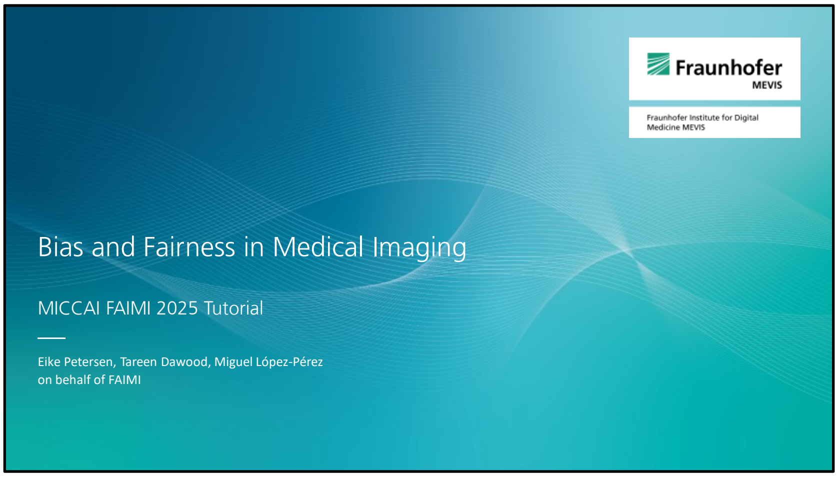


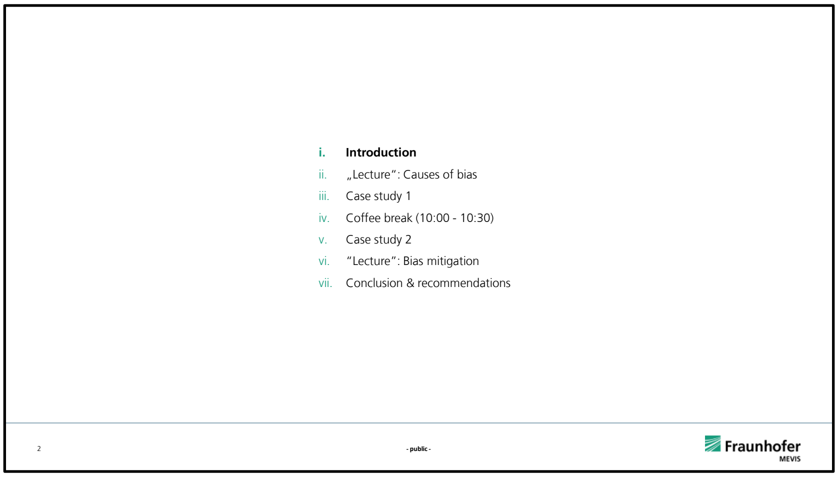
Fraunhofer MEVIS

Fraunhofer Institute for Digital Medicine MEVIS

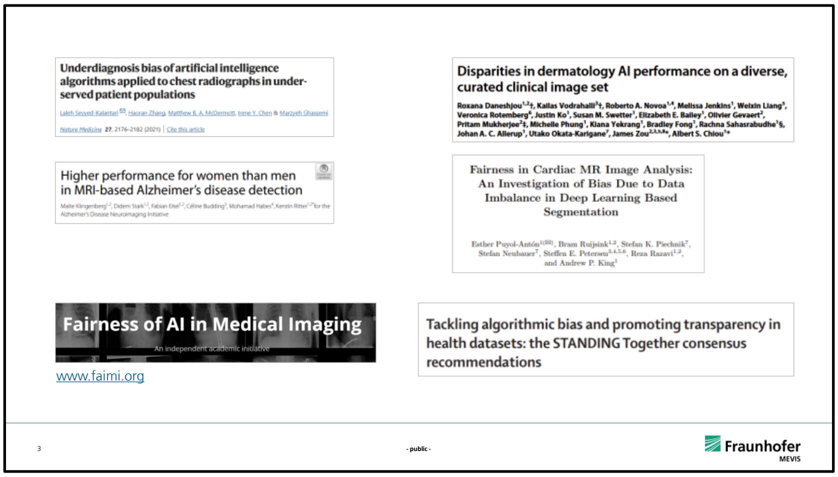
Bias and Fairness in Medical Imaging

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

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

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]*

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]

Fairness of AI in Medical Imaging

An independent academic initiative

www.faimi.org

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

3 - public - Fraunhofer MEVIS

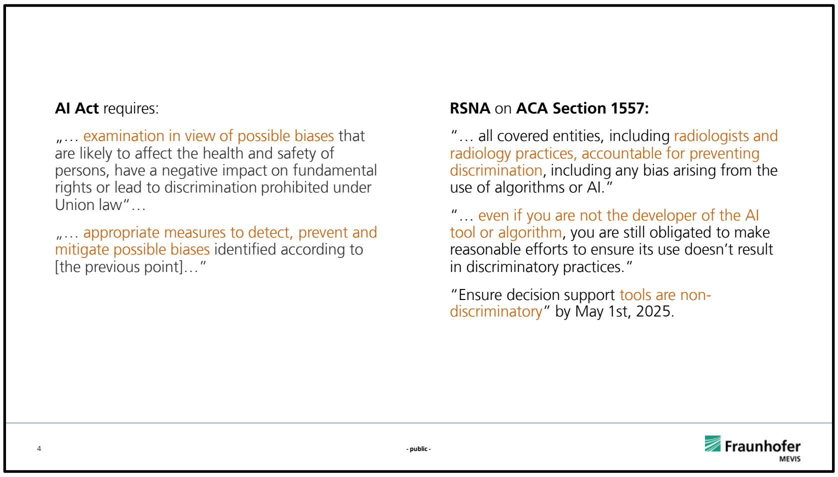
Bias has been observed in all areas of medical image analysis: Chest X-Ray analysis (most of the existing literature is on this), MRI, CT, dermatology, histopathology, classification/segmentation, etc.

Most bias analyses cover biases w.r.t. sex/gender, race/ethnicity, sometimes age. All other characteristics much less studied.

Many initiatives + guidelines to raise awareness and propose best practices.

STANDING Together: large consensus effort & detailed guideline, published in & endorsed by The Lancet in Jan 2025.

FAIMI: That's us! ☺ Independent academic initiative, now also a MICCAI SIG. We organize events / workshops / special issues etc. Sign up for our newsletter (on faimi.org)!



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]…“

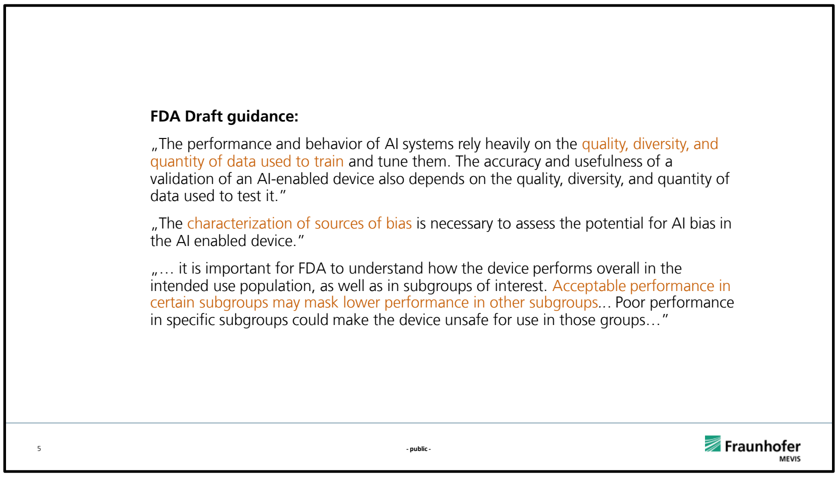
RSNA on **ACA Section 1557:**

"… 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.

AI Act: https://eur-lex.europa.eu/eli/reg/2024/1689/oj
RSNA on ACA Section 1557: https://www.rsna.org/-/media/files/rsna/practice-tools/faq-for-section-1557-aca.pdf



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."

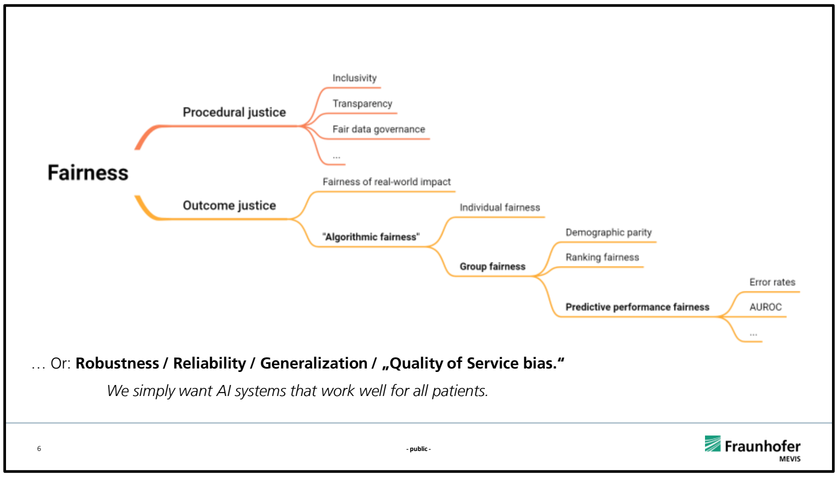
„The characterization of sources of bias is necessary to assess the potential for AI bias in the AI enabled device."

„… 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…"

- public -

Fraunhofer
MEVIS

https://www.fda.gov/media/184856/download
Document issued on January 7, 2025. (There was a public comment period; not entirely sure what happened to this.)



Fairness

Procedural justice

Inclusivity

Transparency

Fair data governance

…

Outcome justice

Fairness of real-world impact

"Algorithmic fairness"

Individual fairness

Group fairness

Demographic parity

Ranking fairness

Predictive performance fairness

Error rates

AUROC

…

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

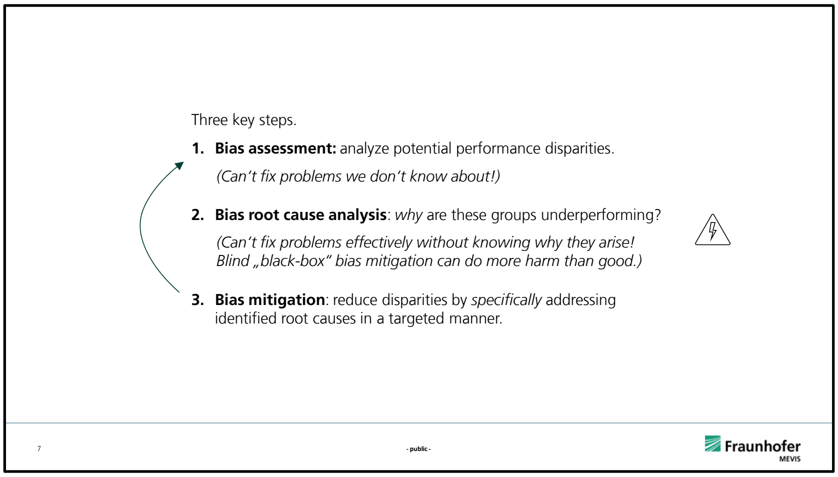
We simply want AI systems that work well for all patients.

6 - public - Fraunhofer MEVIS

Fairness/ethics encompass many very different ethical considerations, many of which are relevant to the medical imaging context. See e.g. D'Antonoli (2020), Ethical considerations for artificial intelligence: an overview of the current radiology landscape, https://pmc.ncbi.nlm.nih.gov/articles/PMC7490024/.

Here, we will focus on one very particular aspect: the „simple" question of whether models work equally well in different groups of patients, or whether there are groups for which models systematically underperform.

We will not care so much about differences between particular performance metrics (accuracy vs. sensitivity vs. specificity vs. AUROC vs. …). This is, again, not to say that the choice of a clinically meaningful and appropriate target metric is not important – it is! – but rather that all our messages in the following will largely focus on simply ***building better models***. This should translate into improvements across most, if not all, such metrics.



Three key steps.

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

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

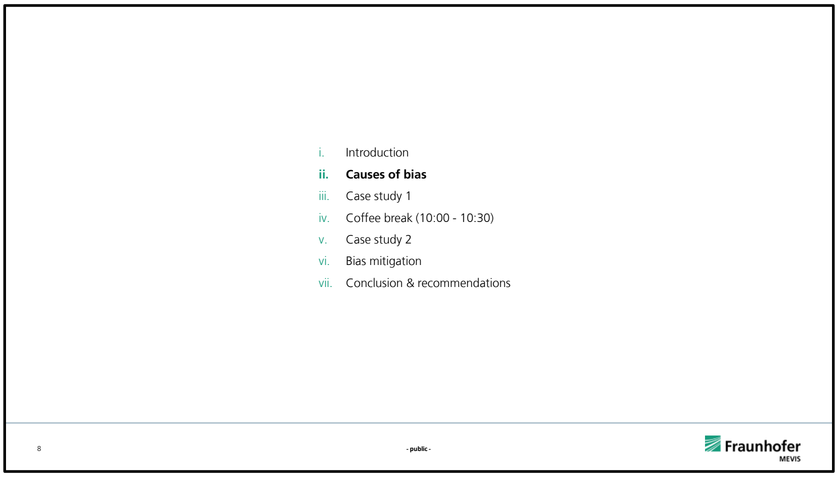
2. **Bias root cause analysis**: *why* are these groups underperforming?

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

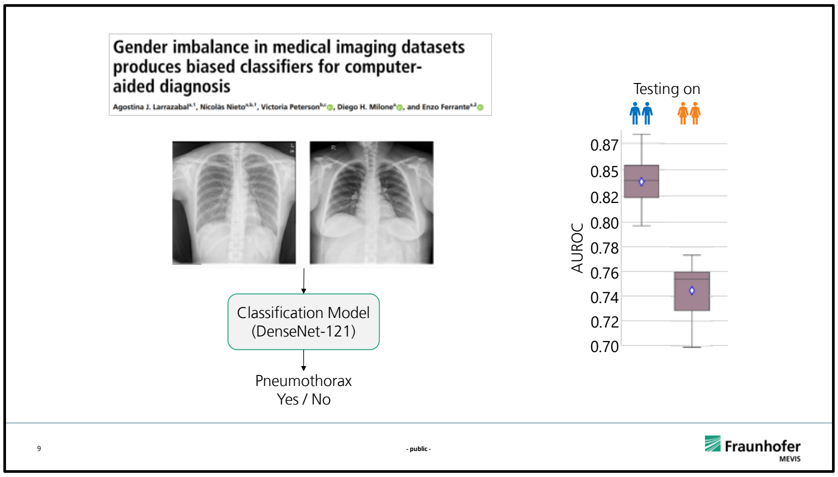
3. **Bias mitigation**: reduce disparities by *specifically* addressing identified root causes in a targeted manner.

7 - public - Fraunhofer MEVIS

Our focus will mostly be on step 2 here (root causes), but we will also briefly talk about the other two steps.



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

9 - public - Fraunhofer MEVIS

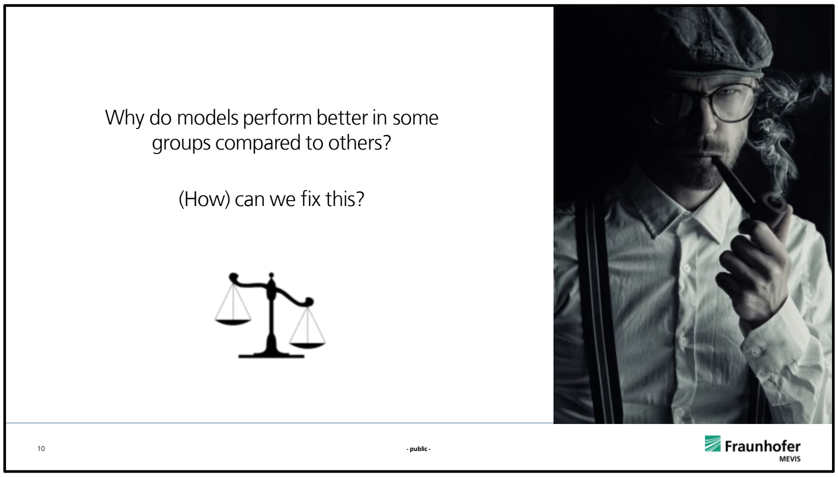
Pneumothorax: gas within the pleural space.

Tips to help to find pneumothoraces include:

- the lung edge: you should not be able to see the lung edge - if you can, the region peripherally is likely a pneumothorax.
- absence of vessels: the lung should have vessels running through it, these are white branching structures on the x-ray. If there are no vessels, there may be a pneumothorax.
- I am to this day incapable of seeing a pneumothorax on an x-ray. ☺

This is the gender-balanced case on the NIH dataset in the mentioned paper.

Larrazabal et al. (2020), Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis, https://www.pnas.org/doi/10.1073/pnas.1919012117.



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

(How) can we fix this?

10

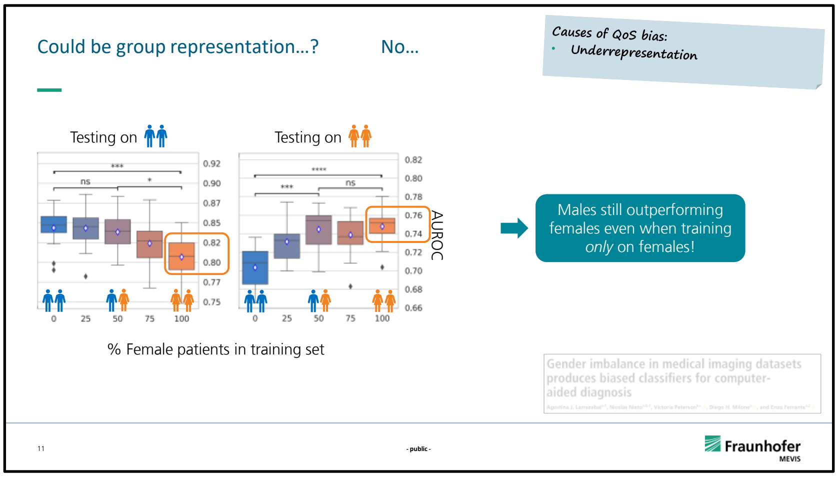
- public -

Fraunhofer MEVIS

Finding the root causes of bias can involve quite a lot of detective work! It's often highly nontrivial to uncover the source of observed performance disparities, as we will see in the following.

However, we need to find out what is *causing* a performance disparity in order to be able to say:

1. Whether or how the disparity can be fixed, and
2. Whether the observed disparity is indeed unfair. (There could be valid reasons for performance disparities, for instance if a task is just intrinsically much harder in one group for biological reasons.)



Could be group representation…?

No…

Causes of QoS bias:
- Underrepresentation

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

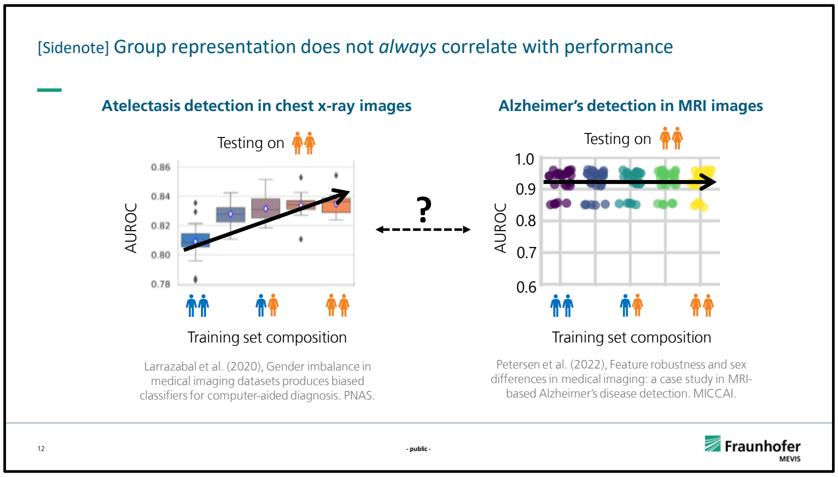
Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis

Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante

The first hypothesis for the previously shown performance disparity is often: were women under-represented in the training set?

This is not the case; that model was trained on a gender-balanced dataset. In fact, as we can see here, the model still performs better on men compared to women even when training *only* on female patients! (Note the different y axes in the two graphs.)

So: while under-representation *is* often a cause of under-performance (and the results in that paper very nicely show the effect of different group representations), this is clearly *not* the explanation for the observed performance gap.



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

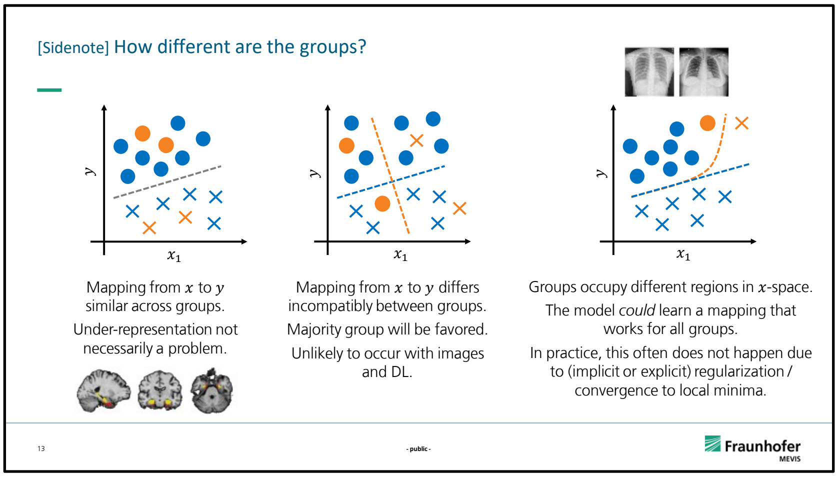
Atelectasis detection in chest x-ray images

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.

In a separate study on brain MRI-based AD classification (using the ADNI datasets), we found *no* effect of training dataset composition (in terms of gender representation) on model performance in both genders. Why is that? What is the difference between the two cases?



[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.

13

- public -

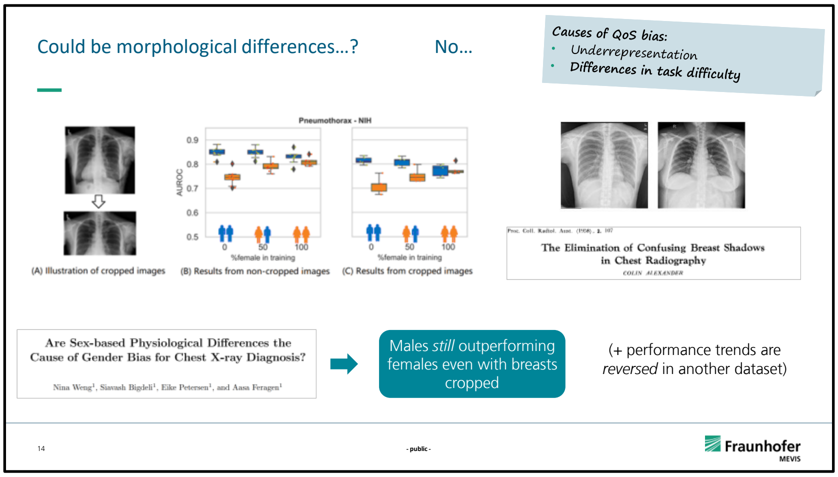
Fraunhofer MEVIS

x: images, y: target labels, e.g. disease labels.

Simplified illustrative 2D illustration to understand the core issues; in reality the x-space (image space) is of course very high-dimensional.

Note: these are *hypotheses* for what might be going on in the CXR / AD cases presented previously.

Also cf. Petersen et al. (2023), The path toward equal performance in medical machine learning, https://www.cell.com/patterns/fulltext/S2666-3899(23)00145-9.



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]

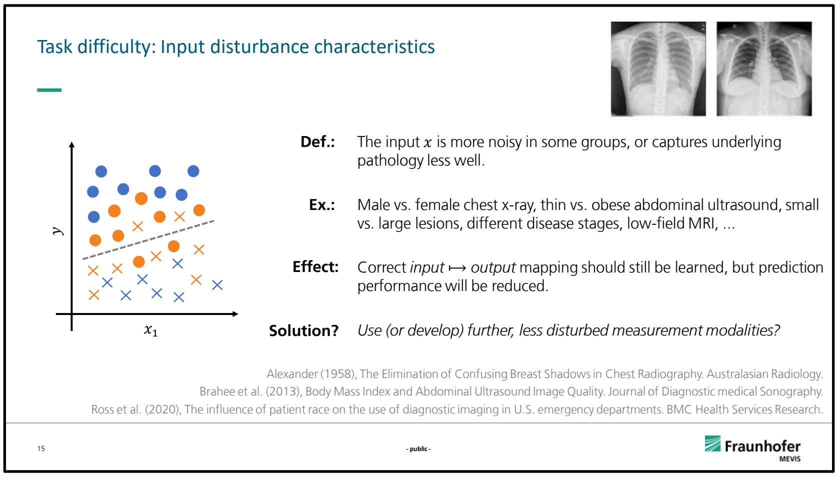
Males *still* outperforming females even with breasts cropped

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

Breast shadows as an obstacle are known issue since at least 1958… =>
Maybe the observed performance differences are simply because breast shadows are obscuring regions of the lungs important for diagnosis, making the classification task intrinsically harder in women?

Apparently not, but it was another plausible hypothesis.

Morphological differences do not *have* to translate into differing task difficulty: here, for example, they did not.



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* ⟼ *output* mapping should still be learned, but prediction performance will be reduced.

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

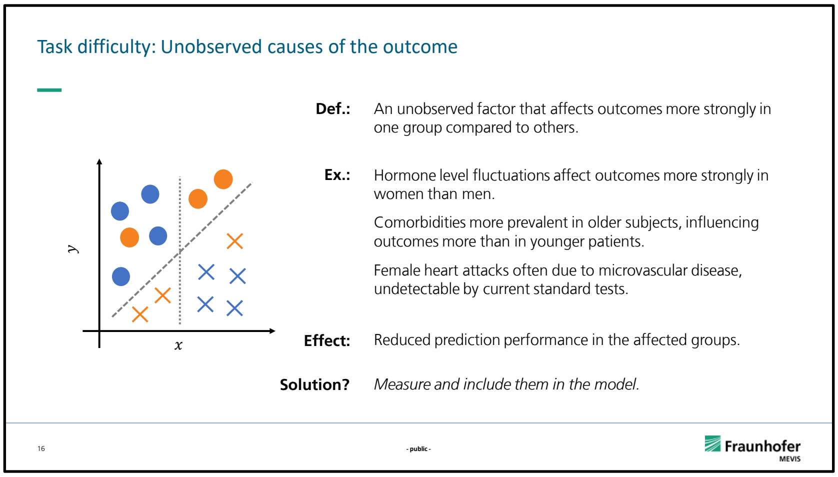
Alexander (1958), The Elimination of Confusing Breast Shadows in Chest Radiography. Australasian Radiology.
Brahee et al. (2013), Body Mass Index and Abdominal Ultrasound Image Quality. Journal of Diagnostic medical Sonography.
Ross et al. (2020), The influence of patient race on the use of diagnostic imaging in U.S. emergency departments. BMC Health Services Research.

- public -

Fraunhofer MEVIS

Potential reason #1 for differences in task difficulty. In the plot, colors correspond to e.g. gender, dots/crosses to diagnosis.

Also cf. Petersen et al. (2023), The path toward equal performance in medical machine learning, https://www.cell.com/patterns/fulltext/S2666-3899(23)00145-9.



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.

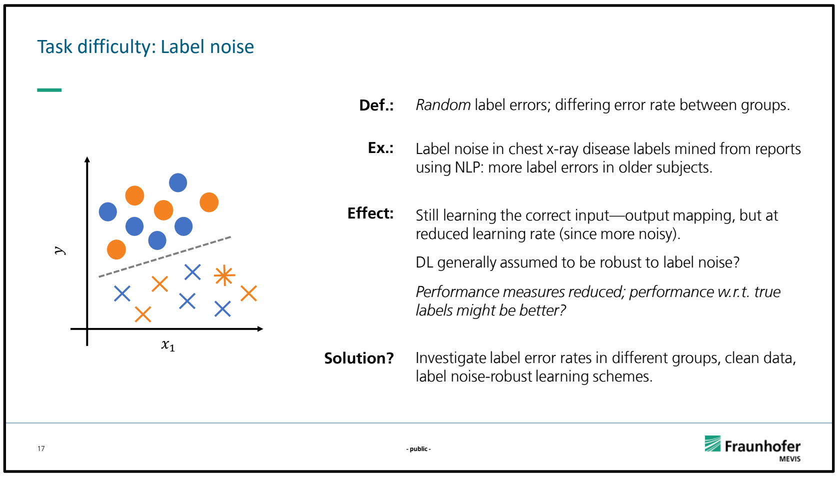
Effect: Reduced prediction performance in the affected groups.

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

16 - public - Fraunhofer MEVIS

Potential reason #2 for differences in task difficulty. Apart from the diagonal dividing line between disease or not, there is a second trend: the female are „high on disease“ for larger x compared to men.

Also cf. Petersen et al. (2023), The path toward equal performance in medical machine learning, https://www.cell.com/patterns/fulltext/S2666-3899(23)00145-9.



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.

17 - public - Fraunhofer MEVIS

Potential reason #3 for differences in task difficulty.

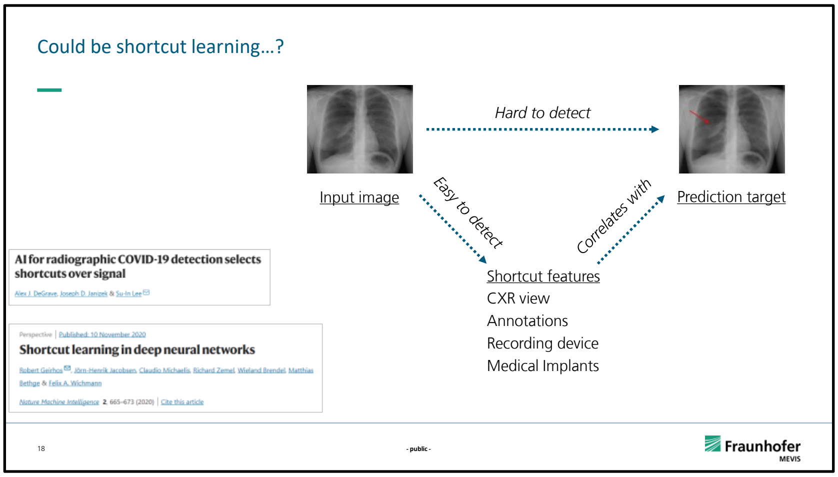
See e.g.

Zhang et al. (2022), Improving the Fairness of Chest X-ray Classifiers. PMLR / CHIL.

Frenay and Verleysen (2014), Classification in the Presence of Label Noise: A Survey. IEEE TNNLS.

Rolnick et al. (2017), Deep Learning is Robust to Massive Label Noise.
Santomartino et al. (2024), Evaluating the Performance and Bias of Natural Language Processing Tools in Labeling Chest Radiograph Reports, https://pubs.rsna.org/doi/10.1148/radiol.232746

Also cf. Petersen et al. (2023), The path toward equal performance in medical machine learning, https://www.cell.com/patterns/fulltext/S2666-3899(23)00145-9.



Could be shortcut learning…?

Input image

Hard to detect

Prediction target

Easy to detect

Correlates with

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

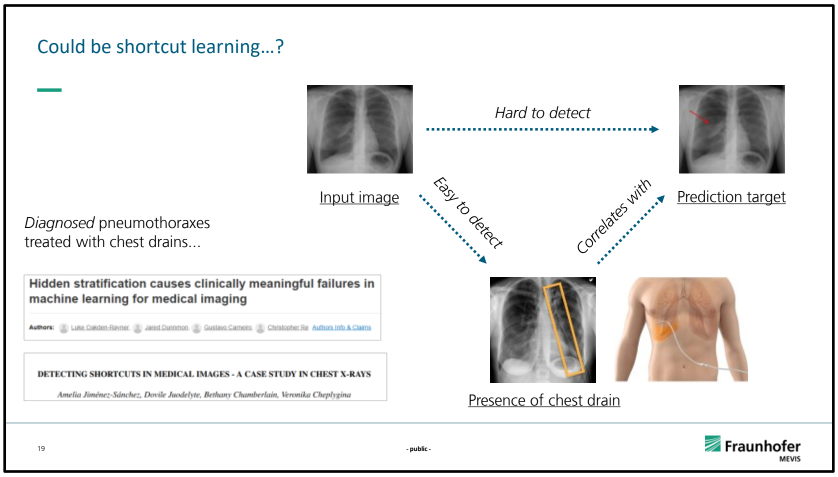
Shortcut learning in deep neural networks

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

Back to our chest x-ray gender performance disparity example…

Shortcut learning is also an issue in segmentation, not only in classification: Lin et al. (2024), Shortcut Learning in Medical Image Segmentation, https://link.springer.com/chapter/10.1007/978-3-031-72111-3_59. The problem appears less severe there, however.



Could be shortcut learning…?

Input image

Hard to detect

Prediction target

Easy to detect

Correlates with

Presence of chest drain

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

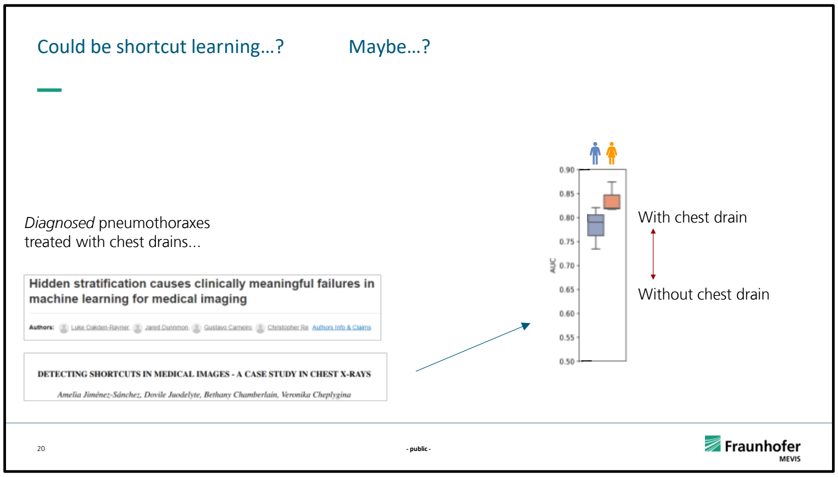
DETECTING SHORTCUTS IN MEDICAL IMAGES - A CASE STUDY IN CHEST X-RAYS

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

19 - public - Fraunhofer MEVIS

Chest drain visualization source Chest Tubes and Indwelling Pleural Catheters | Thoracic Key

It had actually been reported earlier (in 2020, see the excellent Oakden-Rayner paper) that CXR pneumothorax diagnosis models are prone to picking up on the chest drain shortcut. (We did not know this when we started working on this, and the vast majority of the academic CXR diagnosis literature also seems to be blissfully unaware of this issue to this day.)



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

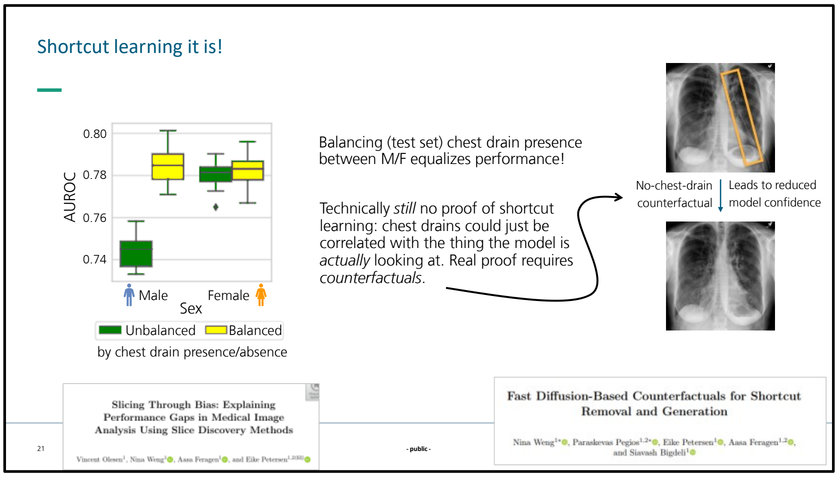
DETECTING SHORTCUTS IN MEDICAL IMAGES - A CASE STUDY IN CHEST X-RAYS

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

The figure is from the second paper (Jiménez-Sanchez et al.) and is on the NIH dataset (like everything until here).

When we split performance estimates by sex/gender *and* the presence/absence of a chest drain in the image, we see opposite disparities between the two genders/sexes!

Does this explain our observed gender performance disparity?



Shortcut learning it is!

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*.

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]

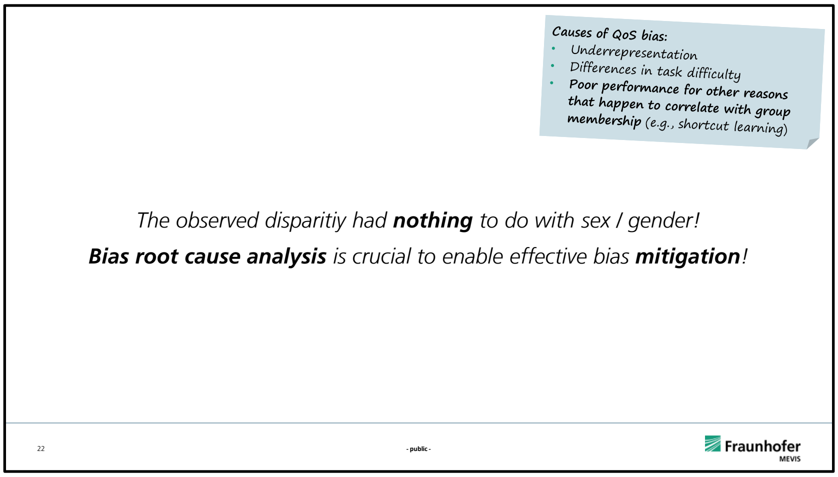
If we balance the test set such that chest drain prevalence is equal in the M/F groups, the previously observed performance disparities disappear!*

Thus, it appears that it is in fact the chest drain shortcut – which is „working differently well" in the two groups – that is causing these observed M/F performance disparities.

Additional evidence in the ECCV 2024 paper mentioned on the right (see https://fastdime.compute.dtu.dk/ or https://arxiv.org/abs/2312.14223): Artificially adding/removing chest drains from images significantly changes model confidence w.r.t. pneumothorax.

Olesen et al., Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis using Slice Discovery Methods, https://arxiv.org/abs/2406.12142

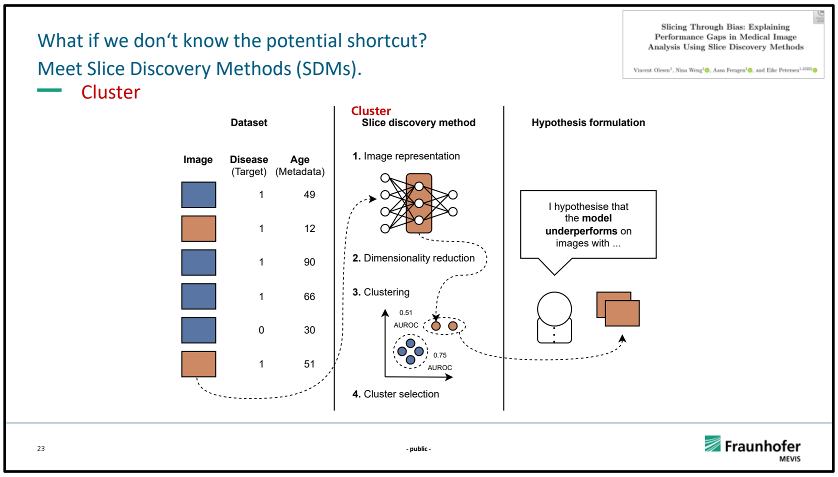
* These results are on the CheXpert dataset, where M/F performance disparities are *reversed* compared to the NIH dataset we were looking at, so far.



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(✉)]

Dataset

Image	Disease (Target)	Age (Metadata)
🟦	1	49
🟧	1	12
🟦	1	90
🟦	1	66
🟦	0	30
🟧	1	51

Cluster
Slice discovery method

1. Image representation
2. Dimensionality reduction
3. Clustering
4. Cluster selection

Hypothesis formulation

I hypothesise that the **model underperforms** on images with ...

What if you *don‘t know* already what might be causing the issue? E.g., if we did not already know about the potential chest drain shortcut, how can we find our about such issues?

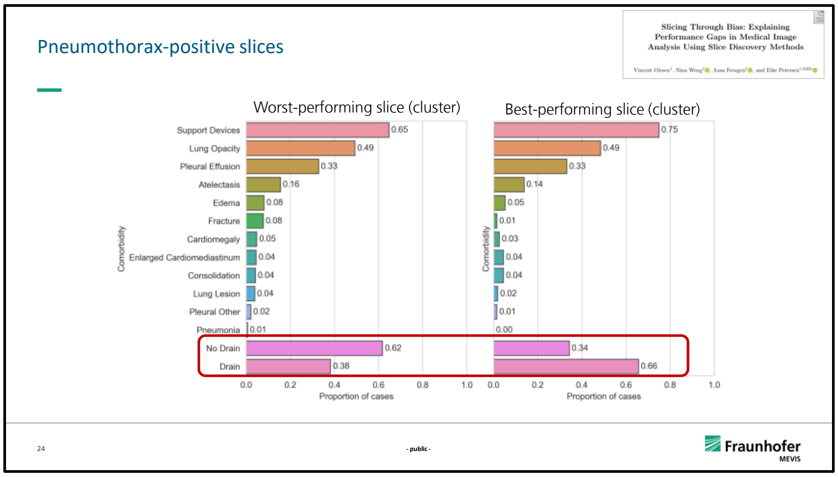
There is a branch of literature (under various different names such as „Automatic Slice Discovery“, „Blind Spot Discovery“, „Failure Mode Detection“, …) that treats this problem.

(The nomenclature is a bit confusing especially in the medical imaging context, unfortunately – „slice“ here does *not* refer to a 2D slice of a 3D image; it refers to a „slice“ of data ≈ a cluster.)

These methods can be summarized quite simply as:

1. Cluster images / volumes based on similarity.
2. Find clusters that are performing unexpectedly well/poorly. (-- statistical significance testing, multiple hypothesis corrections, …)
3. Investigate these manually (visually = look at samples from them) or use multimodal models to automatically suggest captions for the different clusters. Try to find the defining feature of these clusters, thereby generating hypothesis regarding potential causes of good/poor performance.
4. Investigate these generated hypotheses further, such as what we did for the chest drain case on the previous slides. (Specific analyses will be hypothesis- and application-specific.)

Cf. Olesen et al., Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis using Slice Discovery Methods, https://arxiv.org/abs/2406.12142



Pneumothorax-positive slices

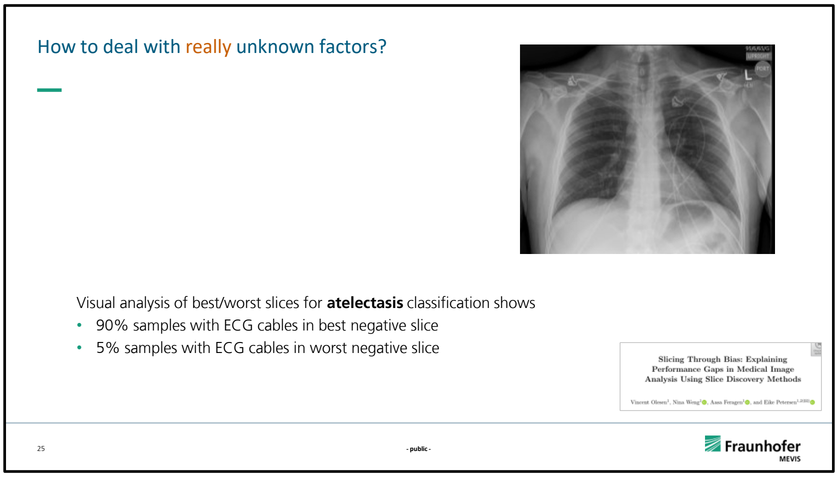
Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods

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

In our case, we already knew about the chest drain shortcut. But let‘s imagine we didn‘t – would our slice discovery method find this?

Indeed, we find that the major difference between the best and worst discovered (pneumothorax-positive) clusters/slices is the prevalence of chest drains.

However, to generate the above plot, we already need chest drain labels / metadata! **What if we don‘t have labels for the potential shortcut factor?**



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

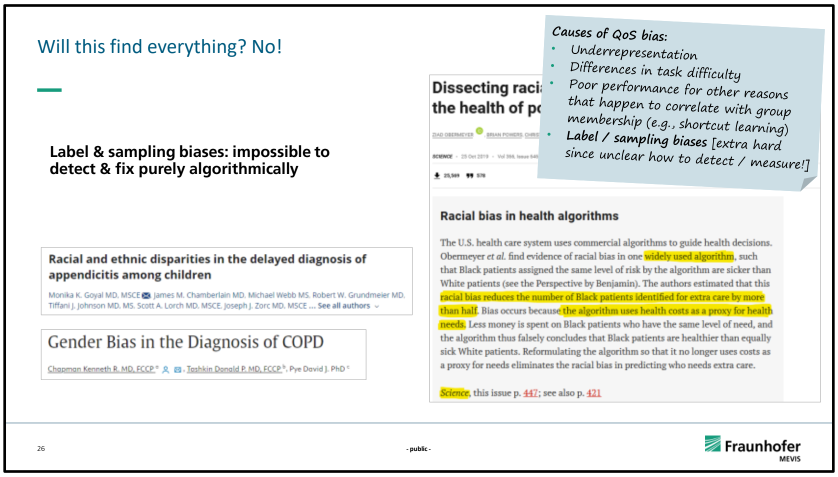
Slicing Through Bias: Explaining Performance Gaps in Medical Image Analysis Using Slice Discovery Methods

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

Here, we had such a case: we really did not know about this shortcut – and, to our knowledge, at least, this had also not been described before.

Based on our slice discovery method, we found several clusters that performed very differently w.r.t. atelectasis classification. Initial visual inspection of the resulting clusters indicated that a differentiating factor might be the presence/absence of ECG cables.

We then manually labeled only the images in the best/worst clusters in terms of presence/absence of ECG cables, and found the shown **very strong differences in ECG cable prevalence**, suggesting another instance of shortcut learning.



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 raci
the health of p

ZIAD OBERMEYER, BRIAN POWERS, CHRIS

SCIENCE • 25 Oct 2019 • Vol 366, Issue 646

25,569 578

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!]*

Racial bias in health algorithms

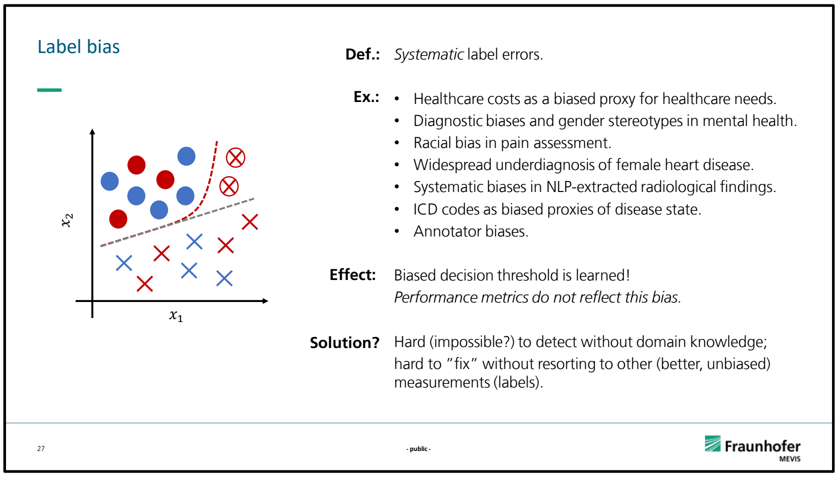
The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The 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

26 - public - Fraunhofer MEVIS

The Paper on the right is Obermeyer et al. (2019), Dissecting racial bias in an algorithm used to manage the health of population, Science.

- Algorithm predicts risk and suggests care
- Observation: At same disease state, black patients are assessed to have lower risk than white, receive less care
- Cause: Less money is spent on black patients, but care cost is used as a proxy (label) for need. (Bias comes from „trained practice“ to under-spend on Black population)



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).

See e.g.

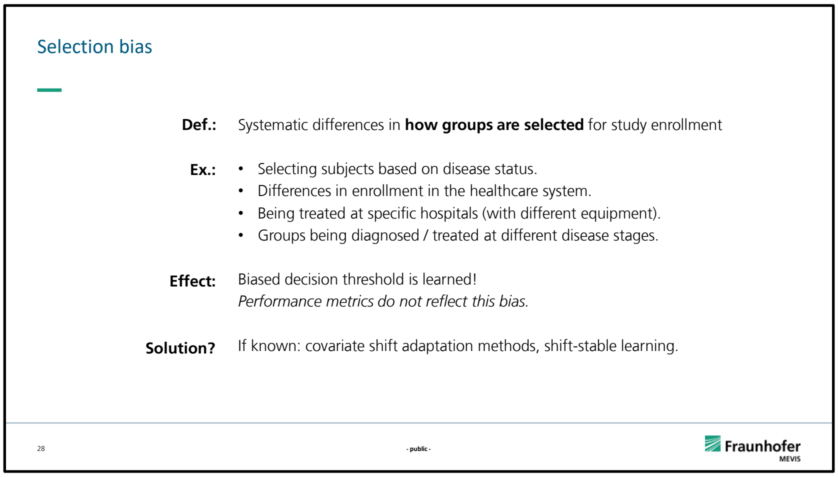
Obermeyer et al. (2019), Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations. Science.

Sigmon et al. (2005), Gender Differences in Self-Reports of Depression: The Response Bias Hypothesis Revisited. Sex Roles.

Hoffman et al. (2016), Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. PNAS.

Wang et al. (2021), Fair Classification with Group-Dependent Label Noise. ACM FaccT.

Santomartino et al. (2024), Evaluating the Performance and Bias of Natural Language Processing Tools in Labeling Chest Radiograph Reports: https://doi.org/10.1148/radiol.232746



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.

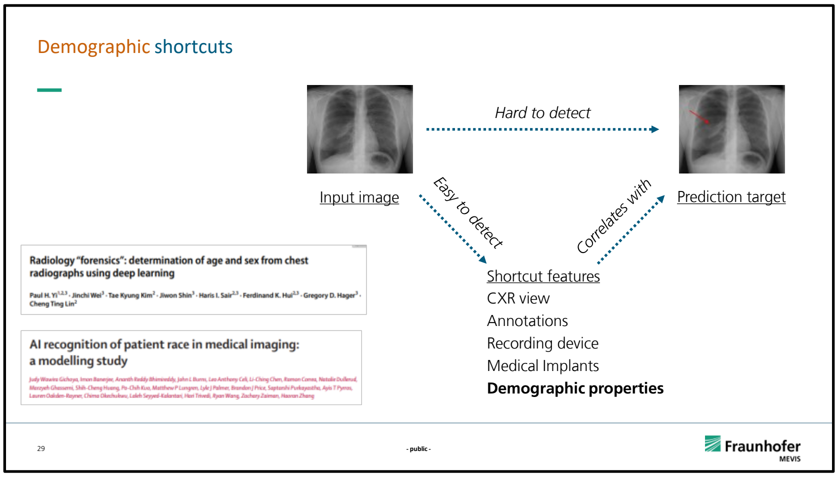
See e.g.

Ellenberg (1994), Selection bias in observational and experimental studies. Statistics in Medicine.

Natanson et al. (1998), The sirens' songs of confirmatory sepsis trials: selection bias and sampling error. Critical Care Medicine.

Zadrozny (2004), Learning and evaluating classifiers under sample selection bias. ICML.

Subbaswamy and Saria (2019), From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics.



Demographic shortcuts

Input image

Hard to detect

Prediction target

Easy to detect

Correlates with

Shortcut features

CXR view

Annotations

Recording device

Medical Implants

Demographic properties

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

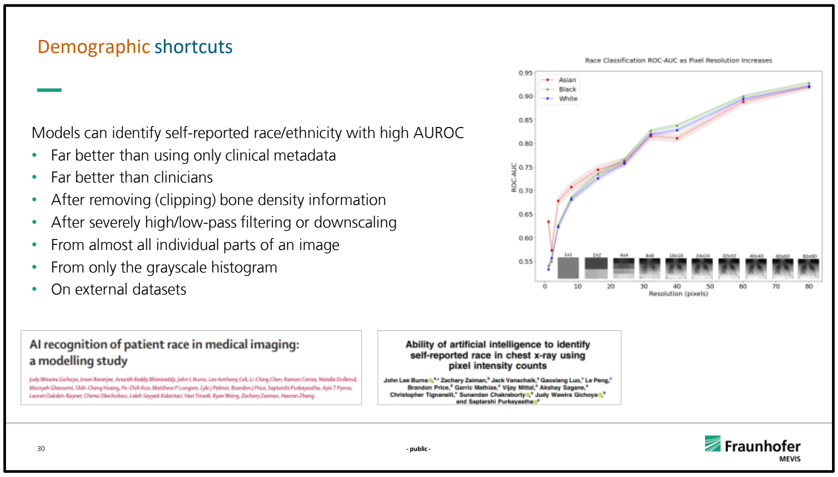
Most/all demographic patient properties can be reconstructed with very high accuracy and robustness from medical images. This raises the potential for models to exploit these potential shortcut features, i.e., learning that membership in a certain group increases or decreases the likelihood of a certain target (disease) label [0].

This is akin to a logistic regression risk prediction model that uses demographic properties as input variables [1,2]

It is also arguably an instance of *direct discrimination* [3-5].

[0] Banerjee et al. (2023), "Shortcuts" Causing Bias in Radiology Artificial Intelligence: Causes, Evaluation, and Mitigation, https://doi.org/10.1016/j.jacr.2023.06.025
[1] Diao et al. (2024), Implications of Race Adjustment in Lung-Function Equations, NEJM
[2] Zink et al. (2024), Race adjustments in clinical algorithms can help correct for racial disparities in data quality, PNAS
[3] Deck et al. (2024), Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness, https://arxiv.org/abs/2403.20089
[4] Gerards and Borgesius (2021), Protected Grounds and the System of Non-Discrimination Law in the Context of Algorithmic Decision-Making and Artificial Intelligence, Colorado Technology Law Journal
[5] Wachter et al. (2021), Bias Preservation in Machine Learning: The Legality of Fairness

Metrics under EU Non-Discrimination Law, West Virginia Law Review
[I am not a lawyer…]



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

Race Classification ROC-AUC as Pixel Resolution Increases

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

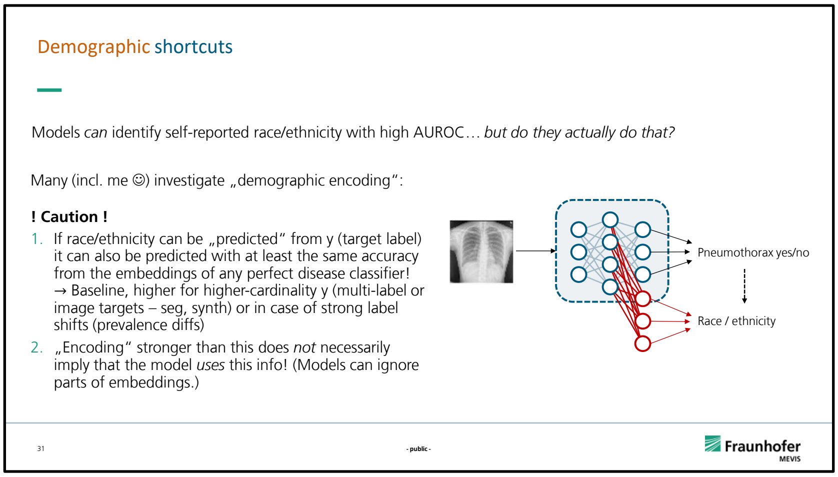
Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima 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]

30 - public - Fraunhofer MEVIS

When looking at the plot, note that the inset images aren‘t at the right x axis positions. This makes you think the 1,2x2,4x4 already work, but they almost don‘t.



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.)

Pneumothorax yes/no

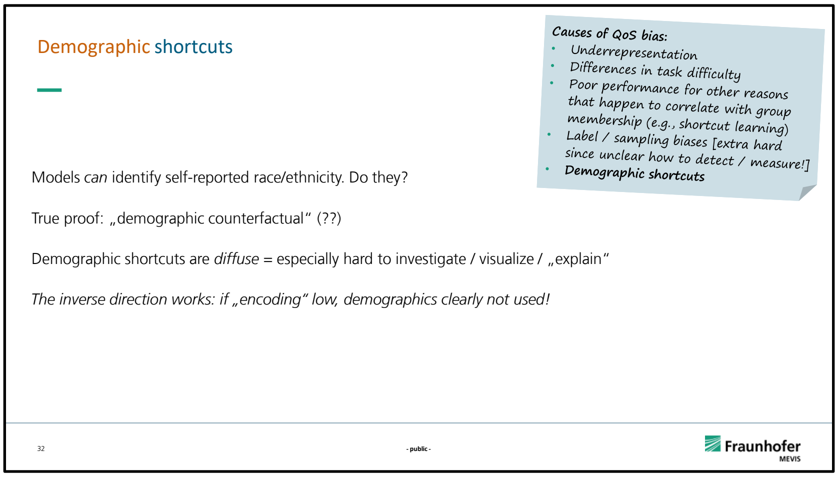
Race / ethnicity

31 - public - Fraunhofer MEVIS

See e.g.

- Petersen et al., „Are demographically invariant models and representations in medical imaging fair?“, https://arxiv.org/abs/2305.01397
- Brown et al., „Detecting shortcut learning for fair medical AI using shortcut testing“, https://www.nature.com/articles/s41467-023-39902-7
- Glocker et al., „Algorithmic encoding of protected characteristics in chest X-ray disease detection models“, https://www.thelancet.com/journals/ebiom/article/PIIS2352-3964(23)00032-4

… among many others.



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“ (??)

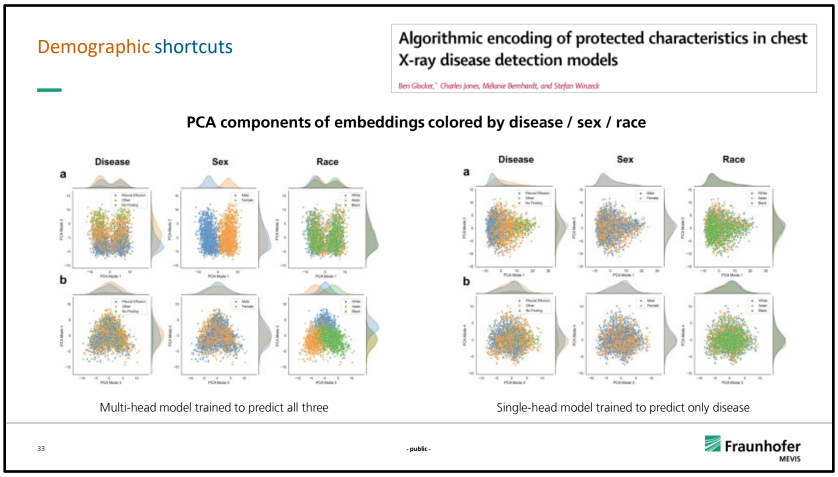
Demographic shortcuts are *diffuse* = especially hard to investigate / visualize / „explain“

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

32 - public - Fraunhofer MEVIS

For a discussion on demographic counterfactuals + pointers to relevant literature, cf. Petersen et al., „Are demographically invariant models and representations in medical imaging fair?“, https://arxiv.org/abs/2305.01397

For the fact that shortcuts are diffuse, cf. Gichoya et al., AI Recognition of patient race …



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

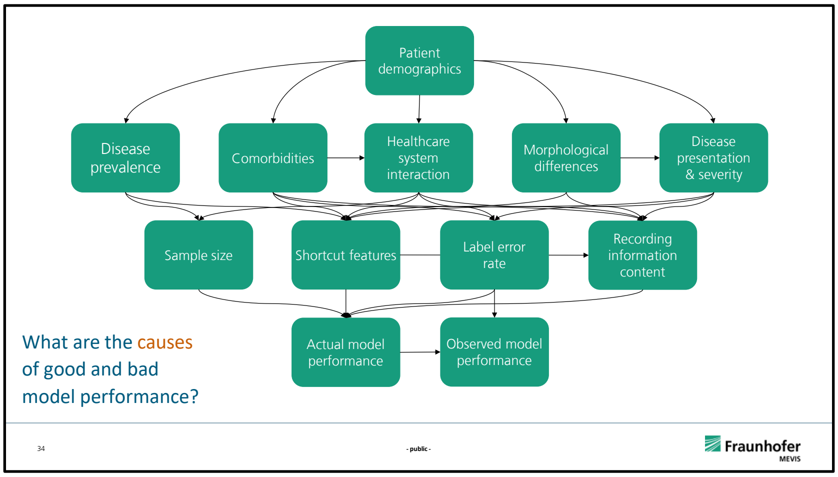
Multi-head model trained to predict all three

Single-head model trained to predict only disease

Empirically, there is actually not so much encoding (this is what the right figure shows) compared to how strongly models *could* encode demographics (this is what the left figure shows)!

I.e., for standard task-specific classifiers, we are clearly not in the „>90% protected attribute classification ability“ space (which would be possible).

That means: the risk is lower than what we may have feared, but it is still non-zero. Do such models exploit demographic properties to *some* degree? We don‘t really know, and I have not seen an approach to investigate this that I personally find very convincing.



Patient demographics

Disease prevalence

Comorbidities

Healthcare system interaction

Morphological differences

Disease presentation & severity

Sample size

Shortcut features

Label error rate

Recording information content

Actual model performance

Observed model performance

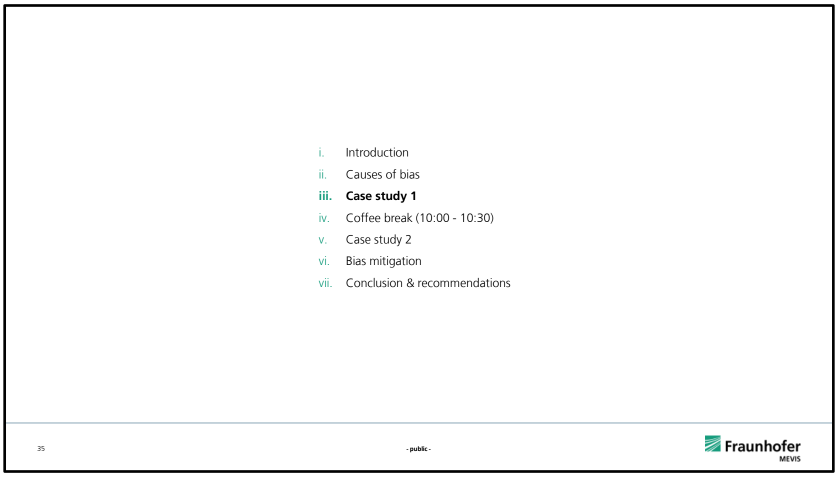
What are the causes of good and bad model performance?

34

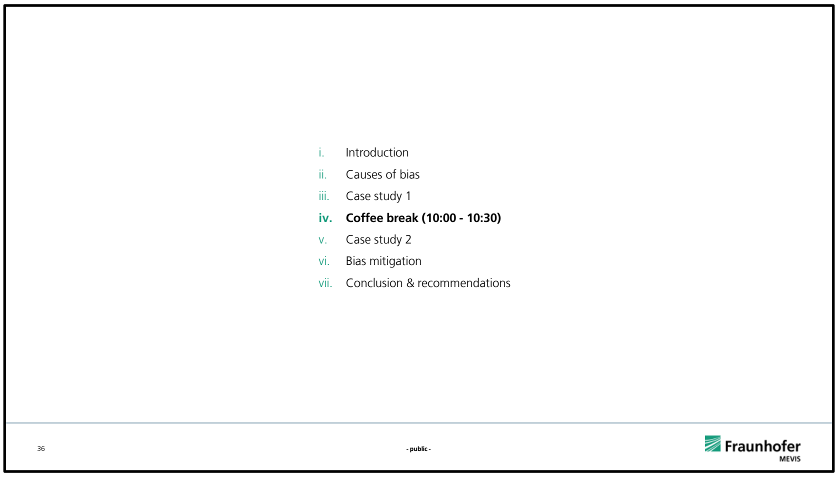
- public -

Fraunhofer MEVIS

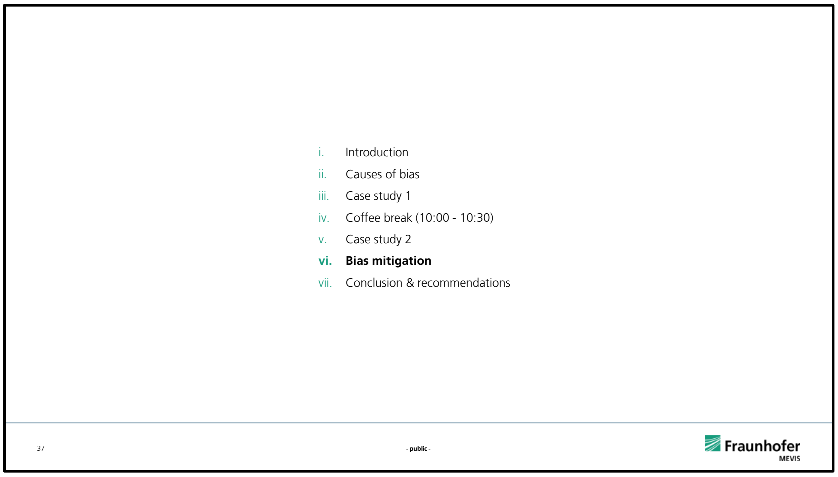
Many things can cause good or bad model performance in a demographic group. Figuring out the relevant cause for an observed disparity is often non-trivial, but it is necessary.



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

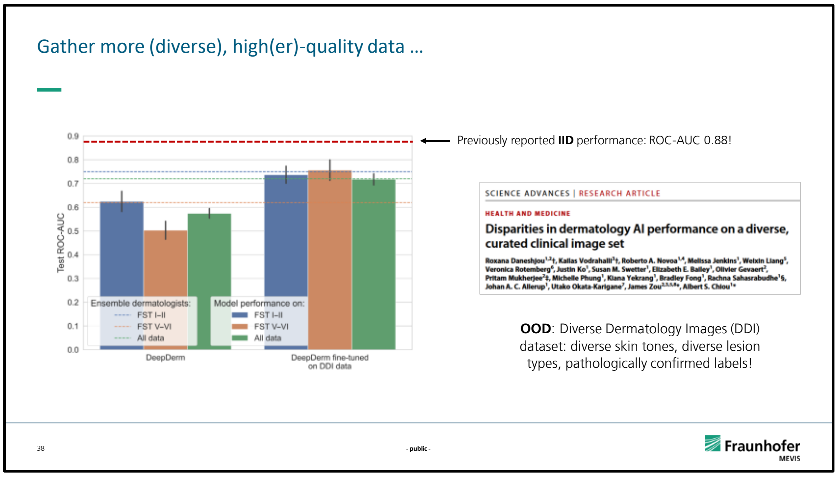
iii. Case study 1

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

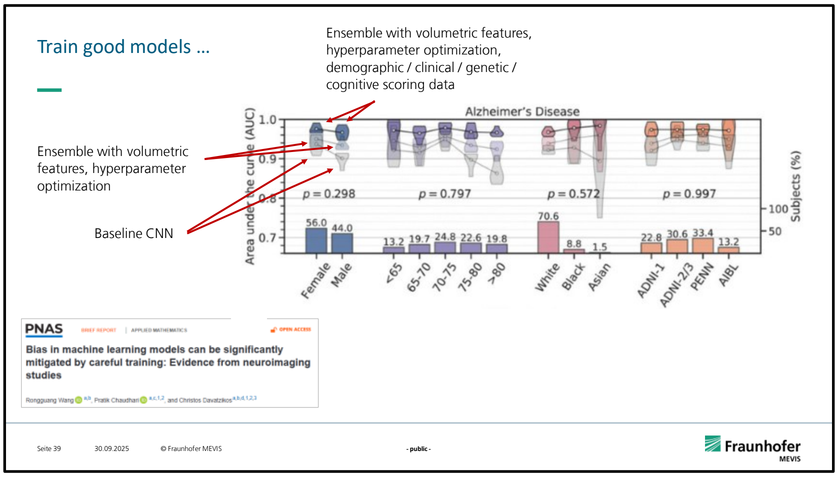
Disparities in dermatology AI performance on a diverse, curated clinical image set

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!

38 - public - Fraunhofer MEVIS

- Diversity matters across *all* dimensions: demographic representation, technical acquisition parameters, disease presentations & subtypes, etc.
- „Higher-quality data“ primarily means: (Diversity and) higher-quality *target labels / outcomes*!
- Clinicians often believe that high-quality image acqusition is of importance, but the opposite is true: models need to see representative real-world images with all the noise, potential confounders, acqusition errors etc. that will inevitably come up in real-world recordings.
- Remark for the plot: The „previous performance“ is on IID data; we are here concerned with the within-OOD-performance comparison.



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

Seite 39 30.09.2025 © Fraunhofer MEVIS - public - Fraunhofer MEVIS

Plot: **Solid colors**: Ensemble using all data. **Translucent Color**: Only images. **Translucent Gray**: Baseline model. **Conclusion:** Preprocessing/HPT improves performance, but only diverse data achieves level performance in subgroups. [1]

Similar results in our MICCAI 2022 paper: https://link.springer.com/chapter/10.1007/978-3-031-16431-6_9

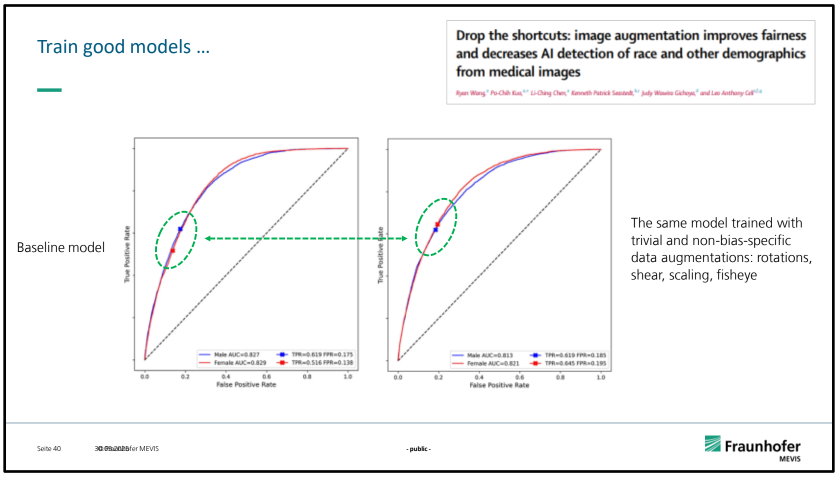
"Normalize away" any **unimportant** group differences wherever possible. (Harmonization, standardization, registration, removal of potential visual confounders, etc.) [2]

Personal, highly unscientific guesstimate: >50% of bias findings reported in the literature are simply a result of models being poorly trained, highly non-robust, and just generally not working very well.

[1] Excerpt from the figure caption: „*For Alzheimer's disease, schizophrenia, and autism spectrum disorder, using data from different studies (e.g., ADNI-1, ADNI-2/3, PENN, and AIBL for Alzheimer's disease), we built an ensemble that uses* ***data from multiple sources*** *(MR imaging features, demographic, clinical variables, genetic factors, and cognitive scores); (SI Appendix). This ensemble was trained using preprocessing and hyperparameter optimization discussed in Materials and Methods. We also trained an ensemble using* ***only imaging features****. Bar plots denote the size of the subgroup/study (%); in many cases,*

there is a strong imbalance in the data. Violin plots denote the test AUC on ***five different held-out subsets of data****. Solid colors indicate that models used all features, while translucent colors indicate that models were trained only on imaging features. Translucent gray denotes the AUC of a baseline deep network (without appropriate preprocessing and hyperparameter tuning). White dots denote the average AUC of each subgroup/study. For the ensemble trained on multisource data, the P-values shown in the figure indicate that we cannot reject the null hypothesis that the AUC for different subgroups has the same mean (at significance level < 0.01). This is not the case for the baseline deep network.*

[2] 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, https://doi.org/10.1016/j.cmpb.2024.108200.



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]

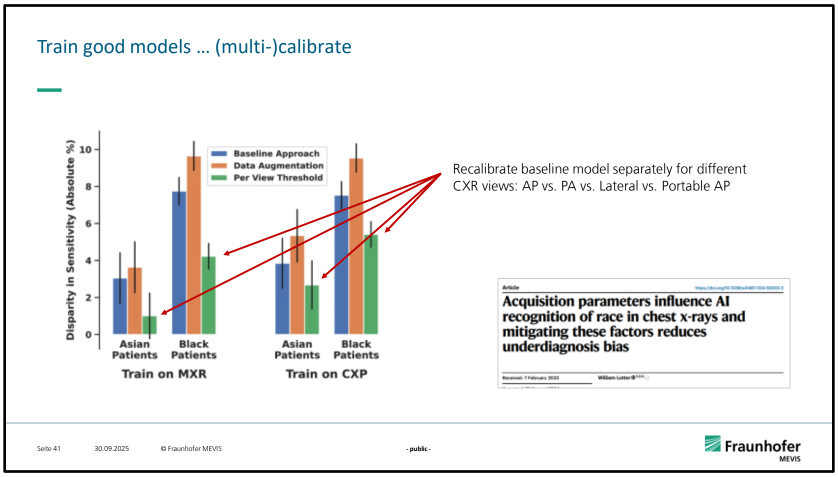
Baseline model

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

This is edema detection from CXR.

In the baseline model, men/women end up at very different TPR/FPR operating points in ROC space at the same decision threshold.

In the model trained with trivial data augmentations, the resulting operating points are very similar.



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

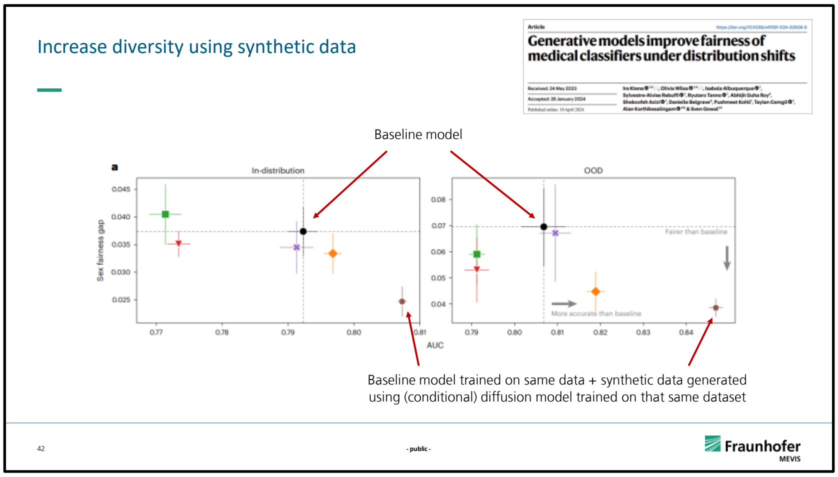
Seite 41 30.09.2025 © Fraunhofer MEVIS - public -

Fraunhofer MEVIS

Ensure robustness to technical variations: scanner type, field strength, view position, lighting conditions, imaging protocol, potential confounders such as medical implants, …

Note: Data augmentation worsens the performance, while here the thresholds help a lot. Neither use any balancing or „fairness“ methods, don‘t even need to know sensitive attributes at train time.

Note that here, lower bars are better.



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, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Taylan Cemgil, Alan Karthikesalingam & Sven Gowal

Baseline model

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

42 - public - Fraunhofer MEVIS

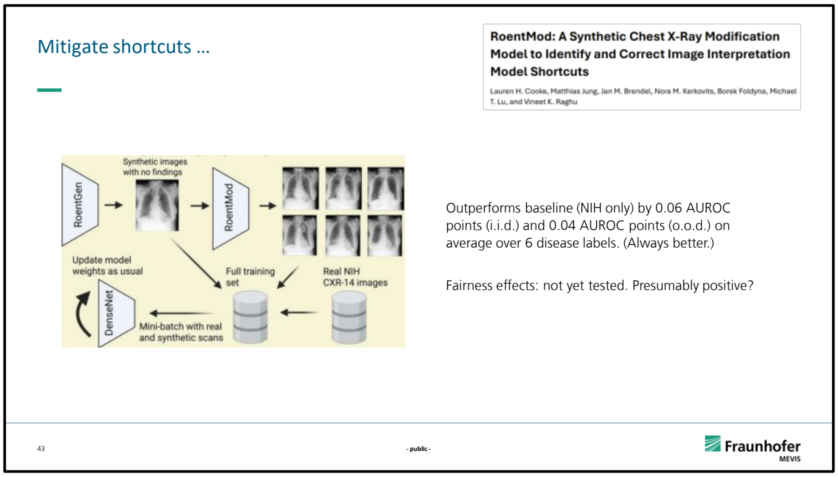
This is just another form of data augmentation, which appears crucial for achieving robustness, but can give very different results in general.

Plot: Average AUC values across several disease labels on CheXpert (In-distribution, left) and ChestX-ray14 (OOD, right). Excellent paper with convincing results in other domains (Histopathology, Dermatology).

This is again **not a specific bias intervention**! The generated synthetic images are only conditioned on the disease label, *not* on demographic group membership.
Also: various works on skin color augmentations in dermatology AI [1,2]

[1] Rezk et al. (2022), Improving Skin Color Diversity in Cancer Detection: Deep Learning Approach, https://derma.jmir.org/2022/3/e39143/
[2] Pakzad et al. (2023), CIRCLe: Color Invariant Representation Learning for Unbiased Classification of Skin Lesions, https://doi.org/10.1007/978-3-031-25069-9_14

Another very promising preprint mentioned in Prof. Langlotz' keynote: Improving Performance, Robustness, and Fairness of Radiographic AI Models with Finely-Controllable Synthetic Data | Abstract



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

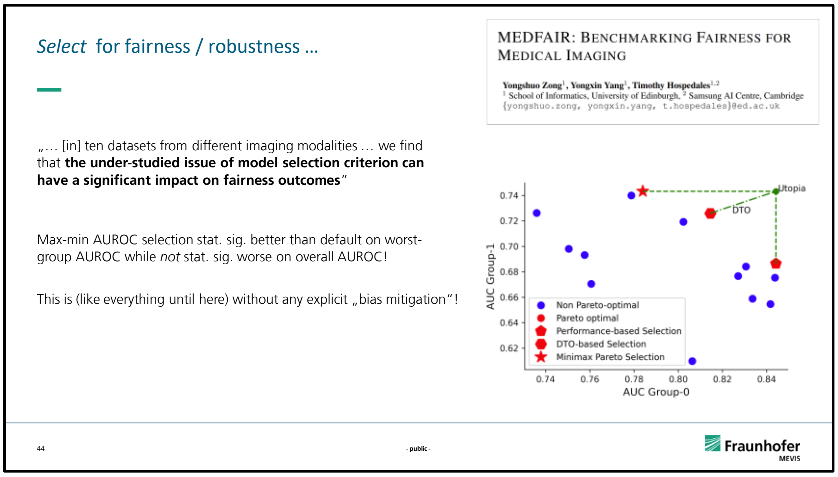
RoentGen
Synthetic images with no findings
RoentMod
Update model weights as usual
Full training set
Real NIH CXR-14 images
DenseNet
Mini-batch with real and synthetic scans

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?

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

This is stable diffusion text-to-image (RoentGen) followed by image-to-image (RoentMod).



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!

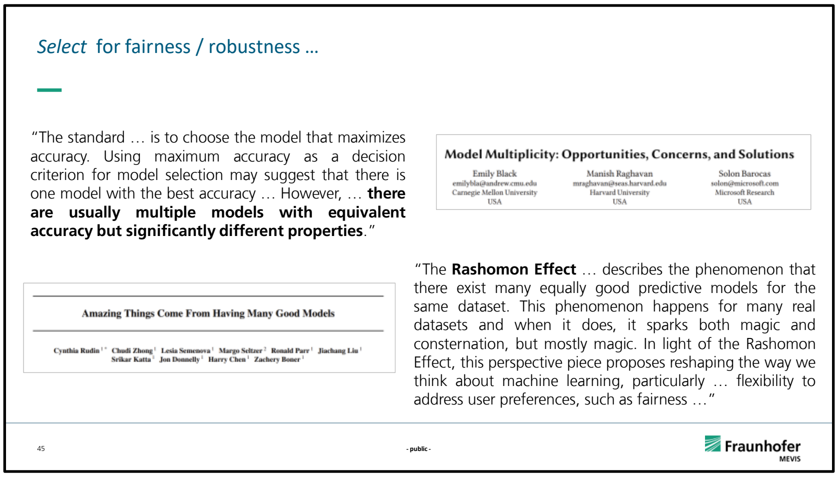
This is (like everything until here) without any explicit „bias mitigation“!

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

44 - public - Fraunhofer MEVIS

„Pareto optimal“ = max min AUROC



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

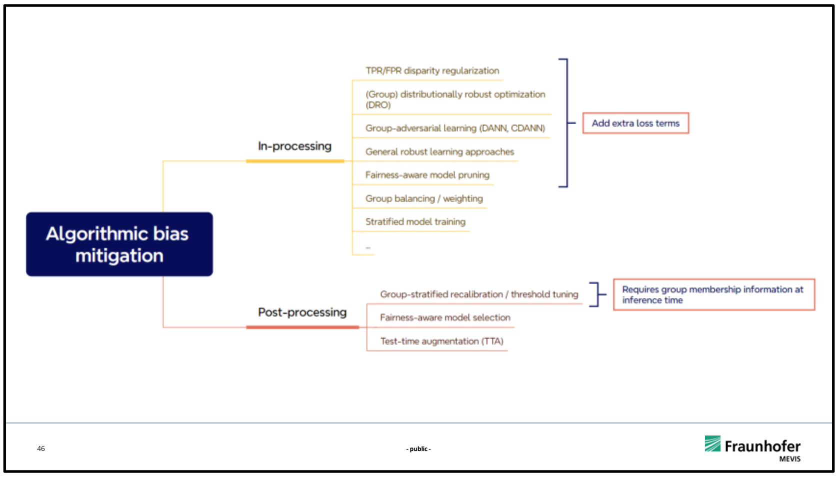
Amazing Things Come From Having Many Good Models

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 …"

45 - public - Fraunhofer MEVIS

Model multiplicity is an emerging and very active field with wide implications for robustness, fairness, interpretability, …



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

46 - public - Fraunhofer MEVIS

Many studies find that none of these consistently outperform standard (baseline) empirical risk minimization (ERM) in terms of resulting model fairness.

Personal opinion: I claim that this should *not* be what you spend most of your time on when dealing with potential bias issues. Focus on improving your data / preprocessing / augmentations / task set-up, "simply" training a good, robust model using standard ML best practices, and bias assessment first. *If* important issues remain, come back here.

The most impactful and practical strategy on this slide is probably fairness-aware *model selection:* simply select the model with the best worst-group performance [1,4]

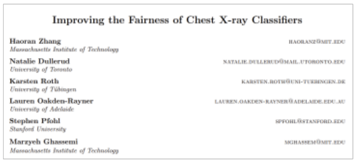
For some overviews and implementations, see e.g.
[1] Zong et al. (2023), MEDFAIR: BENCHMARKING FAIRNESS FOR MEDICAL IMAGING, https://arxiv.org/pdf/2210.01725
[2] Yang et al. (2024), The limits of fair medical imaging AI in real-world generalization, https://www.nature.com/articles/s41591-024-03113-4
[3] Hort et al. (2024), Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey, https://dl.acm.org/doi/full/10.1145/3631326
https://fairlearn.org/
[4] Black et al. (2022), *Model Multiplicity: Opportunities, Concerns, and Solutions*, https://doi.org/10.1145/3531146.3533149.

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

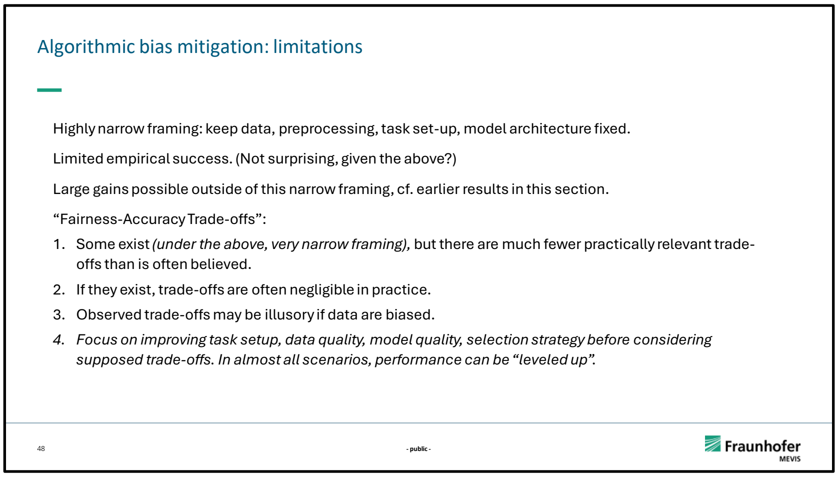
“No method outperforms ERM with statistical significance”



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".*

48

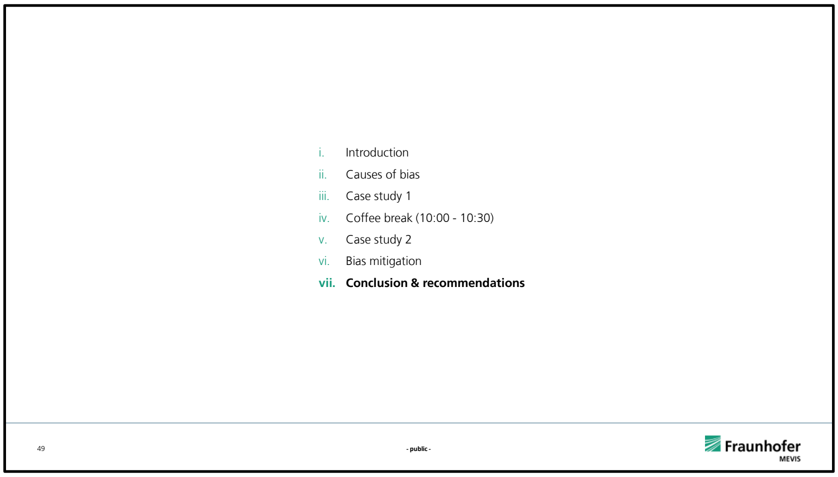
- public -

Fraunhofer MEVIS

On „limited empirical success", cf. e.g. Zong et al and Yang et al cited on the previous slide, + Zhang et al. (2022), Improving the Fairness of Chest X-ray Classifiers, https://proceedings.mlr.press/v174/zhang22a.html.

On "empirically negligible trade-offs", cf. Rodolfa et al. (2021), Empirical observation of negligible fairness–accuracy trade-offs in machine learning for public policy, https://www.nature.com/articles/s42256-021-00396-x.

On "fewer trade-offs than often believed", see e.g.
Lazar Reich and Vijaykumar (2021), A Possibility in Algorithmic Fairness: Can Calibration and Equal Error Rates Be Reconciled? https://arxiv.org/abs/2002.07676
Wick et al. (2019), Unlocking Fairness: a Trade-off Revisited, https://proceedings.neurips.cc/paper/2019/hash/373e4c5d8edfa8b74fd4b6791d0cf6dc-Abstract.html



i. Introduction

ii. Causes of bias

iii. Case study 1

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

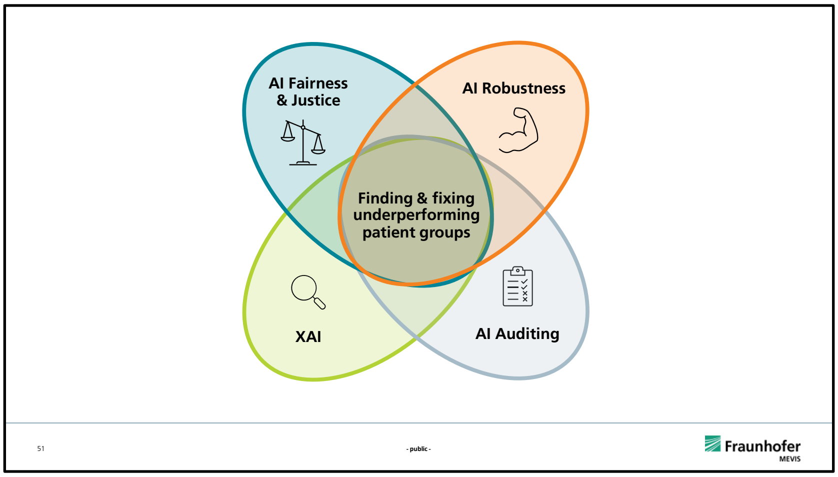
v. Case study 2

vi. Bias mitigation

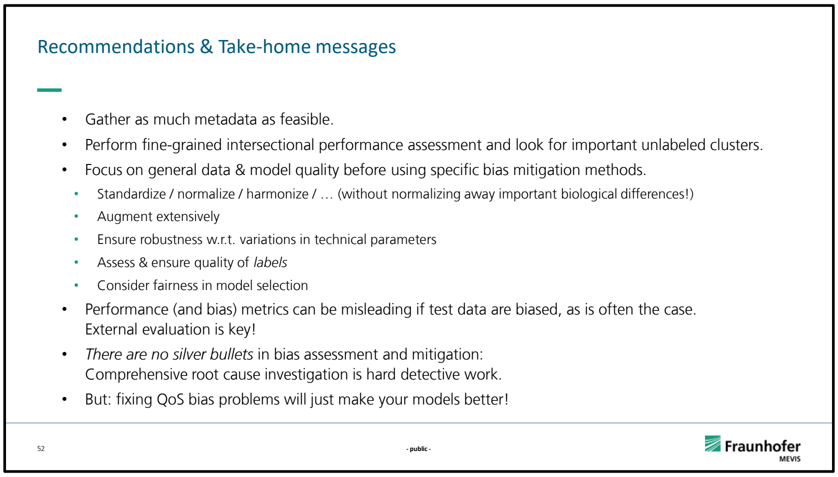
vii. Conclusion & recommendations

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

There are no silver bullets in bias assessment and mitigation. Finding and fixing the causes of underperformance is a nontrivial endeavor but worthwhile.



You *can* be interested in ethics, fairness, etc., but you don‘t have to be.
This is still for you if you just want to build good models!
Models that perform well on all patient groups are robust, trustworthy, and easier to get regulatory approval for.



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!

52

- public -

Intersectional = considering sub-(sub-)groups, e.g., dark-skinned women aged 20-30y with comorbidity X.

In bias assessments, do not forget to:

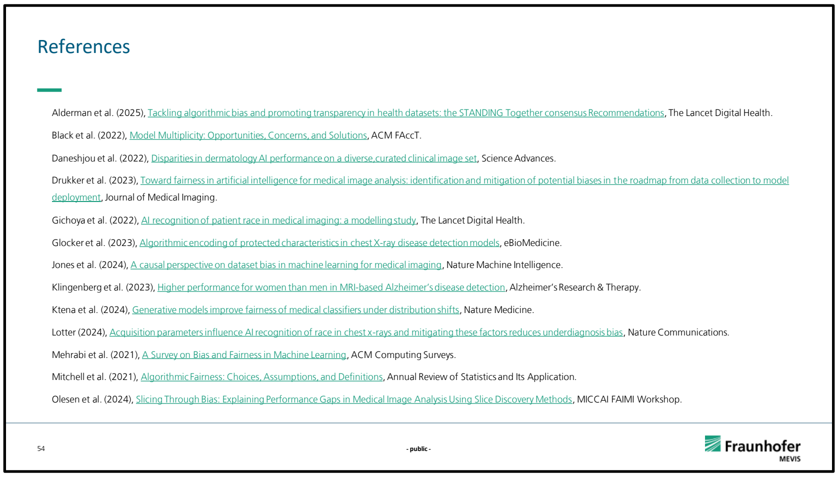
- Quantify metric uncertainty: groups become small and performance measures thus more uncertain.
- Correct for multiple hypotheses testing: we are comparing many different groups.
- Correct for known confounding factors that might differ between demographic groups: this will help disentangle the *causes* of performance discrepancies.
- Also compare groups defined by non-demographic properties (e.g., technical acquisition parameters): this will help find other issues related to (a lack of) model robustness.

Final recommendation

Fairness of AI in Medical Imaging
An independent academic initiative

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.

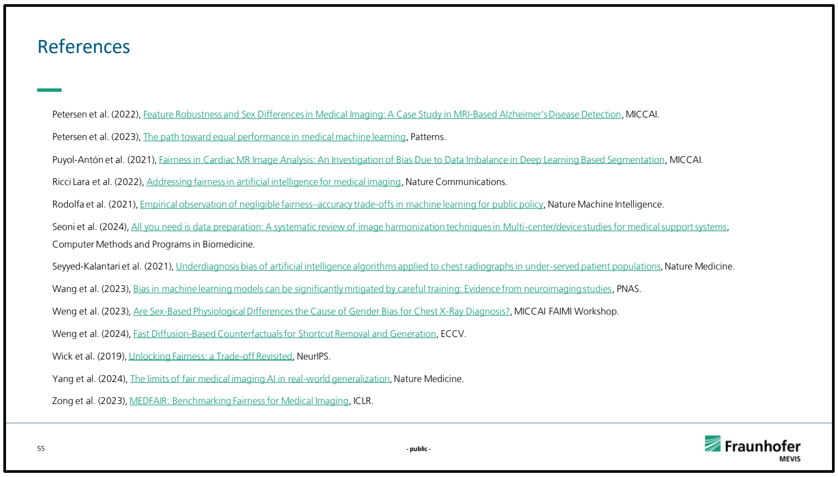
Mitchell et al. (2021), Algorithmic Fairness: Choices, Assumptions, and Definitions, Annual Review of Statistics and Its Application.

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

- public -

A reading list. Some that we referenced here, some that are good introductory reading material, some on more specific topics, some of our own prior work in this area.

Please reach out if looking for a reference on a particular topic!



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.

A reading list. Some that we referenced here, some that are good introductory reading material, some on more specific topics, some of our own prior work in this area.

Please reach out if looking for a reference on a particular topic!