Mitigating Racial Bias in Clinical Prediction Models: A D3M-Inspired Data Selection Approach

DBDS Student Talks Presentation

Rebecca Hurwitz

May 5, 2025

Stanford University

Agenda

1. Research Question Formulation
2. Experimental Outline
3. Background
4. Results
5. Analysis
6. Future Directions
7. Acknowledgments

Research Question Formulation

What are computational strategies to mitigate race-based outputs in LLMs?

- Machine learning models often reflect biases present in training data.
- Models inadvertently learn and amplify these biases, leading to unfair or inaccurate predictions.
- The sheer size of datasets makes it impossible to manually check and correct every data point.
- This necessitates **automated methods for debiasing datasets.**

What are computational strategies to mitigate race-based outputs in LLMs?

What are computational strategies to mitigate race-based outputs in LLMs?

What is the application area in which I will ask and answer this question?

Experimental Outline

- **Objective**: Improve worst subgroup performance for racial, protected, and underresourced groups using debiased fine-tuning.

Experimental Outline

- **Objective**: Improve worst subgroup performance for racial, protected, and underresourced groups using debiased fine-tuning.

- **Method**: Apply data selection method (D3M-inspired) to balance training datasets.

Experimental Outline

- **Objective**: Improve worst subgroup performance for racial, protected, and underresourced groups using debiased fine-tuning.
- **Method**: Apply data selection method (D3M-inspired) to balance training datasets.
- **Expected Outcome**: Boost accuracy in underperforming subgroups and achieve fairer, more robust models.

Background

How do we quantify data's impact in training?

- **Data attribution:** the task of predicting model outputs/behavior at test-time as a function of the input training data.

- In other words: *What would happen if I trained the model on a given subset of my training set?*

Background

- **TRAK**: data attribution method giving us coefficients (scores) to help identify examples that exacerbate discrepancies in group performance

- **D3M**: allows us to actually remove the examples and retrain on a dataset without the harmful examples

we set $\boldsymbol{\beta}$ = 1 (hyperparameter controlling the smoothness of the maximum)

loss of a base classifier $\boldsymbol{\theta}(S)$ on group g (evaluated on the validation set)

$$A_i = \frac{\sum_{g \in \mathcal{G}} \exp(\beta \ell_g) \cdot \tau(g)_i}{\sum_{g' \in \mathcal{G}} \exp(\beta \ell_{g'})}$$

group alignment score: the impact of training sample i on the overall worst-group performance

the i-th coefficient for group g

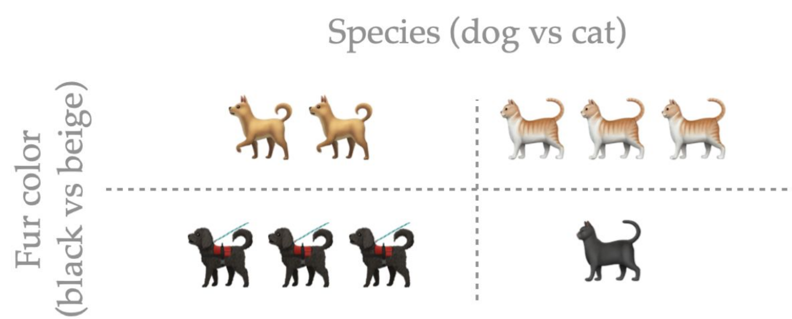
TRAK: Tracing with the Randomly-projected After Kernel

Algorithm 1 TRAK for multi-class classifiers (as implemented)

Input: Learning algorithm $\mathcal{A}$, dataset S of size n, sampling fraction $\alpha \in (0,1]$, correct-class likelihood function $p(z;\theta)$, projection dimension $k \in \mathbb{N}$
Output: Matrix of attribution scores $\mathbf{T} \in \mathbb{R}^{n \times n}$
$f(z;\theta) := \log\left(\frac{p(z;\theta)}{1-p(z;\theta)}\right)$ ▷ Margin function f_θ
for $m \in \{1, \dots, M\}$ **do**
    Sample random $S' \subset S$ of size $\alpha \cdot n$
    $\theta^\star_m \leftarrow \mathcal{A}(S')$ ▷ Train a model on S'
    $\mathbf{P} \sim \mathcal{N}(0,1)^{p \times k}$ ▷ Sample projection matrix
    $\mathbf{Q}^{(m)} \leftarrow \mathbf{0}_{n \times n}$
    for $i \in \{1, \dots, n\}$ **do**
        $\phi_i \leftarrow \mathbf{P}^\top \nabla_\theta f(z_i;\theta^\star_m)$ ▷ Compute gradient at $\theta^\star_m$ and project to k dimensions
        $\mathbf{Q}^{(m)}_{ii} \leftarrow 1 - p(z_i;\theta^\star)$ ▷ Compute weighting term
    end for
    $\Phi_m \leftarrow [\phi_1; \cdots; \phi_n]^\top$
end for
$\mathbf{T} \leftarrow \left[\frac{1}{m} \sum_{m=1}^{M} \Phi_m (\Phi_m^\top \Phi_m)^{-1} \Phi_m^\top\right] \left[\frac{1}{m} \sum_{m=1}^{M} \mathbf{Q}^{(m)}\right]$
return SOFT-THRESHOLD($\mathbf{T}$)

D3M: Data Debiasing via Datamodeling

Training data correlation between **class (species)** and **extra feature (color)** leads to disparate performance.



Goal: "debias" dataset to improve *worst-group accuracy* (WGA):

$$\text{WGA} = \min_{\text{group} \in \{\text{🐕}, \text{🐕‍🦺}, \text{🐈}, \text{🐈‍⬛}\}} \text{Acc(group)}$$

D3M: Data Debiasing via Datamodeling

Our approach: Data Debiasing via Datamodeling (D3M)

Compute impact of each training sample on WGA by predicting WGA as a function of dataset selection.

Worst-group accuracy on a (small) validation set

"Group alignment score"
Learned coefficient; impact of point i on WGA

$$\text{WGA} \approx \sum D_i \cdot A_i$$

Binary; whether we select the i-th training sample

Find selection that maximizes worst-group accuracy by removing most harmful examples.

Group alignment scores

0.3 0.2 0.1 0.0 -1.2 1.2 0.1 0.2 -0.4

+ Only changes data
+ Less accuracy gap
+ Competitive accuracy
+ Only needs test set labels

D3M: Data Debiasing via Datamodeling

- Given a training dataset S_{train} and a validation dataset S_{val}, produce a classifier *f* to minimize **worst case loss** over predefined subpopulations

data distribution $\mathscr{D}$ over triplets (x_i, y_i, g_i)

loss function that quantifies the error of the classifier *f* in predicting the label *y* for input *x*

$$\max_{g' \in \mathcal{G}} \mathbb{E}_{(x,y,g)\sim\mathcal{D}}\left[\ell(f(x),y)\,|\,g=g'\right]$$

set of distinct subpopulations in the data $\mathscr{G}$

input $x_i \in X$
label $y_i \in Y$
subgroup label $g_i \in \mathscr{G}$

(loss function is being evaluated conditional on the data belonging to subgroup g')

What are computational strategies to mitigate race-based outputs in LLMs?

What is the application area in which I will ask and answer this question?

What are computational strategies to mitigate race-based outputs in LLMs?

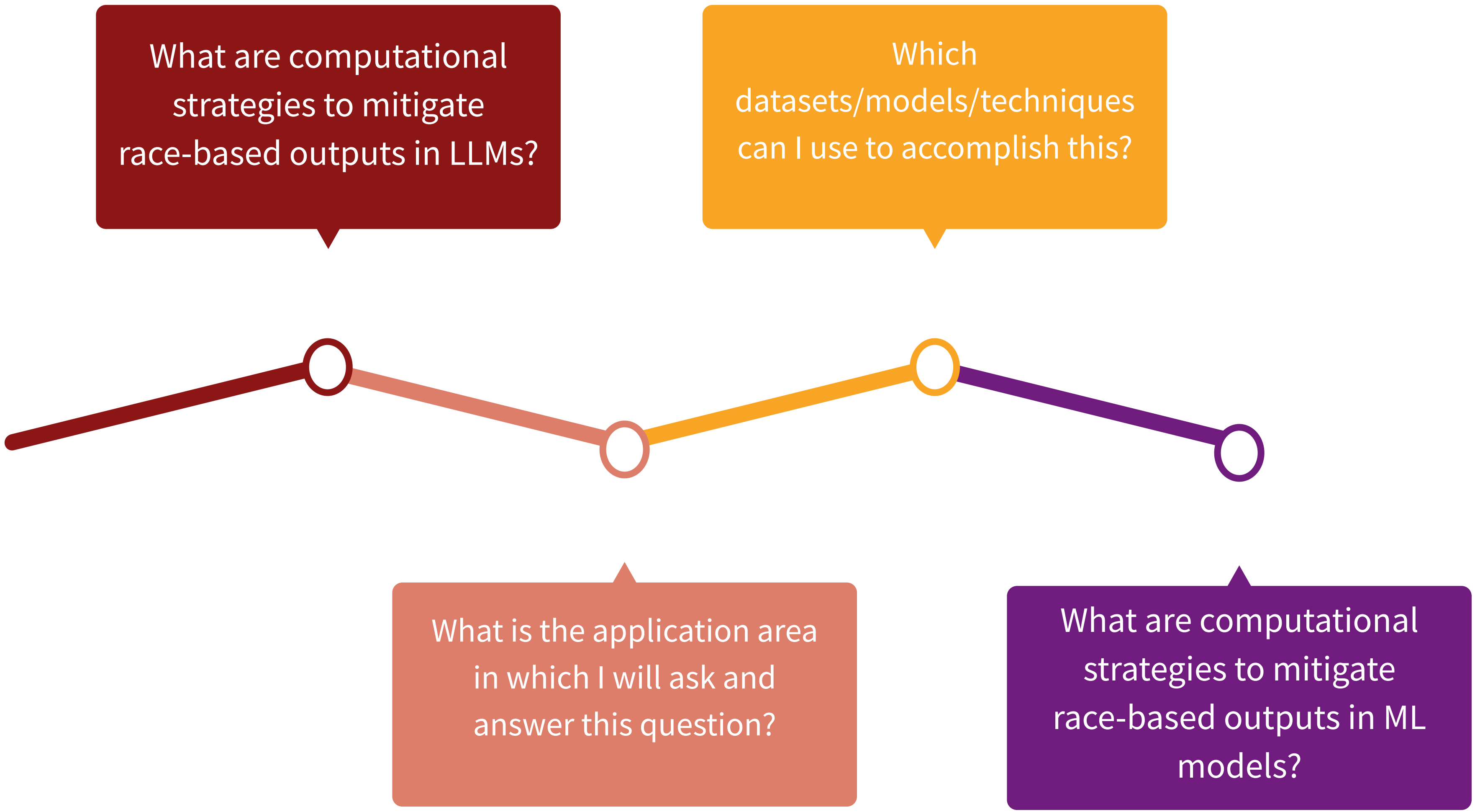
Which datasets/models/techniques can I use to accomplish this?

What is the application area in which I will ask and answer this question?

Experimental Plan

Experimental Plan

- **Datasets**: MIMIC-CXR (metadata) + MIMIC-CXR-JPG (images and CheXpert labels) + MIMIC-IV (demographic/racial data)
- **Models**: ResNet-9, ResNet-50
- **Procedure**: Break down 14 classes into binary datasets. Find the classes with the most discrepancies between groups. Train model - after 5 epochs training, apply 10 epochs of D3M + separately 10 epochs of normal training (15 each).


What are computational strategies to mitigate race-based outputs in LLMs?
What is the application area in which I will ask and answer this question?
Which datasets/models/techniques can I use to accomplish this?
What are computational strategies to mitigate race-based outputs in ML models?

Experimental Plan

Preprocessed various MIMIC data to produce dataset of images with chexpert labels and demographic data. Started with 10,000 chest x-rays, reduced to 2,635 in PA position.

Created 14 datasets (one for each CheXpert image class) to train models to predict yes/no if condition is present. Split into 80:20 train:test. We chose ResNet-50.

Control

Trained model with 15 epochs

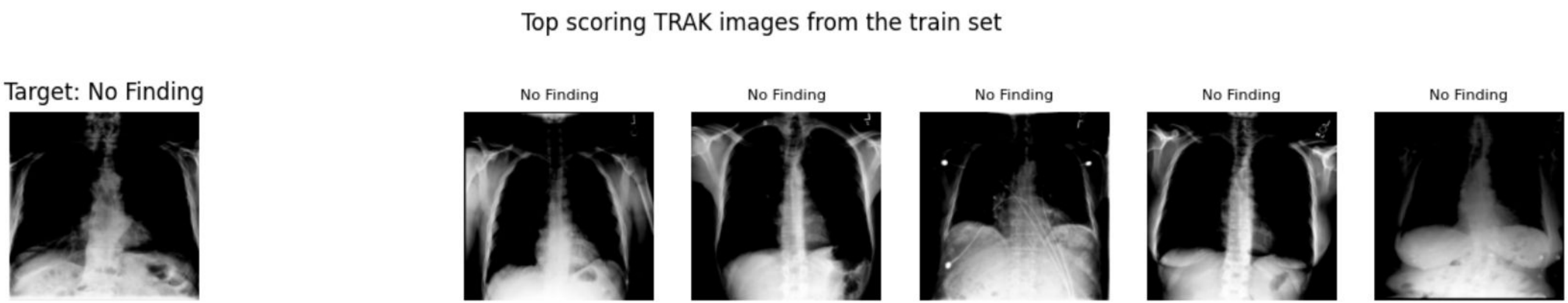
Experiment

Trained model with 5 epochs, changed training data with D3M, trained for 10 more epochs (15 total)

Evaluated on test set and stratified by race. Compared performance between control and experiment.

Results

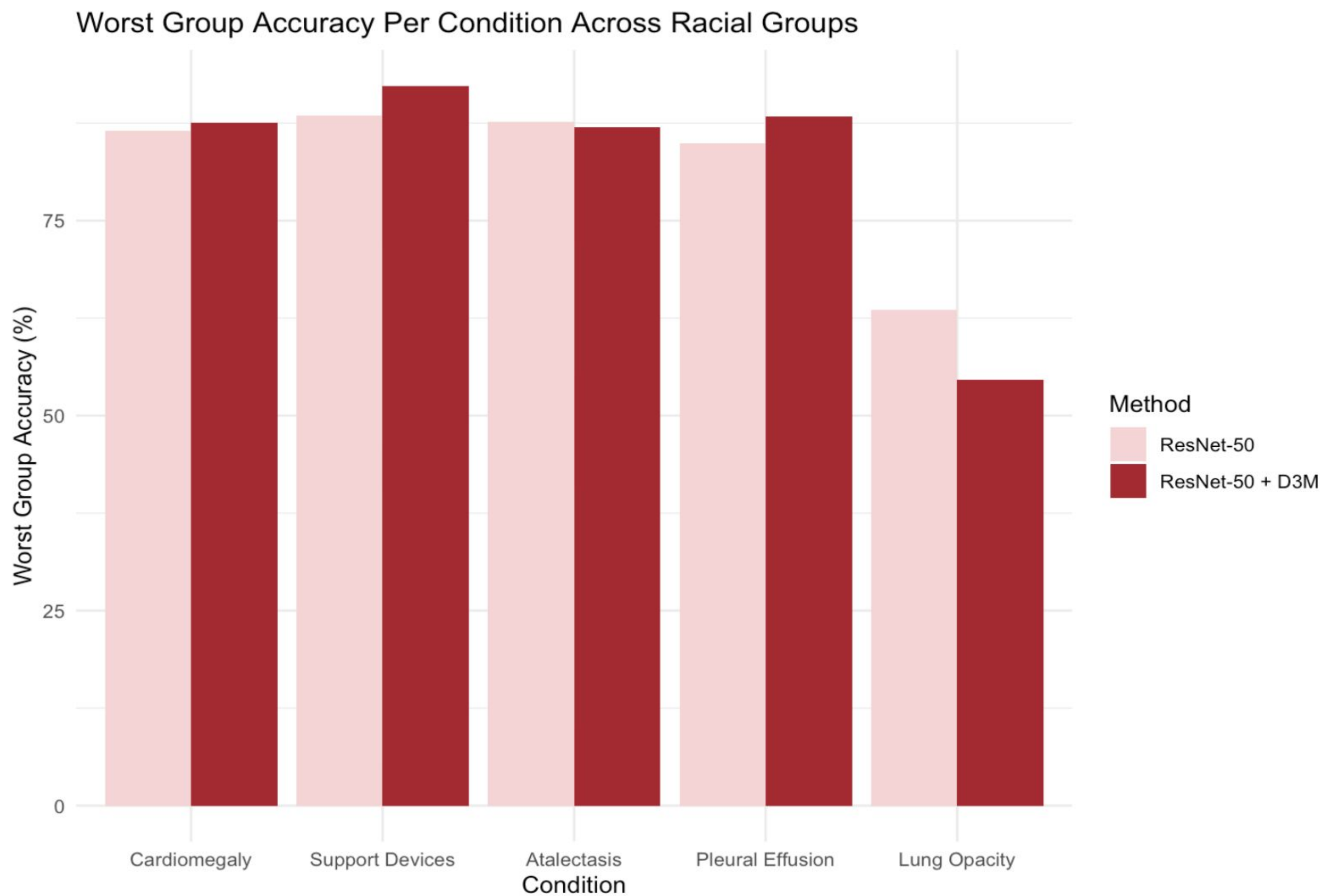
TRAK


Top scoring TRAK images from the train set
Target: No Finding
No Finding
No Finding
No Finding
No Finding
No Finding

Worst Group Accuracies

Method	Worst Group Accuracy (%)				
	Cardiomegaly	Atalectasis	Support Devices	Pleural Effusion	Lung Opacity
ResNet-50	86.5%	**87.7%**	88.5%	84.9%	**63.6%**
ResNet-50 + D3M	**87.6%**	87.0%	**92.3%**	**88.4%**	54.6%

Worst Group Accuracies



Cardiomegaly

Race	Accuracy (%)			Δ in n
	ResNet-50 (n)	ResNet-50 + D3M (n)	Δ	
Asian	100.0 (65)	100.0 (51)	0.0	-14
★ Black or African American	86.5 (336)	**87.6 (291)**	+1.1	-45
Hispanic/Latino	92.1 (148)	92.1 (132)	0.0	-16
Other	95.2 (60)	95.2 (52)	0.0	-8
Unknown	93.9 (425)	93.9 (393)	0.0	-32
White	89.4 (1072)	88.3 (976)	-1.1	-96
Overall	**90.5 (2108)**	**90.1 (1897)**	**0.0**	**-211**

Support Devices

Race	Accuracy (%)			Δ in n
	ResNet-50 (n)	ResNet-50 + D3M (n)	Δ	
Asian	100.0 (65)	100.0 (48)	0.0	-17
Black or African American	98.9 (336)	98.9 (293)	0.0	-43
★ Hispanic/Latino	88.5 (148)	**92.3 (139)**	+3.8	-9
Other	100.0 (60)	100.0 (64)	0.0	4
Unknown	97.2 (425)	**97.4 (377)**	+0.2	-48
White	88.9 (1072)	**89.3 (976)**	+0.4	-96
Overall	**93.0 (2108)**	**93.4 (1897)**	**0.4**	**-211**

Atalectasis

Race	Accuracy (%)			Δ in n
	ResNet-50 (n)	ResNet-50 + D3M (n)	Δ	
Asian	93.8 (65)	93.8 (52)	0.0	-13
Black or African American	92.8 (336)	92.8 (299)	0.0	-37
Hispanic/Latino	91.4 (148)	91.4 (128)	0.0	-20
Other	91.7 (60)	**95.8 (50)**	+4.1	-10
Unknown	90.8 (425)	89.1 (373)	-1.7	-52
★ White	87.7 (1072)	87.0 (971)	-0.7	-101
Overall	**89.8 (2108)**	**89.2 (1897)**	-0.6	**-211**

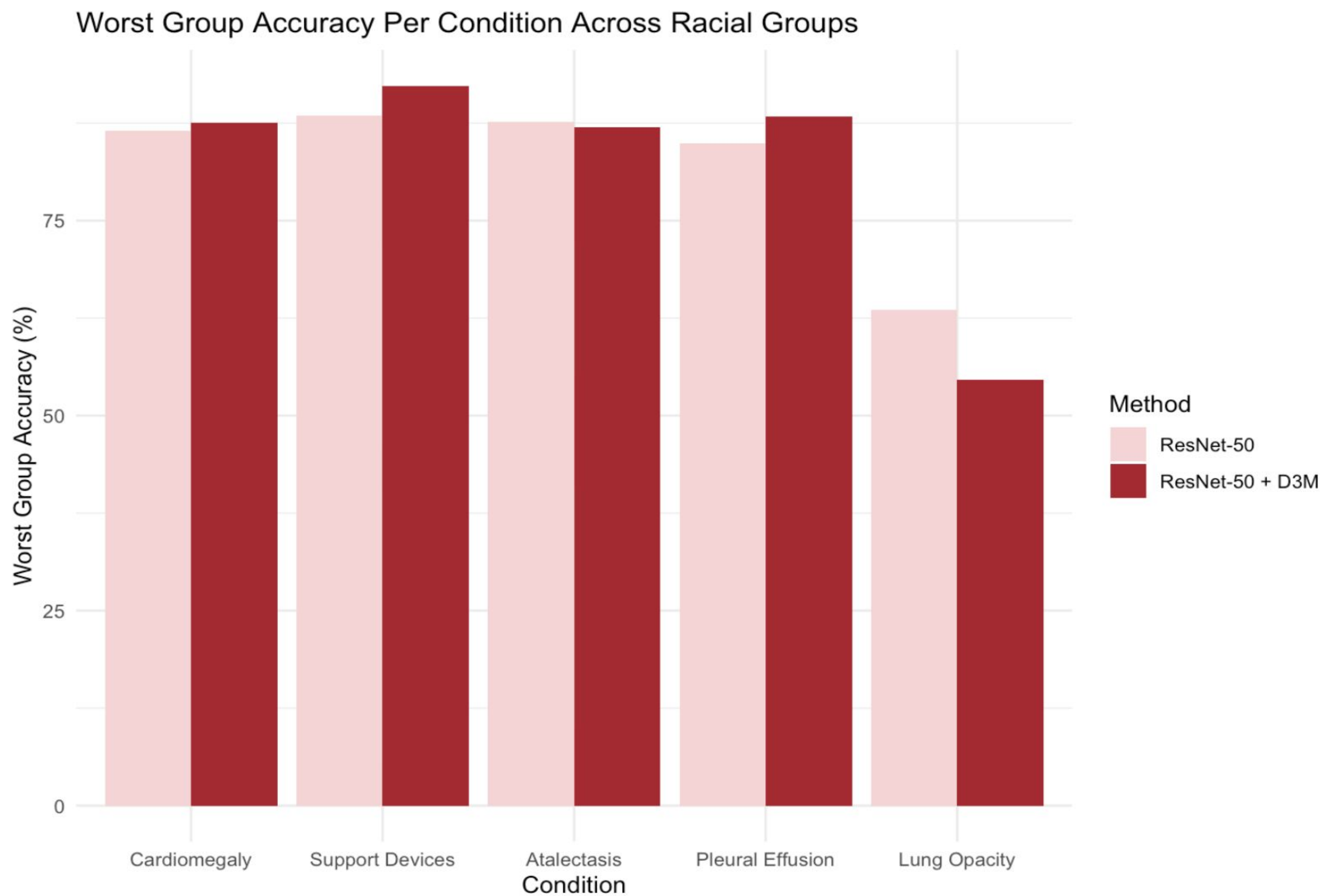
Pleural Effusion

Race	Accuracy (%)			Δ in n
	ResNet-50 (n)	ResNet-50 + D3M (n)	Δ	
Asian	100.0 (65)	100.0 (59)	0.0	-6
★ Black or African American	84.9 (336)	**88.4 (303)**	+3.5	-33
Hispanic/Latino	91.9 (148)	86.5 (133)	-5.4	-15
Other	100.0 (60)	100.0 (54)	0.0	-6
Unknown	95.9 (425)	94.9 (355)	-1.0	-70
White	88.4 (1072)	**89.1 (971)**	+0.7	-101
Overall	**90.1 (2108)**	**90.5 (1897)**	**0.4**	**-211**

Lung Opacity

Race	Accuracy (%)			Δ in n
	ResNet-50 (n)	ResNet-50 + D3M (n)	Δ	
★ Asian	63.6 (65)	54.6 (60)	-9.1	-5
Black or African American	87.3 (336)	82.3 (299)	-5.0	-37
Hispanic/Latino	87.6 (148)	87.5 (132)	-0.1	-16
Other	89.5 (60)	89.5 (52)	0.0	-8
Unknown	92.6 (425)	92.3 (379)	-0.3	-46
White	85.2 (1072)	85.2 (975)	0.0	-97
Overall	**86.7 (2108)**	**85.8 (1897)**	-0.9	**-211**

Worst Group Accuracies



Quick note on D3M examples removed

```python
def get_debiased_train_indices(
    self, group_alignment_scores, use_heuristic=True, num_to_discard=None
):
    """
    If use_heuristic is True, training examples with negative score will be discarded,
    and the parameter num_to_discard will be ignored
    Otherwise, the num_to_discard training examples with lowest scores will be discarded.
    """
    if use_heuristic:
        return [i for i, score in enumerate(group_alignment_scores) if score >= 0]

    if num_to_discard is None:
        raise ValueError("num_to_discard must be specified if not using heuristic.")

    sorted_indices = sorted(
        range(len(group_alignment_scores)),
        key=lambda i: group_alignment_scores[i],
    )
    return sorted_indices[num_to_discard:]
```

Analysis

Analysis

- **Did it work**? D3M increased accuracy of some subgroups but decreased accuracy of others.

- **Considerations:** Dataset size, domain, type of classification.

- **Purpose:** The intended purpose of D3M in our context was specifically to improve accuracy of minority groups.

What are computational strategies to mitigate race-based outputs in LLMs?

What is the application area in which I will ask and answer this question?

Which datasets/models/techniques can I use to accomplish this?

What are computational strategies to mitigate race-based outputs in ML models?

What are computational strategies to mitigate race-based outputs in LLMs?

What is the application area in which I will ask and answer this question?

Which datasets/models/techniques can I use to accomplish this?

What are computational strategies to mitigate race-based outputs in ML models?

Future Directions

- More data!
- Error analysis - what examples contributed to misclassification, specifically for which race?
- Other tasks beyond chest x-rays

Thank You!

Huge thank you to Nigam!

Acknowledgements

National Library of Medicine

Research funded by T15LM007033

Stanford MEDICINE
Department of Biomedical Data Science

Dept of Biomedical Data Science:
Sylvia Plevritis
Dennis Wall
Karen Matthys
Iffat Ahmed
Asmaa Eljibahi
Erica Peterson

Shah Lab

Shah Lab:
Nigam Shah
Alyssa Unell

Citations

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., & Madry, A. (2023). TRAK: Attributing Model Behavior at Scale. arXiv preprint arXiv:2303.14186.

Jain, S., Hamidieh, K., Georgiev, K., Ilyas, A., Ghassemi, M., & Madry, A. (2024). Data Debiasing with Datamodels (D3M): Improving Subgroup Robustness via Data Selection. arXiv preprint arXiv:2406.16846.

I learned SO MUCH this rotation!

- Cuda
- Semaphore
- Not to use the login node
- Research Park
- Srun
- Importance of word choice

Background

- **TRAK**: data attribution method giving us coefficients (scores) to help identify examples that exacerbate discrepancies in group performance

- **D3M**: allows us to actually remove the examples and retrain on a dataset without the harmful examples