



CIM CENTRE FOR INTELLIGENT MACHINES



On Learning Fairness and Accuracy on Multiple Subgroups

Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li,
Charles X. Ling, Tal Arbel, Boyu Wang, Christian Gagné



Institut
intelligence
et données



AI in sociotechnical system



Candidate evaluations for job positions

Driven by AI algorithms



Health risk assessment

Algorithmic Discrimination

Medical AI

The screenshot shows the top portion of a Science journal article page. At the top left is the Science logo in red. To its right are navigation links: 'Current Issue', 'First release papers', 'Archive', and 'About' with a dropdown arrow. A 'Submit manuscript' button is on the far right. Below the navigation is a 'RESEARCH ARTICLE' label. To the right of this label are social media icons for Facebook, Twitter, LinkedIn, YouTube, and Email. The main title of the article is 'Dissecting racial bias in an algorithm used to manage the health of populations'. Below the title, the authors are listed: 'ZIAD OBERMEYER', 'BRIAN POWERS', 'CHRISTINE VOGELI, AND', 'SENDHIL MULLAINATHAN', with ORCID iD icons and a link to 'Authors Info & Affiliations'. Below the authors, the journal information is provided: 'SCIENCE • 25 Oct 2019 • Vol 366, Issue 6464 • pp. 447-453 • DOI: 10.1126/science.aax2342'. At the bottom of this section, there are icons for downloads (14,249), citations (576), a notification bell, a bookmark icon, a quote icon, and a red 'GET ACCESS' button. Below this is a grey box containing the text 'Racial bias in health algorithms'.

Obermeyer et al., 366 Science 447 (2019)

Group Fairness

No prediction disparities in different demographics.

- *Age, gender, race, hospital.....*
- *No unified definitions.*

Trivial Fair Decision

Coin flipping can trivially achieve fair prediction.

- For any job application, the offer is random.

The prediction should be **informative!**



Source: https://en.wikipedia.org/wiki/Coin_flipping

Desiderata in fair learning

- Informative.
 - *Learning the utility of the data*
- Fair
 - *No prediction disparities*

Possibility to simultaneously achieve these two?
Depending on fairness notion.

Group sufficiency

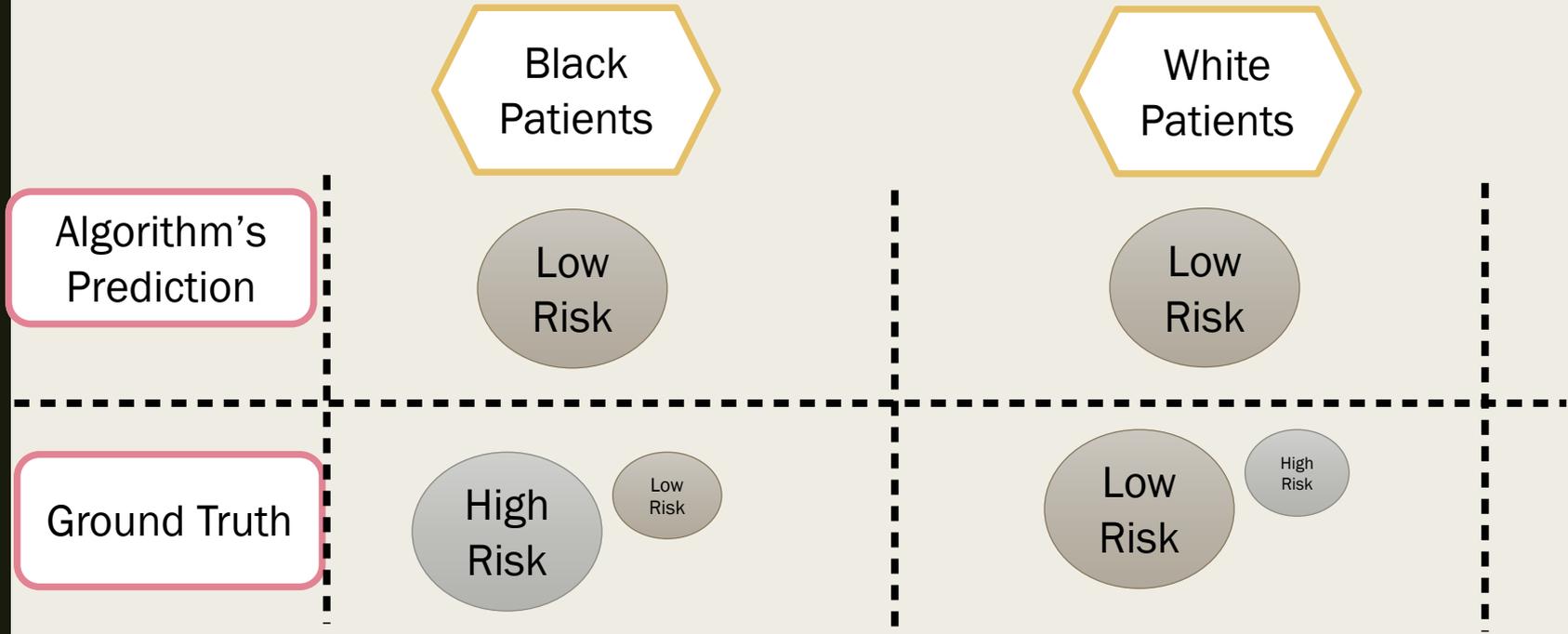
Example in Health AI

- AI algorithms predict the **health-care score** for each patient.
- Higher score -> Sicker

(need to transfer to ICU)

Obermeyer et al., 366 *Science* 447 (2019)

Calibration Bias (Example in Health)



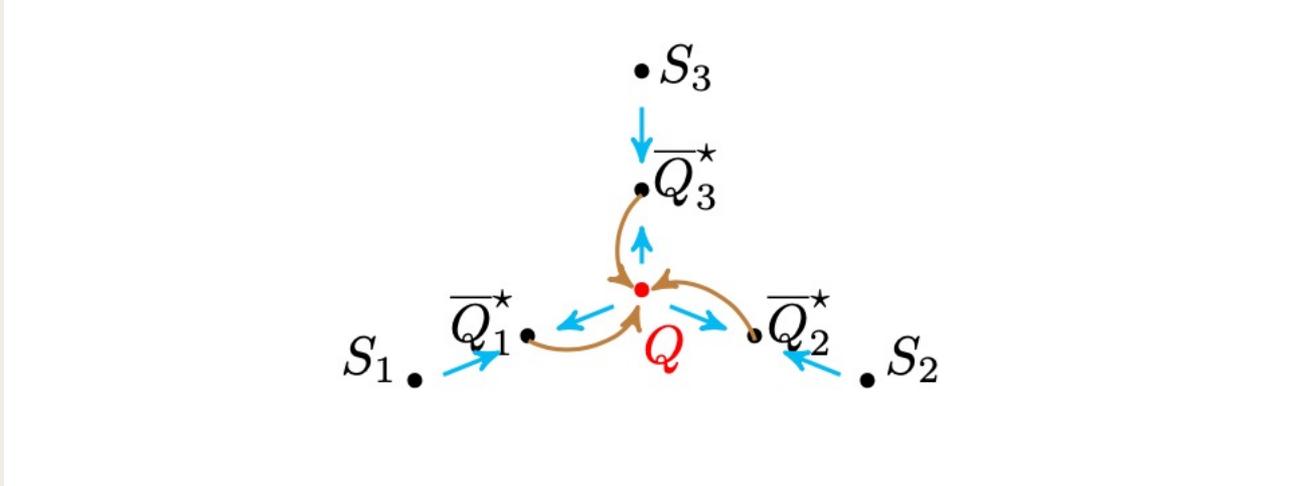
Severity of Black patients is **under-estimated**.

Obermeyer et al., 366 *Science* 447 (2019)

Formal definition

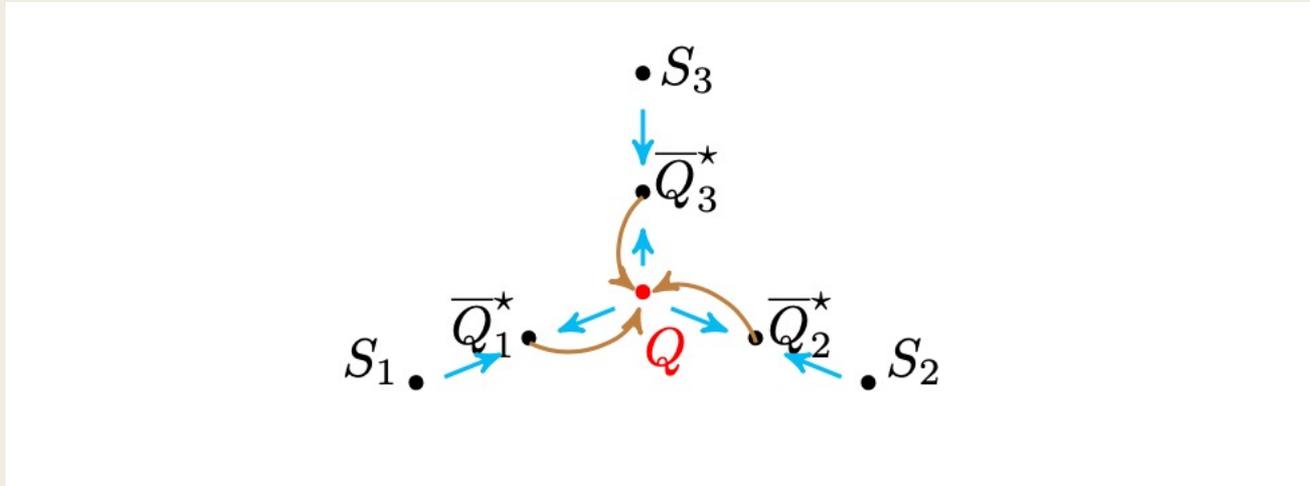
- Group sufficiency: $E[Y|f(X)] = E[Y|f(X), A]$
- Mitigate **bias** across multiple (or many) subgroups
- Limited data within each subgroup
- Learning **data utility with comparable accuracy**

Proposed algorithm (informal)



1. Q : fair and informative predictor.
2. S_1, S_2, S_3 : different subgroup (e.g., data from different ages)
3. Q_1, Q_2, Q_3 : subgroup specific predictors

Step One

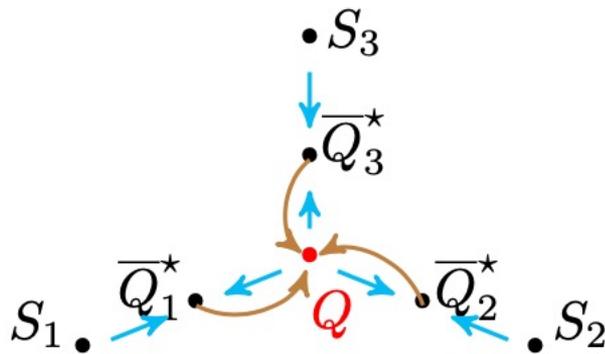


Fix Q (the fair and informative predictor.)

-> Learn Q_1, Q_2, Q_3 from S_1, S_2, S_3 and Q (shared prior information).

Return: Q_1, Q_2, Q_3

Step Two



Fix Q_1, Q_2, Q_3 (the subgroup specific predictor.)

-> Learn Q (shared prior information) from Q_1, Q_2, Q_3 to be closed as possible.

Return: Q

Theoretically prove a fair and informative predictor (see paper).

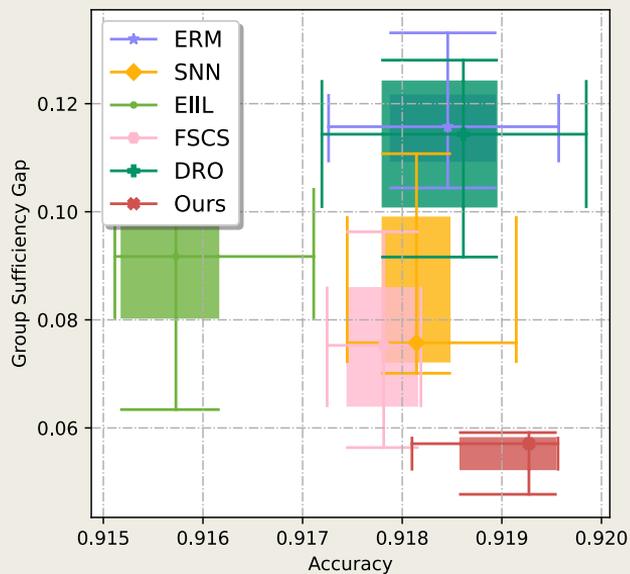
Application (Toxic comments recognition)

Toxic	Comment Text
0	I applaud your father. He was a good man! We need more like him.
0	As a Christian, I will not be patronizing any of those businesses.
0	What do Black and LGBT people have to do with bicycle licensing?
0	Government agencies track down foreign baddies and protect law-abiding white citizens. How many shows does that describe?
1	Maybe you should learn to write a coherent sentence so we can understand WTF your point is.

- Predict the comment being toxic or not.
- Machine Learning algorithms showed biased results on the different **rac**es.

Source: <https://arxiv.org/pdf/2012.07421.pdf>

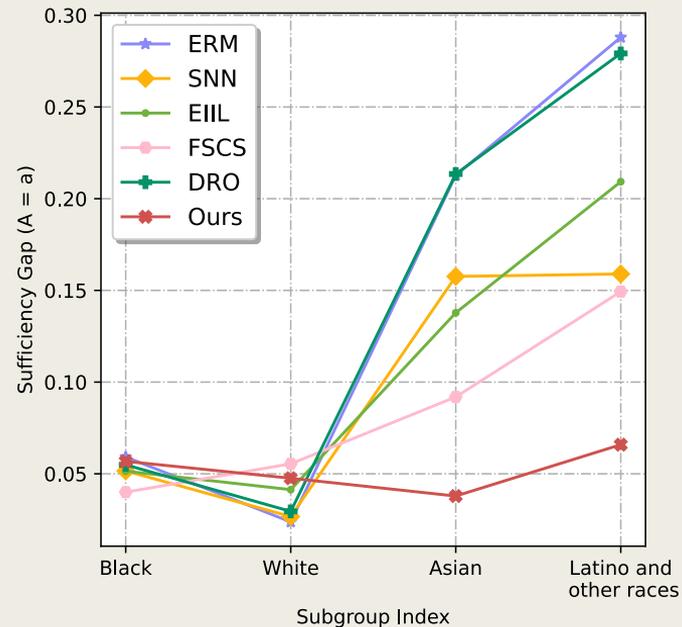
Application (Toxic comments recognition)



Accuracy

Our Framework

higher accuracy
lower sufficiency gap



Different demographics

small sufficiency gap for each group

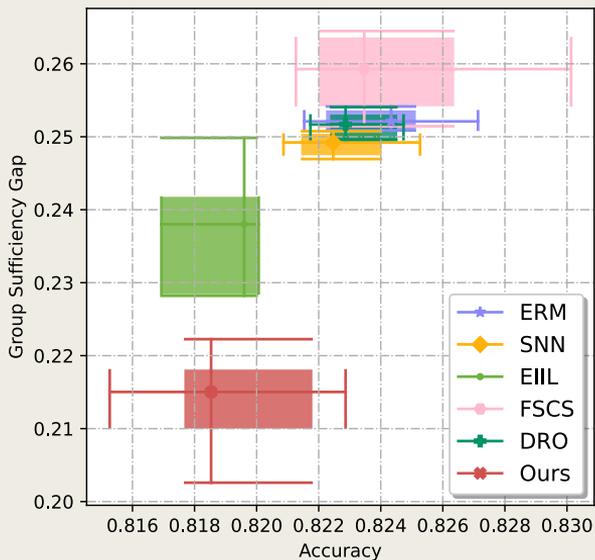
Application (Amazon reviews)

Reviewer ID (d)	Review Text (x)	Stars (y)
Train	Reviewer 1 They are decent shoes. Material quality is good but the color fades very quickly. Not as black in person as shown. Super easy to put together. Very well built.	5
	Reviewer 2 This works well and was easy to install. The only thing I don't like is that it tilts forward a little bit and I can't figure out how to stop it. Perfect for the trail camera	4
	...	5
	Reviewer 10,000 I am disappointed in the quality of these. They have significantly deteriorated in just a few uses. I am going to stick with using foil.	1
	Very sturdy especially at this price point. I have a memory foam mattress on it with nothing underneath and the slats perform well.	5

- Predict the *star* from the review.
- Machine Learning algorithms showed biased results on different **clients**.

Source: <https://arxiv.org/pdf/2012.07421.pdf>

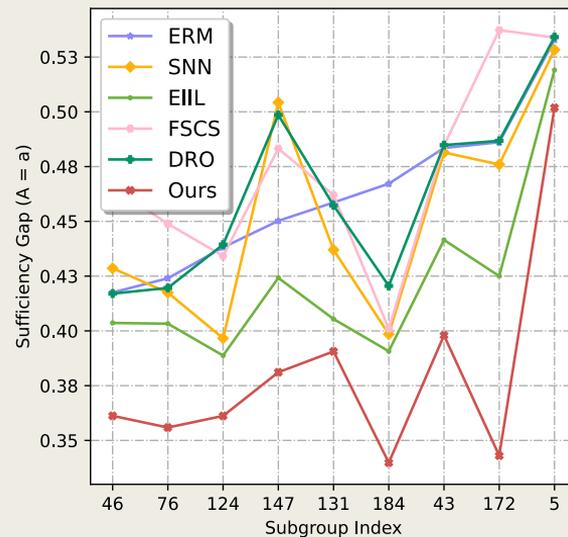
Application (Amazon reviews)



Accuracy

Our Framework

comparable accuracy
lower group sufficiency gap



Different clients

small group sufficiency gap for each client

Conclusions

- A novel provable framework:
 - Mitigate group sufficiency bias;
 - Preserve the utility of data;

Thank you!