Exclusion criteria:

- Psychiatric visits
- Non-White/non-Black patients (incl. unknown/missing race)
- Patients with unknown biological sex

Before/after exclusion criteria:

Michigan Medicine: 602,650 —> 541,274

BIDMC: 447,109 —> 336,824

Before/after 1:1 exact matching

Michigan Medicine: 541,274 —> 141,510 (26.1% matched)

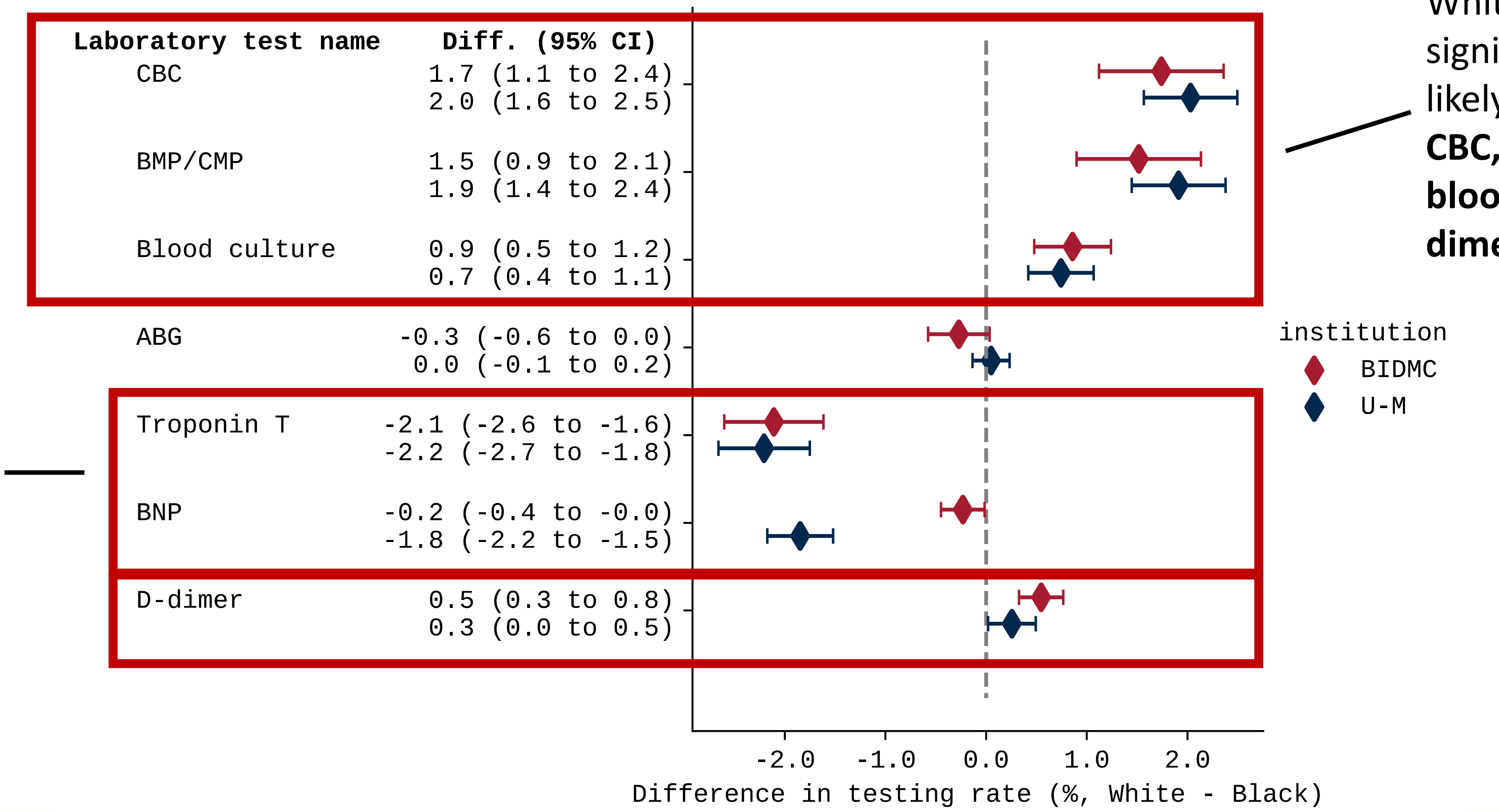
BIDMC: 336,824 —> 94,320 (28.0% matched)

Summary of cohort characteristics (pre-matching)

- **Age:** Black patients were significantly younger than White patients on average (**U-M:** 55 vs. 46 years, p<.001; **BIDMC:** 52 vs. 43 years, p<.001)
- **Biological sex:** Black patients were significantly more likely to be female (**U-M:** 52.0% vs. 62.0%; p<.001, **BIDMC:** 53.1% vs. 57.0%, p<.001)
- **ED triage scores:** Black patients were assessed as less ill on average (lower score; **U-M:** 2.6 vs. 2.7, **BIDMC:** 2.6 vs. 2.8). Chi-sq. test: p<.001.

Significant testing disparities in the ED

Difference in testing rates by race, matched analysis



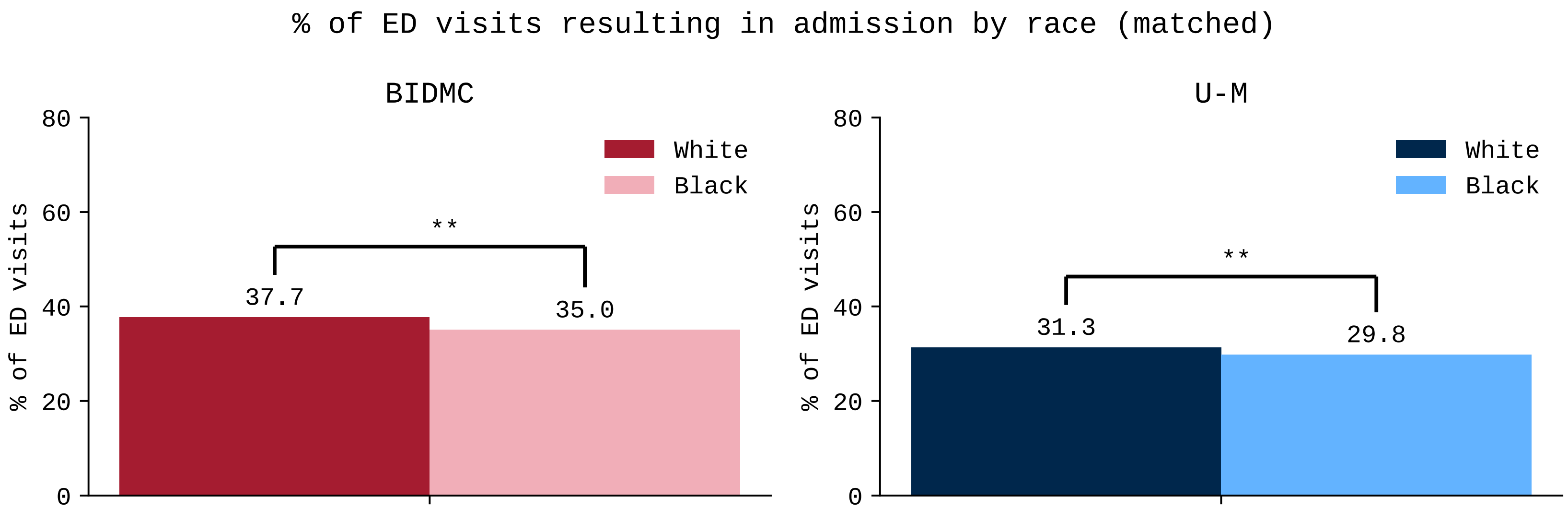
Laboratory test name	Diff. (95% CI)
CBC	1.7 (1.1 to 2.4) 2.0 (1.6 to 2.5)
BMP/CMP	1.5 (0.9 to 2.1) 1.9 (1.4 to 2.4)
Blood culture	0.9 (0.5 to 1.2) 0.7 (0.4 to 1.1)
ABG	-0.3 (-0.6 to 0.0) 0.0 (-0.1 to 0.2)
Troponin T	-2.1 (-2.6 to -1.6) -2.2 (-2.7 to -1.8)
BNP	-0.2 (-0.4 to -0.0) -1.8 (-2.2 to -1.5)
D-dimer	0.5 (0.3 to 0.8) 0.3 (0.0 to 0.5)

White patients significantly more likely to be tested: **CBC, BMP/CMP, blood culture, d-dimer**

Black patients significantly more likely to be tested: **Troponin T, BNP**

UNIVERSITY OF MICHIGAN

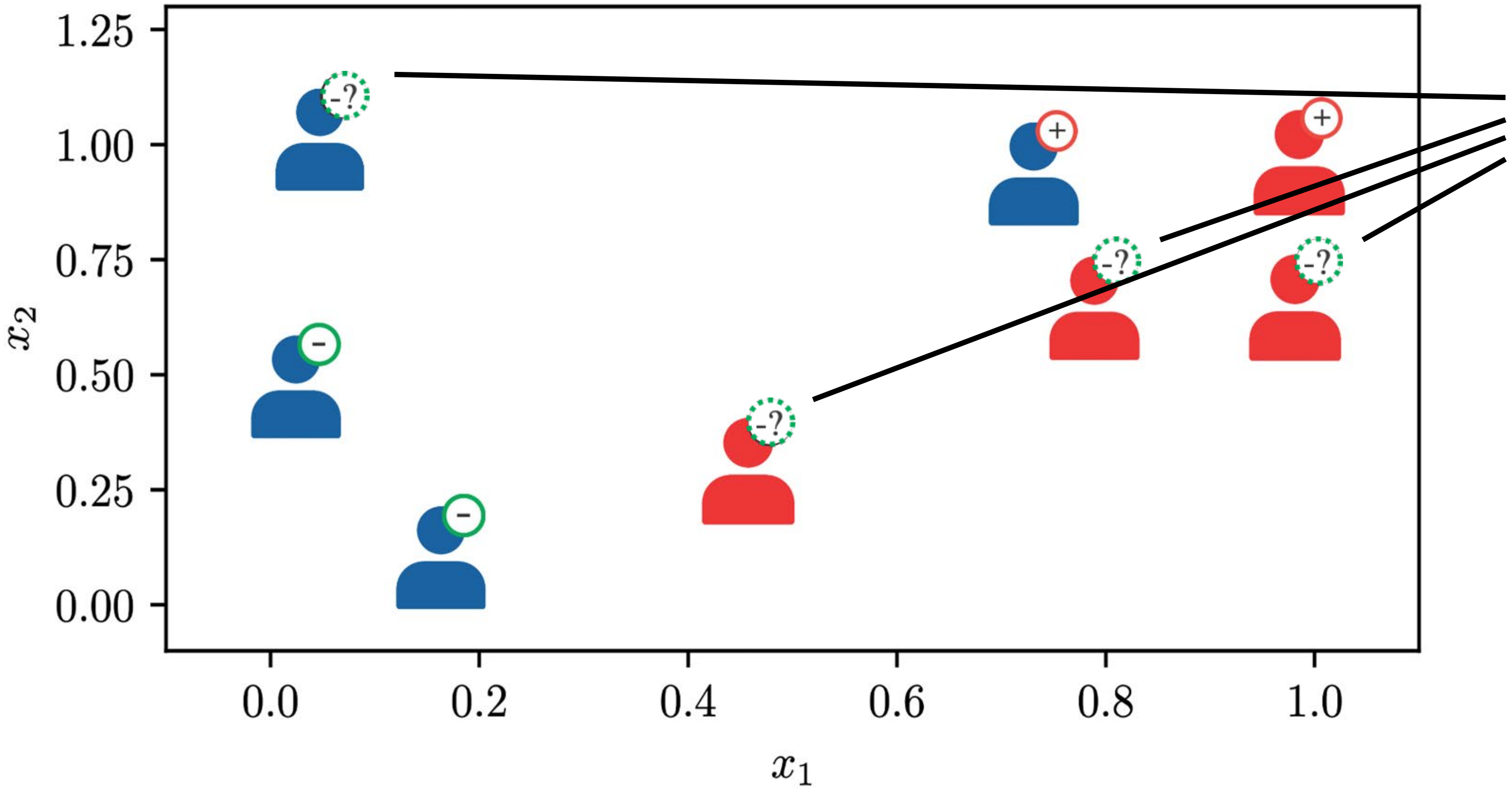
Hospital admission rate disparities



After exact 1:1 matching, racial differences in hospital admission rate following an ED visit also persisted.

A method for mitigating the impacts

- We can interpret predicting missing laboratory test results as a ***missing outcome problem*** —well-studied area in machine learning

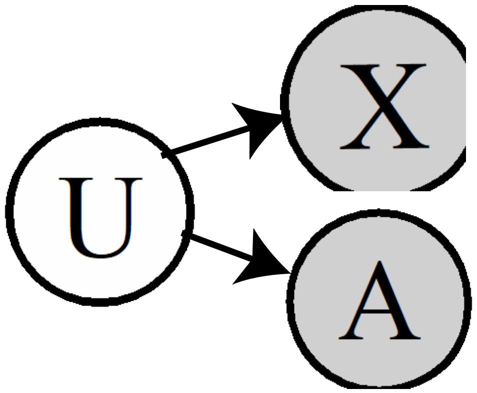


Can we “fill in the blanks?”

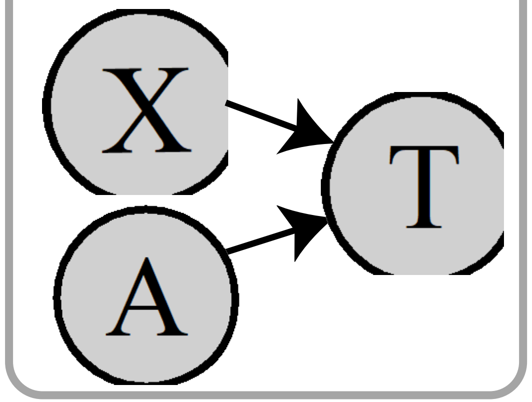
Overview of our approach

- We propose a probabilistic model for bias in laboratory testing and use an *expectation-maximization* algorithm to impute the missing test results

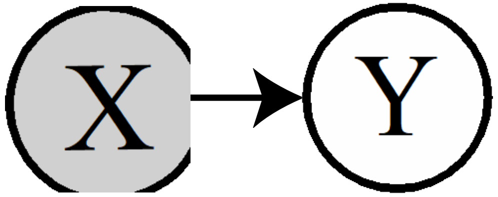
Demographic **groups (A)** might have different observed **features (X)**



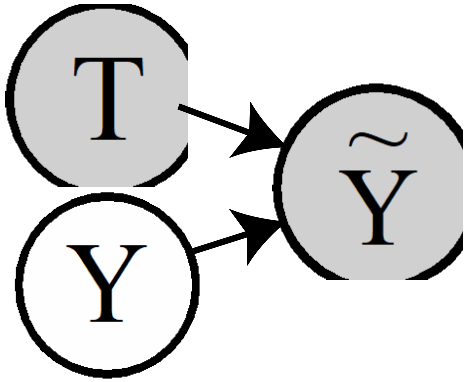
Testing decisions (T) can be **biased (depend on A)**



Ground truth (Y) does **not** directly depend on **demographics (A)**



Observed label is negative if **untested;** equal to **Y** if tested



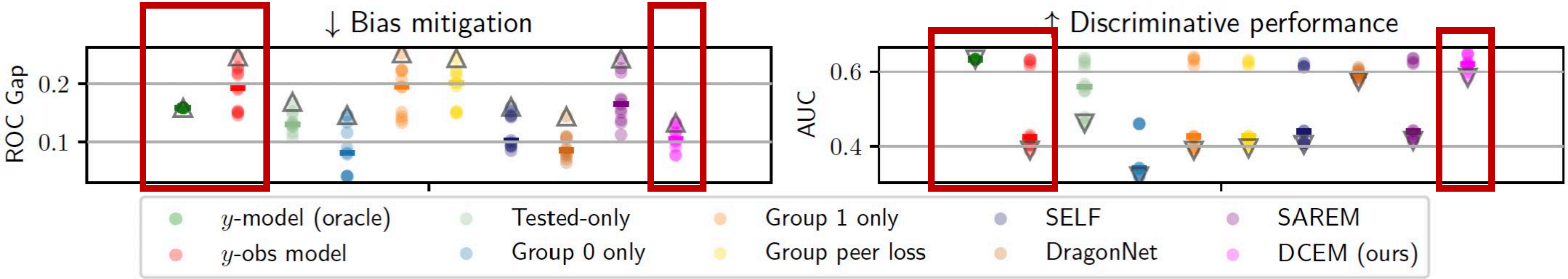
Case study: sepsis classification

- Many sepsis definitions (*e.g.,* Sepsis-3) are dependent on laboratory test results (blood culture) — no test = no diagnosis
- We aim to predict whether a patient will ever develop sepsis during a hospital stay
- We simulate multiple hypothetical testing decisions based on features used by the qSOFA score + report results across all replications
- We evaluate bias mitigation (similar performance across patient groups) and discriminative performance (can ”separate” positive vs. negative) with respect to true sepsis labels

Empirical results

Key methods:

- **green** = train on actual labels (best possible discriminative performance)
- **red** = *default (assume untested = negative)*
- **magenta** = our imputation-based method



Compared to baselines, our method *mitigates bias* and *improves discriminative performance.*

Future Work

- **Improved methods.** The proposed approach eventually fails when testing rates are too low — can we improve the robustness of our method to low testing rates?
- **Evaluation.** Data is often missing in biased ways. Can we design a benchmark/dataset that allows us to evaluate modeling approaches in practice?