Fraunhofer MEVIS

Fraunhofer Institute for Digital Medicine MEVIS

Bias and Fairness in Medical Imaging

MICCAI FAIMI 2025 Tutorial

Eike Petersen, Tareen Dawood, Miguel López-Pérez
on behalf of FAIMI

i. **Introduction**

ii. „Lecture“: Causes of bias

iii. Case study 1

iv. Coffee break (10:00 - 10:30)

v. Case study 2

vi. “Lecture”: Bias mitigation

vii. Conclusion & recommendations

Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew B. A. McDermott, Irene Y. Chen & Marzyeh Ghassemi

Nature Medicine **27**, 2176–2182 (2021) | Cite this article

Disparities in dermatology AI performance on a diverse, curated clinical image set

Roxana Daneshjou[1,2]†, Kailas Vodrahalli[3]†, Roberto A. Novoa[1,4], Melissa Jenkins[1], Weixin Liang[5], Veronica Rotemberg[6], Justin Ko[1], Susan M. Swetter[1], Elizabeth E. Bailey[1], Olivier Gevaert[2], Pritam Mukherjee[2]‡, Michelle Phung[1], Kiana Yekrang[1], Bradley Fong[1], Rachna Sahasrabudhe[1]§, Johan A. C. Allerup[1], Utako Okata-Karigane[7], James Zou[2,3,5,8]*, Albert S. Chiou[1]*

Higher performance for women than men in MRI-based Alzheimer's disease detection

Malte Klingenberg[1,2], Didem Stark[1,2], Fabian Eitel[1,2], Céline Budding[3], Mohamad Habes[4], Kerstin Ritter[1,2]*for the Alzheimer's Disease Neuroimaging Initiative

Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation

Esther Puyol-Antón[1(✉)], Bram Ruijsink[1,2], Stefan K. Piechnik[7], Stefan Neubauer[7], Steffen E. Petersen[3,4,5,6], Reza Razavi[1,2], and Andrew P. King[1]

www.faimi.org

Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus recommendations

AI Act requires:

„… examination in view of possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law“…

„… appropriate measures to detect, prevent and mitigate possible biases identified according to [the previous point]…“

RSNA on **ACA Section 1557:**

“… all covered entities, including radiologists and radiology practices, accountable for preventing discrimination, including any bias arising from the use of algorithms or AI.”

“… even if you are not the developer of the AI tool or algorithm, you are still obligated to make reasonable efforts to ensure its use doesn’t result in discriminatory practices.”

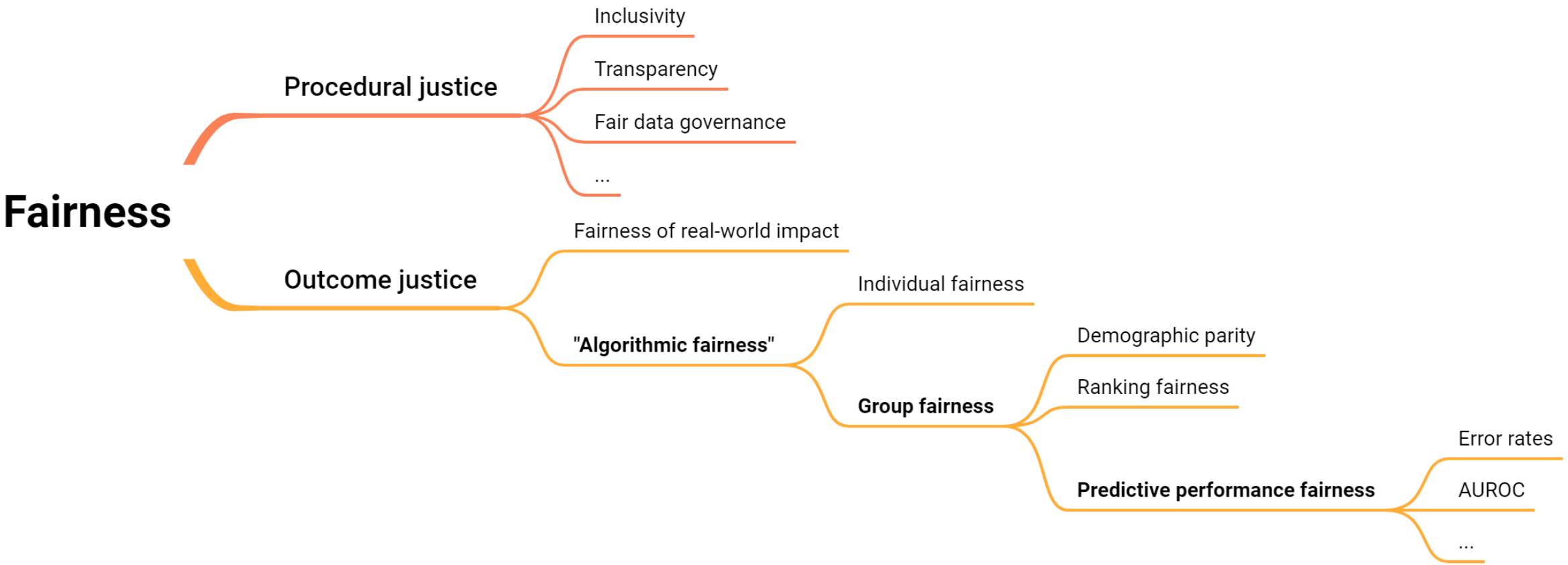
“Ensure decision support tools are non-discriminatory” by May 1st, 2025.

FDA Draft guidance:

„The performance and behavior of AI systems rely heavily on the quality, diversity, and quantity of data used to train and tune them. The accuracy and usefulness of a validation of an AI-enabled device also depends on the quality, diversity, and quantity of data used to test it."

„The characterization of sources of bias is necessary to assess the potential for AI bias in the AI enabled device."

„… it is important for FDA to understand how the device performs overall in the intended use population, as well as in subgroups of interest. Acceptable performance in certain subgroups may mask lower performance in other subgroups… Poor performance in specific subgroups could make the device unsafe for use in those groups…"



… Or: **Robustness / Reliability / Generalization / „Quality of Service bias.“**

We simply want AI systems that work well for all patients.

Three key steps.

1. **Bias assessment:** analyze potential performance disparities.

   (Can't fix problems we don't know about!)

2. **Bias root cause analysis**: *why* are these groups underperforming?

   (Can't fix problems effectively without knowing why they arise!
   Blind „black-box" bias mitigation can do more harm than good.)

3. **Bias mitigation**: reduce disparities by *specifically* addressing identified root causes in a targeted manner.

i. Introduction

ii. Causes of bias

iii. Case study 1

iv. Coffee break (10:00 - 10:30)

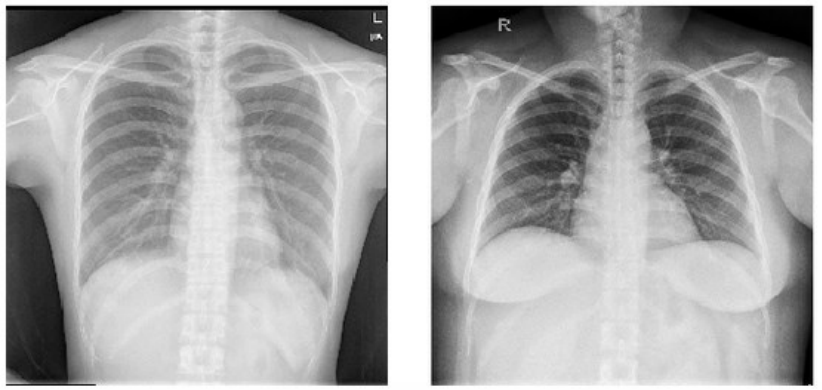
v. Case study 2

vi. Bias mitigation

vii. Conclusion & recommendations

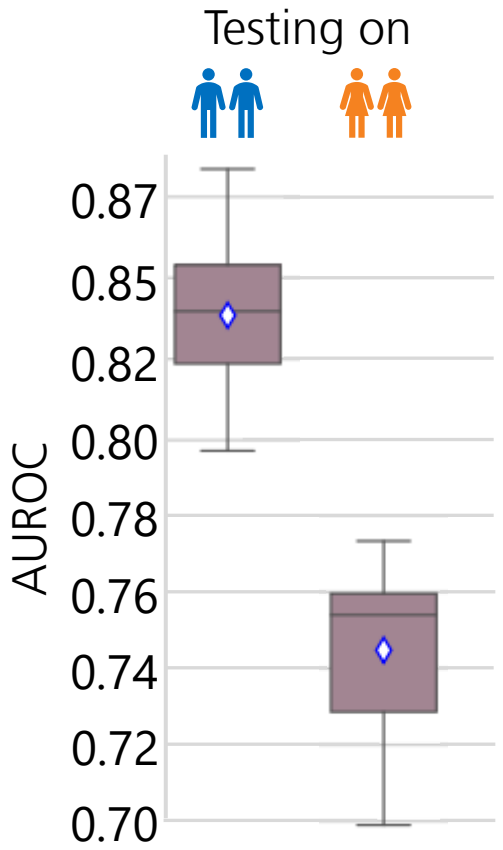
Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal[a,1], Nicolás Nieto[a,b,1], Victoria Peterson[b,c], Diego H. Milone[a], and Enzo Ferrante[a,2]



Classification Model (DenseNet-121)

Pneumothorax Yes / No

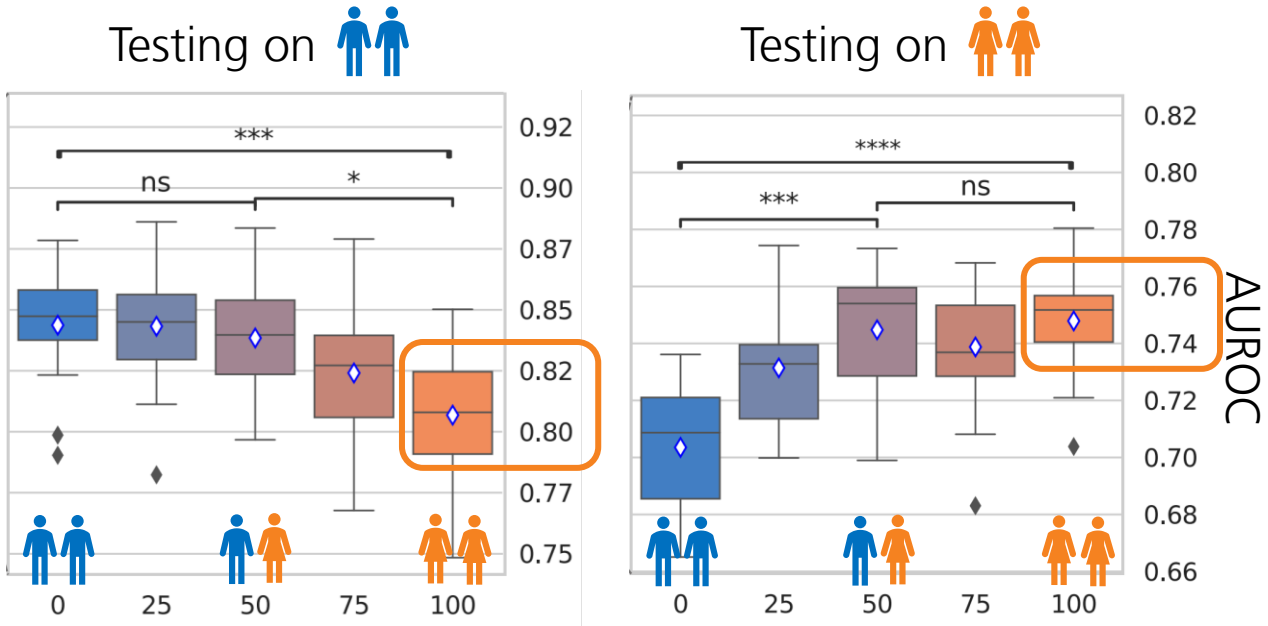


Why do models perform better in some groups compared to others?

(How) can we fix this?

Could be group representation…? No…

Causes of QoS bias:
- Underrepresentation



% Female patients in training set

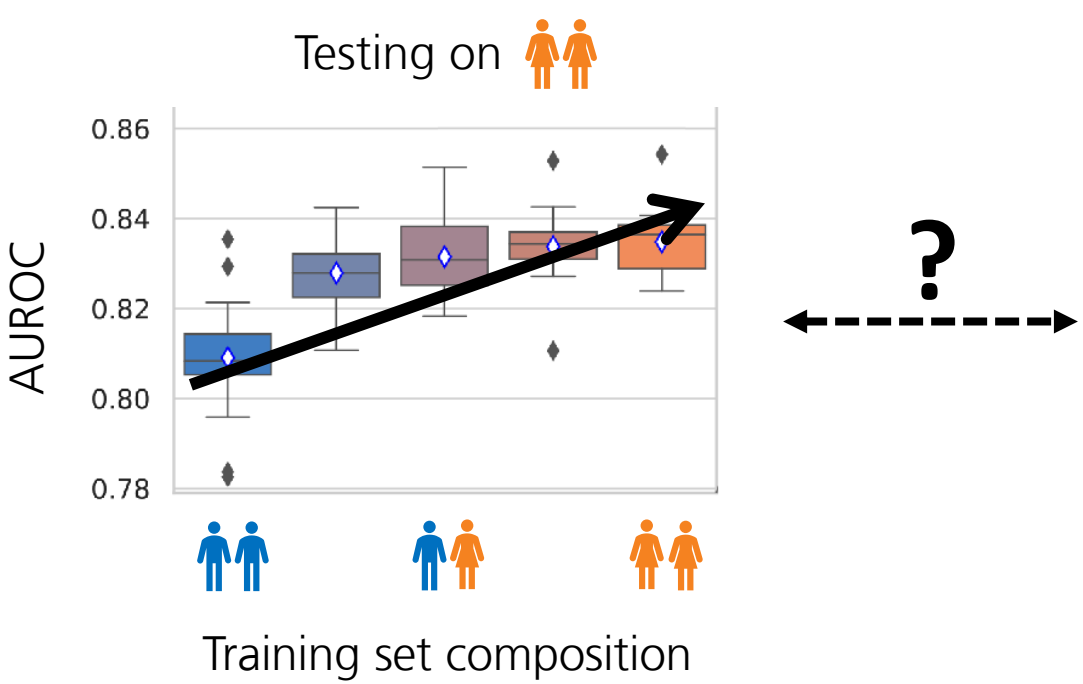
Males still outperforming females even when training *only* on females!

Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal[a,1], Nicolás Nieto[a,b,1], Victoria Peterson[b,c], Diego H. Milone[a], and Enzo Ferrante[a,2]

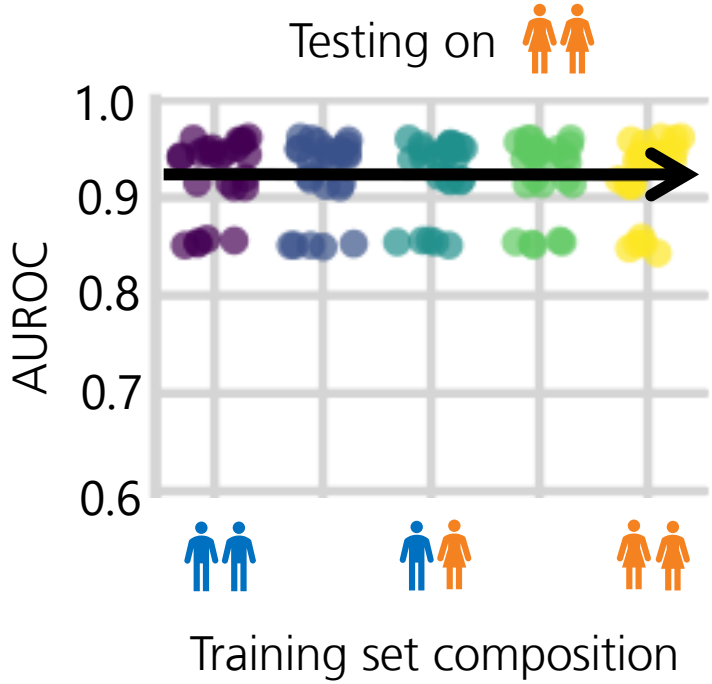
[Sidenote] Group representation does not *always* correlate with performance

Atelectasis detection in chest x-ray images



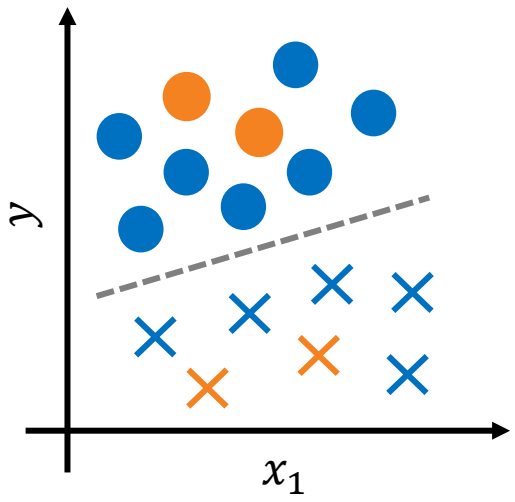
Larrazabal et al. (2020), Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. PNAS.

Alzheimer's detection in MRI images



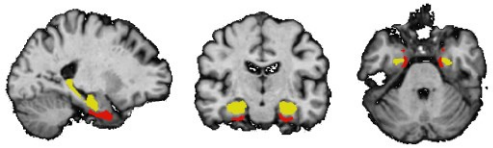
Petersen et al. (2022), Feature robustness and sex differences in medical imaging: a case study in MRI-based Alzheimer's disease detection. MICCAI.

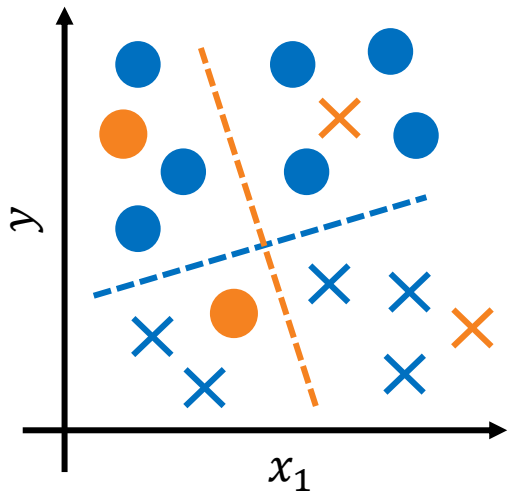
[Sidenote] How different are the groups?



Mapping from x to y similar across groups.

Under-representation not necessarily a problem.

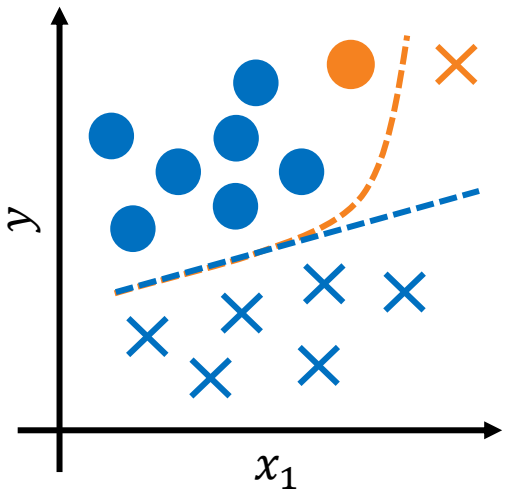




Mapping from x to y differs incompatibly between groups.

Majority group will be favored.

Unlikely to occur with images and DL.



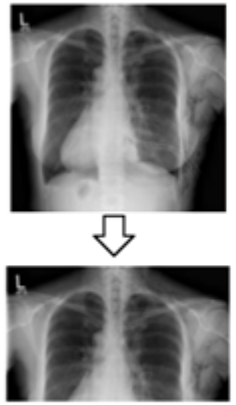
Groups occupy different regions in x-space.

The model *could* learn a mapping that works for all groups.

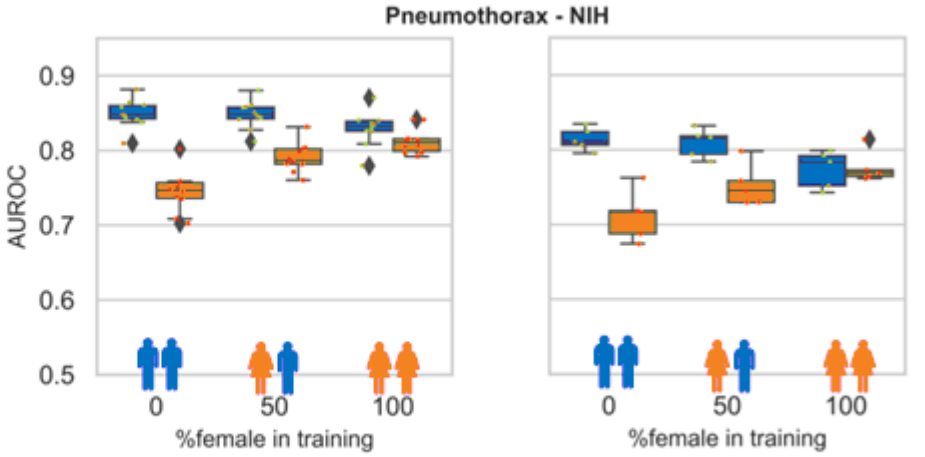
In practice, this often does not happen due to (implicit or explicit) regularization / convergence to local minima.

Could be morphological differences…? No…

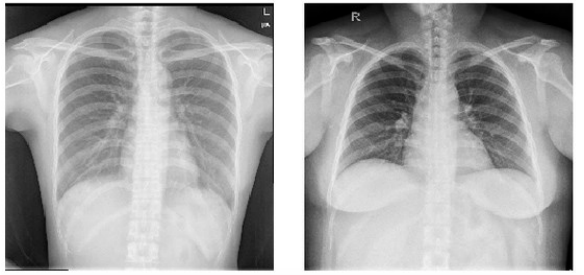
Causes of QoS bias:
- Underrepresentation
- Differences in task difficulty



(A) Illustration of cropped images



(B) Results from non-cropped images (C) Results from cropped images



Proc. Coll. Radiol. Aust. (1958), 2, 107

The Elimination of Confusing Breast Shadows in Chest Radiography

COLIN ALEXANDER

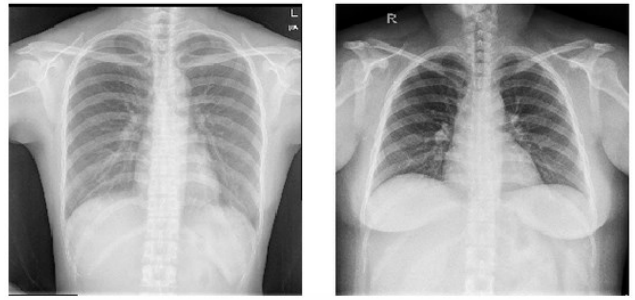
Are Sex-based Physiological Differences the Cause of Gender Bias for Chest X-ray Diagnosis?

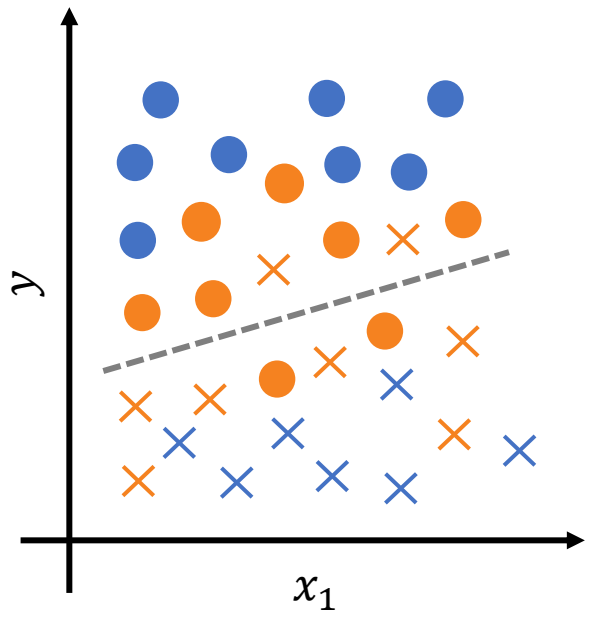
Nina Weng[1], Siavash Bigdeli[1], Eike Petersen[1], and Aasa Feragen[1]

→ Males *still* outperforming females even with breasts cropped

(+ performance trends are *reversed* in another dataset)

Task difficulty: Input disturbance characteristics





Def.: The input x is more noisy in some groups, or captures underlying pathology less well.

Ex.: Male vs. female chest x-ray, thin vs. obese abdominal ultrasound, small vs. large lesions, different disease stages, low-field MRI, ...

Effect: Correct *input* $\longmapsto$ *output* mapping should still be learned, but prediction performance will be reduced.

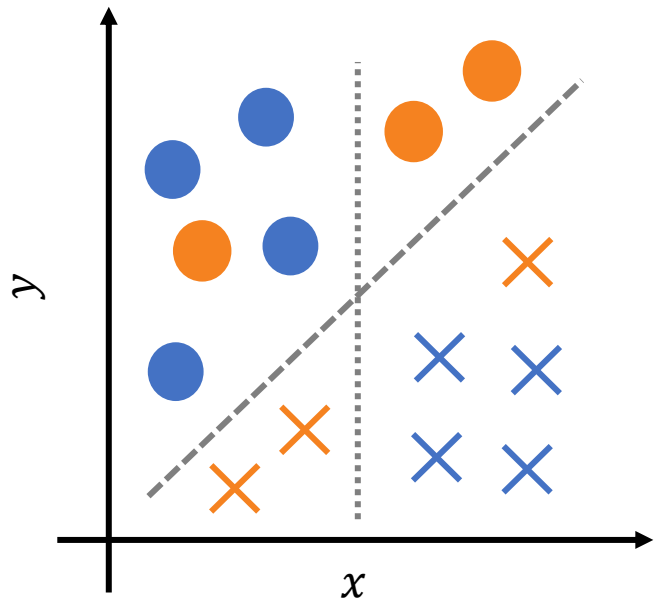
Solution? *Use (or develop) further, less disturbed measurement modalities?*

Alexander (1958), The Elimination of Confusing Breast Shadows in Chest Radiography. Australasian Radiology.

Brahee et al. (2013), Body Mass Index and Abdominal Ultrasound Image Quality. Journal of Diagnostic medical Sonography.

Ross et al. (2020), The influence of patient race on the use of diagnostic imaging in U.S. emergency departments. BMC Health Services Research.

Task difficulty: Unobserved causes of the outcome



Def.: An unobserved factor that affects outcomes more strongly in one group compared to others.

Ex.: Hormone level fluctuations affect outcomes more strongly in women than men.

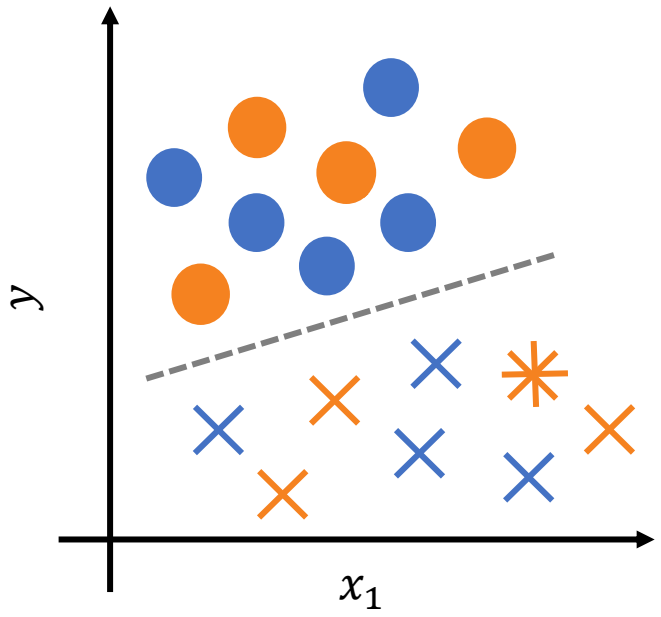
Comorbidities more prevalent in older subjects, influencing outcomes more than in younger patients.

Female heart attacks often due to microvascular disease, undetectable by current standard tests.

Effect: Reduced prediction performance in the affected groups.

Solution? *Measure and include them in the model.*

Task difficulty: Label noise



Def.: *Random* label errors; differing error rate between groups.

Ex.: Label noise in chest x-ray disease labels mined from reports using NLP: more label errors in older subjects.

Effect: Still learning the correct input—output mapping, but at reduced learning rate (since more noisy).

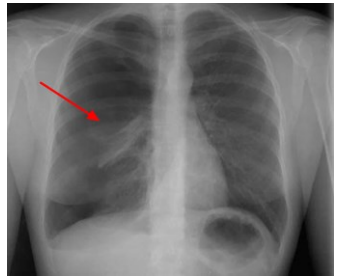
DL generally assumed to be robust to label noise?

Performance measures reduced; performance w.r.t. true labels might be better?

Solution? Investigate label error rates in different groups, clean data, label noise-robust learning schemes.

Could be shortcut learning…?

Hard to detect



Input image

Easy to detect

Correlates with

Prediction target

Shortcut features

CXR view

Annotations

Recording device

Medical Implants

AI for radiographic COVID-19 detection selects shortcuts over signal

Alex J. DeGrave, Joseph D. Janizek & Su-In Lee

Perspective | Published: 10 November 2020

Shortcut learning in deep neural networks

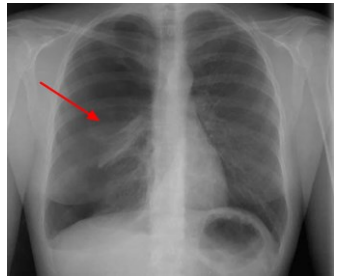
Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge & Felix A. Wichmann

Nature Machine Intelligence **2**, 665–673 (2020) | Cite this article

Could be shortcut learning…?

Input image

Hard to detect



Prediction target

Easy to detect

Correlates with

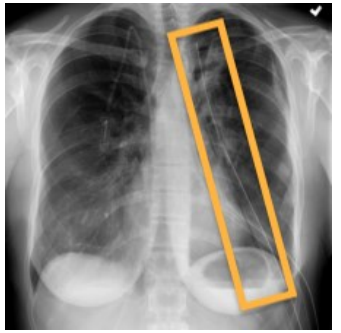
Diagnosed pneumothoraxes treated with chest drains…

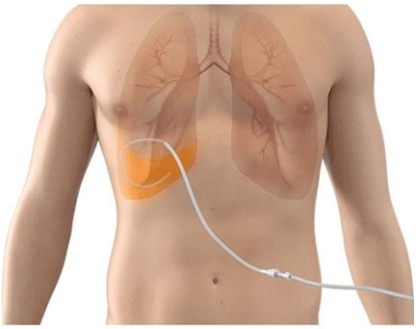
Hidden stratification causes clinically meaningful failures in machine learning for medical imaging

Authors: Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, Christopher Re Authors Info & Claims

DETECTING SHORTCUTS IN MEDICAL IMAGES - A CASE STUDY IN CHEST X-RAYS

Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, Veronika Cheplygina





Presence of chest drain

Could be shortcut learning…? Maybe…?

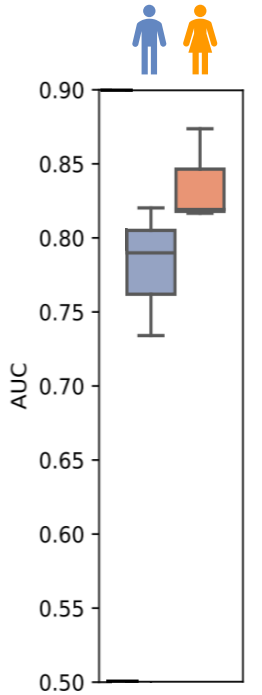
Diagnosed pneumothoraxes
treated with chest drains…

Hidden stratification causes clinically meaningful failures in machine learning for medical imaging

Authors: Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, Christopher Re Authors Info & Claims

DETECTING SHORTCUTS IN MEDICAL IMAGES - A CASE STUDY IN CHEST X-RAYS

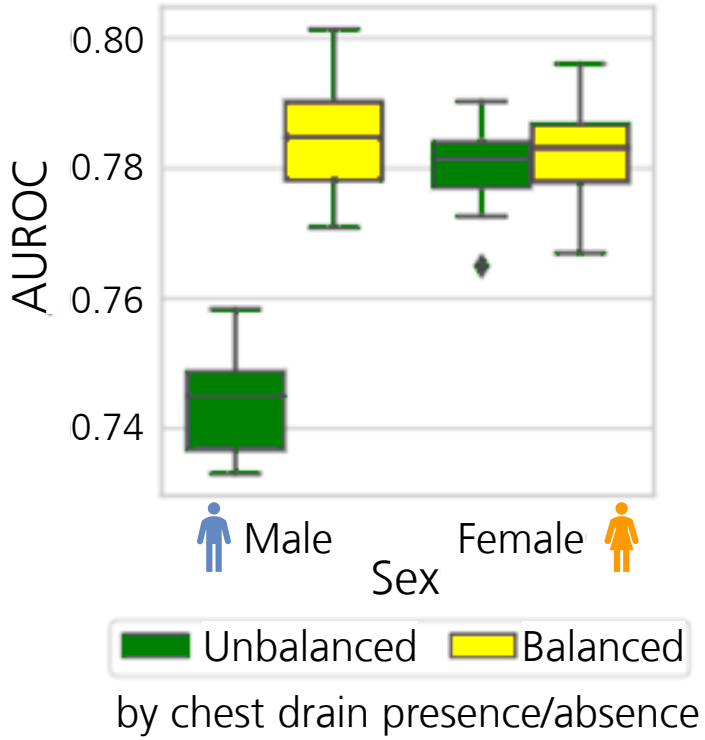
Amelia Jiménez-Sánchez, Dovile Juodelyte, Bethany Chamberlain, Veronika Cheplygina



With chest drain

Without chest drain

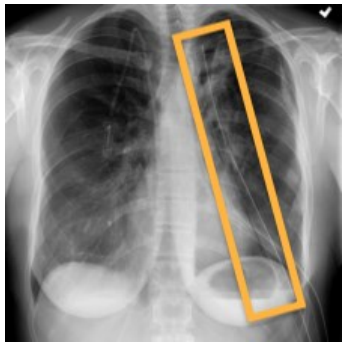
Shortcut learning it is!



by chest drain presence/absence

Balancing (test set) chest drain presence between M/F equalizes performance!

Technically *still* no proof of shortcut learning: chest drains could just be correlated with the thing the model is *actually* looking at. Real proof requires *counterfactuals*.



No-chest-drain counterfactual ↓ Leads to reduced model confidence

Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods

Vincent Olesen[1], Nina Weng[1], Aasa Feragen[1], and Eike Petersen[1,2(✉)]

Fast Diffusion-Based Counterfactuals for Shortcut Removal and Generation

Nina Weng[1*], Paraskevas Pegios[1,2*], Eike Petersen[1], Aasa Feragen[1,2], and Siavash Bigdeli[1]

Causes of QoS bias:

- Underrepresentation
- Differences in task difficulty
- **Poor performance for other reasons that happen to correlate with group membership** (e.g., shortcut learning)

*The observed disparitiy had **nothing** to do with sex / gender!*

***Bias root cause analysis** is crucial to enable effective bias **mitigation**!*

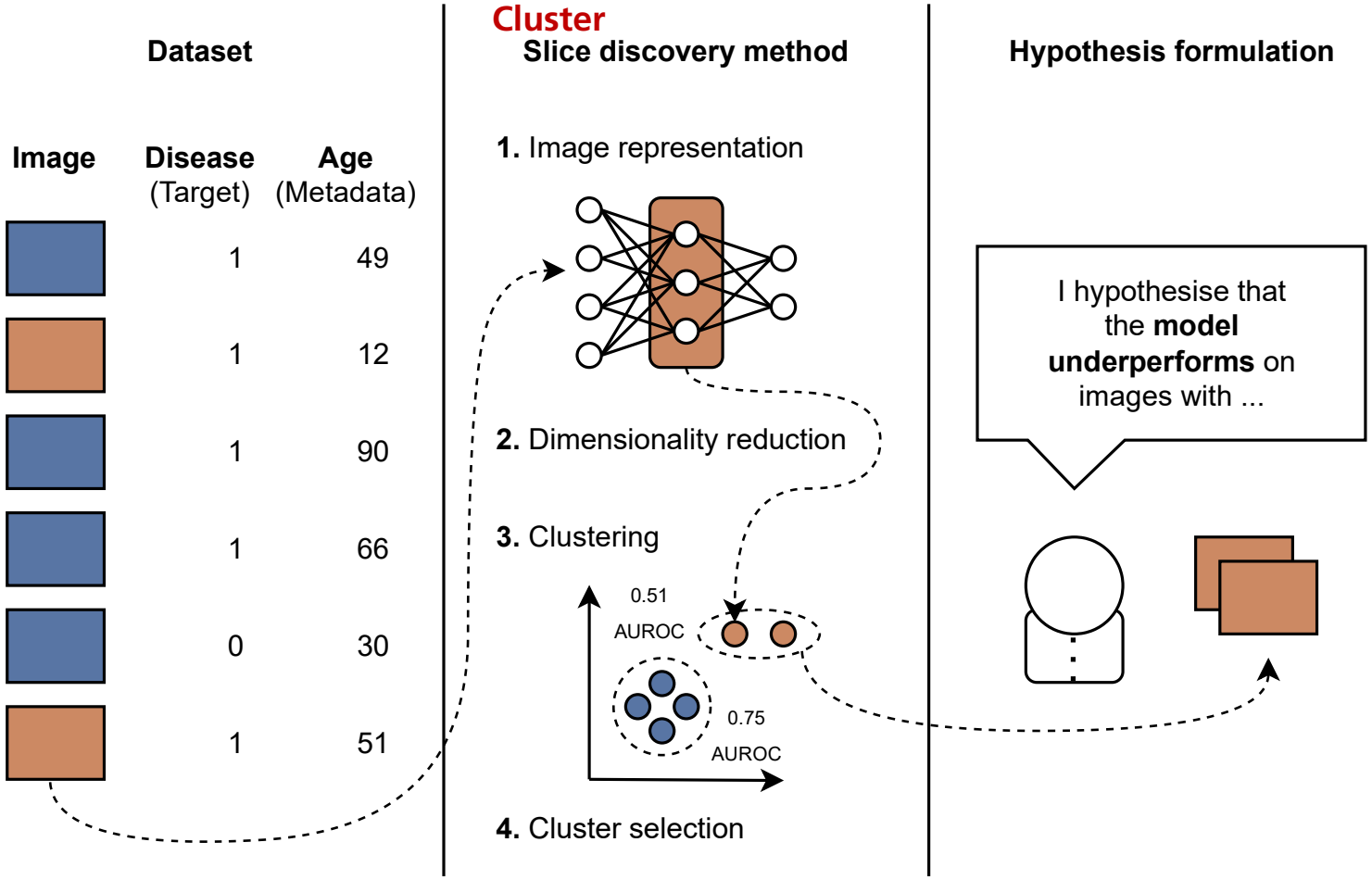
What if we don‘t know the potential shortcut?

Meet Slice Discovery Methods (SDMs).

Cluster

Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods

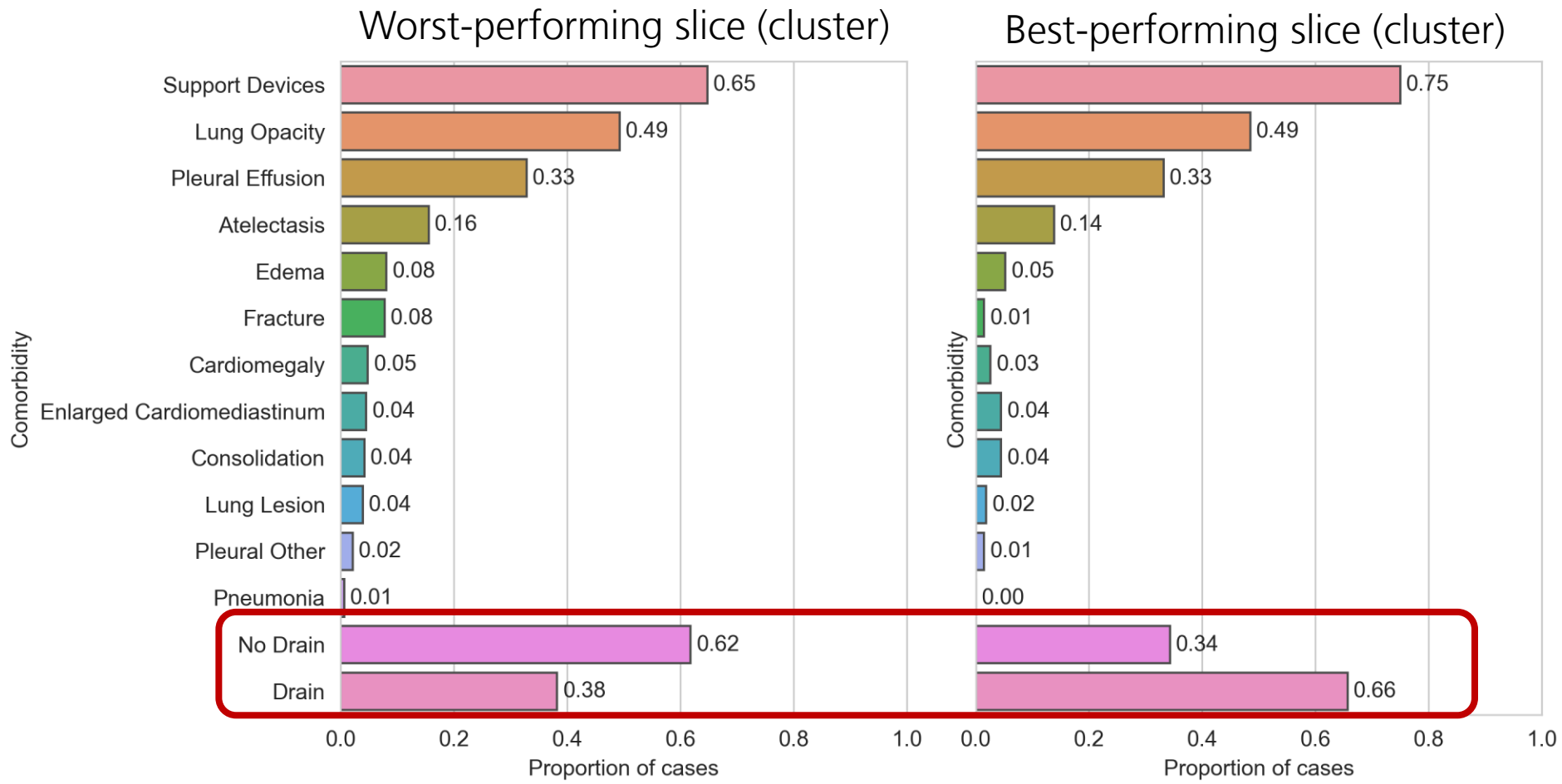
Vincent Olesen[1], Nina Weng[1], Aasa Feragen[1], and Eike Petersen[1,2(✉)]



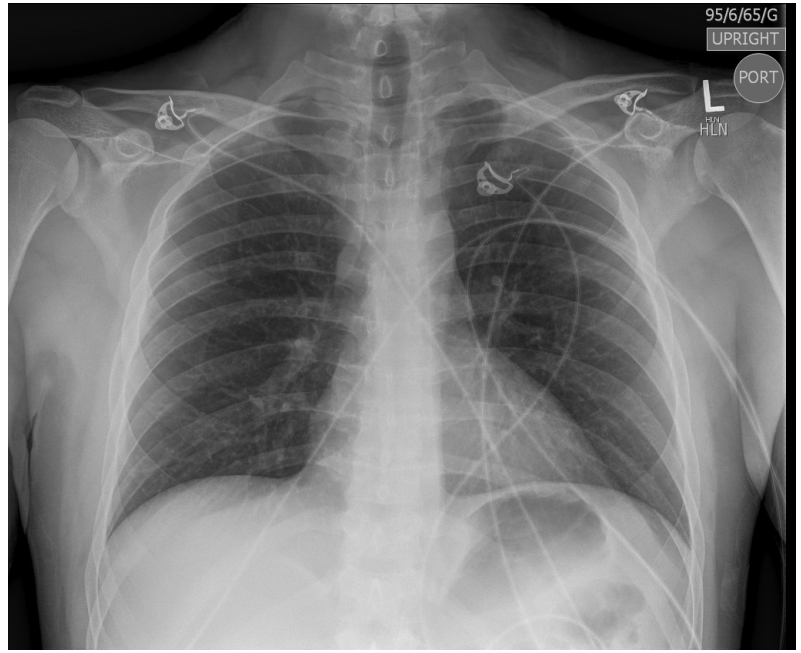
Pneumothorax-positive slices

Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods

Vincent Olesen[1], Nina Weng[1], Aasa Feragen[1], and Eike Petersen[1,2(✉)]



How to deal with really unknown factors?



Visual analysis of best/worst slices for **atelectasis** classification shows

- 90% samples with ECG cables in best negative slice
- 5% samples with ECG cables in worst negative slice

Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods

Vincent Olesen[1], Nina Weng[1], Aasa Feragen[1], and Eike Petersen[1,2(✉)]

Will this find everything? No!

Label & sampling biases: impossible to detect & fix purely algorithmically

Racial and ethnic disparities in the delayed diagnosis of appendicitis among children

Monika K. Goyal MD, MSCE, James M. Chamberlain MD, Michael Webb MS, Robert W. Grundmeier MD, Tiffani J. Johnson MD, MS, Scott A. Lorch MD, MSCE, Joseph J. Zorc MD, MSCE ... See all authors

Gender Bias in the Diagnosis of COPD

Chapman Kenneth R. MD, FCCP [a], Tashkin Donald P. MD, FCCP [b], Pye David J. PhD [c]

Dissecting racia… the health of pc…

ZIAD OBERMEYER, BRIAN POWERS, CHRIS…

SCIENCE · 25 Oct 2019 · Vol 366, Issue 646…

25,569 578

Racial bias in health algorithms

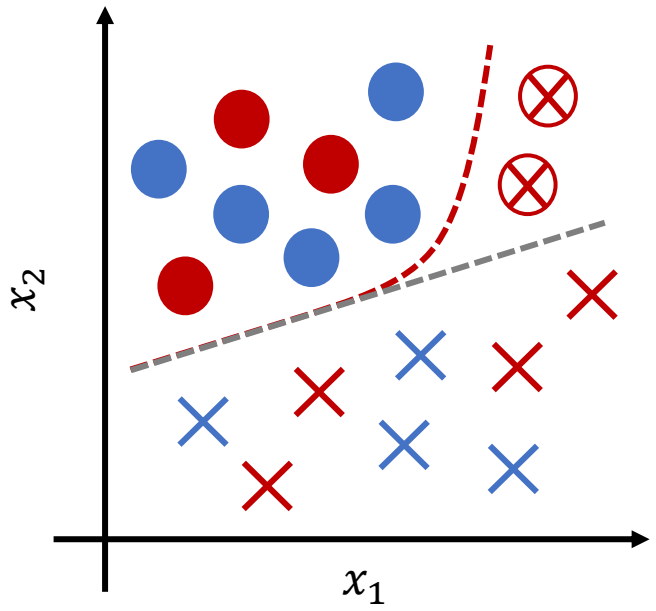
The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.

Science, this issue p. 447; see also p. 421

Causes of QoS bias:

- *Underrepresentation*
- *Differences in task difficulty*
- *Poor performance for other reasons that happen to correlate with group membership (e.g., shortcut learning)*
- ***Label / sampling biases** [extra hard since unclear how to detect / measure!]*

Label bias



Def.: *Systematic* label errors.

Ex.:
- Healthcare costs as a biased proxy for healthcare needs.
- Diagnostic biases and gender stereotypes in mental health.
- Racial bias in pain assessment.
- Widespread underdiagnosis of female heart disease.
- Systematic biases in NLP-extracted radiological findings.
- ICD codes as biased proxies of disease state.
- Annotator biases.

Effect: Biased decision threshold is learned!
Performance metrics do not reflect this bias.

Solution? Hard (impossible?) to detect without domain knowledge; hard to "fix" without resorting to other (better, unbiased) measurements (labels).

Selection bias

Def.: Systematic differences in **how groups are selected** for study enrollment

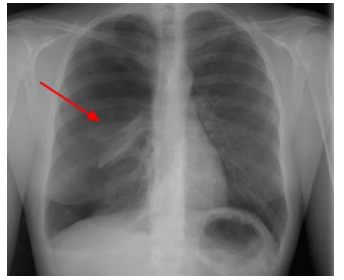
Ex.:
- Selecting subjects based on disease status.
- Differences in enrollment in the healthcare system.
- Being treated at specific hospitals (with different equipment).
- Groups being diagnosed / treated at different disease stages.

Effect: Biased decision threshold is learned!
Performance metrics do not reflect this bias.

Solution? If known: covariate shift adaptation methods, shift-stable learning.

Demographic shortcuts

Hard to detect



Input image

Easy to detect

Correlates with

Prediction target

Radiology "forensics": determination of age and sex from chest radiographs using deep learning

Paul H. Yi[1,2,3] · Jinchi Wei[3] · Tae Kyung Kim[2] · Jiwon Shin[3] · Haris I. Sair[2,3] · Ferdinand K. Hui[2,3] · Gregory D. Hager[3] · Cheng Ting Lin[2]

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang

Shortcut features

CXR view

Annotations

Recording device

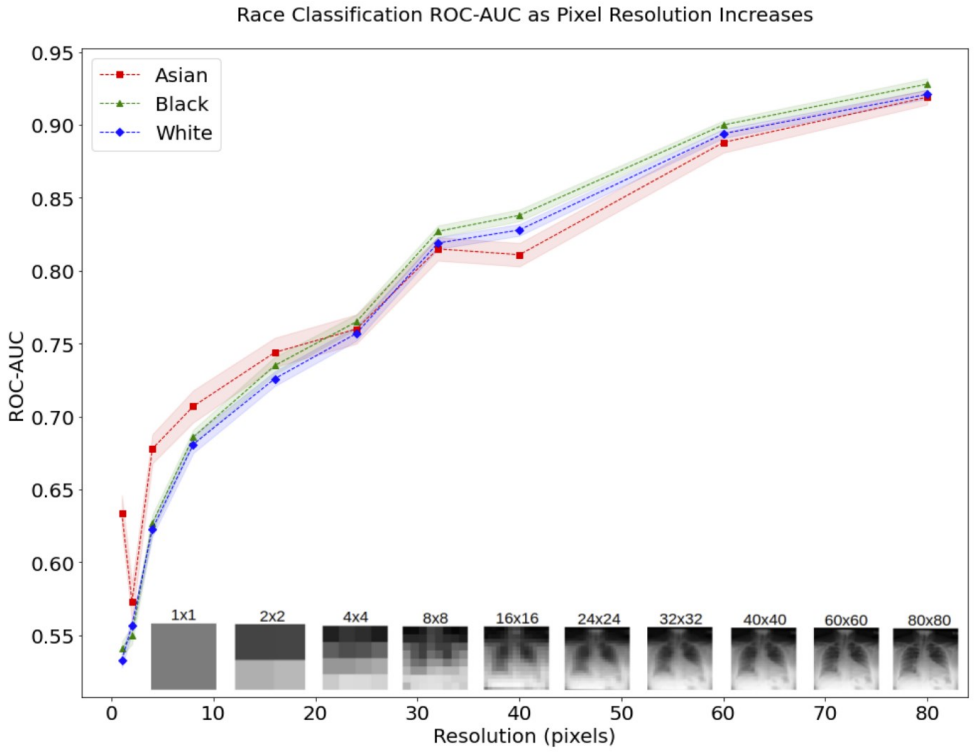
Medical Implants

Demographic properties

Demographic shortcuts

Models can identify self-reported race/ethnicity with high AUROC

- Far better than using only clinical metadata
- Far better than clinicians
- After removing (clipping) bone density information
- After severely high/low-pass filtering or downscaling
- From almost all individual parts of an image
- From only the grayscale histogram
- On external datasets



AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, Haoran Zhang

Ability of artificial intelligence to identify self-reported race in chest x-ray using pixel intensity counts

John Lee Burns[a,*] Zachary Zaiman,[b] Jack Vanschaik,[a] Gaoxiang Luo,[c] Le Peng,[c] Brandon Price,[d] Garric Mathias,[a] Vijay Mittal,[a] Akshay Sagane,[a] Christopher Tignanelli,[c] Sunandan Chakraborty[a] Judy Wawira Gichoya[b] and Saptarshi Purkayastha[a]

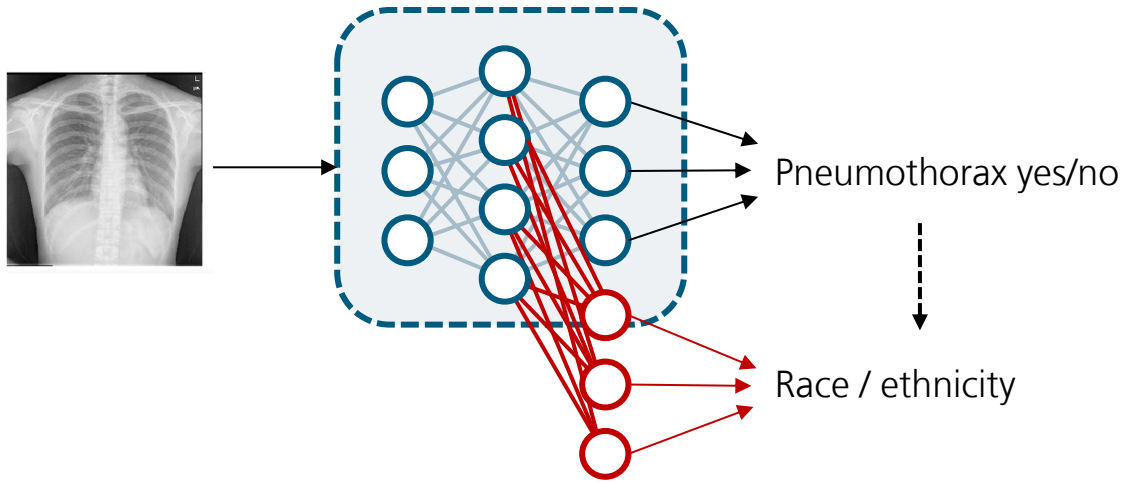
Demographic shortcuts

Models *can* identify self-reported race/ethnicity with high AUROC… *but do they actually do that?*

Many (incl. me ☺) investigate „demographic encoding“:

! Caution !

1. If race/ethnicity can be „predicted“ from y (target label) it can also be predicted with at least the same accuracy from the embeddings of any perfect disease classifier!
   → Baseline, higher for higher-cardinality y (multi-label or image targets – seg, synth) or in case of strong label shifts (prevalence diffs)
2. „Encoding“ stronger than this does *not* necessarily imply that the model *uses* this info! (Models can ignore parts of embeddings.)



Demographic shortcuts

Causes of QoS bias:

- Underrepresentation
- Differences in task difficulty
- Poor performance for other reasons that happen to correlate with group membership (e.g., shortcut learning)
- Label / sampling biases [extra hard since unclear how to detect / measure!]
- **Demographic shortcuts**

Models *can* identify self-reported race/ethnicity. Do they?

True proof: „demographic counterfactual“ (??)

Demographic shortcuts are *diffuse* = especially hard to investigate / visualize / „explain“

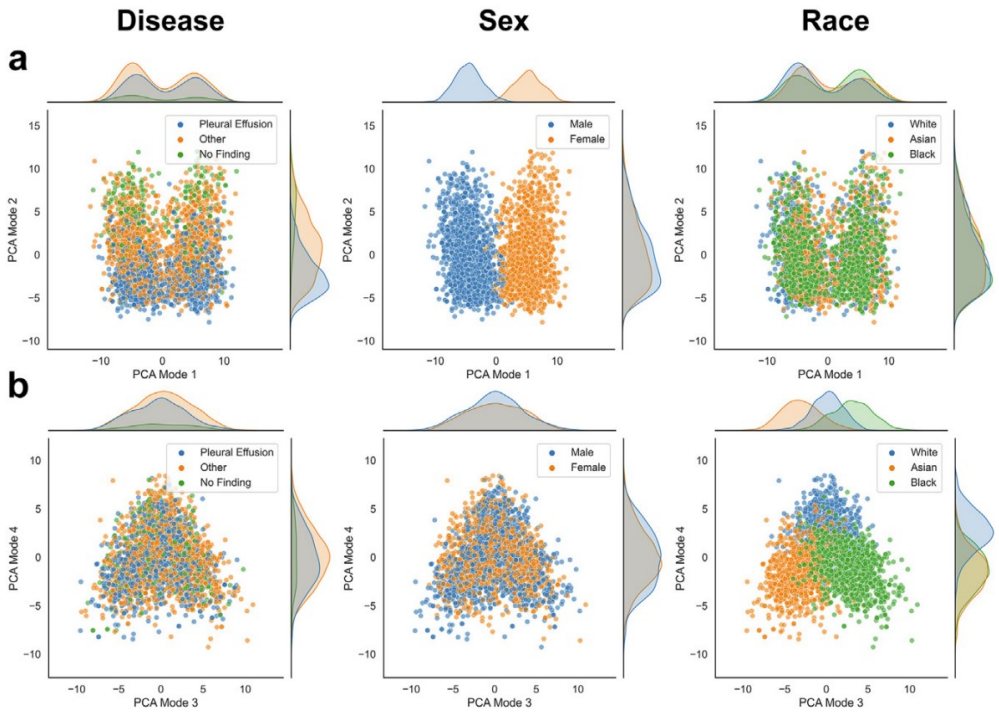
The inverse direction works: if „encoding“ low, demographics clearly not used!

Demographic shortcuts

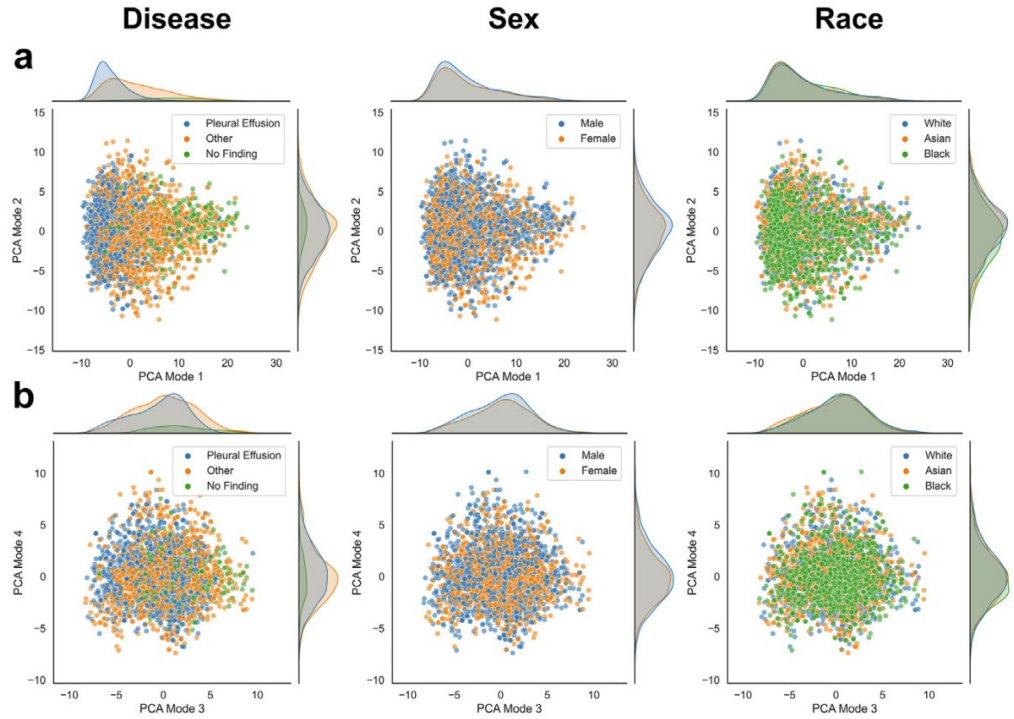
Algorithmic encoding of protected characteristics in chest X-ray disease detection models

Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck*

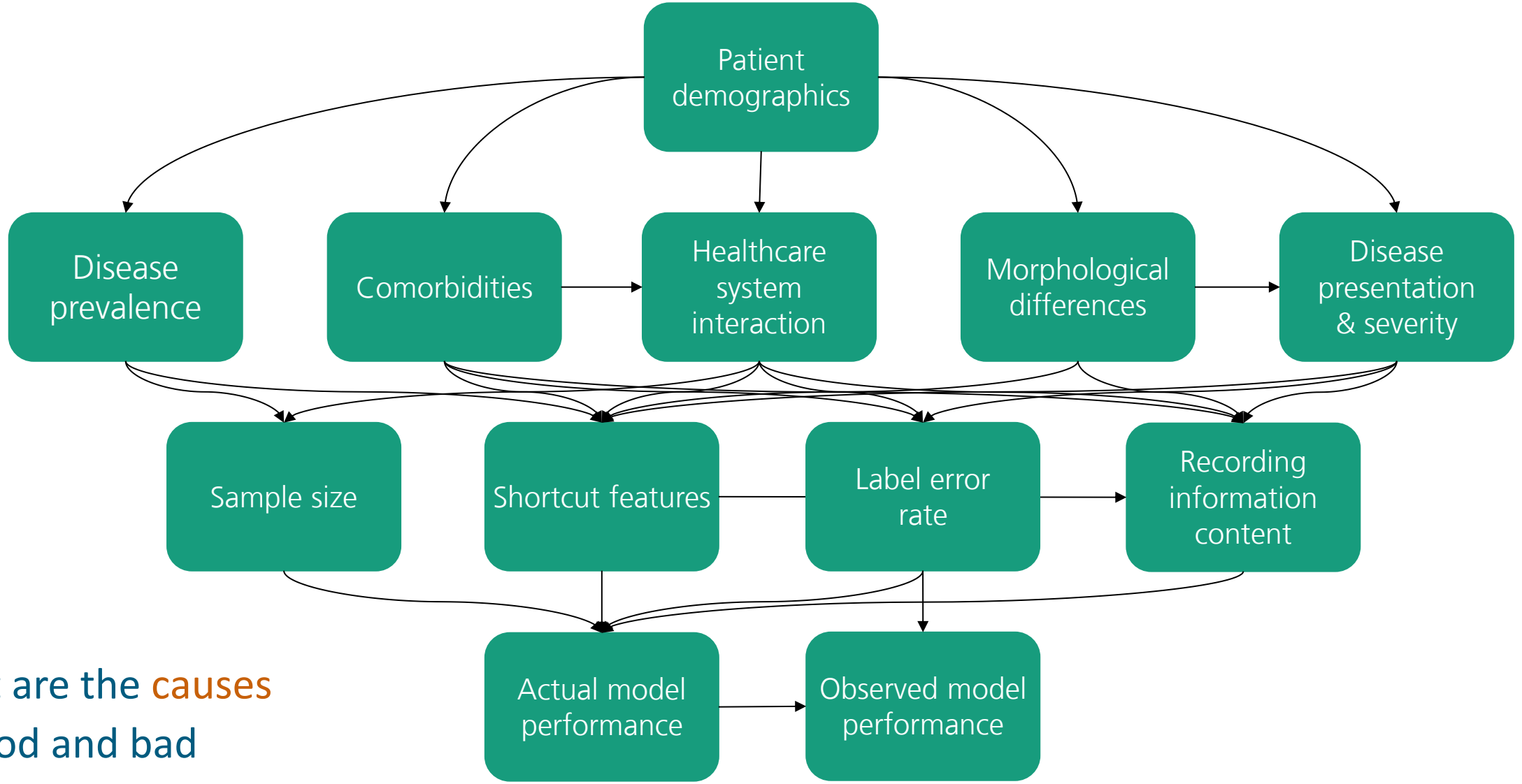
PCA components of embeddings colored by disease / sex / race



Multi-head model trained to predict all three



Single-head model trained to predict only disease



What are the causes of good and bad model performance?

i. Introduction

ii. Causes of bias

iii. Case study 1

iv. Coffee break (10:00 - 10:30)

v. Case study 2

vi. Bias mitigation

vii. Conclusion & recommendations

i. Introduction

ii. Causes of bias

iii. Case study 1

iv. Coffee break (10:00 - 10:30)

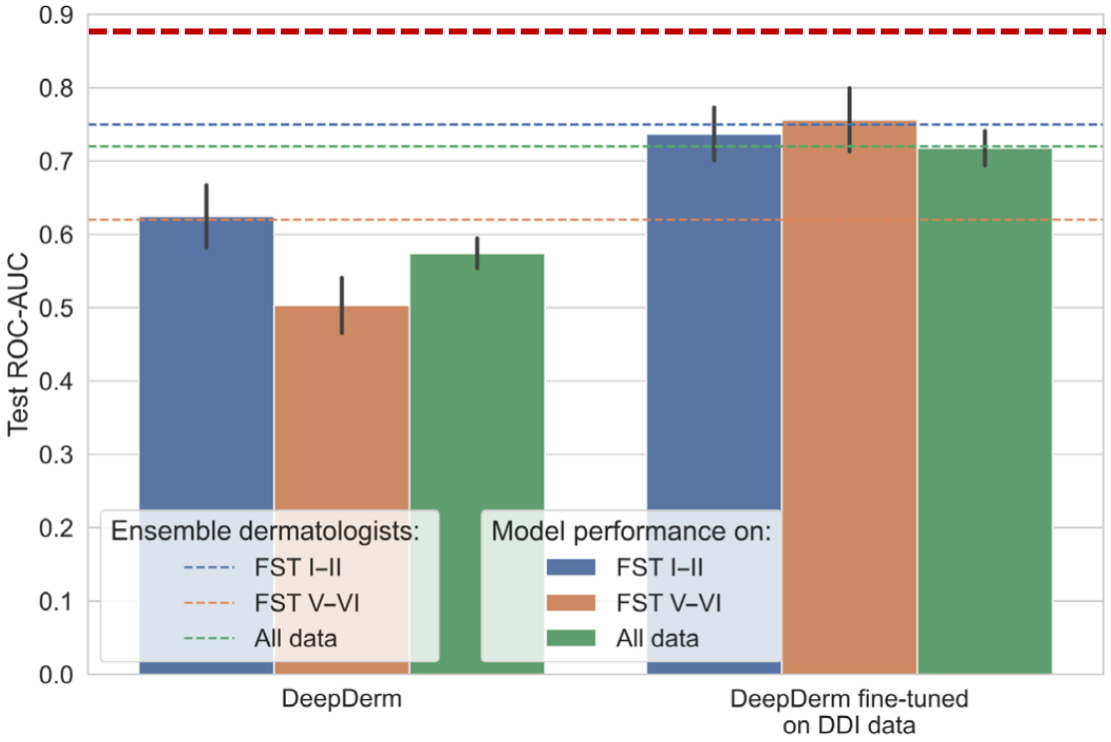
v. Case study 2

vi. Bias mitigation

vii. Conclusion & recommendations

i. Introduction
ii. Causes of bias
iii. Case study 1
iv. Coffee break (10:00 - 10:30)
v. Case study 2
vi. Bias mitigation
vii. Conclusion & recommendations

Gather more (diverse), high(er)-quality data …



← Previously reported **IID** performance: ROC-AUC 0.88!

SCIENCE ADVANCES | RESEARCH ARTICLE

HEALTH AND MEDICINE

Disparities in dermatology AI performance on a diverse, curated clinical image set

Roxana Daneshjou[1,2]†, Kailas Vodrahalli[3]†, Roberto A. Novoa[1,4], Melissa Jenkins[1], Weixin Liang[5], Veronica Rotemberg[6], Justin Ko[1], Susan M. Swetter[1], Elizabeth E. Bailey[1], Olivier Gevaert[2], Pritam Mukherjee[2]‡, Michelle Phung[1], Kiana Yekrang[1], Bradley Fong[1], Rachna Sahasrabudhe[1]§, Johan A. C. Allerup[1], Utako Okata-Karigane[7], James Zou[2,3,5,8]*, Albert S. Chiou[1]*

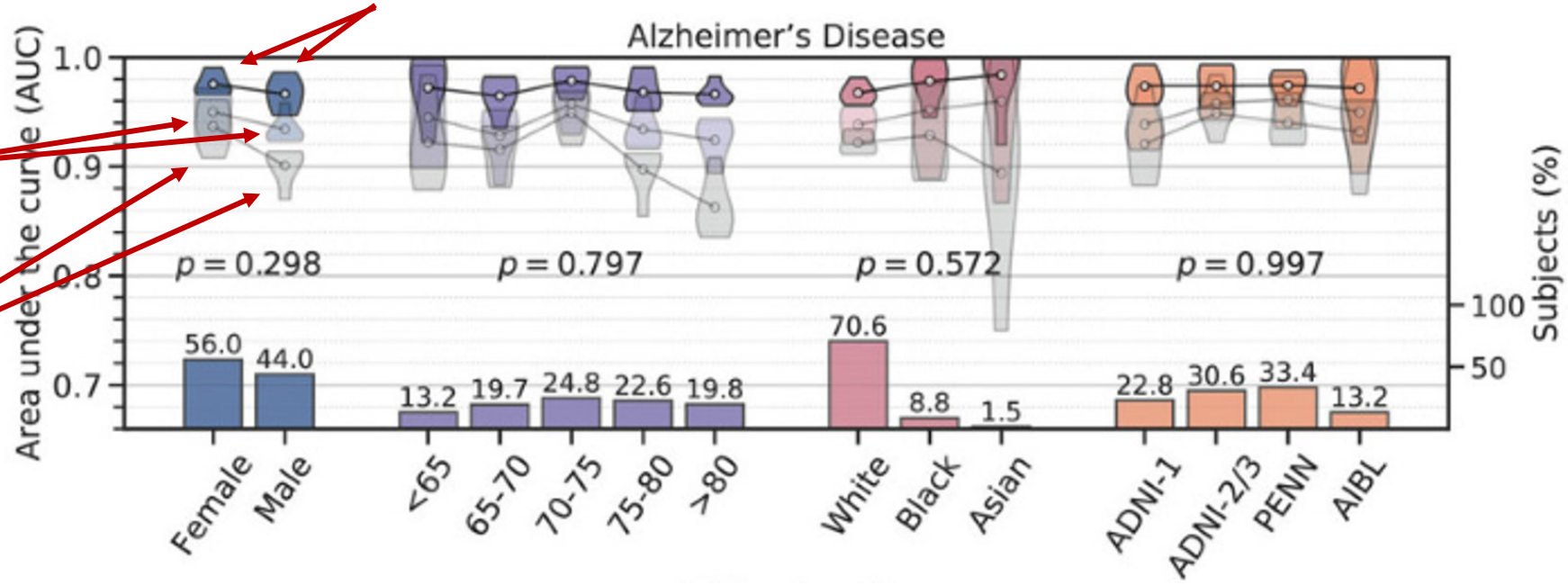
OOD: Diverse Dermatology Images (DDI) dataset: diverse skin tones, diverse lesion types, pathologically confirmed labels!

Train good models …

Ensemble with volumetric features, hyperparameter optimization, demographic / clinical / genetic / cognitive scoring data

Ensemble with volumetric features, hyperparameter optimization

Baseline CNN



PNAS BRIEF REPORT | APPLIED MATHEMATICS OPEN ACCESS

Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies

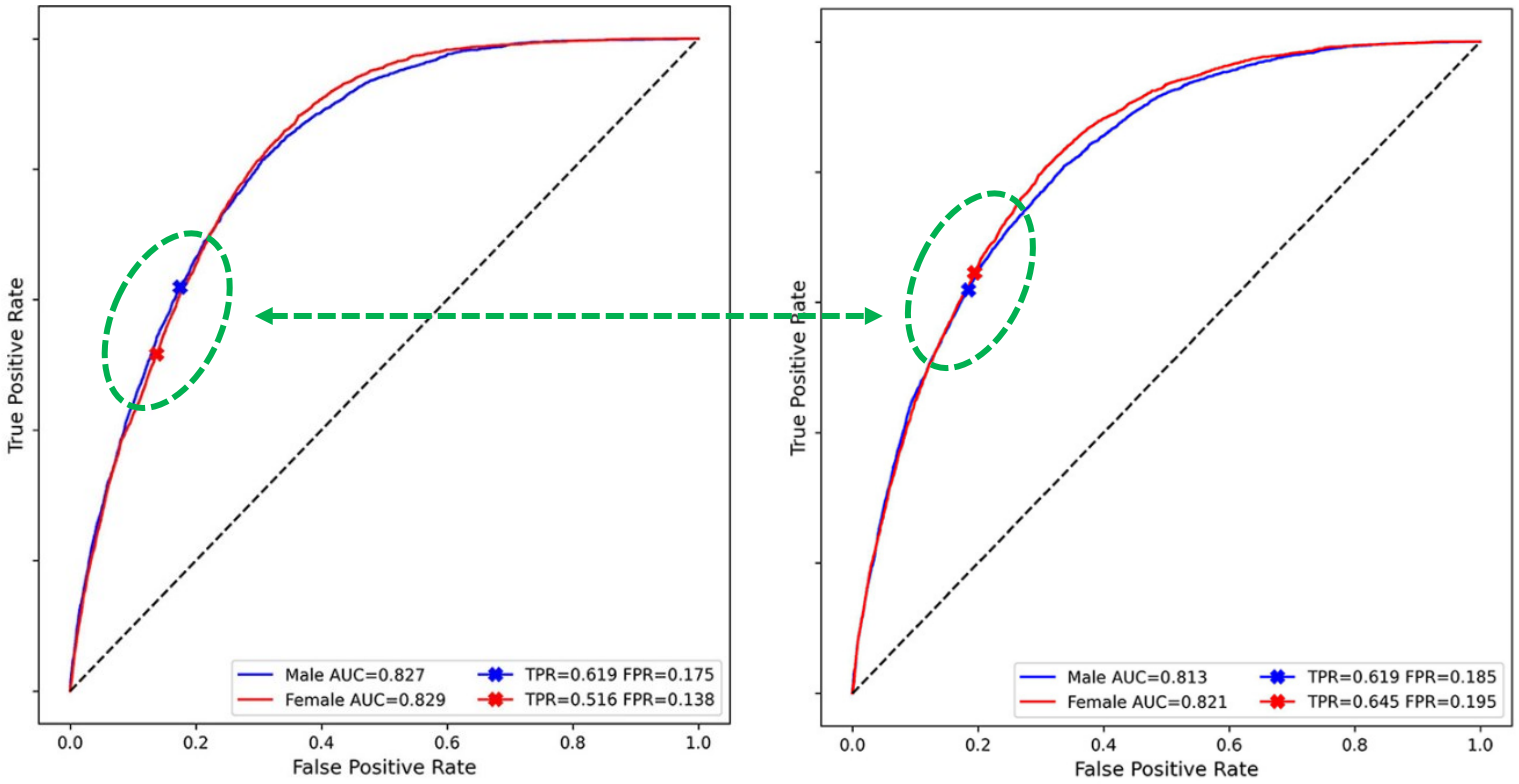
Rongguang Wang [a,b], Pratik Chaudhari [a,c,1,2], and Christos Davatzikos [a,b,d,1,2,3]

Train good models …

Drop the shortcuts: image augmentation improves fairness and decreases AI detection of race and other demographics from medical images

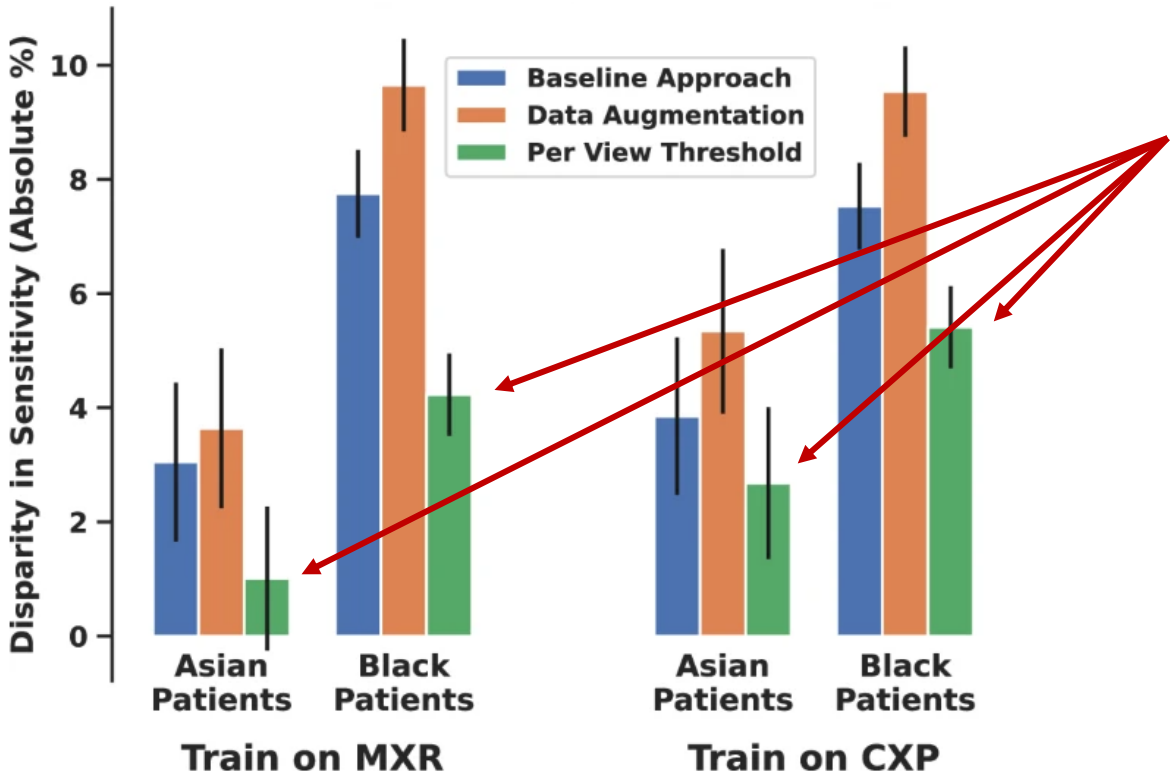
Ryan Wang,[a] Po-Chih Kuo,[a,*] Li-Ching Chen,[a] Kenneth Patrick Seastedt,[b,c] Judy Wawira Gichoya,[d] and Leo Anthony Celi[e,f,g]

Baseline model



The same model trained with trivial and non-bias-specific data augmentations: rotations, shear, scaling, fisheye

Train good models … (multi-)calibrate



Recalibrate baseline model separately for different CXR views: AP vs. PA vs. Lateral vs. Portable AP


Article https://doi.org/10.1038/s41467-024-52003-3

Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias

Received: 7 February 2023 William Lotter [1,2,3]

Fraunhofer MEVIS

Increase diversity using synthetic data

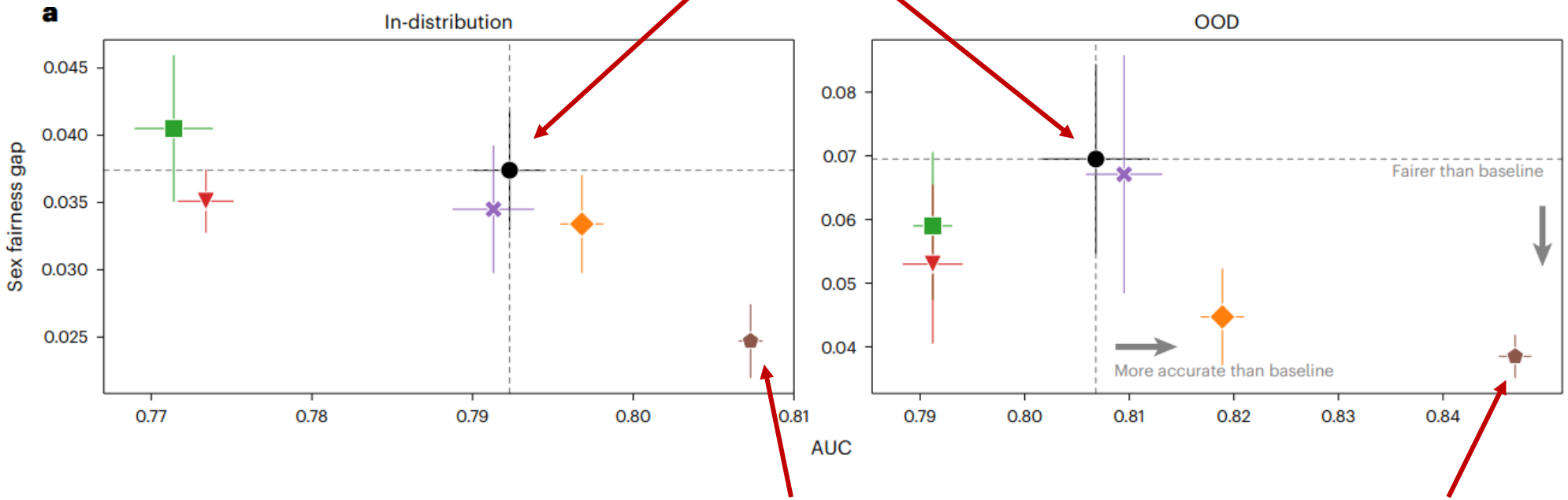
Article
https://doi.org/10.1038/s41591-024-02838-6

Generative models improve fairness of medical classifiers under distribution shifts

Received: 24 May 2023
Accepted: 26 January 2024
Published online: 10 April 2024

Ira Ktena[1,4] ✉, Olivia Wiles[1,4] ✉, Isabela Albuquerque[1], Sylvestre-Alvise Rebuffi[1], Ryutaro Tanno[1], Abhijit Guha Roy[2], Shekoofeh Azizi[1], Danielle Belgrave[3], Pushmeet Kohli[1], Taylan Cemgil[1], Alan Karthikesalingam[2,5] & Sven Gowal[1,5]

Baseline model

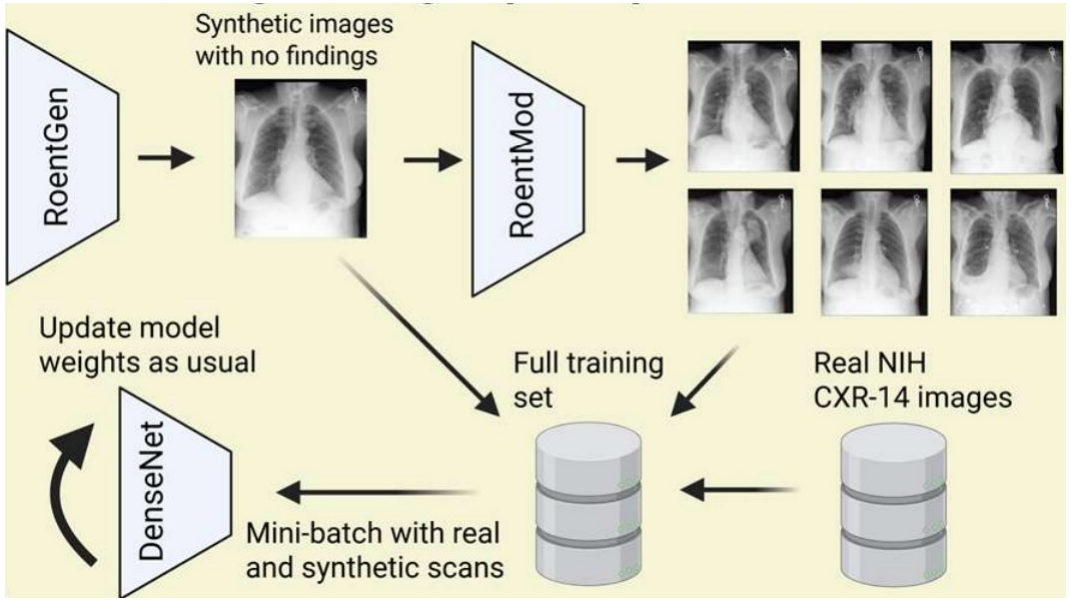


Baseline model trained on same data + synthetic data generated using (conditional) diffusion model trained on that same dataset

Mitigate shortcuts …

RoentMod: A Synthetic Chest X-Ray Modification Model to Identify and Correct Image Interpretation Model Shortcuts

Lauren H. Cooke, Matthias Jung, Jan M. Brendel, Nora M. Kerkovits, Borek Foldyna, Michael T. Lu, and Vineet K. Raghu



Outperforms baseline (NIH only) by 0.06 AUROC points (i.i.d.) and 0.04 AUROC points (o.o.d.) on average over 6 disease labels. (Always better.)

Fairness effects: not yet tested. Presumably positive?

Select for fairness / robustness …

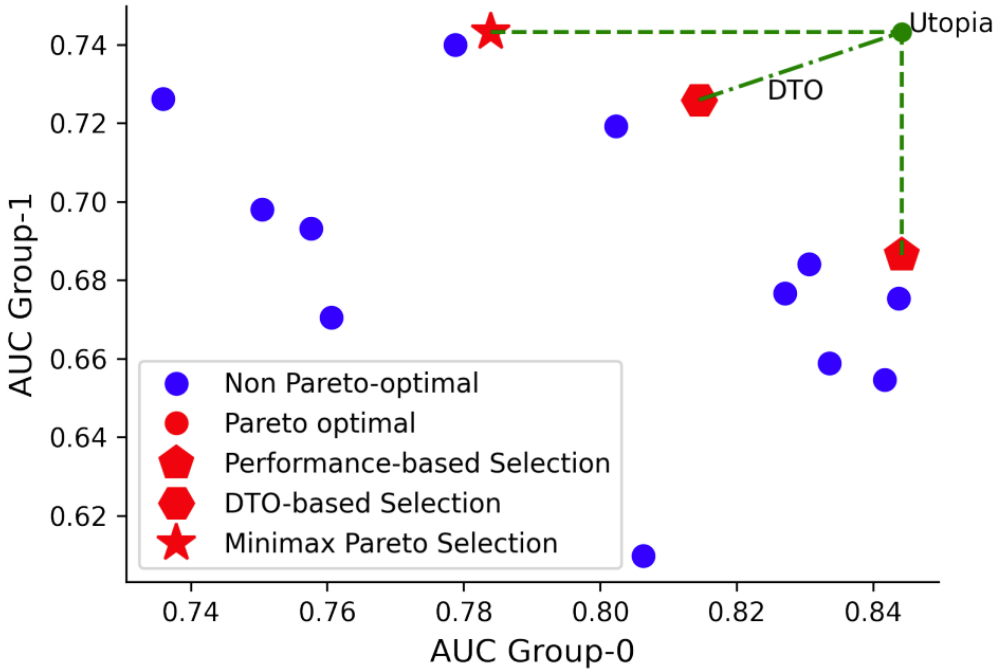
„… [in] ten datasets from different imaging modalities … we find that **the under-studied issue of model selection criterion can have a significant impact on fairness outcomes**"

Max-min AUROC selection stat. sig. better than default on worst-group AUROC while *not* stat. sig. worse on overall AUROC!

This is (like everything until here) without any explicit „bias mitigation"!

MEDFAIR: BENCHMARKING FAIRNESS FOR MEDICAL IMAGING

Yongshuo Zong[1], **Yongxin Yang**[1], **Timothy Hospedales**[1,2]
[1] School of Informatics, University of Edinburgh, [2] Samsung AI Centre, Cambridge
{yongshuo.zong, yongxin.yang, t.hospedales}@ed.ac.uk



Select for fairness / robustness …

"The standard … is to choose the model that maximizes accuracy. Using maximum accuracy as a decision criterion for model selection may suggest that there is one model with the best accuracy … However, … **there are usually multiple models with equivalent accuracy but significantly different properties**."

Model Multiplicity: Opportunities, Concerns, and Solutions

Emily Black
emilybla@andrew.cmu.edu
Carnegie Mellon University
USA

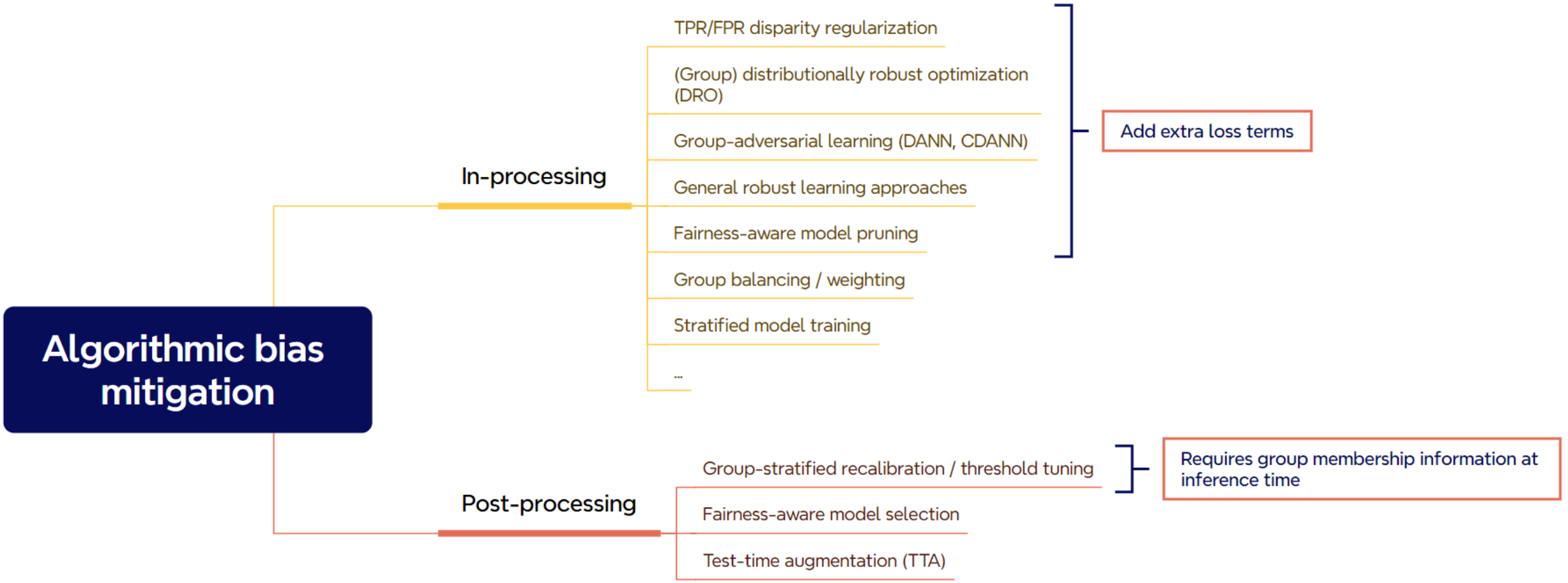
Manish Raghavan
mraghavan@seas.harvard.edu
Harvard University
USA

Solon Barocas
solon@microsoft.com
Microsoft Research
USA

Amazing Things Come From Having Many Good Models

Cynthia Rudin [1 *] Chudi Zhong [1] Lesia Semenova [1] Margo Seltzer [2] Ronald Parr [1] Jiachang Liu [1]
Srikar Katta [1] Jon Donnelly [1] Harry Chen [1] Zachery Boner [1]

"The **Rashomon Effect** … describes the phenomenon that there exist many equally good predictive models for the same dataset. This phenomenon happens for many real datasets and when it does, it sparks both magic and consternation, but mostly magic. In light of the Rashomon Effect, this perspective piece proposes reshaping the way we think about machine learning, particularly … flexibility to address user preferences, such as fairness …"


Algorithmic bias mitigation
In-processing
TPR/FPR disparity regularization
(Group) distributionally robust optimization (DRO)
Group-adversarial learning (DANN, CDANN)
General robust learning approaches
Fairness-aware model pruning
Group balancing / weighting
Stratified model training
...
Add extra loss terms
Post-processing
Group-stratified recalibration / threshold tuning
Fairness-aware model selection
Test-time augmentation (TTA)
Requires group membership information at inference time

Limited effectiveness of (some?) „bias mitigation“ methods

MEDFAIR: BENCHMARKING FAIRNESS FOR MEDICAL IMAGING

Yongshuo Zong[1], **Yongxin Yang**[1], **Timothy Hospedales**[1,2]
[1] School of Informatics, University of Edinburgh, [2] Samsung AI Centre, Cambridge
{yongshuo.zong, yongxin.yang, t.hospedales}@ed.ac.uk

“No method outperforms ERM with statistical significance”

Improving the Fairness of Chest X-ray Classifiers

Haoran Zhang HAORANZ@MIT.EDU
Massachusetts Institute of Technology
Natalie Dullerud NATALIE.DULLERUD@MAIL.UTORONTO.EDU
University of Toronto
Karsten Roth KARSTEN.ROTH@UNI-TUEBINGEN.DE
University of Tübingen
Lauren Oakden-Rayner LAUREN.OAKDEN-RAYNER@ADELAIDE.EDU.AU
University of Adelaide
Stephen Pfohl SPFOHL@STANFORD.EDU
Stanford University
Marzyeh Ghassemi MGHASSEM@MIT.EDU
Massachusetts Institute of Technology

“We find, consistent with prior work on non-clinical data, that methods which strive to achieve better worst-group performance do not outperform simple data balancing. We also find that methods which achieve group fairness do so by worsening performance for all groups.”

Algorithmic bias mitigation: limitations

Highly narrow framing: keep data, preprocessing, task set-up, model architecture fixed.

Limited empirical success. (Not surprising, given the above?)

Large gains possible outside of this narrow framing, cf. earlier results in this section.

“Fairness-Accuracy Trade-offs”:

1. Some exist *(under the above, very narrow framing),* but there are much fewer practically relevant trade-offs than is often believed.
2. If they exist, trade-offs are often negligible in practice.
3. Observed trade-offs may be illusory if data are biased.
4. *Focus on improving task setup, data quality, model quality, selection strategy before considering supposed trade-offs. In almost all scenarios, performance can be “leveled up”.*

i. Introduction

ii. Causes of bias

iii. Case study 1

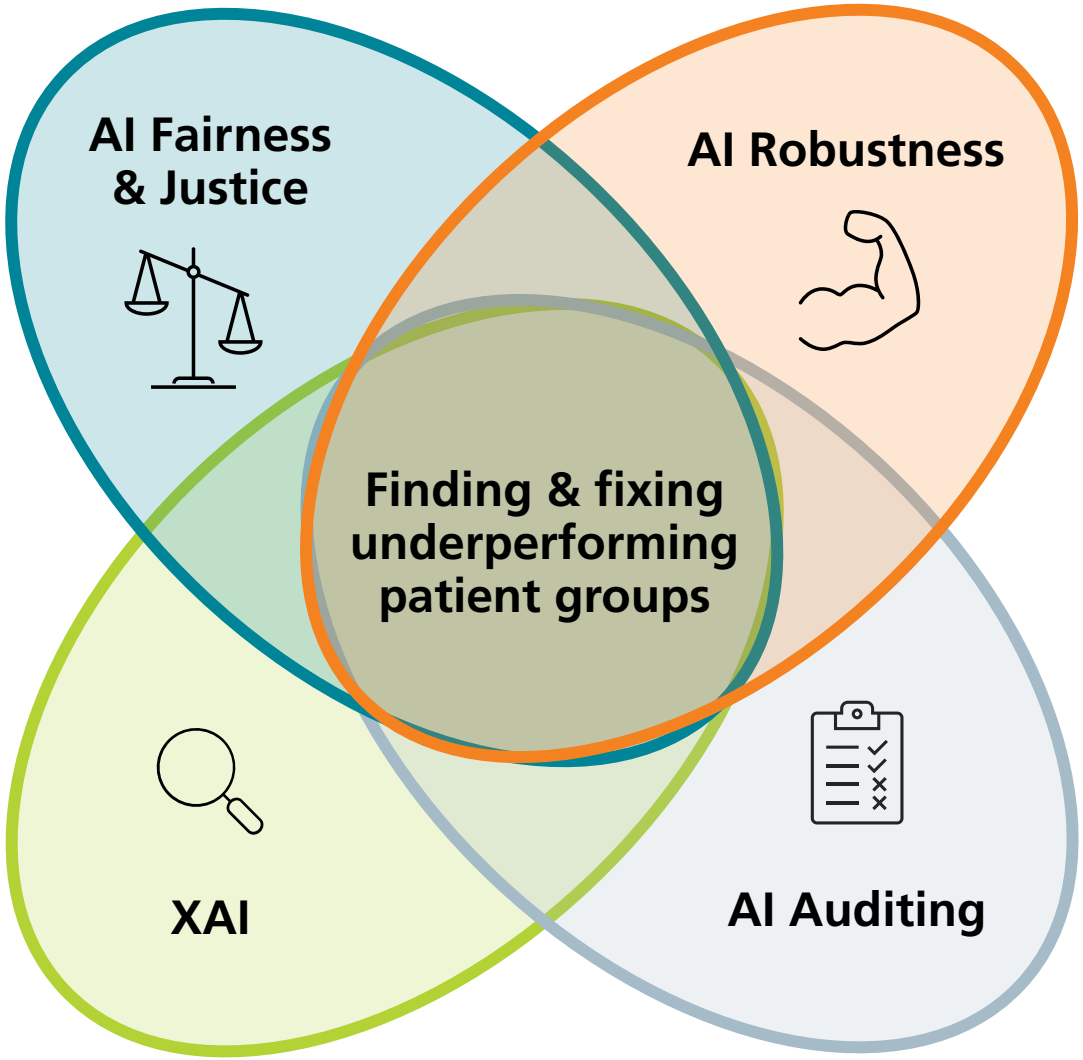
iv. Coffee break (10:00 - 10:30)

v. Case study 2

vi. Bias mitigation

vii. Conclusion & recommendations


Label noise analyses
Additional labeling efforts
Confounder adjustment
Finding hidden clusters
XAI methods
Data augmentation
Prior knowledge about potential shortcuts
The „bias root cause analysis“ toolbox


AI Fairness & Justice
AI Robustness
Finding & fixing underperforming patient groups
XAI
AI Auditing

Recommendations & Take-home messages

- Gather as much metadata as feasible.
- Perform fine-grained intersectional performance assessment and look for important unlabeled clusters.
- Focus on general data & model quality before using specific bias mitigation methods.
  - Standardize / normalize / harmonize / … (without normalizing away important biological differences!)
  - Augment extensively
  - Ensure robustness w.r.t. variations in technical parameters
  - Assess & ensure quality of *labels*
  - Consider fairness in model selection
- Performance (and bias) metrics can be misleading if test data are biased, as is often the case.
  External evaluation is key!
- *There are no silver bullets* in bias assessment and mitigation:
  Comprehensive root cause investigation is hard detective work.
- But: fixing QoS bias problems will just make your models better!

Final recommendation

www.faimi.org

- Newsletter!
- Free virtual online symposium in Nov
- These slides ☺
- Resources on FAIMI topics

References

Alderman et al. (2025), Tackling algorithmic bias and promoting transparency in health datasets: the STANDING Together consensus Recommendations, The Lancet Digital Health.

Black et al. (2022), Model Multiplicity: Opportunities, Concerns, and Solutions, ACM FAccT.

Daneshjou et al. (2022), Disparities in dermatology AI performance on a diverse,curated clinical image set, Science Advances.

Drukker et al. (2023), Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment, Journal of Medical Imaging.

Gichoya et al. (2022), AI recognition of patient race in medical imaging: a modelling study, The Lancet Digital Health.

Glocker et al. (2023), Algorithmic encoding of protected characteristics in chest X-ray disease detection models, eBioMedicine.

Jones et al. (2024), A causal perspective on dataset bias in machine learning for medical imaging, Nature Machine Intelligence.

Klingenberg et al. (2023), Higher performance for women than men in MRI-based Alzheimer's disease detection, Alzheimer's Research & Therapy.

Ktena et al. (2024), Generative models improve fairness of medical classifiers under distribution shifts, Nature Medicine.

Lotter (2024), Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias, Nature Communications.

Mehrabi et al. (2021), A Survey on Bias and Fairness in Machine Learning, ACM Computing Surveys.

Mitchell et al. (2021), Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application.

Olesen et al. (2024), Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods, MICCAI FAIMI Workshop.

References

Petersen et al. (2022), Feature Robustness and Sex Differences in Medical Imaging: A Case Study in MRI-Based Alzheimer's Disease Detection, MICCAI.

Petersen et al. (2023), The path toward equal performance in medical machine learning, Patterns.

Puyol-Antón et al. (2021), Fairness in Cardiac MR Image Analysis: An Investigation of Bias Due to Data Imbalance in Deep Learning Based Segmentation, MICCAI.

Ricci Lara et al. (2022), Addressing fairness in artificial intelligence for medical imaging, Nature Communications.

Rodolfa et al. (2021), Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy, Nature Machine Intelligence.

Seoni et al. (2024), All you need is data preparation: A systematic review of image harmonization techniques in Multi-center/device studies for medical support systems, Computer Methods and Programs in Biomedicine.

Seyyed-Kalantari et al. (2021), Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations, Nature Medicine.

Wang et al. (2023), Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies, PNAS.

Weng et al. (2023), Are Sex-Based Physiological Differences the Cause of Gender Bias for Chest X-Ray Diagnosis?, MICCAI FAIMI Workshop.

Weng et al. (2024), Fast Diffusion-Based Counterfactuals for Shortcut Removal and Generation, ECCV.

Wick et al. (2019), Unlocking Fairness: a Trade-off Revisited, NeurIPS.

Yang et al. (2024), The limits of fair medical imaging AI in real-world generalization, Nature Medicine.

Zong et al. (2023), MEDFAIR: Benchmarking Fairness for Medical Imaging, ICLR.